Advances in Continuous
and Discrete Models

RESEARCH Open Access

Bivariate Bayesian joint modeling of *CEA* and *E2* values for female breast cancer patients in Xinjiang

Tao Ma[1], Chunjie Gao[1], Yipala Yilihamu[1], Jing Liu[1], Ting Zhao[2] and Lei Wang[3,4*]

*Correspondence: wlei81@126.com
[3]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China
[4]Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi 830017, China
Full list of author information is available at the end of the article

Abstract

We investigate the comprehensive impact of dynamic changes in *CEA* (carcinoembryonic antigen) and *E2* (estradiol) values on the survival prognosis of breast cancer patients in Xinjiang, as well as predict their long-term mortality probabilities. This work is based on the longitudinal and survival data of female breast cancer patients followed up by the Affiliated Tumor Hospital of Xinjiang Medical University. Firstly, the Boruta algorithm was used to screen the independent prognostic factors that related with the breast cancer patients in Xinjiang. Moreover, a bivariate Bayesian joint model for longitudinal and time-to-event data was constructed to investigate how the dynamical changes of *CEA* and *E2* values collectively affect the survival prognosis of breast cancer patients in Xinjiang. The predictive performance of the model was assessed by using ROC curves and calibration curves. As a result, the variable screen results of the Boruta algorithm indicated that *CEA*, *E2*, clinical stage, received neoadjuvant treatment, etc., were identified as independent prognostic factors of breast cancer patients in Xinjiang. In addition, it was shown that the association coefficients of the joint model α_1 and α_2 were statistically significant. When all other baseline variables were unchanged, patients' death risk separately increases by approximately 1.577 times ($HR = 2.577$, 95%CI: (1.803, 3.563)) and 0.887 times ($HR = 1.887$, 95%CI: (1.472, 2.454)) with the one-unit collective increase in the log (*CEA*) and log (*E2*). Moreover, the joint model had good discrimination and calibration with an AUC value of 0.778. So the collective increase of *CEA* and *E2* would be associated with the breast cancer patients' poor survival prognosis. It would be essential to monitor the dynamical changes of *CEA* and *E2* values of breast cancer patients in clinical practice in order to provide more accurate individualized treatment for breast cancer patients.

Keywords: Breast cancer; Carcinoembryonic antigen; Estradiol; Longitudinal data; Survival prognosis; Bivariate Bayesian joint model

Breast cancer is a malignant tumor, dependent on sex hormone receptors formed by the abnormal proliferation of mammary epithelial cells. According to Global Cancer Statistics report in 2022, approximately 2.3 million women were diagnosed with breast cancer, with 66,000 deaths [1]. In 2022, the number of new female breast cancer cases in China was about 350,000, accounting for 16.51% of new female cancer cases [2]. In recent years, the incidence of female breast cancer in Xinjiang has been significantly increased, which is higher than the national average [3].

The occurrence and development of breast cancer were closely related to the continuous increase of estradiol (*E2*) values [4]. *E2* is a kind of steroid hormone that could penetrate the cell membrane and interact with estrogen receptors, leading estrogen receptors to dimerize and promote the proliferation of breast cancer cells [5]. On the other hand, carcinoembryonic antigen (*CEA*) plays an important role in tumor cell proliferation and differentiation, and its overexpression could lead to the progression of epithelial cancers [6]. When breast cancer tumor cells lose their polarity, *CEA* may be shed from the plasma membrane to form vesicles and enter the blood circulation. Moreover, more *CEA* will accumulate in the blood as the tumor size increases [7].

At present, patients' biomarker values are usually measured longitudinally to analyze the impact of dynamic changes of biomarkers on patients' disease progression and in order to observe the time trend of disease progression in clinical practice [8]. Survival data are common in clinical research, which could estimate patients' survival probability within the follow-up times and identify the independent prognostic factors that related with the breast cancer patients in Xinjiang [9]. However, analyzing longitudinal and survival data separately and ignoring their potential association will lead to bias [10]. A joint model for longitudinal and time-to-event data was widely used in the research of survival prognosis of breast cancer patients, which could overcome the problem that a single model could not adapt to diverse data types, avoid biases caused by separately modeling, accurately describe the potential association between longitudinal and survival processes, and enhance the predictive accuracy and interpret ability of the model [11–14]. However, it is clear that *CEA* and *E2* measures different aspects of the disease progression of breast cancer patients, and they are biologically interrelated. In addition, there is poor significance to investigate the breast cancer patients' survival prognosis when *E2* is used as a single biomarker [15]. Moreover, *CEA* as a biomarker does not have the high specificity and sensitivity required for breast cancer diagnosis. In clinical practice, better sensitivity and specificity could be achieved by combining *CEA* with other biomarkers [6]. Current research shows that *CEA* combined with *E2* can independently predict the survival prognosis of breast cancer patients with a high diagnostic value [16]. Bivariate Bayesian joint model could overcome the problem that a univariate joint model cannot be applied to multiple types of longitudinal data and could comprehensively analyze the association between dynamic changes of multiple longitudinal data and the breast cancer patients' survival prognosis, as well as to describe the correlation within multiple longitudinal data, which improves predictions compared to the separate analysis per marker [17].

On the other hand, there were many studies based on a single biomarker to study its impact on the survival and prognosis of breast cancer patients in Xinjiang. For instance, Zhang et al. [18] found that the continuous increase of *CEA* values was the independent prognostic factor of breast cancer patients' recurrence and metastasis after radical operation in Xinjiang. Wu et al. [14] proposed a Bayesian joint model to analyze the association between *E2* and the breast cancer patients' survival prognosis in Xinjiang. There were relatively few studies which investigated the comprehensive impact of dynamic changes in *CEA* and *E2* values on the survival prognosis of breast cancer patients in Xinjiang.

Motivated by the aforementioned research, based on information from patients who were diagnosed with breast cancer at the Affiliated Cancer Hospital of Xinjiang Medical between January, 2015 and December, 2019, first, the independent prognostic factors for breast cancer patients were identified by using Boruta algorithm. Moreover, linear mixed effects model and Cox proportional risk model were used to fit the longitudinal data(longitudinal submodel) and survival data(survival submodel) of breast cancer patients, respectively, a bivariate Bayesian joint model of longitudinal and survival data which based on longitudinal measurements of *CEA* and *E2* were further established to investigate how the dynamical changes of *CEA* and *E2* values collectively influence the survival prognosis of breast cancer patients in Xinjiang. The predictive performance of the joint model was assessed using ROC and calibration curves. Therefore, the joint model could provide a theoretical basis for the prevention and treatment of breast cancer patients in Xinjiang.

1 Objects and methods

1.1 Patient information extraction and selection

In this paper, 1241 patients diagnosed with breast cancer were followed up by the Affiliated Cancer Hospital of Xinjiang Medical University between January 2015 and December 2019 by using the follow-up and electronic medical record system. In addition, this hospital is the only cancer research center in Xinjiang. The starting point of the follow-up was the time when patients were diagnosed with breast cancer and the outcome event was the death due to breast cancer. The follow-up time was measured in days, and the deadline of the follow-up was December 31, 2019.

Basic demographic, clinicopathological, and survival information for patients was gathered, including *CEA* values, *E2* values, age, clinical stage, T-stage, N-stage, M-stage, the type of operation, preoperative chemotherapy, preoperative targeted therapy, preoperative radiotherapy, and received neoadjuvant therapy. The serum *E2* values (pg/ml) and serum *CEA* (ug/L) were measured longitudinally. Inclusion and exclusion criteria were as follows:

Patients were included if they met all three criteria: 1) diagnosed through biopsy or other clinical diagnostic methods; 2) without other serious disease; and 3) with complete clinicopathological and follow-up information.

Exclusion criteria were: 1) breast cancer patients whose longitudinal observations of CAE or *E2* were fewer than two because of the loss to follow-up; 2) breast cancer patients who had no critical information; and 3) breast cancer patients who were pregnant or breastfeeding.

Therefore, 971 patients were included with a total of 2690 *CEA* and *E2* person-times. These values were longitudinally measured 3 times per person on average. The follow-up rate was 92.98%. The follow-up duration for each patient was 26–2389 days with a median follow-up time of 690 days. This study involving breast cancer participants was censored and approved by the Medical Ethics Committee of the Affiliated Cancer Hospital of Xinjiang Medical University (approval number: K2023001). The instances fulfilling the criteria were gradually screened out in accordance with the inclusion and exclusion criteria (see Fig. 1). Baseline data chart for female breast cancer patients in Xinjiang are shown in Table 1.

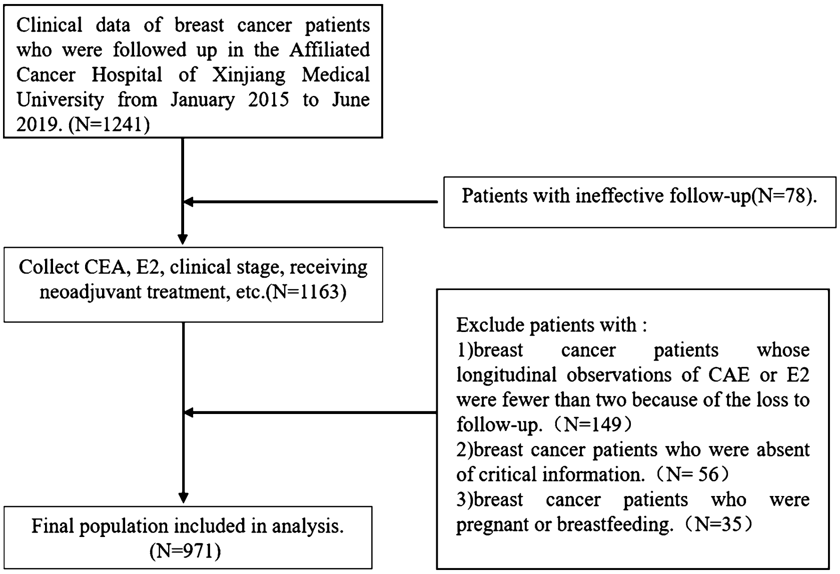
Clinical data of breast cancer patients who were followed up in the Affiliated Cancer Hospital of Xinjiang Medical University from January 2015 to June 2019. (N=1241)
Patients with ineffective follow-up(N=78).
Collect CEA, E2, clinical stage, receiving neoadjuvant treatment, etc.(N=1163)
Exclude patients with :
1)breast cancer patients whose longitudinal observations of CAE or E2 were fewer than two because of the loss to follow-up.（N=149）
2)breast cancer patients who were absent of critical information.（N= 56）
3)breast cancer patients who were pregnant or breastfeeding.（N=35）
Final population included in analysis. (N=971)

Figure 1 Breast Cancer Patient Screening Flowchart

1.2 A joint model of longitudinal and time-to-event data

The change trends of *CEA* and *E2* values over time were fitted by different linear mixed-effects model, respectively. Let $y_{1i}(t)$ and $y_{2i}(t)$ represent the longitudinal measurements of *CEA* and *E2* of the patient i $(i = 1,\dots,n)$ at a specific time point t, respectively. The models are then listed below:

$$y_{1i}\left(t\right) = \boldsymbol{x}_{1i}^{T}\left(t\right)\boldsymbol{\beta}_{1} + \boldsymbol{z}_{1i}^{T}\left(t\right)\boldsymbol{b}_{1i} + \varepsilon_{1i}\left(t\right) = \boldsymbol{m}_{1i}\left(t\right) + \varepsilon_{1i}\left(t\right),$$

$$y_{2i}(t) = \boldsymbol{x}_{2i}^{T}\left(t\right)\boldsymbol{\beta}_{2} + \boldsymbol{z}_{2i}^{T}\left(t\right)\boldsymbol{b}_{2i} + \varepsilon_{2i}\left(t\right) = \boldsymbol{m}_{2i}(t) + \varepsilon_{2i}\left(t\right),$$

where $\boldsymbol{x}_{1i}^{T}\left(t\right)$ and $\boldsymbol{x}_{2i}^{T}$ are the design vectors for the fixed effects regression coefficients $\boldsymbol{\beta}_{1}$ and $\boldsymbol{\beta}_{2}$; $\boldsymbol{z}_{1i}^{T}\left(t\right)$ and $\boldsymbol{z}_{2i}^{T}\left(t\right)$ are the design vectors for the random effects $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$; the complete vector of random effects $\boldsymbol{b}_{i} = (\boldsymbol{b}_{1i}^{T}, \boldsymbol{b}_{2i}^{T})$ is assumed to follow a multivariate normal distribution with mean zero and variance–covariance matrix D, that is, $\boldsymbol{b}_{i} \sim MVN\left(0, D\right)$; $\varepsilon_{1i}\left(t\right)$ and $\varepsilon_{2i}\left(t\right)$ are the random errors following normal distribution, i.e., $\varepsilon_{1i}\left(t\right) \sim N(0,\sigma_{1}^{2})$ and $\varepsilon_{2i}\left(t\right) \sim N(0,\sigma_{2}^{2})$; $\boldsymbol{m}_{1i}(t)$ and $\boldsymbol{m}_{2i}(t)$ represent the true longitudinal processes of *CEA* and *E2* for the patient i $(i = 1,\dots,n)$ at a specific time point t, respectively. We assume a proportional hazard model for the risk of the event as follows:

$$\begin{aligned} h\left(t \mid \mathcal{M}_{i}\left(t\right), \boldsymbol{\omega}_{i};\right) &= \lim_{\Delta t \to 0} \frac{P\left(t \le T_{i}^{*} \le t + \Delta t \mid T_{i}^{*} \ge t, \mathcal{M}_{i}, \boldsymbol{\omega}_{i}\right)}{\Delta t} \\ &= \lambda_{0}\left(t\right)\exp\left(\boldsymbol{\omega}_{i}^{T}\boldsymbol{\gamma} + \alpha_{1}\boldsymbol{m}_{1i}\left(t\right) + \alpha_{2}\boldsymbol{m}_{2i}(t)\right), \end{aligned}$$

where $\boldsymbol{\mathcal{M}}_{i} = \{\boldsymbol{\mathcal{M}}_{1i}, \boldsymbol{\mathcal{M}}_{2i}\}$; $\boldsymbol{\mathcal{M}}_{1i} = \{\boldsymbol{m}_{1i}\left(s\right), 0 < s < t\}$; $\boldsymbol{\mathcal{M}}_{2i} = \{\boldsymbol{m}_{2i}\left(s\right), 0 < s < t\}$; $T_{i} = \min\left(T_{i}^{*}, C_{i}\right)$ represents the true observation time when the outcome event occurs for patient i, and C_{i} represents the censoring time of patient i; $\boldsymbol{\omega}_{i}$ is the time-independent q-dimensional fixed effect covariates, $\boldsymbol{\gamma}$ was the coefficient vector of $\boldsymbol{\omega}_{i}^{T}$; $\lambda_{0}\left(t\right)$ is the baseline

Table 1 Baseline Data Chart for Female Breast Cancer Patients in Xinjiang from January 2015 to December 2019

Variables	Total (n = 971)	Survival (n = 855)	Death (n = 116)
Follow up-time	690.52 ± 50.98	725.12 ± 60.98	670.52 ± 45.28
CEA	5.66 ± 0.87	3.64 ± 0.21	21.01 ± 7.28
E2	154.56 ± 4.55	132.71 ± 4.04	320.39 ± 22.07
Age	48.17 ± 1.87	48.58 ± 0.17	45.09 ± 0.432
Neoadjuvant therapy			
yes	720 (74.15%)	609 (71.23%)	111 (95.69%)
no	251 (25.85%)	246 (28.77%)	5 (4.31%)
Clinical stage			
I	477 (49.12%)	438 (51.23%)	39 (33.62%)
II	293 (30.18%)	261 (30.53%)	32 (27.59%)
III	181 (18.64%)	145 (16.96%)	36 (31.03%)
IV	20 (2.06%)	11 (1.29%)	9 (7.76%)
Preoperative chemotherapy			
yes	701 (72.19%)	617 (72.16%)	84 (72.41%)
no	270 (27.81%)	238 (27.84%)	32 (27.59%)
Preoperative targeted therapy			
yes	896 (92.28%)	787 (92.05%)	109 (93.97%)
no	75 (7.72%)	68 (7.95%)	7 (6.03%)
Preoperative radiotherapy			
Yes	697 (71.78%)	627 (73.33%)	70 (60.34%)
No	274 (28.21%)	228 (26.66%)	46 (39.65%)
T-stage			
I	387 (39.86%)	347 (40.58%)	40 (34.48%)
II	460 (47.37%)	403 (47.13%)	57 (49.14%)
III	64 (6.59%)	53 (6.20%)	11 (9.48%)
IV	60 (6.18%)	52 (6.08%)	8 (6.90%)
N-stage			
I	425 (43.77%)	384 (44.91%)	41 (35.34%)
II	347 (35.74%)	302 (35.32%)	45 (38.79%)
III	96 (9.89%)	85 (9.94%)	11 (9.48%)
IV	103 (10.61%)	84 (9.82%)	19 (16.38%)
M-stage			
I	923 (95.06%)	821 (96.02%)	102 (87.93%)
II	48 (4.94%)	34 (3.98%)	14 (12.07%)
Operations			
nonsurgical	128 (13.18%)	100 (11.70%)	28 (24.14%)
radical surgery	684 (70.44%)	612 (71.58%)	72 (62.07%)
breast conserving surgery	146 (15.04%)	133 (15.56%)	13 (11.21%)
breast reconstruction surgery	13 (1.34%%)	10 (1.17%)	3 (2.59%)

hazard function. The parameter α_1 and α_2 quantify the strength of association between $\boldsymbol{m_{1i}}(t)$ and $\boldsymbol{m_{2i}}(t)$ and the death risk at the same time point.

1.3 Parameter estimation method

MCMC is widely applied for parameter estimation of Bayesian methods. The basic idea of MCMC is to establish a stable posterior distribution by using Markov chain, to obtain multiple samples through random sampling, and to infer the posterior expectation of parameters based on these samples. The posterior distribution of the parameters is as follows:

$$p\left(\boldsymbol{\theta}, \boldsymbol{b_i} \mid T_i, y_{1i,} y_{2i}, \delta_i\right) \propto \prod_{i=1}^{n} p\left(y_{1i,} y_{2i} \mid \boldsymbol{b}_i ; \boldsymbol{\theta_y}\right) p\left(T_i, \delta_i \mid \boldsymbol{b_i} ; \boldsymbol{\theta_t}\right) p\left(\boldsymbol{b_i} ; \boldsymbol{D}\right) p\left(\boldsymbol{\theta}\right),$$

where $\boldsymbol{\theta}$ represents the set of all the unknown parameters in the joint model, i.e., $\boldsymbol{\theta} = \left\{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \alpha_1, \alpha_2, \lambda_0, \gamma, D, \sigma_1, \sigma_2\right\}$; $\boldsymbol{\theta_y}$ represents the set of all the unknown parameters in lon-

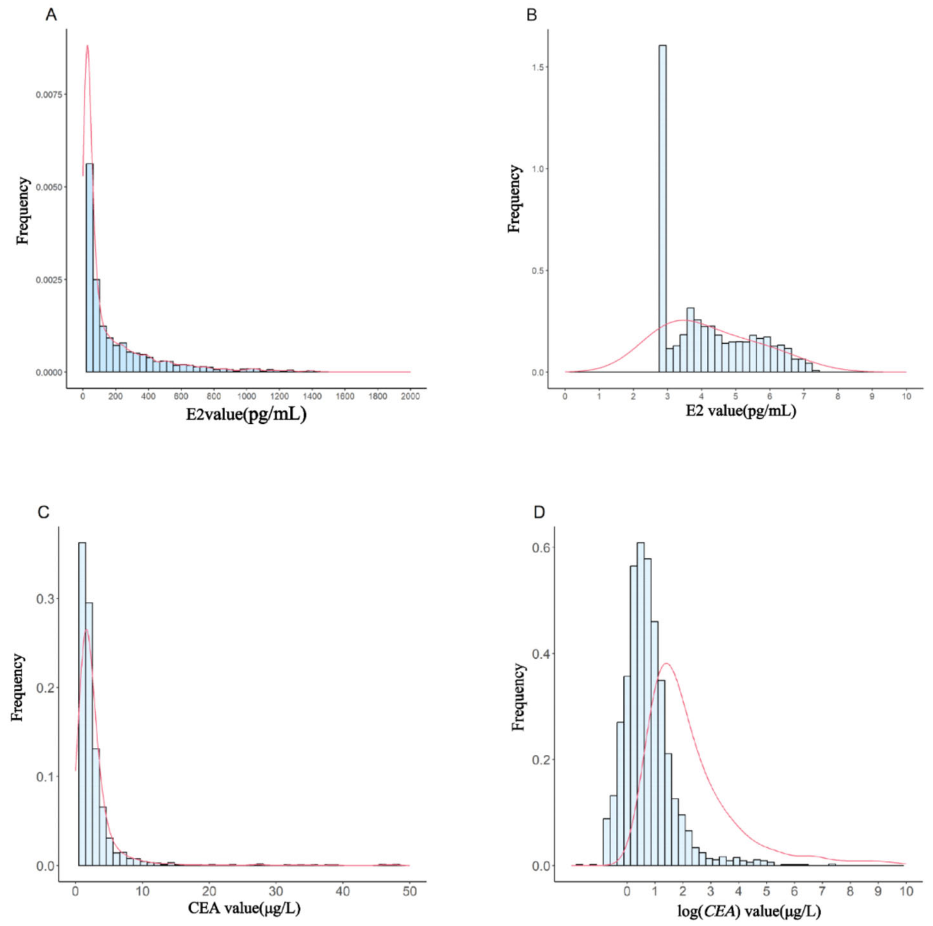

Figure 2 Histograms and probability density plot of *CEA* and *E2* of breast cancer patients

gitudinal processes, i.e., $\boldsymbol{\theta_y} = \left\{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, D, \sigma_1, \sigma_2\right\}$; $\boldsymbol{\theta_t}$ represents the set of all the unknown parameters in the survival processes, i.e., $\boldsymbol{\theta_t} = \{\lambda_0, \gamma, \alpha_1, \alpha_2\}$; $p\,(\boldsymbol{\theta})$ represent the prior distribution of $\boldsymbol{\theta}$.

2 Research results

2.1 The results of the descriptive research

The results of the normality test indicated that *CEA* and *E2* values of breast cancer patients showed a skewed distribution (Fig. 2, (A) and (C)). However, longitudinal data in a linear mixed model was required to (approximately) follow the normal distribution. Thus, by logarithmically transforming the original *CEA* and *E2* values of patients, as shown in Figs. 2(B) and 2(D), to log(CEA) and log($E2$), respectively, we obtained approximately normal distributions.

2.2 Variable screening based on Boruta algorithm

Nine independent prognostic markers, including *E2, CEA*, T-stage, neoadjuvant treatment, age, N-stage, clinical stage, M-stage, and operations, were identified by applying Boruta algorithm (see Fig. 3). Here “shadowmean”, “shadowmin”, and “shadowmax” are the mean, minimum, and maximum importance scores of all shadow features, respec-

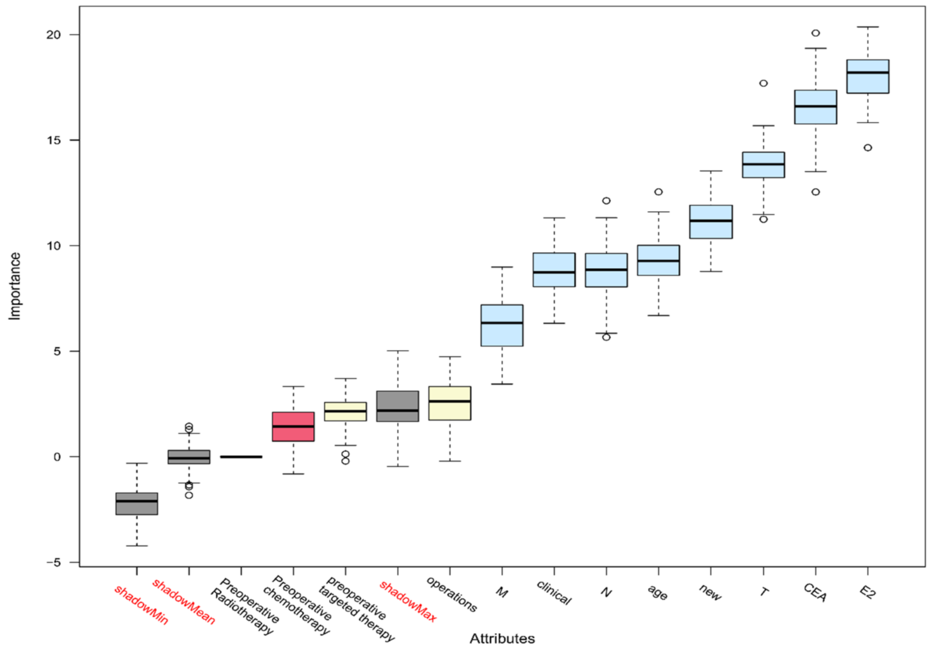

Figure 3 The results of variable screening based on Boruta algorithm

tively. When the feature variables' score was greater than the shadow max, the variable was considered significant and related to the survival prognosis of breast cancer patients. Particularly, within the independent prognostic factors of breast cancer patients, *E2* was the strongest, followed by *CEA*, while operations was the weakest variable.

2.3 Bivariate Bayesian joint model of longitudinal and time-to-event data

2.3.1 Longitudinal submodel outcomes

The dynamical changes of patients' log(*CEA*) and log(*E2*) values were respectively fitted by two linear mixed models, and the fitting results are shown in Table 2. The final longitudinal submodel was identified as follows:

$$\begin{cases} y_{1i}(t) = \beta_0 + \beta_1 X_{\text{Follow up-time}} + \beta_2 X_{\text{age}} + \beta_3 X_{\text{operations}} + b_0 \\ \qquad\qquad + b_{1i} X_{\text{Follow up-time}} + \varepsilon_{1i}(t), \\ y_{2i}(t) = \beta_4 + \beta_5 X_{\text{Follow up-time}} + \beta_6 X_{\text{age}} + \beta_7 X_{\text{clinical stage}} \\ \qquad\qquad + b_1 + b_{2i} X_{\text{Follow up-time}} + \varepsilon_{2i}(t). \end{cases}$$

2.3.2 Survival submodel

As shown in Table 3, neoadjuvant therapy, clinical stage, and age were significant to the survival prognosis of breast cancer patients in Xinjiang by using univariate and multivariate Cox regression. The optimal survival submodel was identified as follows:

$$h_i(t) = \lambda_0(t)\exp\{\gamma_1\omega_{\text{Neoadjuvant therapy}} + \gamma_2\omega_{\text{clinical stage}} + \gamma_3\omega_{\text{age}} + \alpha_1\boldsymbol{m_{1i}}(t) + \alpha_2\boldsymbol{m_{2i}}(t)\}.$$

Table 2 Parameter estimation results for the longitudinal submodel

Variables	*CEA*			*E2*		
	Coeff	*Std*	*P*	*Coeff*	*Std*	*P*
Follow up-time	1×10^{-4}	5×10^{-5}	<0.05*	3×10^{-4}	6×10^{-5}	<0.01*
T-stage	–	–	–	–	–	–
II	0.042	0.052	0.418	0.018	0.038	0.632
III	0.116	0.097	0.231	−0.244	0.071	<0.05*
IV	0.062	0.102	0.541	−0.354	0.076	<0.05*
N-stage	–	–	–	–	–	–
II	−0.041	0.051	0.424	−0.097	0.038	<0.05*
III	−0.049	0.077	0.527	−0.473	0.057	<0.01*
IV	0.259	0.080	<0.05*	−0.383	0.059	<0.01*
M-stage	–	–	–	–	–	–
II	0.230	0.222	0.300	−0.113	0.164	0.490
Neoadjuvant therapy	–	–	–	–	–	–
yes	0.087	0.053	0.099	−0.069	0.039	<0.05*
Age	0.007	0.002	<0.05*	−0.022	0.001	<0.01*
Clinical stage	–	–	–	–	–	–
II	−0.001	0.040	0.968	1.171	0.033	<0.05*
III	−0.050	0.046	0.277	2.294	0.038	<0.01*
IV	0.088	0.085	0.301	2.962	0.072	<0.01*
Operations	–	–	–	–	–	–
radical surgery	−0.201	0.066	<0.05*	0.029	0.049	0.542
breast conserving surgery	−0.252	0.084	<0.05*	0.016	0.062	0.789
breast reconstruction surgery	−0.226	0.098	<0.05*	0.144	0.148	0.331

*Indicates that the variable is statistically significant at $P < 0.05$.

Table 3 Parameter estimation results for the survival submodel

Variables	Univariate Cox regression			Multivariate Cox regression		
	Coeff	*HR*	*P*	*Coeff*	*HR*	*P*
Neoadjuvant therapy	–	–	–	–	–	–
yes	−2.132	0.118	<0.05*	−1.545	0.213	<0.05*
age	−0.020	0.979	0.091	−0.024	0.975	<0.05*
Clinical stage	–	–	–	–	–	–
II	0.416	1.516	<0.01*	0.460	1.585	<0.05*
III	1.369	3.935	<0.01*	0.980	2.665	<0.05*
IV	1.336	3.805	<0.01*	1.617	5.042	<0.05*
T-stage	–	–	–	–	–	–
II	0.440	1.553	<0.05*	–	–	–
III	0.194	1.215	0.567	–	–	–
IV	0.037	1.038	0.096	–	–	–
N-stage	–	–	–	–	–	–
II	0.704	2.021	0.844	–	–	–
III	−0.210	1.635	0.258	–	–	–
IV	0.561	1.601	0.165	–	–	–
M-stage	–	–	–	–	–	–
II	0.933	2.542	0.167	–	–	–
Operations	–	–	–	–	–	–
radical surgery	−1.180	0.307	0.069	–	–	–
breast conserving surgery	−1.148	0.317	<0.05*	–	–	–
breast reconstruction surgery	−1.151	0.501	0.056			

*Indicates that the variable is statistically significant at $P < 0.05$.

Table 4 Parameter estimation for the Bayesian multivariate joint model

Variables	*Coeff*	*HR* (95%*CI*)	*P*
Neoadjuvant therapy	–	–	–
yes	−1.966	0.136 (0.048, 0.35)	<0.05*
Age	0.0042	1.004 (0.98, 1.03)	0.773
Clinical stage	–	–	–
II	0.130	1.133 (1.032, 2.201)	<0.05*
III	0.724	2.051 (1.152, 3.593)	<0.05*
IV	1.560	4.758 (1.993, 11.02)	<0.05*
Association coefficients	–	–	–
α_1	0.947	2.577 (1.803, 3.563)	<0.01*
α_2	0.635	1.887 (1.472, 2.454)	<0.01*

*Indicates that the variable is statistically significant at $P < 0.05$.

2.4 Joint model

A bivariate Bayesian joint model was established to investigate the effect of dynamical changes of *CEA* and *E2* values on the breast cancer patients' survival prognosis. It was

$$\begin{cases} y_{1i}(t) = \beta_0 + \beta_1 X_{\text{Follow up-times}} + \beta_2 X_{\text{age}} + \beta_3 X_{\text{operations}} + b_0 \\ \qquad\qquad + b_{1i} X_{\text{Follow up-times}} + \varepsilon_{1i}(t), \\ y_{2i}(t) = \beta_4 + \beta_5 X_{\text{Follow up-times}} + \beta_6 X_{\text{age}} + \beta_7 X_{\text{clinical stage}} + b_1 \\ \qquad\qquad + b_{2i} X_{\text{Follow up-times}} + \varepsilon_{2i}(t), \\ h_i(t) = \lambda_0(t) \exp\{\gamma_1 \omega_{\text{Neoadjuvant therapy}} + \gamma_2 \omega_{\text{clinical stage}} \\ \qquad\qquad + \gamma_3 \omega_{\text{age}} + \alpha_1 \boldsymbol{m_{1i}}(t) + \alpha_2 \boldsymbol{m_{2i}}(t). \end{cases}$$

As shown in Table 4, the association coefficients α_1 and α_2 are statistically significant and patients' death risk increases by approximately 1.577 times ($HR = 2.577$, 95%CI: $(1.803, 3.563)$) and 0.887 times ($HR = 1.887$, 95%CI: $(1.472, 2.454)$), respectively, with the one-unit increase of log (*CEA*) and log(*E2*).

2.5 The predictive performance and goodness of fit of the joint model

Finally, four patients were randomly selected to predict their long-term mortality probability. As shown in Fig. 4, four patients' death risk would increase gradually when their *CEA* and *E2* values increased. In particular, with the increase of the measure times for *CEA* and *E2* values, the confidence intervals of survival curves became narrower, which indicates that the estimation of patients' long-term mortality probability was more accurate and the prediction of patients' survival status was more reliable (see Fig. 4B). As shown in Fig. 4C, the *CEA* and *E2* outlier values could have a certain impact on the predicted precision of Bayesian joint model. Moreover, the patient's death risk sharply increased and the long-term mortality was higher than 80% when the patient's *CEA* and *E2* values had a significant upward trend (see Fig. 4D).

The ROC and calibration curves were plotted to assess the predictive performance of the joint model. As shown in Fig. 5A, the predictive performance of the joint model was within a reasonable range with an AUC value of 0.778. The model's prediction value was more closely aligned with the 45° reference line, which indicated that the model had a high degree of calibration performance (see Fig. 5B).

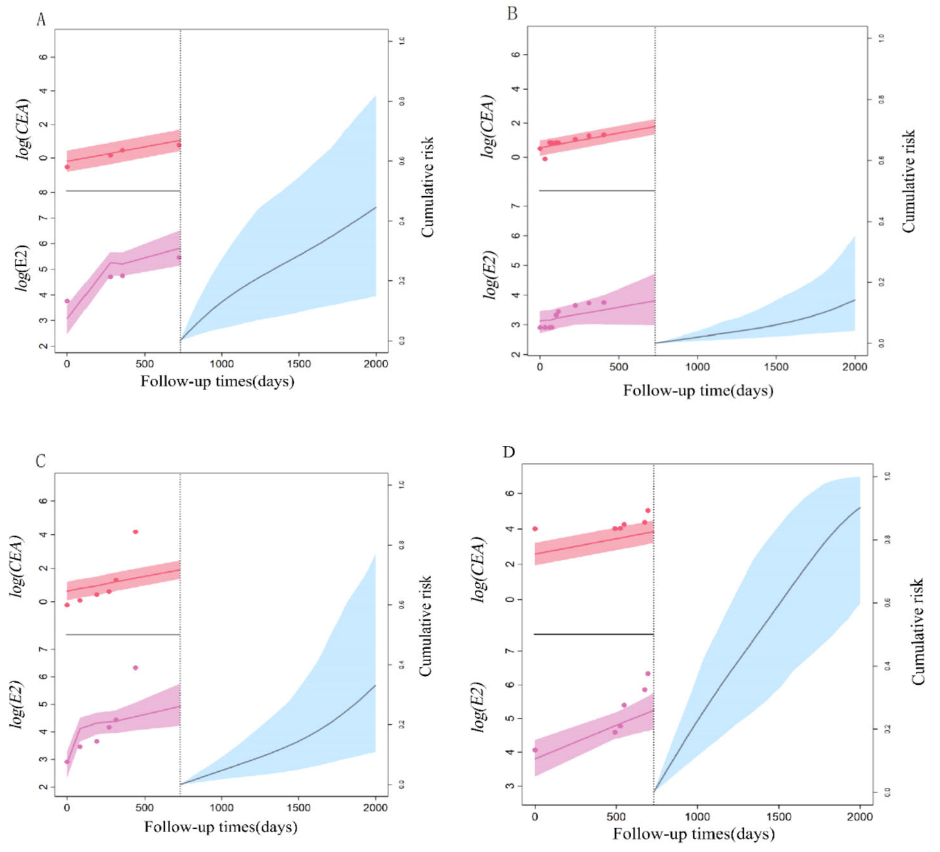

Figure 4 Prediction of four patients' long-term mortality probability

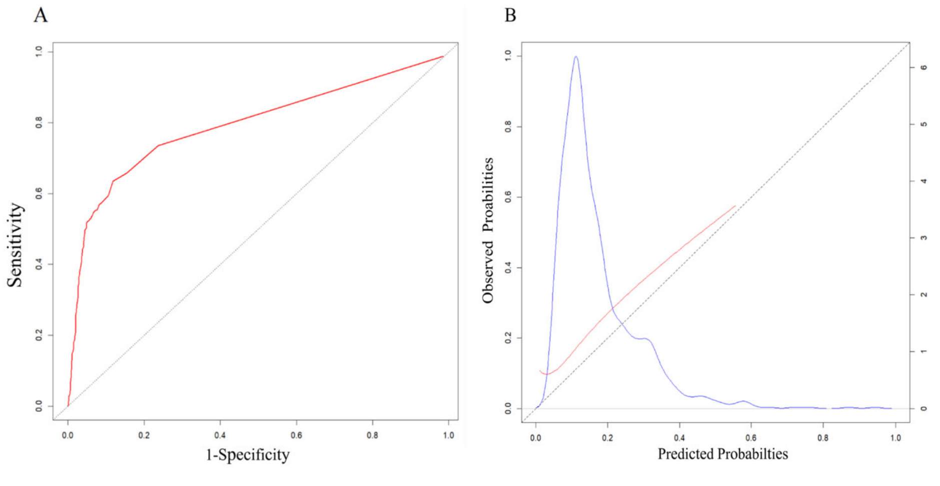

Figure 5 ROC and calibration curves of the joint model

3 Discussion

Bivariate Bayesian joint models for longitudinal and time-to-event data could comprehensively analyze the association between dynamic changes of multiple longitudinal data and the breast cancer patients' survival prognosis, as well as describe the correlation within multiple longitudinal data. In this paper, based on the independent prognostic factors screened by Boruta algorithm, a bivariate Bayesian joint model for longitudinal and time-to-event data was proposed to investigate the impact of dynamical changes of *CEA* and *E2* values of breast cancer patients' survival prognosis in Xinjiang.

First, it was shown from parameter estimation results of the joint model that the collective increase of *CEA* and *E2* values was an independent factor of breast cancer patients' survival prognosis in Xinjiang. And when all other baseline variables were unchanged, patients' death risk separately increased by approximately 1.577 times ($HR = 2.577$, $95\%CI$: $(1.803, 3.563)$) and 0.887 times ($HR = 1.887$, $95\%CI$: $(1.472, 2.454)$) with the one-unit collective increase in log (*CEA*) and log (*E2*). This result was similar to the conclusions of Shao et al. [19] and Li et al. [20]. This may be due to the reason that *CEA* is a member of the cell surface glycoprotein family and it is one of the tumor markers for various glandular cancers in clinical practice [7]. With the increase of *CEA* values, the body's inhibition on immune and T cells would be enhanced, which could lead to the weakening of the inhibitory effect of the immune system on cancer cellular oncogenesis, proliferation, differentiation, and metastasis. And it would start up a vicious circle, further increasing the expression values of *CEA* [21]. In addition, *E2* could affect cellular immune response processes by facilitating the polarization of macrophages toward an immune-suppressive state in the tumor microenvironment, leading to immune cell dysfunction and development of breast cancer [22]. Resent research found that *E2* could also decrease the number of peripheral eosinophils and tumor-associated tissue eosinophilia by inhibiting the proliferation and survival of maturing eosinophils, and it would lead to tumor growth of breast cancer [23]. In addition, breast cancer patients' survival prognosis was closely related to the expression of estrogen in their body. And this may due to the imbalance between *E2* and antiestrogen, which promotes the further proliferation of breast cancer cells, leading to the deterioration of the patients' condition [24]. Patients' survival prognosis would be greatly improved by using selective estrogen receptor modulators (tamoxifen, raloxifene, toremifene, etc.), elective estrogen-receptor downregulators and aromatase inhibitors such as anastrozole, letrozole, and astmestane [25]. Therefore, continuously and simultaneously measuring *CEA* and *E2* would be beneficial for monitoring breast cancer patients' process, and could assess the risk of early recurrence or metastasis of breast cancer patients.

Moreover, it was shown from parameter estimation results of the joint model that the death risks of the patients in clinical stage II ($HR = 1.133$, $95\%CI$: $(1.032, 2.201)$), stage III ($HR = 2.051$, $95\%CI$: $(1.152, 3.593)$), and stage IV ($HR = 4.758$, $95\%CI$: $(1.993, 11.02)$) were higher than the death risk in stage I (we used stage I as the reference category). This result was similar to the conclusions of DeSantis et al. [26] and Walters et al. [27]. The higher the patient's clinical stage, the larger the tumor. And then, it would lead to the high probability of lymph node and distant metastasis, and finally cause the high risk of death. Therefore, it was suggested that breast cancer patients should regularly assess their clinical stage to enhance treatment efficacy and improve survival outcomes.

Last but not least, parameter estimation results also indicated that the neoadjuvant therapy for breast cancer patients could significantly reduce their death risk ($HR = 0.136$, $95\%CI$: $(0.048, 0.35)$), which is consistent with the results in [28–30]. The reasons for this result might be that the neoadjuvant therapy could reduce the tumor volume, make those patients temporarily unable to undergo surgery become eligible for surgery, and make those patients not suitable for breast-conserving surgery become feasible for breast-conserving surgery, thereby improving breast cancer patients' survival prognosis [6].

There are some limitations in this paper. First of all, due to the small sample size of this study, the 95% confidence intervals of the association coefficients in the model were relatively wider, and the probabilities of finding the parameters within these intervals were relatively lower [31, 32]. Therefore, the sample size would need to be expanded in our future study. In addition, α-fetoprotein, human chorionic gonadotrophin, and casein in human body, which would interact with each other, play vital roles in the occurrence and development of breast cancer [33]. Hence, multivariate Bayesian joint models could be applied to explore the comprehensive impact of biomarkers such as α-fetoprotein, etc., on the prognosis of breast cancer patients [34, 35]. On the other hand, although the traditional Bayesian joint model can investigate the effect of the conditional mean of the longitudinal outcomes on the survival prognosis of patients, it cannot determine the effect of the median or lower/higher quantile of the longitudinal outcome on the survival prognosis of patients. Therefore, the Bayesian joint model based on quantile regression could be established to analyze the effect of different quantile of biomarkers on the survival prognosis of breast cancer patients in a further study [36].

4 Conclusions

In conclusion, *CEA*, *E2*, etc., were identified as independent prognostic factors of breast cancer patients in Xinjiang based on Boruta algorithm. Moreover, a bivariate Bayesian joint model of longitudinal and time-to-event data was established to investigate the impact of dynamical changes in *CEA* and *E2* values on the survival prognosis of breast cancer patients in Xinjiang. It was indicated that a collective increase of *CEA* and *E2* values could increase the risk of early recurrence or metastasis of breast cancer patients.

Acknowledgements
Not applicable.

Author contributions
TM: Conceptualization; Software; Writing – original draft; Writing – review & editing. CG: Methodology; Writing – original draft. YY: Software. JL: Software. TZ: Data curation; Resources; Writing – review & editing; Funding acquisition. LW: Conceptualization; Methodology; Writing – original draft; Writing – review & editing; Funding acquisition. All authors read and approved the final manuscript.

Funding
The work is supported by the National Natural Science Foundation of China (Grant No. 12061079), Natural Science Foundation of Xinjiang (Grant No. 2022D01C287), the Project of Topnotch Talents of Technological Youth of Xinjiang (Grant No. 2022TSYCCX0108).

Data availability
The data that support the findings of this study are available from the Affiliated Cancer Hospital of Xinjiang Medical University, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Requests to access the datasets should be directed to T. Z, zhaoting0557@163.com.

Declarations

Ethics approval and consent to participate
The studies involving human participants were reviewed and approved by Medical Ethics Committee of the Affiliated Cancer Hospital of Xinjiang Medical University (approval number: K-2023001).

Consent for publication
Consent was obtained from the participants (or legal parents or guardians of the children) for the release of the reported individual patient data.

Competing interests

The authors declare no competing interests.

Author details

[1]College of Public Health, Xinjiang Medical University, Urumqi, Xinjiang 830017, China. [2]Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830011, China. [3]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China. [4]Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi 830017, China.

Received: 23 October 2024 Accepted: 12 January 2025 Published online: 05 February 2025

References

1. Bray, F., et al.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **74**(3), 229–263 (2024)
2. Tan, N., et al.: Burden of female breast and five gynecological cancers in China and worldwide. Chin. Med. J. **137**(18), 2190–2201 (2024)
3. Li, H.F., Guo, C.M., Wang, H.Y., et al.: Epidemiological analysis of 1701 cases of breast cancer in a third-grade hospital of Urumqi of Xinjiang province. Pract. J. Cancer **37**(06), 975–979 (2022)
4. Chakraborty, B., Byemerwa, J., Shepherd, J., et al.: Inhibition of estrogen signaling in myeloid cells increases tumor immunity in melanoma. J. Clin. Invest. **131**, e151347–e151347 (2021)
5. Yuan, M., Yu, X.: Endocrine therapy resistance mechanism and targeted therapy strategy for estrogen receptor-positive breast cancer. Chin. Bull. Life Sci. **34**(12), 1559–1568 (2022) (in Chinese)
6. Hao, C., Zhang, G., Zhang, L.: CEA levels in 49 different types of cancer and noncancer diseases. Prog. Mol. Biol. Transl. Sci. **162**, 213–227 (2019)
7. Hammarstrom, S.: The carcinoembryonic antigen (CEA) family: structures, suggested functions and expression in normal and malignant tissues. Cancer Biol. **9**, 67–81 (1999)
8. Bolker, B.M.: Linear and generalized linear mixed models. In: Ecological Statistics: Contemporary Theory and Application, pp. 309–333 (2015)
9. Crowther, M.J., Abrams, K.R., Lambert, P.C.: Joint modeling of longitudinal and survival data. Stata J. **13**(1), 165–184 (2013)
10. Wu, L.: Mixed Effects Models for Complex Data. CRC Press, Boca Raton (2009)
11. Self, S., Pawitan, Y.: Modeling a marker of disease progression and onset of disease. In: AIDS Epidemiology: Methodological Issues, pp. 231–255 (1992)
12. De Gruttola, V., Tu, X.M.: Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics **50**(4), 1003–1014 (1994)
13. Azarbar, A., Wang, Y., Nadarajah, S.: Bayesian modeling of longitudinal and survival data in breast cancer patients. Commun. Stat., Theory Methods **50**(2), 400–414 (2021)
14. Wu, M., Zhang, T., Gao, C., et al.: Joint-modeling of estradiol levels and survival data of breast cancer patients in the case-cohort design. Chin. Gen. Pract. **13**(6), 941–953 (2024)
15. Li, J., et al.: Effect of estradiol as a continuous variable on breast cancer survival by menopausal status: a cohort study in China. Breast Cancer Res. Treat. **194**(1), 103–111 (2022)
16. Guo, X., Liu, Z., Li, C., et al.: Analysis of the diagnostic value of sex hormone combined with carcinoembryonic antigen in lymph node metastasis of breast cancer patients. Chin. J. Endocrin. Surg., 162–165 (2023)
17. Verbeke, G., Fieuws, S., Molenberghs, G., et al.: The analysis of multivariate longitudinal data: a review. Stat. Methods Med. Res. **23**(1), 42–59 (2014)
18. Zhang, W., Xu, C.L., The, Z.L.: Correlation between CEA, CA153, CA199 and breast cancer pathological classification. Cardiovasc. Dis. Electron. J. Integr. Tradit. Chin. West. Med. **7**(23), 176–177 (2019)
19. Shao, Y., Sun, X., He, Y., Liu, C., Liu, H.: Elevated levels of serum tumor markers CEA and CA15-3 are prognostic parameters for different molecular subtypes of breast cancer. PLoS ONE **10**(7), e0133830ee0133830 (2015)
20. Li, J., Li, C., Feng, Z., et al.: Effect of estradiol as a continuous variable on breast cancer survival by menopausal status: a cohort study in China. Breast Cancer Res. Treat. **194**(1), 103–111 (2022). https://doi.org/10.1007/s10549-022-06593-5
21. Zhang, S.J., Hu, Y., Qian, H.L., et al.: Expression and significance of ER, PR, VEGF, CA15-3, CA125 and CEA in judging the prognosis of breast cancer. Asian Pac. J. Cancer Prev. **14**(6), 3937–3940 (2013)
22. Chakraborty, B., Byemerwa, J., Shepherd, J., et al.: Inhibition of estrogen signaling in myeloid cells increases tumor immunity in melanoma. J. Clin. Invest. **131**(23), e151347 (2021)
23. Artham, S., Juras, P.K., Goyal, A., et al.: Estrogen signaling suppresses tumor-associated tissue eosinophilia to promote breast tumor growth. Sci. Adv. **10**(39), eadp2442–eadp2442 (2024)
24. Santen, R.J., Stuenkel, C.A., Yue, W.: Mechanistic effects of estrogens on breast cancer. Cancer J. **28**(3), 224–240 (2022)
25. Farkas, S., Szabó, A., Hegyi, A.E., et al.: Estradiol and estrogen-like alternative therapies in use: the importance of the selective and non-classical actions. Biomedicines **10**(4), 861 (2022)
26. DeSantis, C.E., Fedewa, S.A., Goding Sauer, A., et al.: Breast cancer statistics, 2015: convergence of incidence rates between black and white women. CA Cancer J. Clin. **66**(1), 31–42 (2016)
27. Walters, S., Maringe, C., Butler, J., et al.: Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000–2007: a population-based study. Br. J. Cancer **108**, 1195–1208 (2013)
28. Wuerstlein, R., Harbeck, N.: Neoadjuvant therapy for HER2-positive breast cancer. Rev. Recent Clin. Trials **12**(2), 81–92 (2017)
29. Teshome, M., Hunt, K.K.: Neoadjuvant therapy in the treatment of breast cancer. Surg. Oncol. Clin. **23**(3), 505–523 (2014)
30. Mauriac, L., Debled, M., Durand, M., et al.: Neoadjuvant tamoxifen for hormone-sensitive non-metastatic breast carcinomas in early postmenopausal women. Ann. Oncol. **13**(2), 293–298 (2002)
31. Silva, I.R., Oliveira, D.W.R.: Confidence-credible intervals. Commun. Stat., Theory Methods **51**(9), 2783–2802 (2022)
32. Amrhein, V., Greenland, S.: Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. J. Inf. Technol. **37**(3), 316–320 (2022)

33. Gray, B.N.: Value of CEA in breast cancer. Aust. N.Z. J. Surg. **54**(1), 1–2 (1984)
34. Chen, J., Huang, Y., Wang, Q.: Semiparametric multivariate joint model for skewed-longitudinal and survival data: a Bayesian approach. Stat. Med. **42**(27), 4972–4989 (2023)
35. Zou, H., Zeng, D., Xiao, L., et al.: Bayesian inference and dynamic prediction for multivariate longitudinal and survival data. Ann. Appl. Stat. **17**(3), 2574–2595 (2023)
36. Farcomeni, A., Viviani, S.: Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. Stat. Med. **34**(7), 1199–1213 (2015)

Publisher's Note

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