Breast Cancer Research

RESEARCH

Open Access

Development and validation of a risk prediction model for premenopausal breast cancer in 19 cohorts

Kristen D. Brantley[1,2*], Michael E. Jones[3], Rulla M. Tamimi[4], Bernard A. Rosner[5,6], Peter Kraft[7], Hazel B. Nichols[8], Katie M. O'Brien[9], Hans-Olov Adami[10,11], Amaia Aizpurua[12,13], Amy Berrington de Gonzalez[14], William J. Blot[15], Tonje Braaten[16], Yu Chen[17], Jessica Clague DeHart[18], Laure Dossus[19], Sjoerd Elias[20], Renée T. Fortner[21,22], Montserrat Garcia-Closas[23], Inger T. Gram[16], Niclas Håkansson[24], Susan E. Hankinson[25], Cari M. Kitahara[26], Woon-Puay Koh[27], Martha S. Linet[26], Robert J. MacInnis[28,29], Giovanna Masala[30], Lene Mellemkjær[31], Roger L. Milne[28,29,32], David C. Muller[33], Hannah Lui Park[34], Kathryn J. Ruddy[35], Sven Sandin[10], Xiao-Ou Shu[36], Sandar Tin Tin[37], Thérèse Truong[38], Celine M. Vachon[39], Lars J. Vatten[40], Kala Visvanathan[41], Elisabete Weiderpass[42], Walter Willett[1,5], Alicja Wolk[24], Jian-Min Yuan[43,44], Wei Zheng[36], Dale P. Sandler[9], Minouk J. Schoemaker[45], Anthony J. Swerdlow[3,46] and A. Heather Eliassen[1,5]

Abstract

Background Incidence of premenopausal breast cancer (BC) has risen in recent years, though most existing BC prediction models are not generalizable to young women due to underrepresentation of this age group in model development.

Methods Using questionnaire-based data from 19 prospective studies harmonized within the Premenopausal Breast Cancer Collaborative Group (PBCCG), representing 783,830 women, we developed a premenopausal BC risk prediction model. The data were split into training (2/3) and validation (1/3) datasets with equal distribution of cohorts in each. In the training dataset variables were chosen from known and hypothesized risk factors: age, age at menarche, age at first birth, parity, breastfeeding, height, BMI, young adulthood BMI, recent weight change, alcohol consumption, first-degree family history of BC, and personal history of benign breast disease (BBD). Hazard ratios (HR) and 95% confidence intervals (CI) were estimated by Cox proportional hazards regression using age as time scale, stratified by cohort. Given that complete information on all risk factors was not available in all cohorts, coefficients were estimated separately in groups of cohorts with the same available covariate information, adjusted to account for the correlation between missing and non-missing variables and meta-analyzed. Absolute risk of BC (in situ or invasive) within 5 years, was determined using country-, age-, and birth cohort-specific incidence rates. Discrimination

Prior presentation: Development of an absolute risk prediction model for premenopausal breast cancer in an international consortium (Selected Oral Abstract). 2023. San Antonio Breast Cancer Symposium, San Antonio, TX.

*Correspondence:
Kristen D. Brantley
kbrantley@g.harvard.edu
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

(area under the curve, AUC) and calibration (Expected/Observed, E/O) were evaluated in the validation dataset. We compared our model with a literature-based model for women < 50 years (iCARE-Lit).

Results Selected model risk factors were age at menarche, parity, height, current and young adulthood BMI, family history of BC, and personal BBD history. Predicted absolute 5-year risk ranged from 0% to 5.7%. The model overestimated risk on average [E/O risk = 1.18 (1.14–1.23)], with underestimation of risk in lower absolute risk deciles and overestimation in upper absolute risk deciles [E/O 1st decile = 0.59 (0.58–0.60); E/O 10th decile = 1.48 (1.48–1.49)]. The AUC was 59.1% (58.1–60.1%). Performance was similar to the iCARE-Lit model.

Conclusion In this prediction model for premenopausal BC, the relative contribution of risk factors to absolute risk was similar to existing models for overall BC. The discriminatory ability was nearly identical (< 1% difference in AUC) to the existing iCARE-Lit model developed in women under 50 years. The inability to improve discrimination highlights the need to investigate additional predictors to better understand premenopausal BC risk.

Keywords Risk prediction model, Premenopausal breast cancer, Young-onset breast cancer

Introduction

Breast cancer (BC) is the leading cancer diagnosis and the leading cause of cancer death among women worldwide [1–4]. While diagnosis prior to menopause is less common than postmenopausal diagnosis, younger age at diagnosis often involves a more aggressive form of disease and is associated with worse prognosis [2, 5–7]. Moreover, younger women are less likely to be diagnosed through screening, as the World Health Organization and most country-specific commissions recommend mammographic screening for average risk women beginning at 50 years of age [8–10], though the US has recently lowered the recommended starting age to 40–45 years [11, 12]. Rising incidence of premenopausal BC globally in recent decades [2, 13] highlights the need to understand risk of BC in young women better, to inform clinical surveillance among younger women.

While several risk prediction models for overall BC with similar sets of risk factors are widely used, including Gail [14], Tyrer-Cuzick [15], and BOADICEA [16, 17] models, most have been developed among primarily postmenopausal women [18]. This limits generalizability to premenopausal women, especially given that associations of some established risk factors, such as adiposity [19, 20], differ by menopausal status, and other risk factors, such as breastfeeding [21] and recent weight change [22], are hypothesized to be more strongly associated with BC risk in premenopausal women. Thus, performance of existing models is generally worse for premenopausal v. postmenopausal women, with lower discriminatory ability and overestimation of risk seen among the premenopausal group [23–25]. Only one model to date has been specifically developed for women < 50 years of age, though relative risk contributions were derived from existing literature and not modeled within cohort data (iCARE 'synthetic' model, hereafter: iCARE-Lit) [26–29].

Here, we leveraged a large international consortium, the Premenopausal Breast Cancer Collaborative Group (PBCCG), to select model risk factors from a set of previously established and hypothesized risk factors that are widely available from standard survey questionnaires and develop a 5-year absolute risk prediction model in premenopausal women. We considered reproductive and lifestyle factors, anthropometrics, and personal and family history.

Methods

Cohort

The PBCCG was created to study risk factors for premenopausal BC by combining data from prospective cohort studies across Europe, North America, Asia, and Australia. Details on cohort design and data harmonization have been described [30]. Briefly, cohorts were eligible to join PBCCG if they had more than 100 cases BC diagnosed before age 55. To date, 22 cohorts have provided baseline questionnaire data, follow-up questionnaire data (where available, N= 17 cohorts), and incident case information to one of two centralized data repositories. Cases (in situ and invasive BC) diagnosed before menopause were identified by self-report, medical record, and/or linkage to cancer registries. Menopausal status was defined using data from multiple questionnaire cycles by (a) self-reported age at menopause (31%), or, where missing, (b) age first known to be postmenopausal if under 50 (1%), or (c) age last known premenopausal if over age 50 (15%), or (d) age 50 if no information on menopausal status was provided (53%). At least one follow-up cycle after age 55 was requested from each cohort to enable retrospective classification of menopausal status if necessary [30, 31]. Variables for future analyses were harmonized using a common protocol at the data centers [30].

Nineteen cohorts were included in this analysis, representing Asia (N= 2), Europe (N= 6), the United States (N= 10), and Australia (N= 1).

Statistical analyses

Individuals were assigned to multiple 5-year risk intervals based on age throughout follow-up, with covariates updated over time based on the most recent questionnaire, with censoring at age of menopause. Within individual cohort studies, if missing for only some individuals, age at menarche, alcohol consumption, BMI at age 18–24 years, and breastfeeding duration were imputed with median values (Table S1). This method was used given low within-cohort missingness (< 10% for each variable within each cohort, details provided in: Supplemental Methods "Within-cohort imputation") and given that low correlations between BC risk factors inhibits usefulness of conditional imputation. Yearly BMI was imputed using linear interpolation to allow for the calculation of recent weight change.

To facilitate internal model validation, the sample was split randomly within cohorts (two-thirds allocated to training dataset and one-third to validation dataset). Model building steps were performed within the training dataset, using STATA version 16 [32]. Cox proportional hazards regression [33, 34], with age as the underlying time-scale and stratified by cohort, was used to estimate hazard ratios (HR) and 95% confidence intervals (CI) for premenopausal BC. Availability of model variables of interest varied by cohort due to either lack of initial cohort data collection or lack of integration into the version of the harmonized dataset used in this analysis. Variables were selected from the following set of established and hypothesized risk factors if $p < 0.05$ by groupwise forward selection: first-degree family history of BC, reproductive factors (age at menarche, parity, age at first birth, breastfeeding), anthropometrics (height, BMI, BMI at age 18–24, recent weight change), current alcohol intake, and personal history of BBD. This method was used to enable incorporation of data from all cohorts in variable selection, despite differences in availability of certain variables. Details on the selection of variables are provided in the Supplemental Methods. To ensure that variables identified as model predictors were not dependent on this selection method, we also performed backwards selection within the set of cohorts with information available on all variables of interest ($N = 8$ cohorts).

After selection, linearity was assessed using Martingale residuals and the assumption of proportional hazards was tested visually by checking plots of Schoenfeld residuals by age in the full model and -ln(-ln(Survival probability) vs. ln(age), and statistically, by interaction tests between each variable and age. No substantial deviations from linearity or proportional hazards were detected (see Supplemental Methods). Because some variables selected into the final model were not available in all cohorts, final risk model coefficients (HRs) were estimated using covariance adjustment and a generalized meta-analysis approach [35]. First, cohorts were grouped based on which of the required regression variables were available. Cox models were run for each group of cohorts. For groups missing the same set of covariates, coefficient estimates for the available covariates were adjusted using the covariance matrix from the studies with complete covariate data. Adjusted estimates were then meta-analyzed with weighting by the inverse variance (see Supplemental Methods). Model coefficients were compared to those derived directly from the subset of cohorts with complete covariate data.

Absolute risk estimates were calculated as:

$$5y\ absolute\ risk = years\ in\ age\ group_j \times \frac{IR_{ijk}}{RR_{ijk}} \times exp^{lp}$$

where $i =$ country, $j = 5\text{-}year$ age groups, $k =$ birth year, and $lp =$ linear predictor from the Cox model. Incidence rates (IR_{ijk})were obtained from the International Agency for Research on Cancer (IARC)'s Global Cancer Observatory (GCO) project (2020), available by region, country, 5-year age groups, and birth cohort from various cancer registries [36, 37]. To align with absolute risk assignment used in existing risk models and ensure incident rates represented average rates within our cohort [14], the obtained IR_{ijk} was divided by RR_{ijk}, the mean relative risk (exp^{lp}) in country i, age group j, and birth year k in our training data sets [14].

Model performance was evaluated in the validation dataset. Discrimination was measured by the area under the curve (AUC) based on (1) linear predictor deciles and (2) absolute risk deciles. Expected and observed relative and absolute risks were calculated and plotted by training data-determined relative or absolute risk deciles, respectively [38]. For calibration within individual studies, deciles were recalculated based on the training data for each study.

We compared performance of our model among the subset of women < 50 y of age with that of the iCARE-Lit model [26–29]. Five-year age group-specific SEER incidence rates were applied when calculating absolute risk for both prediction models. Details are provided in Supplemental Methods.

Invasive-only model

Given that knowledge of risk of in situ or invasive BC can inform screening in younger women, we focused on BC overall as our primary outcome. However, because etiologies of invasive and in situ BC may differ, it is important to separately consider invasive cases. We repeated all analyses using invasive BC as the outcome to account for potential biological differences, to ensure accurate

assignment of global incidence rates, and to facilitate comparison of this model with iCARE-Lit, which was developed among invasive cases [29, 38]. Individuals who developed in situ BC were censored at time of diagnosis.

Among invasive BC cases, we further included a sensitivity analysis testing model variable selection and model performance for ER + and ER- BC separately, censoring at time of BC diagnosis of opposite or unknown subtype.

Results

The analytic cohort included 783,830 premenopausal women, with 9,618 incident premenopausal BC cases followed for 8.1 years on average (max = 25.2 y). The majority of participants were from North America and Western Europe (Table 1) and were White (58%) and non-Hispanic (99%) (Table 2). The average age at baseline questionnaire was 39.9 years (standard deviation, SD = 6.9 y). Most (79%) were parous, with mean age at first birth of 25 years, and 2 births and 15 months of breastfeeding on average. Approximately 10% of the cohort reported a first-degree family history of BC and 17% reported a personal history of BBD. Average BMI at most current questionnaire cycle was 24.0 kg/m^2 (SD = 4.5 kg/m^2), with a lower BMI reported in young adulthood [mean (SD) = 21.2 (3.1) kg/m^2]. Mean four-year weight change prior to the most current questionnaire was + 0.5 kg (Table 2). Approximately 50% of the cohort were current alcohol drinkers. Aside from variables missing by design (Table S2), missingness was low within-cohorts (< 10%) (Supplemental Methods). Compared to the full cohort, in the subset of cohorts with no model variables missing by design (N = 426,128, cases = 5,704), BBD history was more common (31% vs. 17%), while means and proportions of other variables were similar (Table S3).

Relative risk model

Variables selected in the model were age at menarche, parity, height, BMI (current and young adulthood), personal history of BBD, and first-degree family history of BC (Table 3). Variables were consistent when selecting among studies with complete covariate data only and coefficients for the prediction model determined from generalized meta-analysis were similar to those calculated for this subset (N = 284,149, cases = 3,777) (Table 3, Table S4). Lower risk of premenopausal BC was observed with increasing age at menarche [adjusted

Table 1 Distribution of 19 cohorts in the PBCCG (N = 783,830, breast cancer cases = 9,618)

Cohort	Acronym	N (%)	Cases (%)
Europe			
BCN Generations Study	BGS	61,303 (7.8%)	458 (4.8%)
Etude Epidemiologique aupres de femmes de la Mutuelle Generale de L'Education Nationale	E3 N	58,360 (7.5%)	1081 (11.2%)
European Prospective Investigation into Cancer and Nutrition	EPIC	89,464 (11.4%)	921 (9.6%)
The HUNT2 Study	HUNT2	16,955 (2.2%)	44 (0.5%)
Norwegian Women and Cancer Study	NOWAC	76,001 (9.7%)	427 (4.4%)
Swedish Mammography Cohort	SMC	26,329 (3.4%)	157 (1.6%)
The Swedish Women's Lifestyle and Health Study	SWLHS	47,371 (6.0%)	243 (2.5%)
North America (United States)			
California Teachers Study	CTS	46,251 (5.9%)	628 (6.5%)
Campaign against Cancer and Heart Disease	CLUEII	4,099 (0.5%)	47 (0.5%)
Mayo Mammography Health Study	MMHS	5,017 (0.6%)	53 (0.6%)
Nurses'Health Study	NHS	95,012 (12.1%)	1937 (20.1%)
Nurses'Health Study II	NHSII	115,623 (14.8%)	2397 (24.9%)
NYU Women's Health Study	NYUWHS	6,688 (0.9%)	197 (2.0%)
The Sister Study	SIS	16,877 (2.2%)	374 (3.9%)
Southern Community Cohort Study	SCCS	14,126 (1.8%)	55 (0.6%)
US Radiologic Technologists Cohort	USRTC	52,559 (6.7%)	412 (4.3%)
Asia			
Shanghai Women's Health Study	SWHS	33,707 (4.3%)	85 (0.9%)
Singapore Chinese Health Study	SCHS	9,989 (1.3%)	38 (0.4%)
Australia			
Melbourne Collaborative Cohort Study	MCCS	8,099 (1.0%)	64 (0.7%)
Total		**783,830**	**9,618**

Table 2 Characteristics of participants from 19 cohort studies within the PBCCG ($N = 783,830$, breast cancer cases = 9,618)

Characteristic	Mean (SD) or N (%)[a]
Race/Ethnicity	
White	455,286 (58%)
Black	18,836 (2.4%)
Asian	49,200 (6.3%)
Hispanic	5,094 (< 1%)
Other/Unknown	251,791 (32%)
Age at baseline questionnaire (years)	39.9 (6.9)
Age at menarche (years)	12.8 (1.5)
Height (cm)	164.2 (6.5)
BMI (kg/m^2)	24.0 (4.5)
Missing	11,310 (1.4%)
BMI age 18–21 (kg/m^2)[b]	21.2 (3.1)
Missing	23,602 (3.0%)
4-y weight change (kg)[b]	0.5 (1.8)
Missing	35,403 (4.5%)
Current alcohol drinker (yes)	393,978 (50%)
Missing current alcohol use	241,439 (31%)
Alcohol (drinks/week, among current drinkers)[b]	4.2 (8.1)
Missing	40,716 (65%)
Family history of BC[b]	77,747 (9.9%)
Missing	55,482 (7%)
History of benign breast disease[b]	134,266 (17%)
Missing	348,889 (44%)
Parous	618,405 (79%)
Parity (among parous)	2.2 (1.1)
Missing	6 (<0.1%)
Age at first birth (among parous)	25.2 (4.5)
Missing	4,670 (0.8%)
Breastfeeding months (among parous)[b]	15 (11.8)
Missing	120,085 (19%)
Breast cancer cases	**9,618 (1.2%)**
Age at diagnosis (years)	46.4 (4.8)
In situ	1673 (17%)
Invasive	7,914 (82%)
Stage I	2393 (25%)
Stage II	1843 (19%)
Stage III	605 (6.3%)
Stage IV	123 (1.3%)
Unknown	3460 (32%)
In situ vs. invasive status missing	31 (0.3%)
ER Status	
Positive	4,735 (49%)
Negative	1,517 (16%)
Borderline/Unknown	3,366 (35%)
PR status	
Positive	4,097 (43%)
Negative	1,717 (18%)
Borderline/Unknown	3,799 (39%)
HER2 Status	
Positive	657 (6.8%)
Negative	2,877 (30%)
Borderline/Unknown	6,084 (63%)

ER Estrogen receptor, *PR* Progesterone receptor, *HER2* Human epidermal growth receptor 2

[a] Mean (SD) and N (%) after within-cohort imputation

[b] Among cohorts without variable missing by design

Table 3 Beta coefficients and hazard ratios (HR) and 95% confidence intervals (CI) for the PBCCG risk model for premenopausal breast cancer

	Full Dataset[a]		Subset with complete covariate data[b]
Variable	Beta	HR (95% CI)[c]	HR (95% CI)[d]
Age at menarche (per year)	−0.046	0.96 (0.94–0.97)	0.95 (0.93–0.98)
Parity (per child)	−0.131	0.88 (0.84–0.91)	0.89 (0.87–0.92)
Height (cm)[e]	0.110	1.12 (1.03–1.21)	1.17 (1.11–1.23)
BMI (kg/m^2)[f]	−0.124	0.88 (0.82–0.95)	0.92 (0.88–0.95)
BMI age 18–24 (kg/m2)[f]	−0.079	0.92 (0.85–1.01)	0.89 (0.83–0.96)
History of BBD	0.440	1.55 (1.45–1.66)	1.55 (1.45–1.67)
Family history of BC	0.541	1.72 (1.59–1.85)	1.72 (1.58–1.87)

BBD Benign breast disease, *BMI* Body mass index, *PBCCG* Premenopausal Breast Cancer Collaborative Group

[a] Estimates from training dataset, $N = 522,700$ Cases = 6,392

[b] Cohorts include those with no model variables missing by design: BCN Generations Study, E3 N, Nurses' Health Study, Nurses' Health Study II, The Sister Study, Women's Lifestyle and Health Study, Shanghai Women's Health Study. Estimates from training data, $N = 284,149$, Cases = 3,777

[c] Beta estimates calculated following generalized meta-analysis with random effects model, mutually adjusted for all listed variables

[d] Beta estimates calculated directly from Cox PH model results among complete case dataset, mutually adjusted for all listed variables

[e] Estimated per 10 cm

[f] Estimated per 5 kg/m2

HR (aHR) per year (95% CI) = 0.96 (0.94–0.97)], parity [aHR per child = 0.88 (0.84–0.91)], current BMI [aHR per 5 kg/m^2 = 0.88 (0.82–0.95)], and BMI in young adulthood (aHR per 5 kg/m^2 = 0.92 (0.85–1.01)], while higher risk was observed with height [aHR per 10 cm = 1.12 (1.03–1.21)], personal history of BBD [aHR = 1.55 (1.45–1.66)], and first-degree family history of BC [aHR = 1.72 (1.59–1.85)]. The magnitude of the linear predictor, which estimates the log relative risk for an individual based on their covariate distribution, varied substantially by cohort (Table S5).

Absolute risk & model calibration

Five-year absolute risk of BC ranged from 0.00017% to 5.65% for all individuals (Table S5). Overall, the model overestimated absolute risk of premenopausal BC [E/O (95% CI = 1.18 (1.14–1.23)] (Table S6), though calibration varied widely across cohorts. Overestimation was similar in the subset with complete covariate data [E/O = 1.19 (1.14–1.25)] (Table S7). The absolute risk of developing BC was underestimated in lower deciles of absolute risk and overestimated in higher deciles of risk, (Table 4, Fig. 1A). Calibration of relative risk based on covariate-determined risk score deciles showed less overestimation in the upper tails (Fig. 1B). In the subset with complete covariate data, absolute risk calibration was similar, and relative risk was well-calibrated (Figure S1 A, S1B).

Model discrimination was modest, with an AUC of 59.1% (58.1–60.1%) for the full model (Table S6) and similar discrimination in the complete case subset (AUC = 60.2%, 95% CI = 58.9–61.4%) (Table S7). Overestimation was particularly high for the Asian cohorts with limited case numbers.

Comparison with iCARE-Lit < 50 model

Most variables are the same between our model and the iCARE-Lit model, with the exception of alcohol consumption and oral contraceptive use (both used in iCARE-Lit). Given the lack of consistent information on current oral contraceptive use in PBCCG, we considered all participants to be missing this variable when calculating the iCARE-Lit risk. When restricting to women < 50 years of age and comparing our model to the iCARE-Lit model (excluding use of OCs given missing data in PBCCG) performance was similar [our model: AUC = 59.8 (58.2–61.4), E/O = 1.20 (1.13–1.27) vs. iCARE-Lit < 50: AUC = 60.7 (59.0–62.3), E/O = 1.20 (1.13–1.27)].

Table 4 Expected over observed (E/O) ratio and 95% confidence interval (CI) by decile of relative or absolute risk score among 19 cohorts within the PBCCG in testing dataset (*N* = 261,130, breast cancer cases = 3,226)

Decile	E/O (95% CI)	
	By linear predictor decile[a,b]	By absolute risk decile[b]
1	1.40 (1.39–1.42)	0.59 (0.58–0.60)
2	1.19 (1.18–1.20)	0.69 (0.68–0.70)
3	1.19 (1.18–1.19)	0.93 (0.92–0.94)
4	1.22 (1.21–1.23)	0.98 (0.98–0.99)
5	1.32 (1.31–1.33)	1.02 (1.01–1.02)
6	1.12 (1.12–1.13)	1.04 (1.04–1.05)
7	1.12 (1.11–1.13)	1.23 (1.22–1.24)
8	1.11 (1.10–1.11)	1.35 (1.34–1.36)
9	1.20 (1.20–1.21)	1.41 (1.41–1.42)
10	1.15 (1.14–1.15)	1.48 (1.48–1.49)

[a] Linear predictor = $\log_e$ relative risk for individual calculated by the prediction model

[b] Decile of relative risk and absolute risk scores determined by distribution in training dataset

Invasive BC model

There were 7,914 invasive BC cases within the PBCCG cohorts (Table S8). Among those with known estrogen receptor (ER) status, most (75%) had ER + BC (Table S9). The relative risk model included all variables selected in the in situ + invasive model plus alcohol consumption. Following meta-analysis, HR estimates were similar to those in the in situ + invasive model. Though current alcohol was selected into the model as a significant predictor for the full cohort using our procedure (see Supplemental Methods), the HR was not significant in the full cohort after meta-analysis accounting for all cohorts [aHR (95% CI) drinks/week = 1.01 (0.99–1.02)] (Table S10). Alcohol intake remained modestly significantly associated with risk in the complete case model [aHR (95% CI) drinks/week = 1.01 (1.00–1.02)] (Table S10).

Five-year absolute risk of invasive BC ranged from 0.00018% to 5.67%. The invasive-only model overestimated risk more than the in situ + invasive model [average E/O = 1.45 (95% CI = 1.39–1.50)]. Risk was underestimated in the lowest deciles of absolute risk to a lesser degree than in the in situ + invasive model, though risk was highly overestimated in upper risk deciles [e.g., in the 10 th decile of absolute risk, the model predicted 87% more cases than observed (Table S11, Figure S2 A)]. Relative risk was well-calibrated (Figure S2B).

Among women < 50 years of age, discrimination and average calibration were similar for our model and the iCARE-Lit model [our model: AUC = 58.2% (56.3–60%), E/O = 1.49 vs. iCARE-Lit < 50: AUC = 59.3% (57.4–61.1%), E/O = 1.49]. While the iCARE-Lit model was better calibrated in lower risk deciles, overestimation was higher in upper risk deciles [e.g., E/O in decile 10 iCARE-Lit = 2.17 (2.14–2.20) vs. our model = 1.66 (1.64–1.68)].

Model by ER status

Availability of ER-status of cases by cohort is provided in Table S12. Due to limited case numbers for ER-negative BC, we conducted a sensitivity analysis for ER-positive BC only. Variables selected for the ER-positive BC specific model were the same as those chosen in our overall model (Table S13), with a similar contribution to risk for age at menarche, height, parity, BMI, and BMI in young adulthood. Family history of BC and personal history of BBD were weaker risk factors in the ER-positive

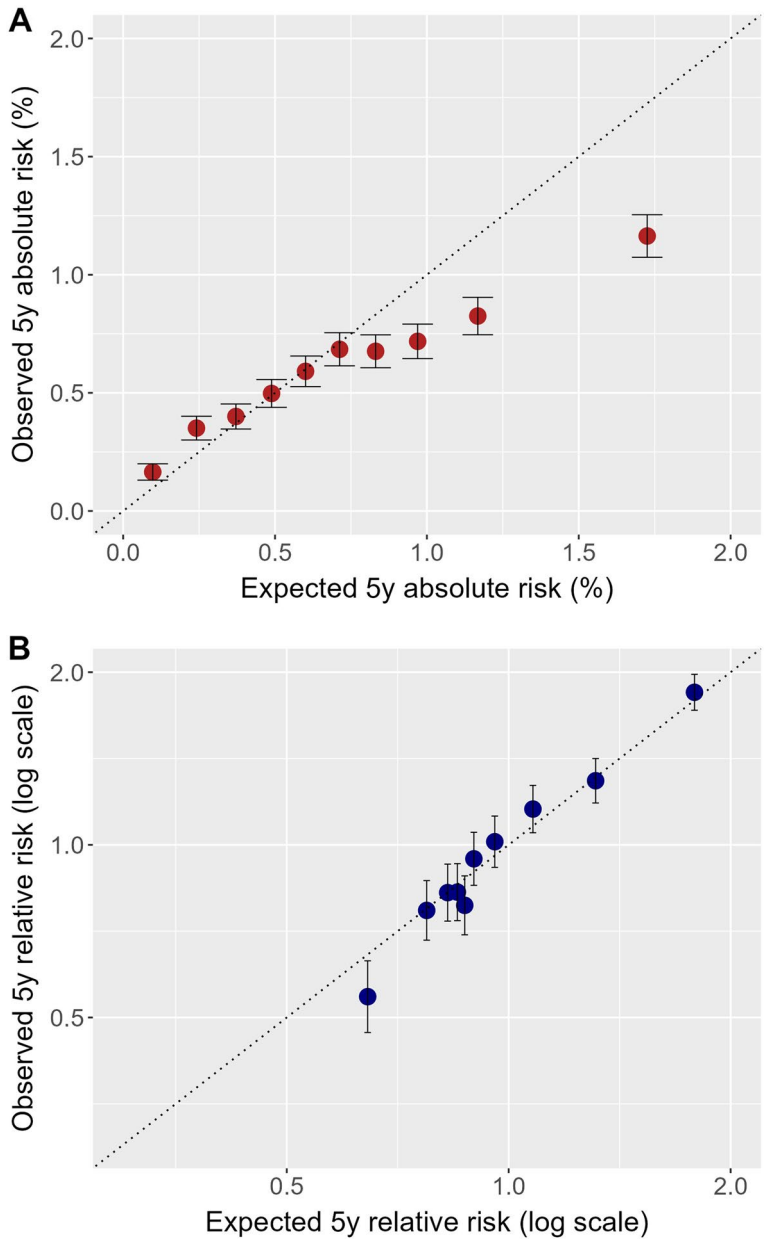

Fig. 1 Calibration for the risk model among 19 cohorts within the PBCCG. E/O within the testing dataset (*N* = 261,130, breast cancer cases = 3,226), for (**A**) absolute risk of BC a by decile of absolute risk in the training dataset and (**B**) relative risk of BC in the testing dataset, by decile of relative risk in the training dataset (log scale)

specific model (Table S13). The ER-positive set overestimated BC risk [E/O overall ER-positive = 2.43 (2.30–2.56)] (Table S14, Figure S3), which may in part to the large amount of missingness for ER-status. Discrimination of the ER-positive model was similar to the overall invasive model [ER-positive: AUC by linear predictor = 58.0% (56.5–59.5%), AUC by absolute risk = 62.2% (60.8–63.6%)].

Discussion

By pooling data from 19 international cohorts, we built an absolute risk prediction model for premenopausal BC informed by the largest set of data on premenopausal women to date. Our model tended to overestimate risk and discrimination remained modest, with performance similar to existing models for overall BC. Other cohorts, which have fewer premenopausal women and/or limited follow-up time compared to PBCCG, are unlikely to

produce better risk predictions using similar variables. To improve premenopausal BC risk prediction, it will be essential to incorporate additional known risk factors (e.g., OC use, polygenic risk score, and mammographic density), to increase granularity of data on existing variables, to test the contribution of novel risk factors such as early life environmental exposures, and to account for gene-environment interactions.

The variables selected in our premenopausal BC risk model matched those included in models for overall BC risk [14, 15, 17, 26]. It is important to note that, while our model included both ER-positive and ER-negative cases, given the much higher proportion of ER + cases, this subtype dominates our results. The strongest risk factors in our model were personal history of BBD and first-degree family history of BC, both of which had HRs consistent with previous models [14, 15, 29]. Personal history of BBD was associated with a 55% increased risk of premenopausal BC in this population, similar to values reported in the Gail model (HR for one biopsy if < 50 years = 1.70) [14], the Tyrer-Cuzick Model (HR for prior biopsy among women < 50 years = 1.26–2.0, depending on biopsy findings) [15], and the iCARE-Lit model (HR for history of BBD = 1.68 for women) [26]. Current estimates for the contribution of family history to overall BC risk vary across the literature, with some using detailed pedigrees [15, 17], and others using simple first-degree family history [39]. We found that first-degree family history of BC was associated with a 72% higher risk of premenopausal BC in this cohort, which is lower than estimates given for iCARE-Lit (< 50 model, aHR = 2.5) [29]. When modeling risk of ER + BC specifically, history of BBD and family history had a lower hazard ratio than in our overall model [aHR BBD = 1.21 (0.61–2.41), aHR family history = 1.67 (1.49–1.87)] suggesting these factors were potentially influenced by risk of ER-negative BC.

Similarities in reproductive factors between our model and existing risk prediction models include inverse associations of age at menarche and parity with premenopausal BC risk. Interestingly, age at first birth was not selected in our model, although a younger age at first birth has been associated with lower risk in cohorts of majority postmenopausal women [14, 15, 29, 40]. The lack of predictive value may be explained by the lag time required after pregnancy to experience the benefit of decreased BC risk following birth; for example, in the PBCCG, pregnancy was positively associated with BC risk up until 24 years after birth [41]. Based on variations in data collection methods across cohorts, we were unable to assess recency of last birth, which may be a more important predictor of risk within premenopausal women [41]. While some studies have reported an inverse association between breastfeeding and risk of premenopausal BC [21, 42–44], breastfeeding was not selected in our model. Heterogeneity across molecular subtypes may have contributed to the lack of selection, as stronger protective associations have been reported for triple negative BC [43], and for luminal Bsubtypes [44]. It is also possible that our results were biased by the high percentage of missingness by design in PBCCG cohorts for breastfeeding.

Our finding that BMI in young adulthood contributed to risk prediction aligns with data suggesting stronger inverse associations between young adult BMI and BC among premenopausal (vs. postmenopausal) women [45–47]. The inverse association between adulthood BMI and premenopausal BC found here follows the paradoxical relationship that is well-established in the literature [47]. While studies have shown increased BC risk among premenopausal women with short-term weight gain [22], and earlier age at BC diagnosis overall with increased annual adulthood weight gain [48], short-term weight change was not selected into our model. Lack of an observed association may be explained by the small average change over four years (+ 0.5 kg), and a narrow distribution of weight changes, precluding our ability to assess more extreme alterations. Further, because the linear interpolation method assumes constancy of weight change over questionnaire cycles, projected values may have been too small.

Our finding that alcohol consumption was selected as a predictor of invasive BC risk but not overall BC aligns with results from a study in the UK Biobank that found alcohol was not a risk factor for ductal carcinoma in situ [49]. Given that intake after in situ BC diagnosis has been associated with development of invasive BC [50], it is plausible that alcohol may play a role in later stages of BC development. However, the exclusion of this variable in our final model may simply be due to chance.

We did not assess the potential contribution of OC use and smoking to risk prediction due to limited detailed information available on these variables in the harmonized dataset, beyond ever vs. never use for OCs and current smoking status, though both remain of interest. A recent meta-analysis found no association for ever vs. never OC use and premenopausal BC risk [51]; however, studies have shown increased premenopausal BC risk for current OC users [52], or recent (< 5 years ago) users [53]. Evaluation is complicated by evidence that the association between OC use and BC differs across molecular subtypes [54], and the wide variation of types and dosages of OCs used across age groups, birth cohorts, and countries. Smoking has been shown to be associated with increased risk of BC only when incorporating information on duration, intensity, and years of quitting smoking [55–57], and, timing of initiation before first childbirth

[57, 58]. In the future, detailed data collection for active and passive [59, 60] smoking and OC use should be prioritized.

Given the small absolute risk of premenopausal BC for most women, the assignment of appropriate population-based incidence rates is essential to accurate risk prediction. We attempted to account for geographic, age, and birth-cohort differences using data collected on invasive BC incidence rates from GCO data [36]. Despite this, absolute risks were overestimated, especially among those with higher risk profiles. While the estimation was imperfect due to our assignment of invasive BC incidence rates to 1,673 in situ BC cases in our cohort, when predicting risk for invasive BC only, estimation was not improved. While this can be explained in part by differences in model specification, which disallows a direct comparison between models, the lack of improvement in the invasive-only model may be indicative of non-representative incidence rates, perhaps due to within-country heterogeneity and varied quality of data for different birth cohorts.

Prior validation of established risk prediction models in the BCN Generations Study demonstrated AUCs based on relative risk scores similar to that of our model [Gail model (AUC = 54.6%), Tyrer-Cuzick model (AUC = 57.0%), and iCARE-Lit < 50y (AUC = 58.8%)] [26]. While the addition of polygenic risk score (PRS), pathogenic variants, and mammographic density have been shown to enhance risk prediction in existing models [61], we did not have data to test these in PBCCG.

Though this is the largest study of premenopausal BC risk modeling to date, there are several limitations to note. First, the PBCCG is still overwhelmingly represented by majority White, Western European and North American Studies, and findings cannot be generalized to diverse populations. The large amount of missing data by design complicated model building and calibration; while we used selection methods and generalized meta-analysis to obtain the most informative set of variables and beta coefficients for our model, coefficient estimates may be biased due to missing data. While the harmonization of menopausal status information was carefully considered in the creation of the PBCCG, it is possible that we have misclassified some individuals missing explicit information on menopausal status, limiting the performance of our risk prediction model.

Overall, the similarity of our model performance to the iCARE-Lit model demonstrates the difficulties of predicting premenopausal BC risk using variables readily obtained from cohort questionnaires and medical records, even with substantial case numbers. The addition of genetic information and breast density may be a first step toward improving risk prediction. However, PRS and RR estimates for BC due to associated variants were developed using data from majority postmenopausal women. Given that interactions between age and PRS, while minimal, have been shown [62], moving forward, it will be important to consider how germline variation uniquely influences premenopausal BC risk. Ultimately, further research into the mechanisms driving premenopausal BC development is necessary to improve clinical risk prediction in younger women.

Conclusions

Here we developed and internally validated the first risk prediction model for premenopausal breast cancer, using questionnaire-based data from over 780,000 women. Our model performed similarly to a literature-derived model for women < 50 years of age, with performance statistics similar to those of existing risk prediction models (developed among primarily postmenopausal women). This indicates that additional variables and/or a deeper level of detail of questionnaire-based variables are needed to better predict premenopausal BC risk.

Abbreviations

BC	Breast cancer
PBCCG	Premenopausal Breast Cancer Collaborative Group
BBD	Benign breast disease
HR	Hazard ratio
CI	Confidence interval
AUC	Area under the curve
E/O	Expected to observed ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-025-02031-8.

Supplementary Material 1: Figure 1. Calibration for the risk model among 19 cohorts within the PBCCG in subset of cohorts with complete covariate data. E/O within the testing dataset, N=141,979, breast cancer cases=1,927) for: (A) absolute risk of BC in testing dataset by decile of absolute risk in the training dataset (B) relative risk of BC in the testing dataset, by decile of relative risk in the training dataset.

Supplementary Material 2: Figure 2. Invasive-only BC model calibration among 19 cohorts within the PBCCG. Observed vs. expected in testing dataset (N=261,130, breast cancer cases=2,640): (A) absolute risk of BC in testing dataset by decile of absolute risk in the training dataset (B) relative risk of BC in the testing dataset, by decile of relative risk in the training dataset.

Supplementary Material 3: Figure 3. ER-positive BC model calibration among 16 cohorts within the PBCCG. Observed vs. expected in testing dataset (Breast cancer cases=1,390): (A) absolute risk of BC in testing dataset by decile of absolute risk in the training dataset (B) relative risk of BC in the testing dataset, by decile of relative risk in the training dataset.

Supplementary Material 4.

Supplementary Material 5.

Acknowledgements

We thank the National Cancer Institute Cohort Consortium for facilitating this collaboration. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease

Control and Prevention's National Program of Cancer Registries (NPCR) and/or the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. Central registries may also be supported by state agencies, universities, and cancer centers. Participating central cancer registries include the following: AL, AR, AZ, CA, CO, CT, DE, DC, FL, GA, HI, IA, IL, IN, KY, LA, MD, MA, MI, MO, MS, NE, NJ, NM, NY, NC, OH, OK, OR, PA, SC, TN, TX, VA, WA, WI. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Department of Health and Human Services, the National Institutes of Health, the National Cancer Institute, or the state cancer registries.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization. The opinions, findings, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the official views of the State of California, Department of Public Health, the National Cancer Institute, the National Institutes of Health, the Centers for Disease Control and Prevention or their Contractors and Subcontractors, or the Regents of the University of California, or any of its programs.

Authors' contributions

KDB wrote the manuscript text. KDB and MEJ conducted the statistical analyses. AHE, BR, and RMT conceptualized the idea and KDB, AHE, BR, RMT, and PK developed the methodology. MJS, DPS, KMO and MEJ supplied the harmonized dataset. All authors reviewed the manuscript.

Funding

This study was supported by the Division of Intramural Research at NIEHS, in the National Institutes of Health, under projects Z01-ES044005 for D.P. Sandler and Z01-ES102245 for C.R. Weinberg. This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health. The Sister Study was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (project Z01-ES044005 to DPS) which also partially supported the PMBCCG. The Singapore Chinese Health Study was supported by NIH grants (No. R01-CA144034 and UM1-CA182876). The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS)—Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology—ICO (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford). (United Kingdom); The BCN Generations Study is supported by Breast Cancer Now and The Institute of Cancer Research, United Kingdom, and acknowledges support by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and the Institute of Cancer Research, London. Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases were ascertained through the Victorian Cancer Registry. The California Teachers Study and the research reported in this publication were supported by the National Cancer Institute of the National Institutes of Health under award number U01-CA199277; P30-CA033572; P30-CA023100; UM1-CA164917; and R01-CA077398. The collection of cancer incidence data used in the California Teachers Study was supported by the California Department of Public Health pursuant to California Health and Safety Code Sect. 103885; Centers for Disease Control and Prevention's National Program of Cancer Registries, under cooperative agreement 5 NU58DP006344; the 3 National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The NHS was supported by the NIH (UM1-CA186107 and P01-CA87969) and the NHS2 was supported by the NIH (UM1-CA176726). Preparation of the manuscript was supported by funding from the National Cancer Institute (T32 CA009001) to A.H. Eliassen.

Data availability

PBCCG data is housed at the Coordinating Centers (Institute of Cancer Research (ICR), London, and The National Institute of Environmental Health Sciences (NIEHS) and the harmonized dataset cannot be transferred to other institutions. Those seeking to conduct an analysis in PBCCG may put in a formal proposal to the Consortium leaders via email and reasonable requests will be considered. Data usage agreements will be required.

Declarations

Ethics approval and consent to participate

Approval from institutional review boards and individual consent for all cohorts in the PBCCG conformed to each study's ethics review requirements.

Consent for publication

The authors give their consent for publication.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, 375 Longwood Ave LW737, Boston, MA 02115, USA. [3]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. [4]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [5]Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [7]Division of Cancer Epidemiology and Genetics, Transdivisional Research Program, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [8]University of North Carolina, Gillings School of Global Public Health, Chapel Hill, NC, USA. [9]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, Durham, NC, USA. [10]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [11]Clinical Effectiveness Research Group, Institute of Health and Society, University of Oslo, Oslo, Norway. [12]Ministry of Health of the Basque Government, Sub Directorate for Public Health and Addictions of Gipuzkoa, San Sebastian, Spain. [13]Biogipuzkoa Health Research Institute, Epidemiology of Chronic and Communicable Diseases Group, San Sebastián, Spain. [14]Clinical Cancer Epidemiology, Institute of Cancer Research, Sutton, London, UK. [15]Department of Medicine, Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA. [16]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway. [17]Department of Population Health, Division of Epidemiology, NYU Grossman School of Medicine, New York, NY, USA. [18]School of Community and Global Health, Claremont Graduate University, Claremont, CA, US. [19]Nutrition and Metabolism Branch, International Agency for Research On Cancer, Lyon, France. [20]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands. [21]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany. [22]Department of Research, Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway. [23]Cancer Epidemiology and Prevention Unit, Institute of Cancer Research, Sutton, London, UK. [24]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [25]Department

of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA. [26]Division of Cancer Epidemiology and Genetics, Radiation Epidemiology Branch, National Cancer Institute, Bethesda, MD, USA. [27]Healthy Longevity Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [28]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC, Australia. [29]Centre for Epidemiology and Biostatics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia. [30]Clinical Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network (ISPRO), Florence, Italy. [31]Danish Cancer Institute, Copenhagen, Denmark. [32]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia. [33]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [34]Department of Pathology and Laboratory Medicine, Department of Epidemiology, UC Irvine School of Medicine, Irvine, CA, USA. [35]Department of Oncology, Mayo Clinic, Rochester, MN, USA. [36]Department of Medicine, Institute for Medicine and Public Health, Division of Epidemiology, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [37]Cancer Epidemiology Unit, Oxford Population Health, University of Oxford, Oxford, UK. [38]Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, 94805 Villejuif, France. [39]Department of Quantitative Health Sciences, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA. [40]Department of Public Health and Nursing, Faculty of Medicine, Norwegian, University of Science and Technology, Trondheim, Norway. [41]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [42]Cancer, World Health Organization, International Agency for Research On , Lyon, France. [43]Cancer Epidemiology and Prevention Program, UPMC Hillman Cancer Center, Pittsburgh, PA, USA. [44]Department of Epidemiology, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA. [45]Real World Solutions, IQVIA, Amsterdam, Netherlands. [46]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK.

Received: 21 January 2025 Accepted: 21 April 2025
Published online: 01 May 2025

References

1. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. 2022;66:15–23. https://doi.org/10.1016/j.breast.2022.08.010.
2. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. Lancet Glob Health. 2020;8(8):e1027–37. https://doi.org/10.1016/S2214-109X(20)30215-1.
3. Dyba T, Randi G, Bray F, et al. The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. Eur J Cancer. 2021;157:308–47. https://doi.org/10.1016/j.ejca.2021.07.039.
4. Koh B, Tan DJH, Ng CH, et al. Patterns in cancer incidence among people younger than 50 years in the US, 2010 to 2019. JAMA Netw Open. 2023;6(8): e2328171. https://doi.org/10.1001/jamanetworkopen.2023.28171.
5. Partridge AH, Hughes ME, Warner ET, et al. Subtype-dependent relationship between young age at diagnosis and breast cancer survival. J Clin Oncol. 2016;34(27):3308–14. https://doi.org/10.1200/JCO.2015.65.8013.
6. Korde LA, Partridge AH, Esser M, Lewis S, Simha J, Johnson RH. Breast cancer in young women: research priorities. A report of the young survival coalition research think tank meeting. J Adolesc Young Adult Oncol. 2015;4(1):34–43. https://doi.org/10.1089/jayao.2014.0049.
7. Azim HA, Partridge AH. Biology of breast cancer in young women. Breast Cancer Res. 2014;16(4):427. https://doi.org/10.1186/s13058-014-0427-5.
8. World Health Organization. WHO Position Paper on Mammography Screening, 2014. https://www.paho.org/cancer.
9. Ren W, Chen M, Qiao Y, Zhao F. Global guidelines for breast cancer screening: a systematic review. Breast. 2022;64:85–99. https://doi.org/10.1016/j.breast.2022.04.003.
10. European guidelines on breast cancer screening and diagnosis. Accessed April 12, 2024. https://cancer-screening-and-care.jrc.ec.europa.eu/en/ecibc/european-breast-cancer-guidelines.
11. Draft Recommendation Statement: Breast Cancer Screening. Published online May 9, 2023. Accessed April 12, 2024. https://www.uspreventiveservicestaskforce.org/uspstf/draft-recommendation/breast-cancer-screening-adults#fullrecommendationstart.
12. Oeffinger KC, Fontham ETH, Etzioni R, et al. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. JAMA. 2015;314(15):1599–614. https://doi.org/10.1001/jama.2015.12783.
13. Sung H, Jiang C, Bandi P, et al. Differences in cancer rates among adults born between 1920 and 1990 in the USA: an analysis of population-based cancer registry data. Lancet Public Health. 2024;9(8):e583–93. https://doi.org/10.1016/S2468-2667(24)00156-7.
14. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81(24):1879–86. https://doi.org/10.1093/jnci/81.24.1879.
15. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004;23(7):1111–30. https://doi.org/10.1002/sim.1668.
16. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer. 2004;91(8):1580–90. https://doi.org/10.1038/sj.bjc.6602175.
17. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med. 2019;21(8):1708–18. https://doi.org/10.1038/s41436-018-0406-9.
18. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132(2):365–77. https://doi.org/10.1007/s10549-011-1818-2.
19. Anderson GL, Neuhouser ML. Obesity and the risk for premenopausal and postmenopausal breast cancer. Cancer Prev Res (Phila). 2012;5(4):515–21. https://doi.org/10.1158/1940-6207.CAPR-12-0091.
20. Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet. 2008;371(9612):569–78. https://doi.org/10.1016/S0140-6736(08)60269-X.
21. Abraham M, Lak MA, Gurz D, Nolasco FOM, Kondraju PK, Iqbal J. A narrative review of breastfeeding and its correlation with breast cancer: current understanding and outcomes. Cureus. 2023;15(8):e44081. https://doi.org/10.7759/cureus.44081.
22. Rosner B, Eliassen AH, Toriola AT, et al. Short-term weight gain and breast cancer risk by hormone receptor classification among pre- and postmenopausal women. Breast Cancer Res Treat. 2015;150(3):643–53. https://doi.org/10.1007/s10549-015-3344-0.
23. Spiegelman D, Colditz GA, Hunter D, Hertzmark E. Validation of the Gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst. 1994;86(8):600–7. https://doi.org/10.1093/jnci/86.8.600.
24. Dartois L, Gauthier É, Heitzmann J, et al. A comparison between different prediction models for invasive breast cancer occurrence in the French E3N cohort. Breast Cancer Res Treat. 2015;150(2):415–26. https://doi.org/10.1007/s10549-015-3321-7.
25 Gabrielson M, Ubhayasekera K, Ek B, et al. Inclusion of plasma prolactin levels in current risk prediction models of premenopausal and postmenopausal breast cancer. JNCI Cancer Spectr. 2018;2(4):pky055. https://doi.org/10.1093/jncics/pky055.
26. Pal Choudhury P, Wilcox AN, Brook MN, et al. Comparative validation of breast cancer risk prediction models and projections for future risk stratification. J Natl Cancer Inst. 2019;112(3):278–85. https://doi.org/10.1093/jnci/djz113.
27. Cintolo-Gonzalez JA, Braun D, Blackford AL, et al. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. Breast Cancer Res Treat. 2017;164(2):263–84. https://doi.org/10.1007/s10549-017-4247-z.
28. Hurson AN, Pal Choudhury P, Gao C, et al. Prospective evaluation of a breast-cancer risk model integrating classical risk factors and polygenic risk in 15 cohorts from six countries. Int J Epidemiol. 2022;50(6):1897–911. https://doi.org/10.1093/ije/dyab036.
29 Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. J Natl Cancer Inst. 2014;106(11):dju305. https://doi.org/10.1093/jnci/dju305.

30. Nichols HB, Schoemaker MJ, Wright LB, et al. The premenopausal breast cancer collaboration: a pooling project of studies participating in the national cancer institute cohort consortium. Cancer Epidemiol Biomarkers Prev. 2017;26(9):1360–9. https://doi.org/10.1158/1055-9965.EPI-17-0246.
31. Schoemaker MJ, Nichols HB, Wright LB, et al. Association of body mass index and age with subsequent breast cancer risk in premenopausal women. JAMA Oncol. 2018;4(11):e181771. https://doi.org/10.1001/jamaoncol.2018.1771.
32. StataCorp. Stata Statistical Software: Release 17. College Station: StataCorp LLC; 2021.
33. van Dijk PC, Jager KJ, Zwinderman AH, Zoccali C, Dekker FW. The analysis of survival data in nephrology: basic concepts and methods of Cox regression. Kidney Int. 2008;74(6):705–9. https://doi.org/10.1038/ki.2008.294.
34. Cox D. Regression models and life-tables. J R Stat Soc B. 1972;34:187–220.
35. Kundu P, Tang R, Chatterjee N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. Biometrika. 2019;106(3):567. https://doi.org/10.1093/biomet/asz030.
36. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49. https://doi.org/10.3322/caac.21660.
37. International Agency for Cancer Research. Cancer Today. https://gco.iarc.fr/today/en.
38. Pal Choudhury P, Maas P, Wilcox A, et al. iCARE: an R package to build, validate and apply absolute risk models. PLoS One. 2020;15(2):e0228198. https://doi.org/10.1371/journal.pone.0228198.
39. Rice MS, Tworoger SS, Hankinson SE, et al. Breast cancer risk prediction: an update to the Rosner-Colditz breast cancer incidence model. Breast Cancer Res Treat. 2017;166(1):227–40. https://doi.org/10.1007/s10549-017-4391-5.
40. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. Am J Epidemiol. 2000;152(10):950–64. https://doi.org/10.1093/aje/152.10.950.
41. Nichols HB, Schoemaker MJ, Cai J, et al. Breast cancer risk after recent childbirth: a pooled analysis of 15 prospective studies. Ann Intern Med. 2019;170(1):22–30. https://doi.org/10.7326/M18-1323.
42. Stuebe AM, Willett WC, Xue F, Michels KB. Lactation and incidence of premenopausal breast cancer, a longitudinal study. Arch Intern Med. 2009;169(15):1364–71. https://doi.org/10.1001/archinternmed.2009.231.
43. Islami F, Liu Y, Jemal A, et al. Breastfeeding and breast cancer risk by receptor status–a systematic review and meta-analysis. Ann Oncol. 2015;26(12):2398–407. https://doi.org/10.1093/annonc/mdv379.
44. Giudici F, Scaggiante B, Scomersi S, Bortul M, Tonutti M, Zanconati F. Breastfeeding: a reproductive factor able to reduce the risk of luminal B breast cancer in premenopausal White women. Eur J Cancer Prev. 2017;26(3):217–24. https://doi.org/10.1097/CEJ.0000000000000220.
45. Yang TO, Cairns BJ, Pirie K, et al. Body size in early life and the risk of postmenopausal breast cancer. BMC Cancer. 2022;22(1):232. https://doi.org/10.1186/s12885-022-09233-9.
46. Warner ET, Hu R, Collins LC, et al. Height and body size in childhood, adolescence, and young adulthood and breast cancer risk according to molecular subtype in the Nurses' Health Studies. Cancer Prev Res (Phila). 2016;9(9):732–8. https://doi.org/10.1158/1940-6207.CAPR-16-0085.
47. van den Brandt PA, Ziegler RG, Wang M, et al. Body size and weight change over adulthood and risk of breast cancer by menopausal and hormone receptor status: a pooled analysis of 20 prospective cohort studies. Eur J Epidemiol. 2021;36(1):37–55. https://doi.org/10.1007/s10654-020-00688-3.
48. Azrad M, Blair CK, Rock CL, Sedjo RL, Wolin KY, Demark-Wahnefried W. Adult weight gain accelerates the onset of breast cancer. Breast Cancer Res Treat. 2019;176(3):649–56. https://doi.org/10.1007/s10549-019-05268-y.
49. Peila R, Arthur R, Rohan TE. Risk factors for ductal carcinoma in situ of the breast in the UK Biobank cohort study. Cancer Epidemiol. 2020;64:101648. https://doi.org/10.1016/j.canep.2019.101648.
50. Jordahl KM, Malone KE, Baglia ML, et al. Alcohol consumption, smoking, and invasive breast cancer risk after ductal carcinoma in situ. Breast Cancer Res Treat. 2022;193(2):477–84. https://doi.org/10.1007/s10549-022-06573-9.
51. Kanadys W, Barańska A, Malm M, et al. Use of oral contraceptives as a potential risk factor for breast cancer: a systematic review and meta-analysis of case-control studies up to 2010. Int J Environ Res Public Health. 2021;18(9):4638. https://doi.org/10.3390/ijerph18094638.
52. Fitzpatrick D, Pirie K, Reeves G, Green J, Beral V. Combined and progestagen-only hormonal contraceptives and breast cancer risk: a UK nested case–control study and meta-analysis. PLoS Med. 2023;20(3):e1004188. https://doi.org/10.1371/journal.pmed.1004188.
53. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Lancet. 1996;347(9017):1713–27. https://doi.org/10.1016/s0140-6736(96)90806-5.
54. Barańska A. Oral contraceptive use and assessment of breast cancer risk among premenopausal women via molecular characteristics: systematic review with meta-analysis. Int J Environ Res Public Health. 2022;19(22):15363. https://doi.org/10.3390/ijerph192215363.
55. He Y, Si Y, Li X, Hong J, Yu C, He N. The relationship between tobacco and breast cancer incidence: a systematic review and meta-analysis of observational studies. Front Oncol. 2022;12:961970. https://doi.org/10.3389/fonc.2022.961970.
56. Scala M, Bosetti C, Bagnardi V, et al. Dose-response relationships between cigarette smoking and breast cancer risk: a systematic review and meta-analysis. J Epidemiol. 2023;33(12):640–8. https://doi.org/10.2188/jea.JE20220206.
57. Gram IT, Park SY, Kolonel LN, et al. Smoking and risk of breast cancer in a racially/ethnically diverse population of mainly women who do not drink alcohol: The MEC Study. Am J Epidemiol. 2015;182(11):917–25. https://doi.org/10.1093/aje/kwv092.
58. Bjerkaas E, Parajuli R, Weiderpass E, et al. Smoking duration before first childbirth: an emerging risk factor for breast cancer? Results from 302,865 Norwegian women. Cancer Causes Control. 2013;24(7):1347–56. https://doi.org/10.1007/s10552-013-0213-1.
59. Gram IT, Wiik AB, Lund E, Licaj I, Braaten T. Never-smokers and the fraction of breast cancer attributable to second-hand smoke from parents during childhood: the Norwegian Women and Cancer Study 1991–2018. Int J Epidemiol. 2022;50(6):1927–35. https://doi.org/10.1093/ije/dyab153.
60. Possenti I, Scala M, Carreras G, et al. Exposure to second-hand smoke and breast cancer risk in non-smoking women: a comprehensive systematic review and meta-analysis. Br J Cancer. 2024;131(7):1116–25. https://doi.org/10.1038/s41416-024-02732-5.
61. Yang X, Eriksson M, Czene K, et al. Prospective validation of the BOADICEA multifactorial breast cancer risk prediction model in a large prospective cohort study. J Med Genet. 2022;59(12):1196–205. https://doi.org/10.1136/jmg-2022-108806.
62. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J Hum Genet. 2019;104(1):21–34. https://doi.org/10.1016/j.ajhg.2018.11.002.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.