Long-Term Cause-Specific Mortality After Surgery for Women With Breast Cancer: A 20-Year Follow-Up Study From Surveillance, Epidemiology, and End Results Cancer Registries

Breast Cancer: Basic and Clinical Research
Volume 11: 1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1178223417711429

Gabriel Escarela[1], Alan Jiménez-Balandra[1], Gabriel Núñez-Antonio[1] and Antonio Gordillo-Moscoso[2]

[1]Departamento de Matemáticas, Universidad Autónoma Metropolitana–Unidad Iztapalapa, Mexico City (CDMX), Mexico. [2]Facultad de Medicina, Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico.

ABSTRACT

BACKGROUND: Research into long-term cause-specific mortality of women diagnosed with breast cancer is important because it allows for the splitting of the population into patients who eventually die from breast cancer and from other causes. The adoption of this approach helps to identify patients with an elevated risk of eventual death from breast cancer.

OBJECTIVE: The primary aim of this study was to examine the associations between both sociodemographic and clinicopathologic characteristics and the underlying risks of death from breast cancer and from other causes for women diagnosed with breast cancer. A second aim was to propose a predictive biomarker of cause-specific mortality in terms of treatment and several important characteristics of a patient.

METHODS: A cohort of 16 511 female patients diagnosed with breast cancer in 1990 was obtained from the Surveillance, Epidemiology, and End Results cancer registries and followed for 20 years. A mixture model for the regression analysis of competing risks was used to identify factors and confounders that affected either the eventual cause-specific mortality or conditional cause-specific hazard rates, or both. Missing data were handled with multiple imputation.

RESULTS: Curvilinear relationships of age at diagnosis along with race, marital status, breast cancer type, tumor size, estrogen receptor status, extension, lymph node status, type of surgery, and radiotherapy status were significant risk factors for the cause-specific mortality, with extension and lymph node status appearing to be confounded with the effects of both type of surgery and radiotherapy status. The score obtained from combining a set of predictors showed to be an accurate predictive biomarker.

CONCLUSIONS: In cause-specific mortality of women diagnosed breast cancer, prognosis appears to depend on both sociodemographic and clinicopathologic factors. The predictive biomarker proposed in this study may help identifying the level of seriousness of the disease earlier than traditional methods, potentially guiding future allocation of resources for better patient care and management strategies.

KEYWORDS: Breast carcinoma, causes of death, hazards models, logistic model, risk assessment, risk factors

RECEIVED: February 8, 2017. **ACCEPTED:** April 21, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 355 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by stipend grants from Conacyt, Universidad Autónoma Metropolitana-Iztapalapa, and Universidad Autónoma de San Luis Potosí, Mexico.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Gabriel Escarela, Departamento de Matemáticas, Universidad Autónoma Metropolitana–Unidad Iztapalapa, AT-351, Av San Rafael Atlixco No.186, Col Vicentina, C.P. 09340, Mexico City (CDMX), México. Email: ge@xanum.uam.mx

Introduction

Breast cancer has been the most common incident form of female cancer worldwide and exerts a considerable economic burden.[1,2] Therefore, both postoperative follow-up and eventual cure of breast cancer represent important epidemiologic, social, and public health issues. Concurrent with the successful survivorship efforts in terms of female breast cancer diagnosis and treatment, many patients with breast cancer are being overtreated for the disease given the lack of sufficiently accurate prognostic and predictive information.[3] Improving triage decisions for individual patients requires precise measures of the time to death from breast cancer versus when the patient would die from a competing cause. However, most studies have focused on either the overall risk of death or breast cancer–specific survival with relatively short-term follow-ups after treatment, leaving uncertainty about the long-term risks of death from breast cancer in the presence of other acting causes.

In this article, a carefully formulated mixture model for competing risks is implemented for the regression analysis of long-term cause-specific mortality with a set of sociodemographic and clinicopathologic concomitant variables within a large-scale population-based cohort of women diagnosed with breast cancer that underwent surgery. In the 2 death outcome settings outlined here, there can be 3 coefficients for each predictor, 1 describing how the predictor affects long-term cause-specific mortality and 2 describing how it affects the conditional latency for each cause of death, which allowed for various useful epidemiologic interpretations for each risk factor. The risk score in the long-term cause-specific mortality component of the most parsimonious mixture model was employed to define a predictive biomarker which helps to both quantify the severity of the disease and discriminate patients with different risk levels of ultimate death from breast cancer, providing a potential decision-support tool for triage based on important characteristics of the patient's health and demographic status.

Materials and Methods

Data

The cohort used in this study was obtained from the US National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program, which is a high-quality epidemiologic surveillance system consisting of population-based cancer incidence and survival data covering up to 26% of the US population. Since records began in 1973, the registries of all newly diagnosed cases of breast cancer found within the defined geographic regions of the SEER program routinely collect information on sociodemographics, clinicopathological characteristics of the tumor, site-specific surgery, postoperative radiation status, date of cancer diagnosis, date of death, and cause of death. The cases of female breast cancer considered in this study were registered in 1990 in the SEER Registries located in Alaska, California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington. Excluded were cases diagnosed by death certificate or autopsy and with surgery of other regional sites, distant sites, or distant lymph nodes.

The competing risks considered in this study were death from breast cancer and death from other causes, whereas the predictors used in the regression model were age at diagnosis, race (white, black, and other), marital status (unmarried: never married, separated, divorced, or widowed and married: married and common-law married), location health status (whether the state the patient was diagnosed in had an America's Health Ranking value under 0.5 in 1990 or not),[4] histologic type (duct, lobular, and other), size of tumor (less than 2 cm and 2 cm and greater), grade (I and II: well or moderately differentiated tumor cells and III and IV: poorly differentiated, undifferentiated, or anaplastic tumor cells), estrogen receptor status (positive or borderline and negative), laterality (right and left), extension (localized: in situ or without underlying tumor or no evidence of it, confined to breast tissue and fat and further extension: invasive components, extensive skin involvement, inflammatory carcinoma, further extension, or metastasis), lymph node status (no node involvement and node positive), surgery (breast preserving: codes 10-38 and mastectomy: codes 40-78), and postoperative radiotherapy status (no radiation and radiation).

Statistical methods

Because a considerable number of patients were still alive at the end of the 20-year follow-up, the use of the conventional logistic regression to estimate the proportion of patients who eventually die from either cause was prohibitive. To estimate such proportions along with the corresponding conditional hazard rates, a mixture model for competing risks was used to simultaneously estimate the underlying cumulative incidence functions (CIFs), which are defined as follows:

$$F_j\left(t\right) = \Pr\left\{\text{death from cause } j \text{ within time } t\right\}$$

where $j = 1$ and $j = 2$ denote breast cancer and other causes, respectively. In the present setting, the CIF for cause j is given by the product of the probability of eventually dying from cause j and the conditional cumulative distribution function of cause j.[5] The former, referred here as *the logistic component*, was specified with a logistic regression model as follows:

$$\Pr\left\{\text{eventual death from breast cancer}; \mathbf{x}\right\} = \frac{\exp\left(\delta + \boldsymbol{\delta}'\mathbf{x}\right)}{\left[1 + \exp\left(\delta + \boldsymbol{\delta}'\mathbf{x}\right)\right]}$$

where δ and **δ** are, respectively, the parameters corresponding to the intercept and the vector of coefficients, and **x** is a vector of predictor variables, whereas the latter, referred here as *the conditional component* for risk *j*, was specified with a parametric Cox proportional hazards model with a Weibull baseline survivor function as follows:

$$1 - \Pr\left\{\begin{array}{l}\text{death after time } t \text{ given that} \\ \text{the patient will eventually} \\ \text{ die from cause } j; \mathbf{z}\end{array}\right\} = 1 - S_{0j}\left(t\right)^{\exp(\boldsymbol{\beta}_j'\mathbf{z})}$$

where $S_{0j}(t)$ is the Weibull survival distribution for risk j, $\boldsymbol{\beta}_j$ is the corresponding vector of coefficients, and **z** is a vector of predictors; here, the Weibull model for each risk was parameterized as in the function pweibull of the **R** language,[6] and **x** may contain some or all of the variables in **z**, as well as other variables not included in **z**. The choice of a Weibull model over other competing survival distributions was based chiefly on its flexibility.[7] The type of estimators used in this study was obtained with maximum likelihood,[5] the corresponding log-likelihood function was maximized using the function nlm of the **R** language,[6] and the resulting hessian was employed to estimate the asymptotic covariance matrix.

To adjust for the fluctuating changes of age at diagnosis as a continuous regressor in both the incidence and the conditional cause-specific mortality, an orthogonal polynomial of second degree of age at diagnosis was included in each component in the mixture model. Also, a propensity score was included in the mixture model to adjust for differences in preoperative patient background; here, the propensity score was estimated using a parsimonious multivariate logistic regression model with indication for treatment as the outcome variable and using as many significant regressors as possible. The variable coding scheme for the factors used in this study was treatment contrasts,[8] which creates a dummy variable for each nonbaseline level of a factor and sets the coefficient of the baseline level in each categorical variable equal to 0; here, the first category of each factor described above was set as the baseline level.

It was assumed that the missing data mechanism was missing at random, which specifies that the probability that a data value is missing depends on values of variables that were actually measured. The method employed here for dealing with the missing data was multiple imputation, and in the model to impute each regressor included as many significant auxiliary variables as possible, which tends to minimize bias and make the missing at random mechanism more plausible.[9] The inference procedure consisted of the generation of multiple stochastically enhanced data sets using the *mice* package in the **R** language,[10] then each completed data set was analyzed using the model described below for complete data, and finally, the results were combined using Rubin rules.[11] Grade had a large proportion of missing values and a χ^2 test showed that it was highly correlated with ER, which is consistent with other cross-sectional findings.[12,13] To improve the quality of the enhanced data sets and avoid multicollinearity, grade was used in the multiple imputation process and then removed from the data analysis.

The main model choice criterion for the identification of the appropriate covariates to be included in both the logistic and Cox models was based on the Bayesian information criterion (BIC). The value of the BIC for a specific mixture model is given by $\text{BIC} = -2 \times l(\boldsymbol{\theta}) + n_p \times \log(n)$, where l denotes the log-likelihood function, $\boldsymbol{\theta}$ is the vector with the maximum likelihood estimators, n_p is the number of parameters in the model, and n is the size of the cohort. The criterion consists on choosing the model for which BIC is the smallest. Backward elimination was employed to arrive at the best fitting model in each imputed data set. The problem of variable selection was addressed with the "impute, then select" strategy, which involves initially performing multiple imputation and subsequently applying Bayesian variable selection to each of the enhanced data sets.[14] The variables included in the final model appeared in at least 50% of the selected models obtained in the imputed data sets.[15] Both the variable selection process and the combined results were based on 100 enhanced data sets.

Patients with worse pathologic features at diagnosis were expected to be more likely to undergo mastectomy or postoperative radiotherapy, or both, and thus, any treatment comparisons is confounded by differences in severity of breast cancer between patients and, therefore, highly susceptible to bias. The methodology used for handling the confounding in this nonrandomized observational study consisted on interpreting the effects of surgery, radiation status, and variables whose 2-way interactions with surgery and radiation turn out to be significant as causal effects and the remaining effects in the regression analysis as risk factors effects.

Results

Patient characteristics

The cohort studied here consists of n = 16 511 female patients who were diagnosed with unilateral primary breast cancer during 1990 and underwent surgery. The end of follow-up was December 31, 2011. Among the 16 511 women, 5670 had breast-preserving surgery (BPS) and 10 841 underwent mastectomy. Patient characteristics according to surgical treatment with or without postoperative radiotherapy are summarized in Table 1.

Although the proportions of patients who died from breast cancer with and without radiotherapy after BPS are similar, patients with radiation after BPS had a better survival rate at the end of the follow-up; however, the proportion of women who died from breast cancer after mastectomy and postoperative radiation exceeded more than 2 times the proportion of women who only had mastectomy and had the lowest survival rate at the end of the study. Women without radiation after surgery were more often 59 years old or above, whereas women with postoperative radiation were less often in the 71+ years age group, which suggests that radiotherapy was administered more discriminatively among elderly patients. Patients were mainly of white ethnicity and most of them were married. Roughly, around half of the patients were diagnosed in each category of the location health status. Ductal carcinoma was the most common type of breast cancer. The proportions of tumor sizes 2 cm or larger were similar between patients with or without radiation after BPS but, among patients who had mastectomy, the proportion was larger for those that underwent postoperative radiation. The tumors of patients with postoperative radiotherapy appeared more likely to be of higher histologic grade and estrogen positive or borderline. There is a negligible excess of left-sided disease in women. Patients were more frequently diagnosed with a confined cancer. Most of patients who underwent BPS had no axillary lymph node involvement, whereas positive node involvements were particularly frequent among patients who used radiation after mastectomy; in addition, BPS patients had relatively frequent missing characteristics of size of tumor, grade, ER, and lymph node involvement, particularly among patients without postoperative radiotherapy.

To explore possible differences between the subtypes of breast cancer according to age at diagnosis, Figure 1 displays age-specific incidence rates for duct, tubular, and other subtypes of breast cancer. Although the plots show similar shapes, they exhibit bimodal densities with inflection points around menopause, suggesting the presence of pathologies with 2 different age-dependent etiologies.

Regression results

In the best fitting mixture model, the interaction between type of surgery and radiotherapy status was significant in both

Table 1. Patient characteristics according to surgical treatment and postoperative radiotherapy status of women diagnosed with breast cancer in the SEER Registries in 1990.

	BREAST PRESERVING		MASTECTOMY	
	(n = 5670, 34.3%)		(n = 10 841, 65.7%)	
	NO RADIATION, NO. (%)	RADIATION, NO. (%)	NO RADIATION, NO. (%)	RADIATION, NO. (%)
Characteristic	2415 (42.6)	3255 (57.4)	9981 (92.1)	860 (7.9)
Vital status				
Alive	816 (33.8)	1490 (45.8)	3513 (35.2)	200 (23.3)
Dead from breast cancer	430 (17.8)	536 (16.5)	2201 (22.0)	455 (52.9)
Dead from other causes	1169 (48.4)	1229 (37.7)	4267 (42.8)	205 (23.8)
Sociodemographic characteristics				
Age at diagnosis				
<48	490 (20.3)	766 (23.5)	1901 (19.1)	254 (29.5)
48 to 59	398 (16.5)	767 (23.6)	1966 (19.7)	200 (23.3)
59 to 71	556 (23.0)	1023 (31.4)	2996 (30.0)	260 (30.2)
71+	971 (40.2)	699 (21.5)	3118 (31.2)	146 (17.0)
Race				
White	2113 (87.5)	2871 (88.2)	8735 (87.5)	711 (82.7)
Black	192 (8.0)	230 (7.1)	694 (7.0)	87 (10.1)
Other	110 (4.5)	154 (4.7)	552 (5.5)	62 (7.2)
Marital status				
Unmarried	1100 (45.5)	1126 (34.6)	4126 (41.3)	339 (39.4)
Married	1197 (49.6)	2074 (63.7)	5641 (56.5)	501 (58.3)
Missing	118 (4.9)	55 (1.7)	214 (2.2)	20 (2.3)
Location health status				
Ranking Value ⩽ 0.5	1124 (46.5)	1576 (48.4)	4725 (47.3)	435 (50.6)
Ranking Value > 0.5	1291 (53.5)	1679 (51.6)	5256 (52.7)	425 (49.4)
Tumor characteristics				
Site				
Duct	1802 (74.6)	2708 (83.2)	8099 (81.1)	662 (77.0)
Lobular	298 (12.3)	175 (5.4)	779 (7.8)	73 (8.5)
Other	315 (13.1)	372 (11.4)	1103 (11.1)	125 (14.5)
Size of tumor				
<2 cm	1350 (55.9)	2172 (66.7)	4583 (45.9)	146 (17.0)
⩾2 cm	647 (26.8)	814 (25.0)	4163 (41.7)	616 (71.6)
Missing	418 (17.3)	269 (8.3)	1235 (12.4)	98 (11.4)
Grade				
I and II	515 (21.3)	1008 (31.0)	2342 (23.5)	168 (19.5)
III and IV	337 (14.0)	698 (21.4)	2281 (22.8)	379 (44.1)
Missing	1563 (64.7)	1549 (47.6)	5358 (53.7)	313 (36.4)

Table 1. (Continued)

	BREAST PRESERVING		MASTECTOMY	
	(n = 5670, 34.3%)		(n = 10 841, 65.7%)	
	NO RADIATION, NO. (%)	RADIATION, NO. (%)	NO RADIATION, NO. (%)	RADIATION, NO. (%)
Estrogen receptor status				
Positive or borderline	879 (36.4)	1834 (56.4)	5218 (52.3)	506 (58.8)
Negative	258 (10.7)	515 (15.8)	1558 (15.6)	197 (22.9)
Missing	1278 (52.9)	906 (27.8)	3205 (32.1)	157 (18.3)
Laterality				
Right	1175 (48.7)	1611 (49.5)	4839 (48.5)	406 (47.2)
Left	1240 (51.3)	1644 (50.5)	5142 (51.5)	454 (52.8)
Extension				
Localized	2134 (88.4)	3078 (94.6)	9075 (90.9)	560 (65.1)
Further extension	220 (9.1)	151 (4.6)	809 (8.1)	277 (32.2)
Missing	61 (2.5)	26 (0.8)	97 (1.0)	23 (2.7)
Lymph node status				
No node involvement	1556 (64.4)	2523 (77.5)	6768 (67.8)	194 (22.6)
Lymph node +	258 (10.7)	532 (16.3)	2912 (29.2)	604 (70.2)
Missing	601 (24.9)	200 (6.2)	301 (3.0)	62 (7.2)

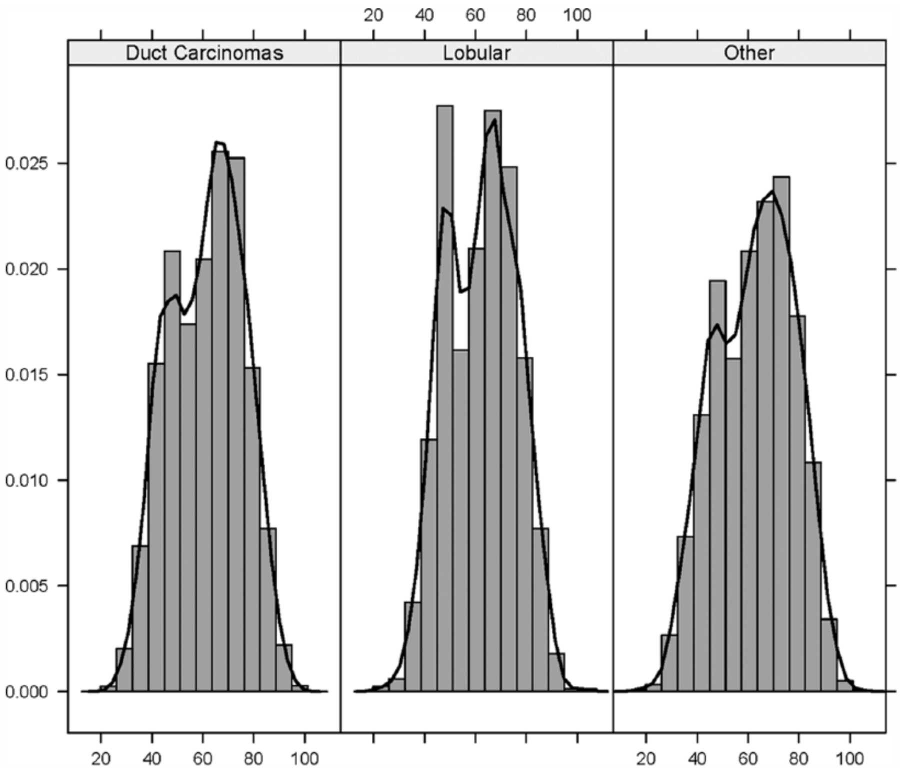

Figure 1. Histogram and density curve of age at diagnosis for duct, lobular, and other subtypes of carcinomas in women diagnosed with breast cancer in 1990.

components in the mixture model, extension had significant 2-way interactions with surgery and radiation in the logistic and conditional components, respectively, whereas the interaction between lymph node status and radiation status was significant in the logistic component only, which may suggest that although the choice of surgery and radiation were mainly based on the staging information, postoperative radiotherapy tended to be given when lymph nodes were present. Both extension and lymph node status appear to be confounded with the effects of type of surgery and radiotherapy status; therefore, type of surgery, radiotherapy status, extension, and lymph node status were regarded as causal effects.

Although the interaction effects between the polynomial of age and either surgery or radiation status had significant effects in the conditional survival component of the best fitting mixture model, such significances turned out to be artificial and negligible at best, which is an indication that the effects are in fact spurious and thus were removed from the best fitting model. Neither the 2-way interactions between the risk factors nor the main effects of the propensity score, location health status, and laterality turned out to be statistically significant in any of the components in the mixture model and were removed from the model as well. The resulting most parsimonious model can then lead to various useful epidemiologic interpretations for each coefficient when the remaining factors and interactions are constant, which enables to better understand the way in which the corresponding risk factor is associated with both the cause-specific proportion and the conditional hazard rates of mortality.

Figure 2 shows fitted and 95% confidence bands curves of the effects of the second-degree polynomial of age at diagnosis in the 3 components of the mixture model. The display at the top shows the effect of age as changes of log odds in the logistic component, *f*(age), which corresponds to the eventual death from breast cancer. It can be noticed that the change in log odds exhibits an ample width of the confidence bands for patients diagnosed before perimenopause that gradually narrows down with age within this range; here, a constant function, that is, a straight line parallel to the *x*-axis, can straightforwardly be superimposed in between the confidence bands, revealing that this group of younger patients share similar risks of eventual death from breast cancer. The fitted change of log odds decreases with age for patients diagnosed after menopause, and the corresponding confidence bands remain relatively narrow until the early 80s and then expand for older ages, always showing a decreasing path and thus indicating that the probability of eventual death from breast cancer decreases as age increases among this group of older patients. The impact of age on the eventual death from breast cancer can be illustrated by comparing the odds ratio of a patient aged 45 years with respect to a patient aged 75 years, which is exp[*f*(45)]/exp[*f*(75)] = 1.50; that is, the odds of a patient in her mid-40s eventually dying from breast cancer is 50% higher than that of a patient in her mid-70s.

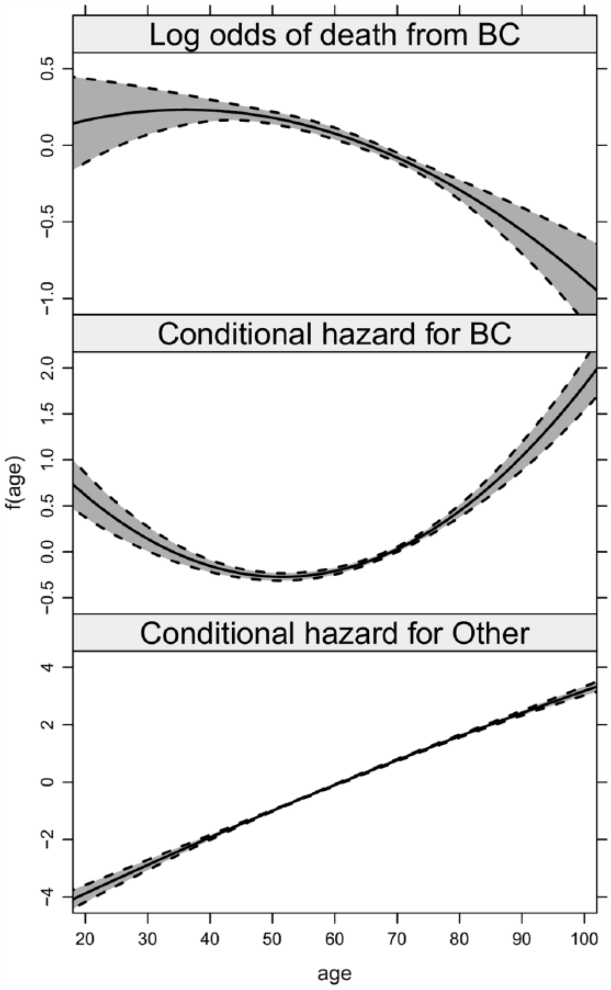

Figure 2. Fitted and 95% confidence bands of the effects of the orthogonal polynomial of age at diagnosis in the logistic component (top), the conditional component of death from breast cancer (BC) (middle), and the conditional component of death from other causes (bottom) for the most parsimonious mixture regression model.

The display in the middle of Figure 2 depicts the effects of age on the conditional hazard rate corresponding to breast cancer. The curve follows a U-shaped pattern reaching the minimum around menopause, which indicates that given that a patient will eventually die from breast cancer, the conditional hazard rate of breast cancer decreases as age increases for patients diagnosed before perimenopause and then it increases for patients diagnosed after menopause. The display at the bottom of Figure 2 shows the effects of age on the conditional hazard rate corresponding to other causes. Much of the curve resembles a straight line, indicating that given that a patient will eventually die from other causes, the corresponding conditional hazard rate increases with age, as expected, which is emphasized by the width of the confidence bands.

Table 2 shows parameter estimates for the most parsimonious mixture model. For the categorical variables, the coefficients in the first column can be interpreted as the change in log odds of eventual death from breast cancer in comparison with the baseline group when the remaining regressors are

Table 2. Coefficients and SEs for the most parsimonious mixture model.

PREDICTORS	LOGISTIC COMPONENT (δ)		BREAST CANCER CONDITIONAL COMPONENT (β_1)		OTHER CAUSES OF DEATH CONDITIONAL COMPONENT (β_2)	
	COEFFICIENT	SE	COEFFICIENT	SE	COEFFICIENT	SE
Intercept	−2.017***	0.071	—	—	—	—
Orthogonal polynomial of age						
First degree	−24.85***	2.780	30.24***	2.824	163.3***	2.846
Second degree	−8.035**	2.711	26.37***	2.486	−4.670*	2.133
Race (white as baseline)						
Black	0.287***	0.073	0.228***	0.065	0.279***	0.050
Other	−0.343***	0.100	0.010	0.101	−0.182**	0.064
Married	—	—	−0.070	0.045	−0.176***	0.026
Histologic type (ductal as baseline)						
Lobular	—	—	−0.229**	0.077	−0.012	0.046
Other	—	—	0.136*	0.065	0.026	0.037
Tumor size ⩾ 2 cm	0.598***	0.049	0.261***	0.052	0.170***	0.028
ER-negative	—	—	0.604***	0.053	−0.009	0.038
Further extension	1.295***	0.124	1.170***	0.111	0.415***	0.111
Lymph node +	0.926***	0.094	0.308***	0.051	0.277***	0.033
Mastectomy	−0.034	0.077	−0.002	0.083	−0.141***	0.035
Radiation	−0.017	0.083	−0.218*	0.095	−0.174***	0.043
Mastectomy:further extension	−0.672***	0.143	−0.602***	0.116	−0.263*	0.116
Radiation:further extension	—	—	−0.268*	0.107	0.054	0.125
Radiation:lymph node +	0.301**	0.107	—	—	—	—
Mastectomy:radiation	0.528***	0.116	0.415***	0.111	0.214*	0.089

Abbreviation: SE, standard error.
":" denotes interaction.
**P* value < .05; ***P* value < .01; ****P* value < .001.

fixed, whereas the coefficients in the last 2 columns can be interpreted as the conditional log hazard ratios for breast cancer (second column) and other causes (third column) in comparison with the baseline group when the remaining regressors are fixed. In the following description, 95% confidence intervals will be used when applicable. The results show that all coefficients associated with the orthogonal polynomials of age at diagnosis were individually significant. The coefficients corresponding to race show evident disparities in breast cancer mortality among the 3 racial groups considered in this study, particularly with respect to black women. The proportion of black women who eventually die of breast cancer is higher than that of white women, with 15% to 54% higher odds than white patients, whereas the proportion of deaths from breast cancer for other ethnicities is lower than that for white women, with 14% to 42% lower odds than white patients. When it comes to comparing conditional hazard rates, black women have worse survival rates than white women, with 10% to 43% and 20% to 46% higher conditional risks from breast cancer and other causes, respectively, than white patients. Women from other ethnicities who eventually die from breast cancer do not seem to have different risks in the corresponding conditional hazard rate from white patients, but their conditional risks for other causes are 5% to 26% lower than white women. Marital status was only significant in the conditional hazard rate of death from other causes and the corresponding coefficient is negative, which indicates that married patients have had longer survival only when the eventual cause of death is other than breast cancer.

Although the individual coefficients of histologic type did not appear to be significant in both the logistic component and the conditional mortality from other causes, the conditional breast cancer–specific mortality for lobular carcinomas and other subtypes was associated with, respectively, 8% to 32% lower and 1% to 30% greater conditional breast cancer mortality risks compared with ductal carcinomas. Tumor size had very significant effects in all 3 components of the mixture model. As one would expect, breast cancer–specific mortality increases for large tumor sizes, with tumors 2 cm or larger having 65% to 100% higher odds of eventual death from breast cancer and 17% to 44% higher conditional risks of breast cancer death than smaller sized tumors; also, tumors 2 cm or larger were associated with 12% to 25% higher conditional risks of other causes of death compared with smaller sized tumors. Estrogen receptor status only turned out to be significant within the conditional hazard rate of breast cancer, with ER-negative being associated with 65% to 103% greater conditional breast cancer mortality risks compared with ER-positive or borderline.

Model and biomarker performance

The prognostic ability of the model is an important feature which can be used as a benchmark of model performance for validation purposes. In the current cause-specific incidence context, a prediction model for an event of interest discriminates well if it is able to define risk groups to individuals according to the probability of eventually experiencing such event. To both define a predictive biomarker and assess the aforementioned discrimination, the risk score in the logistic component of the most parsimonious mixture model, defined as $R = d + \boldsymbol{d}'\mathbf{x}$, where d and $\boldsymbol{d}$ are, respectively, the maximum likelihood estimators of δ and $\boldsymbol{\delta}$, was taken as the biomarker of cause-specific mortality risks; here, multiple imputation was employed to obtain the scores when the predictors were partially observed. In this analysis, it was possible to rank the subjects in the data according to the risk score R and then select those with the 10% lowest of values of R, which were assigned to the *low-risk* group, those with the following 75% lowest of values of R, which were assigned to the *medium-risk* group, and the remaining 15% of values of R, which were assigned to the *high-risk* group.

To assess the predictive fit of the model, it is possible to estimate the parameters of the mixture model from a randomly chosen half of the data, the *training sample*, and then to validate the CIFs on the other half of the data, *the validation sample*; here, the risk groups are defined using the risk score R in the validation sample. Figure 3 displays empirical estimates of the CIFs in each risk group in the validation sample, which were based on Aalen nonparametric event estimates[16,17] and computed using the **R** language function cuminc of the *cmprsk* package,[18] along with the estimated model-based CIFs, which were calculated averaging over the linear combination between the vector of coefficients and the vector of predictors in each component of the most parsimonious mixture model for each risk group. The estimators of the 2 methods are comparable for each risk group, which suggests that both the fitted mixture model and the classification criterion are adequate.

To accurately assess the prediction performance of the model for the 3 risk groups, a calibration plot can be used.[19] In this plot, the nonparametric estimate of the CIF of cause of death j at a chosen time t for each risk group is plotted against the mean of the predicted CIF of the jth cause of death $[F_j(t; \mathbf{x}_1) + F_j(t; \mathbf{x}_2) + \ldots + F_j(t; \mathbf{x}_k)]/k$, where k is the number of subjects within the risk group, $\mathbf{x}_i$ is the vector of predictors for the ith individual in the group, $j = 1, 2$ and $F_j(t; \mathbf{x}_i)$ was obtained using multiple imputation when the corresponding predictors were partially observed. Figure 4 shows the calibration plot of the CIFs corresponding to follow-up times 2, 5, 10, 15, and 20 for the 3 risk groups. For death from breast cancer, the cumulative incidence estimates from the most parsimonious mixture model correspond rather closely to the empiric cumulative incidence estimates for each risk group, whereas for death from other causes such relationship is close for the medium-risk and high-risk groups and for times 2, 5, and 10 in the low-risk group.

In general, it can be concluded that the mixture model gives a good fit to the breast cancer data set and that the predictive biomarker-based discrimination method proposed here achieves a useful ability to both identify and separate women with different breast cancer–specific risks.

Discussion

In the present competing risks analysis of 16 511 patients, curvilinear relationships of age at diagnosis, race, marital status, histologic type, tumor size, ER status, extension, and lymph node status were significant risk factors for the cause-specific mortality, whereas extension and lymph node status appeared to be significant confounders for both type of surgery and radiotherapy status, which had significant effects as well. Location health status did not show significant effects in any of the components in the mixture model, which suggests that both overall lifestyles and living conditions in 1990 did not appear to change the risk for any cause-specific mortality. Also, there was no evidence that laterality is a risk factor for any cause-specific death, which is consistent with previous reports[13]; of particular interest is the fact that the interaction between laterality and radiotherapy status did not appear to be important, which is consistent with recent research in the sense that radiation therapy for left-sided breast cancer, as delivered in the 1990s, has not increased the risk of death from any cause when compared with right-sided tumors.[20,21]

The curvilinear effects of age showed that although patients diagnosed before menopause appeared to have similar odds of eventual death from breast cancer with respect to age at diagnosis, the proportion of women dying from the illness tended to decrease with age at diagnosis after menopause; in addition, the age at diagnosis effect in the conditional hazard rate of breast cancer–specific death showed a parabolic shape with its minimum around menopause. Thus, young patients diagnosed before menopause had the poorest breast cancer prognosis, which is consistent with previous studies[22–25]; however, despite

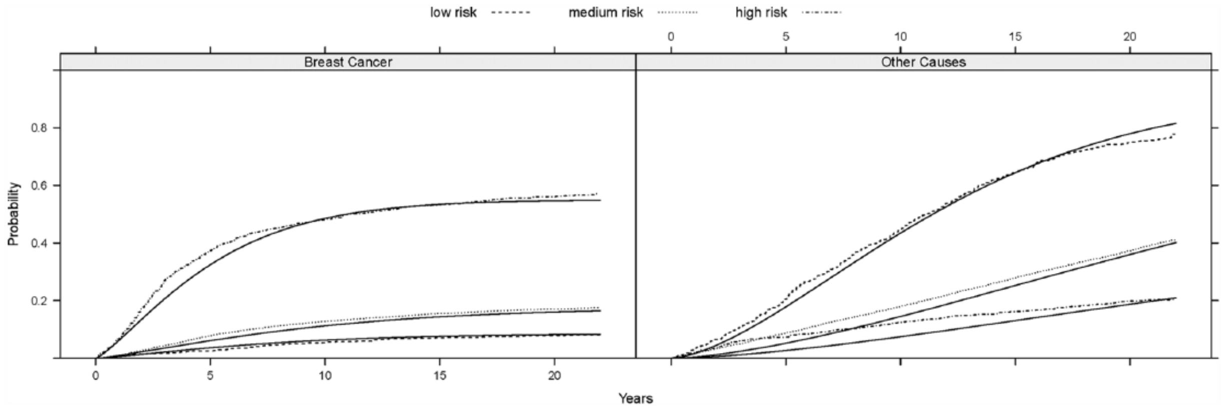

Figure 3. Observed (dotted lines) and fitted (solid lines) cumulative incidence functions of death from breast cancer and other causes for 3 risk groups in the validation sample.

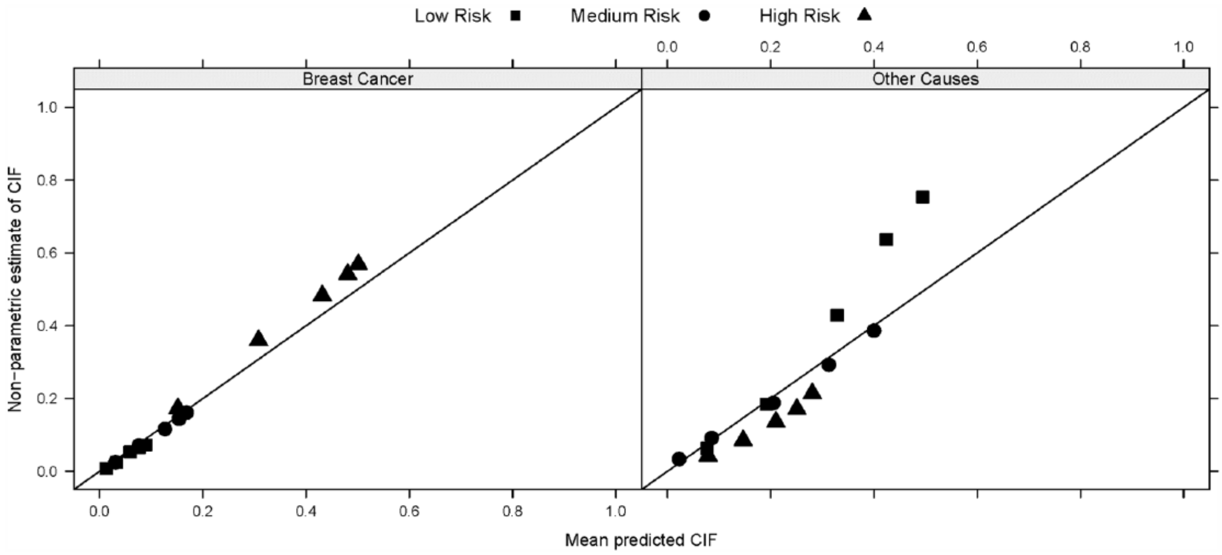

Figure 4. Calibration plot for the low-risk, medium-risk, and high-risk groups in the validation sample at follow-up times 2, 5, 10, 15, and 20.

the fact that older patients are less susceptible to die from their cancer, when they die from it, they do so at hazard rates as high as their younger counterparts or higher for patients diagnosed after 85 years.

Because age-at-diagnosis did not significantly interact with other risk factors and recent research has found evidence of low breast cancer risk around menopause,[23,26–28] the difference in outlook between young and old patients can lead to the speculation that menopause is an important threshold between 2 age-related etiologic mechanisms, suggesting the need for a closer look at age at diagnosis taking due account of menopausal status. As a matter of fact, assessing the risk effects of age at diagnosis as yet has to be properly addressed given the conflicting findings regarding age-related breast cancer mortality, particularly for older patients.[26–32] Such discrepancies may have arisen not only because of differences in methodologies but also because age cutoff points have been defined arbitrarily.[33]

This study confirms previous epidemiologic findings regarding race, particularly when it comes to comparing black and white women.[34–36] Black women had the worst overall prognosis because they were the ethnicity most susceptible to die from breast cancer and had the highest conditional hazard rates in both causes of death. Although marital status turned out to be an important predictor in the mixture model, it failed to show any significant association with breast cancer–specific mortality. It can then be claimed that although marriage and the social support it brings can have a positive impact on the survival when the cause of death is different from breast cancer, it does not seem to influence the breast cancer outcome, challenging the findings of previous cancer-specific survival studies that claim that unmarried patients are at significantly higher risk of death resulting from their cancer.[37]

The odds of eventual death from breast cancer did not seem to differ both among the 3 different cancer subtypes and among the 2 ER statuses; however, among patients who ultimately die from breast cancer, those with a lobular carcinoma and those with an ER-positive or borderline status had considerable better survival, which is somehow consistent with previous research.[38,39] Size of tumor is clearly an important predictor, with tumors 2 cm or larger not only being associated with worse breast cancer outcomes than smaller tumors, as expected, but also with higher hazard rates when the cause of death is different from cancer, which could reflect the endogenous long-term side effects of aggressive adjuvant treatments given to patients with large tumors in the 1990s.[40]

The biomarker of breast cancer–specific mortality defined here depends on age, race, tumor size, extension, and lymph node status, which are readily available at diagnosis and on surgery type and radiotherapy status. Although based on simple measures, the biomarker offers an accurate tool to determine a patient's prognosis when she is diagnosed with breast cancer and therapy scenarios are considered. Adding the biomarker to clinical criteria has the potential to improve risk-based triage and can have a great impact on the individualization of the care and management strategies, thus optimizing throughput and resources.

The strengths of the data set are the comparatively large number of subjects involved and the reasonably long follow-up period. These strengths enabled the analysis to investigate the data in more detail and with rather more confidence than studies with a more modest size and a shorter follow-up period would allow; however, although cohort studies offer several important advantages over other forms of observational studies, they tend to be susceptible to both selection and exposure bias.[41]

With long-term follow-up studies, there are other important tensions to confront. Although the increasing length of a follow-up produces confidence to identify the proportion of subjects who will eventually die from the disease, a longer follow-up also has the outcome of the cohort originally being recruited at a time much more remote from contemporary conditions. Accordingly, it can be questioned whether an analysis of patients diagnosed with breast cancer in 1990 provides useful information relevant to those diagnosed in the present year. Over the past 3 decades, there has been a considerable development of new technologies and improved surgical techniques against breast cancer. In addition, given that there have been changes in lifestyles that have had influence in delaying menopause in the United States,[42] it seems reasonable to speculate that the effects of age at diagnosis on either cause of death have changed as well.

Although there have been some adjustments in the existing regimens and changes in women's lifestyles, nevertheless, it is somehow difficult to believe that anything has happened to challenge the relationships among clinicopathologic and sociodemographic variables that have been probed in this study. However, in the final analysis, that question can be answered only by conducting similar research with later cohorts and in other countries to chart the boundaries of the results identified here.

Author Contributions

GE and AJ-B contributed to the conception of the study. GE, AJ-B, and GN-A analyzed the data. GE, AJ-B, GN-A, and AG-M wrote the first draft of the manuscript; contributed to the writing of the manuscript; agreed with manuscript results and conclusions; and made critical revisions and approved final version. GE, GN-A, and AG-M jointly developed the structure and arguments for the paper. All authors reviewed and approved the final manuscript.

Disclosures and Ethics

Access to the SEER database to retrieve the data for this study was granted by the SEER program under the reference number 12899-Nov2011 after a Data-Use Agreement for the SEER 1973-2011 Research Data File was signed. Because this study is a database-dependent analysis rather than experimental research on humans, and cancer is a reportable disease in the participating SEER Registries, approval of ethics committees and consent from participating patients are not needed. Patient records/information were anonymized and de-identified prior to analysis.

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010;127:2893–2917.
2. Boyle P, Levin B. *World Cancer Report 2008*. Lyon, France: IARC Press; 2008.
3. Alvarado M, Ozanne E, Esserman L. Overdiagnosis and overtreatment of breast cancer. *Am Soc Clin Oncol Educ Book*. 2012:e40–e45.
4. Peppard PE, Kindig DA, Dranger E, Jovaag A, Remington PL. Ranking community health status to stimulate discussion of local public health issues: the Wisconsin County Health Rankings. *Am J Public Health*. 2008;98:209–212.
5. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. *Appl Statist*. 1985;34:201–211.
6. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/. Published 2014.
7. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Control Clin Trials*. 2003;24:682–701.
8. Chambers JM, Hastie TJ. Statistical models. In: Chambers JM, Hastie TJ, eds. *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole; 1992:13–44.
9. Schafer JL. *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall; 1997.
10. VanBuuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1–67.
11. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons; 1987.
12. Fisher ER, Osborne CK, McGuire WL, et al. Correlation of primary breast cancer histopathology and estrogen receptor content. *Breast Cancer Res Treat*. 1981;1:37–41.
13. Kamby C, Andersen J, Ejlertsen B, et al. Pattern of spread and progression in relation to the characteristics of the primary tumour in human breast cancer. *Acta Oncol*. 1991;30:301–308.
14. Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. *Biometrics*. 2005;61:498–506.
15. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med*. 2008;27:3227–3246.
16. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann Stat*. 1978;6:534–545.
17. Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat*. 1978;6:701–726.
18. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26:2389–2430.
19. Kattan MW, Heller G, Brennan MF. A competing risks nomogram for sarcoma-specific death following local recurrence. *Stat Med*. 2003;22:3515–3525.
20. Rutter CE, Chagpar AB, Evans SB. Breast cancer laterality does not influence survival in a large modern cohort: implications for radiation-related cardiac mortality. *Int J Radiat Oncol Biol Phys*. 2014;90:329–334.
21. Bao J, Yu KD, Jiang YZ, Shao ZM, Di GH. The effect of laterality and primary tumor site on cancer-specific mortality in breast cancer: a SEER population-based study. *PLoS ONE*. 2014;9:e94815.
22. Anders CK, Hsu DS, Broadwater G, et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J Clin Oncol*. 2008;26:3324–3330.

23. Fei C, DeRoo LA, Sandler DP, Weinberg CR. Menopausal symptoms and the risk of young-onset breast cancer. *Eur J Cancer*. 2013;49:798–804.
24. Partridge AH, Goldhirsch A, Gelber S, et al. Breast cancer in younger women. In: Harris JR, Lippman ME, Morrow M, Osborne CK, eds. *Diseases of the Breast*. 5th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2014:1101–1111.
25. Thangjam S, Laishram RS, Debnath K. Breast carcinoma in young females below the age of 40 years: a histopathological perspective. *South Asian J Cancer*. 2014;3:97–100.
26. Yankaskas BC. Epidemiology of breast cancer in young women. *Breast Dis*. 2005–2006;23:3–8.
27. Huang Y, Malone KE, Cushing-Haugen KL, Daling JR, Li CI. Relationship between menopausal symptoms and risk of postmenopausal breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2011;20:379–388.
28. Assi HA, Khoury KE, Dbouk H, Khalil LE, Mouhieddine TH, El Saghir NS. Epidemiology and prognosis of breast cancer in young women. *J Thorac Dis*. 2013;5:S2–S8.
29. Gnerlich JL, Deshpande AD, Jeffe DB, Sweet A, White N, Margenthaler JA. Elevated breast cancer mortality in women younger than age 40 years compared with older women is attributed to poorer survival in early-stage disease. *J Am Coll Surg*. 2009;208:341–347.
30. Barchielli A, Balzi D. Age at diagnosis, extent of disease and breast cancer survival: a population-based study in Florence, Italy. *Tumori*. 1999;86:119–123.
31. Yancik R, Wesley MN, Ries LA, et al. Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. *JAMA*. 2001;285:885–892.
32. Brandt J, Garne JP, Tengrup I, et al. Age at diagnosis in relation to survival following breast cancer: a cohort study. *World J Surg Oncol*. 2015;13:33.
33. Beadle BM, Woodward WA, Buchholz TA. The impact of age on outcome in early-stage breast cancer. *Semin Radiat Oncol*. 2011;21:26–34.
34. Akinyemiju T, Moore JX, Ojesina AI, Waterbor JW, Altekruse SF. Racial disparities in individual breast cancer outcomes by hormone-receptor subtype, area-level socio-economic status and healthcare resources. *Breast Cancer Res Treat*. 2016;157:575–586.
35. Ademuyiwa FO, Gao F, Hao L, et al. US breast cancer mortality trends in young women according to race. *Cancer*. 2015;121:1469–1476.
36. Schinkel JK, Zahm SH, Jatoi I, et al. Racial/ethnic differences in breast cancer survival by inflammatory status and hormonal receptor status: an analysis of the Surveillance, Epidemiology, and End Results data. *Cancer Causes Control*. 2014;25:959–968.
37. Aizer AA, Chen MH, McCarthy EP, et al. Marital status and survival in patients with cancer. *J Clin Oncol*. 2013;31:3869–3876.
38. Wasif N, Maggard MA, Ko CY, Giuliano AE. Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. *Ann Surg Oncol*. 2010;17:1862–1869.
39. Munoz D, Near AM, van Ravesteyn NT, et al. Effects of screening and systemic adjuvant therapy on ER-specific US breast cancer mortality. *J Natl Cancer Inst*. 2014;106:dju289.
40. Fornier MN, Modi S, Panageas KS, Norton L, Hudis C. Incidence of chemotherapy-induced, long-term amenorrhea in patients with breast carcinoma age 40 years and younger after adjuvant anthracycline and taxane. *Cancer*. 2005;104:1575–1579.
41. Thadhani R, Tonelli M. Cohort studies: marching forward. *Clin J Am Soc Nephrol*. 2006;1:1117–1123.
42. Nichols HB, Trentham-Dietz A, Hampton JM, et al. From menarche to menopause: trends among US women born from 1912 to 1969. *Am J Epidemiol*. 2006;164:1003–1011.