Am J Cancer Res 2023;13(7):3203-3220
www.ajcr.us /ISSN:2156-6976/ajcr0150536

Original Article

Risk, molecular subtype and prognosis of second primary breast cancer: an analysis based on first primary cancers

Chaofan Li[1*], Chong Du[1*], Yusheng Wang[2*], Mengjie Liu[1], Fang Zhao[1], Jia Li[1], Weiwei Wang[1], Xinyu Wei[1], Jingkun Qu[1], Zhangjian Zhou[1], Yinbin Zhang[1], Shuqun Zhang[1]

[1]*Department of Oncology, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157 West Fifth Street, Xi'an, Shaanxi, P. R. China;* [2]*Department of Otolaryngology, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157 West Fifth Street, Xi'an, Shaanxi, P. R. China.* [*]*Equal contributors and co-first authors.*

Received April 4, 2023; Accepted July 8, 2023; Epub July 15, 2023; Published July 30, 2023

Abstract: Second primary breast cancer (SPBC) was potentially related to other cancers, which may impact its incidence, prognosis and therapeutic approaches. Nevertheless, few studies have characterized this relationship and analyzed the subtypes of SPBC. Our study intended to investigate the occurrence and prognosis of SPBC. We analyzed the patterns, clinical characteristics, standardized incidence ratio (SIR) and standardized mortality ratio (SMR) of patients with SPBC. The propensity score matching (PSM) approach was further used to balance the differences in clinical features between patients with primary breast cancer (PBC) and SPBC, then Kaplan-Meier (KM) survival analysis was used to compare their overall survival and breast cancer-specific survival. Finally, a predictive model was constructed to estimate the 3- and 5-year survival rates of SPBC patients. We found that the SIR of individuals with SPBC was significantly higher in cancer survivors than in the general population (SIR=1.16, 95% CI=1.15-1.17, P<0.05). SPBC patients with first primary lung/bronchus cancer had a much higher SMR (SMR=1.71, 95% CI=1.58-1.85, P<0.05) compared with survivors of other malignancies. Individuals with SPBC had a larger proportion of the HR-/HER2- subtype than those with PBC. Particularly among survivors of estrogen-dependent ovarian and breast cancer, the proportion of the HR-/HER2- subtype of SPBC considerably rose. After propensity score matching, we discovered that SPBC patients' overall survival remained poorer than that of PBC patients (HR=1.43, 95% CI=1.39-1.47, P<0.001). However, the prognosis of SPBC in first primary thyroid cancer survivors was better than PBC patients (HR=0.64, 95% CI=0.55-0.75, P<0.001). Also, an extreme gradient boosting (XGBoost) model was developed to evaluate the 3-year (AUC=0.817) and 5-year survival (AUC=0.825) of SPBC patients. Our data demonstrated the distinct biological performance of SPBC with various first primary cancers. Furthermore, our findings revealed an indispensable role of first primary cancer (FPC) in the development of SPBC and provided an additional theoretical basis for the clinical follow-up and identification of SPBC.

Keywords: Second primary breast cancer, standardized incidence ratio, standardized mortality ratio, molecular subtype, propensity score matching, machine learning

Introduction

The second primary cancer (SPC) has emerged as a major public health concern since an increasing number of cancer patients have long-term survival as a result of better therapeutic outcomes [1]. It is estimated there will be 20.3 million cancer survivors in the United State by 2026, according to the 5-year relative survival rate being 67.4% for all cancer patients [2]. Long-term cancer survivors have a great chance to develop SPC [3].

In 2022, there were 287,850 new cases of female breast cancer (BC) in the United States, and an estimated death of 43,250 cases [4]. And it showed that BC also has higher incidence rate as SPC [1]. Understanding the epidemiology of the second primary breast cancer (SPBC) is incredibly crucial for cancer survivors looking for treatment and prevention solutions. BC is potentially associated with other cancers in terms of family genes and epigenetic variables, which may impact the incidence, prognosis, and molecular of SPBC [5]. Carriers of familial

cancer genes, such as BRCA1/2, were shown to have a greater cumulative risk of BC and ovarian cancer [6, 7]. Once diagnosed with BC, 40% of patients will develop contralateral breast cancer within 20 years [7], whereas SPBC will occur in 12% of ovarian cancer patients who are BRCA1 carriers and 2% of those who are BRCA2 carriers within 10 years [8]. Other groups of germ line mutations, including ATM, TP53, CHEK2, BARD1, PTEN, NF1 and MSH6, are highly related with BC and other familial cancers, and it is believed that these mutations can raise the risk of SPBC [9-11]. Other risk factors, including therapies, lifestyle, and environmental variables, were also been proven to influence the development of SPBC [12]. However, there is a paucity of data examining the effect of various first primary cancers on the overall survival (OS), molecular subtype, and incidence of SPBC.

Clinically, patients with multiple tumors frequently concern about their prognosis. For such intricate issues, nomograms often have low prediction accuracy [13-15]. Therefore, a powerful learner is required to build an artificial intelligence (AI) model that can predict the 3- and 5-year survival rates of SPBC patients with multiple cancers. As part of this retrospective cohort analysis, we investigated the incidence rate, mortality rate, prognosis, and subtypes of SPBC in a population derived from the Surveillance, Epidemiology, and End Results (SEER) database.

Our research team conducted a rigorous comparison of various first primary cancer locations. After that, we created a prognostic prediction model using extreme gradient boosting (XGBoost). The results of this study reveal that SPBC behaves biologically differently in distinct first primary cancer locations. Additionally, it will serve as a reference for SPBC patients with multiple cancers on their prognosis and follow-up.

Materials and methods

Data source and study design

Figure 1 shows the workflow of our study's design and analyses. This study collected the US women diagnosed with breast cancer from 2000-2016 from the the Surveillance, Epidemiology, and End Results (SEER) database [SEER 18 Regs Study Data (2000-2016 changes); version 8.3.9]. This retrospective cohort study was approved by the Institutional Review Committee of the Second Affiliated Hospital of Xi'an Jiaotong University, which decided to waive informed consent because the data are available to the public and do not contain personally identifiable patient information.

From this database, data of women with BC alone and BC combined with a previous first primary cancer (FPC) were collected. In accordance with the morphological and histopathological diagnosis criteria in the International Classification of Cancer Diseases Edition III (ICD-O-3), all cancer patients had proof of their diagnoses. The prerequisites for inclusion include information on race, age, stage, hormone receptor (HR), treatment and marital status are prerequisites for inclusion. Patients with blank information mentioned above were excluded. SPC was ascertained in accordance with the SEER guidelines for determining multiple primaries: 1) the new lesion had the same histological type as the original lesion and occurred at the same time within 2 months, so it was regarded as a single lesion and not as a primary new lesion; 2) the histological type of the new lesion was different from that of the original lesion, and it occurred at the same site at the same time (within 2 months), which was regarded as the new primary tumor; 3) the presence of an achromatic lesion at the same site (2 months or more after the initial diagnosis) is considered a new primary tumor regardless of histological type, unless it is confirmed as a metastatic lesion; 4) new lesions at different sites with the same or different histological types should always be considered as new primary tumors, unless they are clearly metastatic; 5) for paired organs, only one histological type of bilateral synchronous tumor was considered as a single primary tumor; Bilateral tumors with two different histological types are considered both primary unless otherwise stated. According to the SEER database, BC will be classified as primary breast cancer (PBC) if the patient had only been diagnosed with BC. And BC will be classified as the SPBC if the patient has any other pre-existing cancer, while the pre-existing tumor is defined as the FPC of the SPBC. SPBC that was diagnosed within 6 months after the development of BC were excluded, so as the patients with

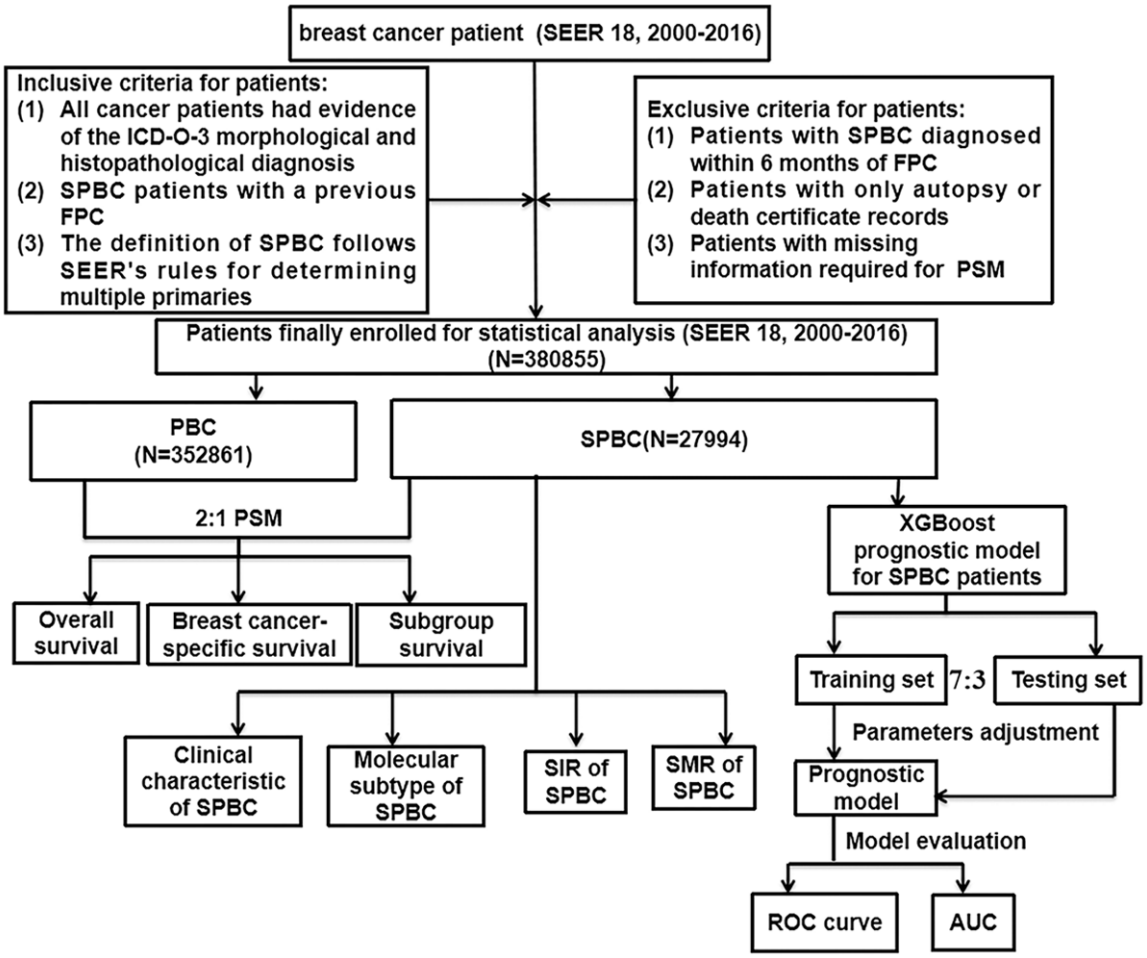


Figure 1. The flowchart describing the process of conducting the study and statistical analysis. SEER: the Surveillance, Epidemiology, and End Results; ICD-O-3: International Classification of Cancer Diseases Edition III; SPBC: second primary breast cancer; FPC: first primary cancer; PSM: propensity score matching; PBC: primary breast cancer; SIR: standardized incidence ratio; SMR: standardized mortality ratio; XGBoost: extreme gradient boosting; ROC: receiver operating characteristic curve; AUC: area under curve.

only autopsy or death certificate records. The occurrence of the FPC was regarded as the exposure factor, the occurrence of the SPBC was regarded as the target event of our observation, and then follow-up was initiated; the interval time between the occurrence of the two primary cancers was regarded as the incubation period of the SPBC. The patient's death, loss to follow-up, or December 31 of 2016 were all considered as the end of the follow-up period.

Statistical analysis

Estimation of standardized incidence/mortality ratio: To compare the relative incidence and mortality risk of SPBC with the general population, we used the SEER*Stat Multiple primary-standardized incidence ratios (MP-SIR) tool (version 8.3.8) to calculate the SIR by dividing the observed number of SPBC by the corresponding total person-years of follow-up and then multiplying by 10,000. All survivors' percentage contributions from each FPC and SPBC combination to the overall incidence of SPBC were calculated, along with the 95% confidence interval (95% CI). The percentage contribution of each FPC and SPBC combination to the overall incidence of SPBC were calculated among all survivors, along with the 95% confidence interval (95% CI). The SIR (95% CI) was calculated as the ratio of the observed to the expected number of SPBC. The expected numbers of SPBC were computed by a weight-

ed sum of stratified incidence rates by latency from the reference population, which may include multiple primary cancers in a single individual. The primary outcomes were incidence (per 10,000 person-years) and relative risk of developing SPBC among FPC survivors (standardized incidence ratio [SIR]). Similar approaches were conducted to calculate the mortality and SMR of SPBC. Subsequently, we stratified the survivors in different groups according to whether they had received radiotherapy or chemotherapy for the FPC to calculate their relative risk of death from SPBC. Results that were statistically different were marked with * for easy viewing.

Propensity score matching: In order to minimize clinical differences and statistical bias when comparing the prognosis of PBC and SPBC patients, propensity score matching (PSM) was performed using a 'nearest' method. Information on clinical characteristics of patients with PBC and SPBC, such as stage, grade, historical type, HR, age, primary site, laterality, marital status, surgery, chemotherapy, radiotherapy and race were added in order of importance and matched at a ratio of 1:2.

Kaplan-Meier (KM) for survival analysis: The PSM-adjusted data for 10-year all-cause and BC-specific survival were collected to compare patients with SPBC and those with PBC by KM analysis. Then, we stratified the SPBC patients into subgroups based on where their FPC was located, and compared their survival to that of PBC patients. We also created PSM-adjusted single-factor COX regression models to further compare the risk of cancer-specific death risk in SPBC patients by sites of FPC. Patients with PBC were assigned to the negative control group.

Molecular subtype analysis of BC: We listed the distribution of molecular subtypes of SPBC with various FPC and compared the differences with PBC by chi-square test. The groups with statistical differences were marked with * in the graph. Since our study was exploratory, we did not use the "Bonferroni" method for correction. The percentages of hormone-receptor positive (HR+) BC at different ages of PBC and SPBC were further examined. And cross-sectional comparisons were made between SPBC with first primary breast or ovarian cancer, which were closely related to the estrogen hormone and genetic factors, and SPBC with other FPC.

XGBoost: XGBoost is a distributed gradient enhancement algorithm optimized based on CART and linear classifiers. The principle of XGBoost algorithm can be summarized as follows: feature vector with the corresponding (output) category yi:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F,$$

Feature selection: univariate COX analysis was performed on the clinical features extracted from the SEER database. To predict 3- and 5-year OS of SPBC, our machine learning model incorporated the clinical characteristics demonstrating statistically significant, including incubation period of SPBC (month since index), primary site of SPBC, histological type of SPBC, hormone-receptor status of SPBC, stage of SPBC, grade of SPBC, age at diagnosis of SPBC and treatment information of SPBC, site of FPC, stage of FPC, grade of FPC, age at diagnosis of FPC, treatment information of FPC, race and marital status. These analyses were performed before the exclusion of patients who survived but lived less than 3 or 5 years at the follow-up cut-off date. Prior to running the training program, a response variable was obtained for survival information, in which 1= survival and 0= death. One-hot encoding was performed for the three multi-classified variables (marital status, race, sites of FPC). Patients were randomly divided into training sets and test sets in a ratio of 7:3. And we compared the performance of random forest (RF), logistic regression (LR), artificial neural network (ANN), the support vector machine (SVM), decision tree (ID3), K-nearest neighbors (KNN) and XGBoost on training sets and test sets. Receiver operating characteristic (ROC) analysis and area under the ROC curve (AUC) were used for the evaluation of the model. All statistical analyses were performed using R software (version 4.0.2). A bilateral tail value of less than 0.05 were considered statistically significant.

Results

Cancer survivors' characteristics

The cohort study included 380,855 female BC patients in the database, including 27,994

Table 1. Standardized incidence ratios (SIR) of second primary breast cancer (SPBC) in cancer survivors by site of first primary cancer

Site of first primary cancer (FPC)	SIR	95% CI	Mean Age at FPC (Year)	Mean Age at SPBC (Year)	Mean interval time between two primaries (Month)
All Sites	1.16*	(1.15-1.17)	62.24	67.02	4.78
Breast	1.39*	(1.37-1.41)	60.76	65.23	4.47
Thyroid	1.19*	(1.13-1.24)	48.61	61.36	12.75
Melanoma of Skin	1.10*	(1.05-1.15)	58.5	66.39	7.89
Uteri	1.08*	(1.05-1.12)	61.89	69.43	7.54
Urinary	1.03	(0.98-1.07)	66.07	71.74	5.67
Ovarian	0.99	(0.92-1.06)	59.26	64.94	5.68
Oral/Pharynx	0.95	(0.87-1.04)	60.63	68.10	7.47
Colon/Rectum/Anus	0.96*	(0.93-0.99)	65.49	73.04	7.55
Lung and Bronchus	0.92*	(0.88-0.97)	67.45	71.14	3.69
Lymphoma/leukemia	0.84*	(0.81-0.88)	58.46	70.33	11.87

*indicates that SIR is statistically significant. SPBC: second primary breast cancer; SIR: Standardized incidence ratios; FPC: first primary breast cancer.

patients with SPBC and 35,2861 patients with PBC. We first investigated the incidence of SPBC in cancer survivors, of all first primary sites in SPBC patients, the breast topped the list (14327; 51.18%), followed by the uteri (2629; 9.39%), colon/rectum/anus (2445; 8.73%), lymphoma/leukemia (1376; 4.92%), urinary (1278; 4.57%), thyroid (1256; 4.49%), melanoma of skin (1234; 4.41%), lung/bronchus (971; 3.47%), ovarian (557; 1.99%), and oral cavity/pharynx (383; 1.37%) (Supplementary Figure 1A). Then we analyzed the causes of death in SPBC patients. BC (44.55%) was the leading cause of death in SPBC patients, followed by heart disease (10.96%), lung/bronchus cancer (4.29%), miscellaneous cancers (3.63%), chronic obstructive pulmonary disease (3.26%), cerebrovascular disease (2.84%), lymphoma/leukemia (2.70%), colon/rectum/anus cancer (2.28%), Alzheimer's disease (1.68%), and ovarian cancer (1.48%) (Supplementary Figure 1B). Patients with thyroid, breast and ovarian cancers have an earlier median age of onset of SPBC. In contrast, survivors of colon/rectum/anus, urinary, lymphoma/leukemia and lung/bronchus cancers appear to develop SPBC much later (Supplementary Figure 2).

SIR of SPBC

There was a considerable increase in SIR at all sites (SIR=1.16, 95% CI=1.15-1.17, P<0.05). Among the top ten FPC, BC (SIR=1.39, 95% CI=1.37-1.41, P<0.05), uteri cancer (SIR=1.08, 95% CI=1.05-1.12, P<0.05), thyroid cancer (SIR=1.19, 95% CI=1.13-1.24, P<0.05), melanoma of skin (SIR=1.10, 95% CI=1.05-1.15, P<0.05) patients demonstrated a higher risk of SPBC compared to the general population, while patients with ovarian cancer (SIR=0.99, 95% CI=0.92-1.06, P>0.05), urinary cancer (SIR=1.03, 95% CI=0.98-1.07, P>0.05) and oral/pharynx cancer (SIR=0.95, 95% CI=0.87-1.04, P>0.05) were not discovered to have an increased risk of developing this cancer. However, the risk was reduced for patients with first primary colon/rectum/anus (SIR=0.96, 95% CI=0.93-0.99, P<0.05), lymphoma/leukemia (SIR=0.84, 95% CI=0.81-0.88, P<0.05) and lung/bronchus (SIR=0.92, 95% CI=0.88-0.97, P<0.05) cancers. In addition, we observed that the mean incubation period of the SPBC was over 12 years for thyroid cancer survivors, while it was shortest in lung/bronchus cancer survivors, only 3.69 years (**Table 1**).

SMR of SPBC

The SMR of SPBC was much higher in lung/bronchus cancer survivors compared to the general population (SMR=1.71, 95% CI=1.58-1.85, P<0.05). In contrast, the SMR of SPBC was much lower in other cancer survivors compared to the general population. The effect of treatment of FPC on SMR of SPBC was also of interest, and further studies revealed that among lung/bronchus cancer survivors, radio-

Table 2. SMR of SPBC in cancer survivors by site and treatment of FPC

Site of FPC	SMR of SPBC (95% CI)	Treatment of first primary cancer			
		No radiation or chemotherapy	Radiation	Chemotherapy	Radiation and chemotherapy
Thyroid	0.52* (0.44-0.62)	0.54* (0.43-0.67)	0.48* (0.40-0.70)	/	/
Melanoma of Skin	0.59* (0.51-0.68)	0.58* (0.51-0.67)	1.04 (0.28-4.03)	/	/
Uteri	0.66* (0.60-0.73)	0.62* (0.56-0.71)	0.98 (0.73-1.26)	1.27 (0.86-1.74)	0.76 (0.50-1.46)
Urinary	0.51* (0.44-0.58)	0.50* (0.46-0.59)	/	0.50* (0.24-0.78)	/
Colon/Rectum/Anus	0.55* (0.51-0.60)	0.56* (0.51-0.62)	0.59 (0.28-1.43)	0.56* (0.45-0.67)	0.49* (0.32-0.64)
Ovarian	0.65* (0.53-0.79)	0.61* (0.39-0.80)	/	0.66* (0.55-0.86)	/
Lymphoma/Leukemia	0.51* (0.46-0.57)	0.49* (0.37-0.60)	0.39* (0.16-0.58)	0.58* (0.44-0.67)	0.44* (0.32-0.76)
Oral/Pharynx	0.77* (0.61-0.95)	0.51* (0.35-0.72)	1.08 (0.72-1.50)	2.57 (0.28-8.42)	1.11 (0.64-1.58)
Lung/Bronchus	1.71* (1.58-1.85)	1.05 (0.91-1.71)	2.25* (1.67-2.59)	2.42* (1.96-2.78)	3.01* (2.49-3.29)

*indicates that SMR is statistically significant. SPBC: second primary breast cancer; SMR: Standardized mortality ratios; FPC: first primary breast cancer.

therapy (SMR=2.25, P<0.05) and chemotherapy (SMR=2.42, P<0.05) greatly increased SMR of SPBC compared to patients who did not receive radiotherapy or chemotherapy (SMR= 1.05, P>0.05). However, radiotherapy for the first primary cancer can't always increase the SMR of breast cancer. The effect of radiotherapy on SPBC diverged in patients with first primary thyroid cancer (SMR: radiation 0.48 vs. no radiation or chemotherapy 0.54) and lymphoma/leukemia (SMR: radiation 0.39 vs. no radiation or chemotherapy 0.49), with a decrease in SMR. Notably, in other cancer survivors such as uteri cancer (SMR: chemotherapy 1.27 vs. no radiation or chemotherapy 0.62), ovarian cancer (SMR: chemotherapy 0.66 vs. no radiation or chemotherapy 0.61), lymphoma/leukemia (SMR: chemotherapy 0.58 vs. no radiation or chemotherapy 0.49), and oral/pharynx cancer (SMR: chemotherapy 2.57 vs. no radiation or chemotherapy 0.51), compared to those who did not receive chemotherapy, SPBC patients with previous chemotherapy had an elevated SMR (**Table 2**).

Survival of SPBC in different FPC patients

Differences in clinical features, pathology, and treatment lead to a significant bias in direct comparisons of prognosis. With PSM, this difference can be virtually eliminated. Overall, the proportion of PBC patients who died from BC, other malignancies and non-oncological disease were 11.45%, 0.92% and 16.7%, respectively. While 11.92% of SPBC patients died from BC, 5.66% died from other cancers and 9.18% died from non-oncology related diseases such as cardiovascular diseases (**Table 3**). In the PSM-adjusted survival study, OS was lower in SPBC patients compared to PBC patients (hazard ratio (HR) =1.43, 95% CI= 1.39-1.47, P<0.001) (**Figure 2A**). And SPBC patients had an increased risk of BC-specific death compared to PBC patients (HR=1.38, 95% CI=1.32-1.44, P<0.001) (**Figure 2B**). Subsequently, we performed a subgroup OS analysis based on the sites of the FPC. OS was better in SPBC patients with first primary thyroid cancer (HR=0.64, 95% CI=0.55-0.75, P<0.001) than in PBC patients. Whereas in melanoma of the skin (HR=0.89, 95% CI= 0.78-1.02, P=0.099) and in uteri cancer survivors (HR=1.01, 95% CI=0.92-1.10, P=0.87) the prognosis of SPBC was not statistically different from that of PBC patients (**Figure 3A-C**). However, the prognosis of SPBC patients got poorer if they carried additional FPC (**Figure 3D-I**). Among them, patients with lung/bronchus cancer (HR=2.77, 95% CI=2.52-3.05, P<0.001) had the worst prognosis (**Figure 3J**).

SPBC patients with FPC had an increased cancer-specific death risk (HR=1.92, 95% CI= 1.85-1.99, P<0.001) compared with PBC patients. All-cause and cancer-specific deaths were compared and analyzed in combination with the stages and causes of death. For thyroid cancer, melanoma of skin and uteri cancer, an earlier stage and better prognosis of the FPC were associated with lower risk of cancer-specific death and better overall prognosis. For lung cancer, ovarian cancer, and lymphoma/leukemia, the later stage of the FPC was, the

Table 3. Baseline characteristics of breast cancer patients included from SEER data cohort and between-group comparisons (primary breast cancer vs. second primary breast cancer)

Characteristics	Unmatched Cohort			1:2 propensity score matched (PSM) Cohort		
	PBC N=352861 (%)	SPBC N=27994 (%)	Unadjusted *P* value	PBC N=55984 (%)	SPBC (N=27994)	PSM-adjusted *P* value
Age			<0.001			0.291
20-39	21732 (6.16)	572 (2.04)		1040 (1.86)	572	
40-59	160508 (45.49)	8045 (28.74)		15923 (28.44)	8045	
60-69	82526 (23.39)	7756 (27.71)		15573 (27.82)	7756	
70-79	55837 (15.82)	7083 (25.30)		14368 (25.66)	7083	
80+	32258 (9.14)	4538 (16.21)		9080 (16.22)	4538	
Race			<0.001			0.096
White	282509 (80.06)	23032 (82.27)		46383 (82.85)	23032	
Black	30590 (8.67)	3003 (10.73)		5860 (10.47)	3003	
Other	39762 (11.27)	1959 (7.00)		3741 (6.68)	1959	
Marriage status			<0.001			0.612
Married	200791 (56.90)	14238 (50.86)		28396 (50.72)	14238	
Others	101765 (28.84)	10280 (36.72)		20731 (37.03)	10280	
Single	50305 (14.26)	3476 (12.42)		6857 (12.25)	3476	
Grade			<0.001			0.318
Well	78435 (22.23)	6403 (22.87)		12586 (22.48)	6403	
Moderately	149643 (42.41)	12302 (43.95)		24582 (43.91)	12302	
Poorly	124783 (35.36)	9289 (33.18)		18816 (33.61)	9289	
Laterality			0.402			0.740
Left	179201 (50.79)	14144 (50.53)		28354 (50.65)	14144	
Right	173660 (49.21)	13850 (49.47)		27630 (49.35)	13850	
Primary site			<0.001			0.961
Upper-outer quadrant	122799 (34.80)	9047 (32.32)		18127 (32.38)	9047	
Upper-inner quadrant	41375 (11.73)	3303 (11.80)		6628 (11.84)	3303	
Lower-inner quadrant	19768 (5.60)	1702 (6.08)		3342 (5.97)	1702	
Lower-outer quadrant	25699 (7.28)	2092 (7.47)		4134 (7.39)	2092	
Central portion	17394 (4.93)	1426 (5.09)		2807 (5.01)	1426	
other site	125826 (35.66)	10424 (37.24)		20946 (37.41)	10424	
Histological type			<0.001			0.485
IDC	265209 (75.16)	20586 (73.54)		41426 (74.00)	20586	
ILC	26061 (7.38)	2634 (9.41)		5156 (9.21)	2634	
Mixed	37538 (10.64)	1763 (6.30)		3521 (6.29)	1763	
Other	24053(6.82)	3011 (10.75)		5881 (10.50)	3011	
Stage			<0.001			0.359
I	166249 (47.11)	15869 (56.69)		31747 (56.71)	15869	
II	127280 (36.07)	8363 (29.87)		16835 (30.07)	8363	
III	45867 (13.00)	2542 (9.08)		5106 (9.12)	2542	
IV	13465 (3.82)	1220 (4.36)		2296 (4.10)	1220	
Surgery			<0.001			0.005
Yes	336993 (95.50)	26272 (93.85)		52811 (94.33)	26272	
No	15868 (4.50)	1722 (6.15)		3173 (5.67)	1722	
Radiation			<0.001			0.239
Yes	191245 (54.20)	10628 (37.97)		21489 (38.38)	10628	
No	161616 (45.80)	17366 (63.03)		34495 (61.62)	17366	
Chemotherapy			<0.001			0.263
Yes	154833 (43.88)	8423 (30.09)		16635 (29.71)	8423	
No	198028 (56.12)	19571 (69.91)		39349 (70.29)	19571	
Hormone receptor			<0.001			0.015
Positive	288292 (81.70)	21817 (77.93)		44039 (78.66)	21817	
Negative	64569 (18.30)	6177 (22.07)		11945 (21.34)	6177	
Cause of death			<0.001			<0.001
By breast cancer	41165 (11.67)	3337 (11.92)		6411 (11.45)	3337 (11.92)	
By other cancer	2365 (0.67)	1584 (5.66)		515 (0.92)	1584 (5.66)	
Other non-tumor-related cause	37655 (10.67)	2569 (9.18)		9348 (16.7)	2569 (9.18)	
Alive	271676 (76.99)	20504 (73.24)		39710 (70.93)	20504 (73.24)	

PBC: primary breast cancer; SPBC: second primary breast cancer; IDC: infiltrating ductal carcinoma; ILC: infiltrating lobular carcinoma.

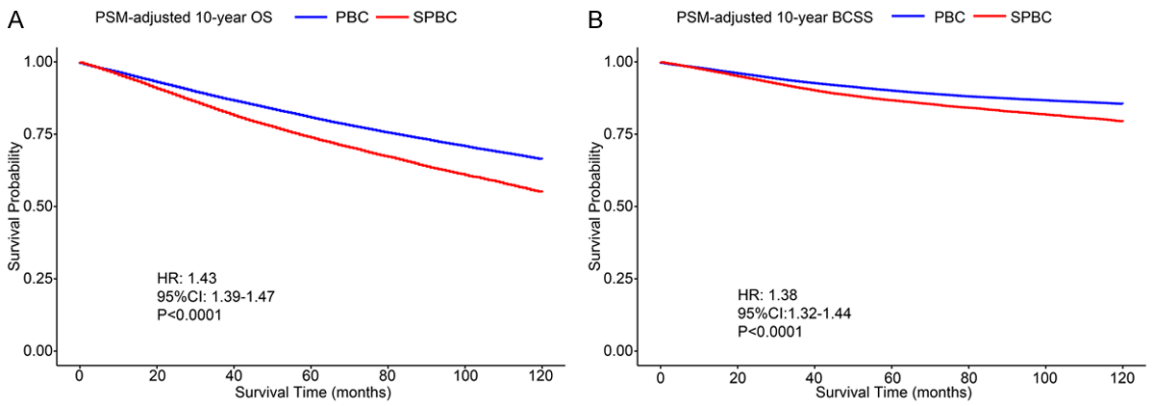


Figure 2. Propensity score matching (PSM) adjusted 10 years OS and BCSS of PBC and SPBC patients. A: OS of PBC and SPBC patients; B: BCSS of PBC and SPBC patients. PBC: primary breast cancer; SPBC: second primary breast cancer; OS: overall survival; BCSS: breast cancer-specific survival; HR: Hazard Ratio; 95% CI: 95% confidence interval.

higher the risk of tumor-related death was, and the poorer the overall prognosis was (**Table 4**).

Molecular subtypes of SPBC in different FPC

We investigated the differences between SPBC and PBC in terms of molecular subtypes (**Figure 4A**). Compared to PBC, there was a lower proportion of HR+/HER2+ subtype (9.30% vs. 11.35%, P<0.05) and a higher proportion of HR-/HER2- subtype (13.31% vs. 11.41%, P<0.05) in SPBC, with little difference in HR+/HER2- and HR-/HER2+ subtypes. Similar proportions of HR-/HER2- subtype of SPBC were observed among survivors of first primary lung/bronchus, lymphoma/leukemia and oral/pharynx cancers, while these proportions were significantly lower in first primary thyroid, melanoma of the skin, uteri, urinary and colon/rectum/anus cancers survivors. This finding was consistent with the results of the survival analysis described above. In particular, we observed a much higher proportion of SPBC with HR-/HER2- subtypes among survivors of first primary breast (16.07% vs. 11.41%, P<0.05) and ovarian cancer (18.48% vs. 11.41%, P<0.05). This discrepancy caught our attention, and as hormone receptor status is somewhat age-related, we further compared the proportion of hormone HR+ subtype of BC by age (**Figure 4B**). The results showed that the percentage of HR+ was much lower in SPBC than in PBC among survivors of first primary breast or ovarian cancers before the age of 65 years, but it was similar to or slightly higher than in PBC after the age of 75 years. In comparison, the percentage of HR+ was higher in SPBC than PBC at all ages in survivors of other FPC. Among breast and ovarian cancer survivors, the percentage of HR+ subtype was considerably higher in SPBC patients aged 45 to 55 years compared to those aged under 45 years, however, there was no such difference among survivors of other FPC. Additionally, the percentage of HR+ subtype increased progressively with age while the growth rate decreased.

Subsequently, we divided the FPC into subgroups according to whether they were treated with radiotherapy or chemotherapy, and studied the proportion of molecular subtypes of SPBC in each group. Notably, the proportion of HR-/HER2- subtype of SPBC was much lower (6.29% vs. 11.20%, P<0.05) following radioisotopes therapy for the first primary thyroid cancer, and the difference was statistically significant (**Table 5**).

Evaluation of machine learning prognostic models

In view of the above findings, we tried to create some predictive models to estimate 3-year and 5-year prognoses of SPBC patients. We started by performing univariate Cox analysis (Supplementary Table 1), followed with test and adjustment of the models repeatedly to determine the available parameters (Supplementary R Code). The ROC curves of the predictions for the training and the validation sets were constructed and the corresponding AUC was calculated. Our XGBoost model performed

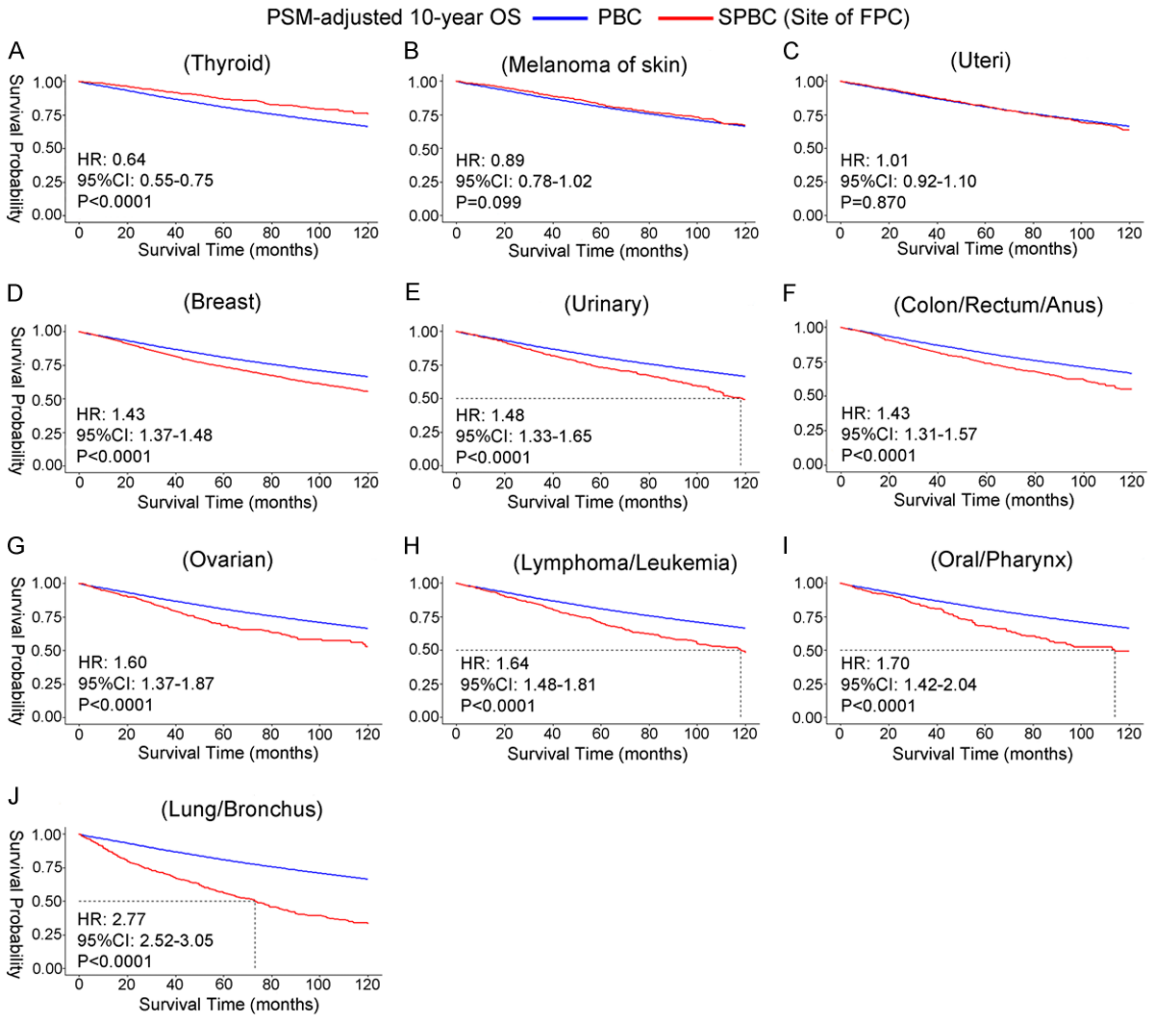


Figure 3. Propensity score matching (PSM) adjusted 10-year OS of PBC and SPBC patients (Stratified by sites of first primary cancer). A: OS of SPBC patients with first primary thyroid cancer; B: OS of SPBC patients with first primary melanoma of skin; C: OS of SPBC patients with first primary uteri cancer; D: OS of SPBC patients with first primary breast cancer; E: OS of SPBC patients with first primary urinary cancer; F: OS of SPBC patients with first primary colon/rectum/anus cancer; G: OS of SPBC patients with first primary ovarian cancer; H: OS of SPBC patients with first primary lymphoma/leukemia; I: OS of SPBC patients with first primary oral/pharynx cancer; J: OS of SPBC patients with first primary lung/bronchus cancer. PBC: primary breast cancer; SPBC: second primary breast cancer; OS: overall survival; HR: Hazard Ratio; 95% CI: 95% confidence interval.

better in predicting the 3-year and 5-year survival of SPBC patients (3-year: AUC=0.817; 5-year: AUC=0.825) (**Figure 5A** and **5B**), compared to RF (3-year: AUC=0.800; 5-year: AUC=0.799), LR (3-year: AUC=0.810; 5-year: AUC=0.789), ANN (3-year: AUC=0.815; 5-year: AUC=0.787), SVM (3-year: AUC=0.623; 5-year: AUC=0.711), ID3 (3-year: AUC=0.643; 5-year: AUC=0.684) and KNN models (3-year: AUC=0.601; 5-year: AUC=0.615) (**Table 6**). We also assessed the ranking of clinical characteristics in terms of their importance in the model (**Figure 5C** and **5D**). It showed that stage of SPBC is the most important factor in patient survival.

Discussion

SPBC has a much higher SIR and mortality rate among all female cancer survivors compared to PBC. And it accounted for 17% of all SPC among women in the United States from 1992 to 2011 [1]. Previous research has shown that early primary tumors have a variety of functions in the morbidity of SPBC patients [1]. However, the available explanations for higher

Table 4. All causes and cancer-related deaths of SPBC at different first primary sites

Site of first primary cancer	All-cause HR (95% CI)	Cancer-specific HR (95% CI)	Stage of first primary cancer			Died of the first primary cancer	Died of breast cancer
			I/II (%)	III (%)	IV (%)	(%)	(%)
All site	1.43 (1.39-1.47)	1.92 (1.85-1.99)	84.6	11.55	3.85	/	/
Thyroid	0.64 (0.55-0.75)	0.89 (0.74-1.09)	78.92	15.54	5.54	0.56	6.05
Melanoma of the skin	0.89 (0.78-1.02)	1.21 (1.02-1.42)	95.19	4.12	0.69	2.02	7.1
Uteri	1.01 (0.92-1.10)	1.22 (1.09-1.37)	90.03	8.19	1.77	1.79	6.85
Breast	1.42 (1.37-1.48)	2.00 (1.91-2.10)	89.79	8.27	1.94	/	15.84
Urinary	1.48 (1.33-1.65)	1.70 (1.48-1.96)	87.31	9.88	2.8	3.68	8.37
Colon/Rectum/Anus	1.48 (1.37-1.60)	1.68 (1.52-1.86)	70.1	26.63	3.28	4.96	7.76
Ovarian	1.61 (1.38-1.88)	2.69 (2.27-3.19)	54.23	32.69	13.08	13.46	6.1
Lymphoma/Leukemia	1.64 (1.48-1.81)	2.13 (1.89-2.42)	53.04	15.39	31.59	8.18	8.69
Oral/Pharynx	1.7 (1.42-2.04)	2.43 (1.95-3.02)	60.28	14.16	25.57	4.04	8.59
Lung/Bronchus	2.78 (2.53-3.06)	3.57 (3.17-4.03)	70.67	19.71	9.62	15.96	9.58

HR: Hazard Ratio.

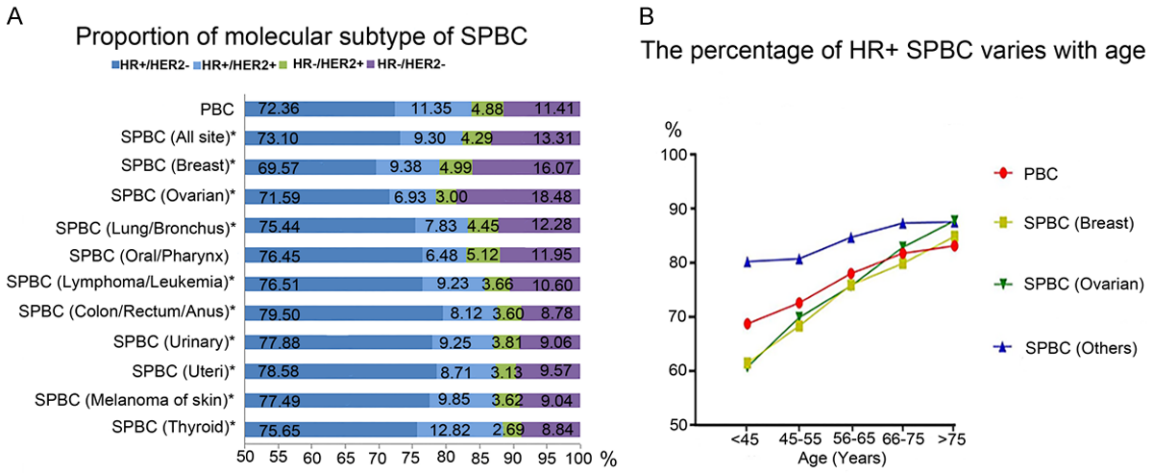


Figure 4. Molecular subtypes of SPBC in different sites of first primary cancer. HR+: Hormone receptor positive; HR-: Hormone receptor negative; HER2+: human epidermal growth factor receptor-2 positive; HER2-: Human epidermal growth factor receptor-2 negative; PBC: Primary breast cancer; SPBC: Second primary breast cancer; SPBC (Site of first primary cancer); * indicates that the differences of the distribution of molecular subtype are statistically significant.

SIR are insufficient, such as the fact that genetic variables increase the risk of SPBC in patients with melanoma of the skin, despite the fact that genetic mutations are uncommon. The interpretation of elevated SIR in thyroid cancer patients is as broad and ambiguous, as the terminology it used [12]. It is crucial to assess the SIR and prognosis of SPBC since it is a clinical problem that is often discussed. The impact of FPC on the molecular subtypes and prognosis of SPBC patients is yet to be studied in depth. Our study has sought to explain the survival disparities between patients with PBC and SPBC by conducting a comprehensive analysis at various levels. For example, SPBC patients are older in age, with a higher proportion of blacks and higher divorce rates in their demographics. Studies have shown that these basic characteristics have a considerable negative impact on their prognosis [16, 17]. In addition to these, SPBC has a 3.73% decrease in the proportion of HR+ BC compared to PBC. Hormone receptor negativity is usually associated with a poor prognosis for

Table 5. Proportion of molecular subtype of second primary breast cancer (SPBC)

Sites of first primary cancer	With/Without Radiotherapy					*P* Value	With/Without Chemotherapy					*P* Value
	Total	HR-/HER2- (%)	HR-/HER2+ (%)	HR+/HER2- (%)	HR+/HER2+ (%)		Total	HR-/HER2- (%)	HR-/HER2+ (%)	HR+/HER2- (%)	HR+/ HER2+ (%)	
Thyroid	556/598	6.29/11.20	2.70/2.68	75.90/75.42	15.11/10.70	0.007	4/1150	0/8.87	25.00/2.61	50.00/75.74	25.00/12.78	/
Melanoma of skin	7/988	28.57/8.91	0/3.64	57.14/77.63	14.29/9.82	/	4/991	0/9.08	0/3.63	100/77.40	0/9.89	/
Uteri	565/1639	10.27/9.33	2.65/3.29	79.47/78.28	7.61/9.09	0.56	289/1915	12.46/9.14	2.08/3.29	77.51/78.75	7.96/8.83	0.227
Urinary	11/1038	18.18/8.96	0/3.85	72.73/77.94	9.09/9.25	/	78/971	6.41/9.27	2.56/3.91	80.77/77.65	10.26/9.17	0.733
Colon/Rectum/Anus	258/1576	8.91/8.76	1.55/3.93	77.91/79.76	11.63/7.55	0.053	582/1252	8.42/8.95	3.26/3.75	78.52/79.95	9.79/7.35	0.332
Ovarian	8/425	12.50/18.59	0/3.06	50.00/72.00	37.50/6.35	/	289/144	21.45/12.50	3.11/2.78	69.20/76.39	6.23/8.33	0.123
Lymphoma/Leukemia	184/910	9.78/10.77	4.89/3.41	73.91/77.03	11.41/8.79	0.494	551/543	11.07/10.13	3.27/4.05	76.59/72.74	9.07/9.39	0.829
Oral/Pharynx	142/151	9.86/13.91	8.45/1.99	76.06/76.82	5.63/7.28	0.063	65/228	7.69/13.16	10.77/3.51	75.38/76.75	6.15/6.58	0.122
Lung/Bronchus	190/551	14.74/11.43	6.32/3.81	70.00/77.31	2.29/7.44	0.199	213/528	16.90/10.42	5.63/3.98	69.95/77.65	7.51/7.95	0.061

Stratified by site of first primary cancer, radiotherapy and chemotherapy.

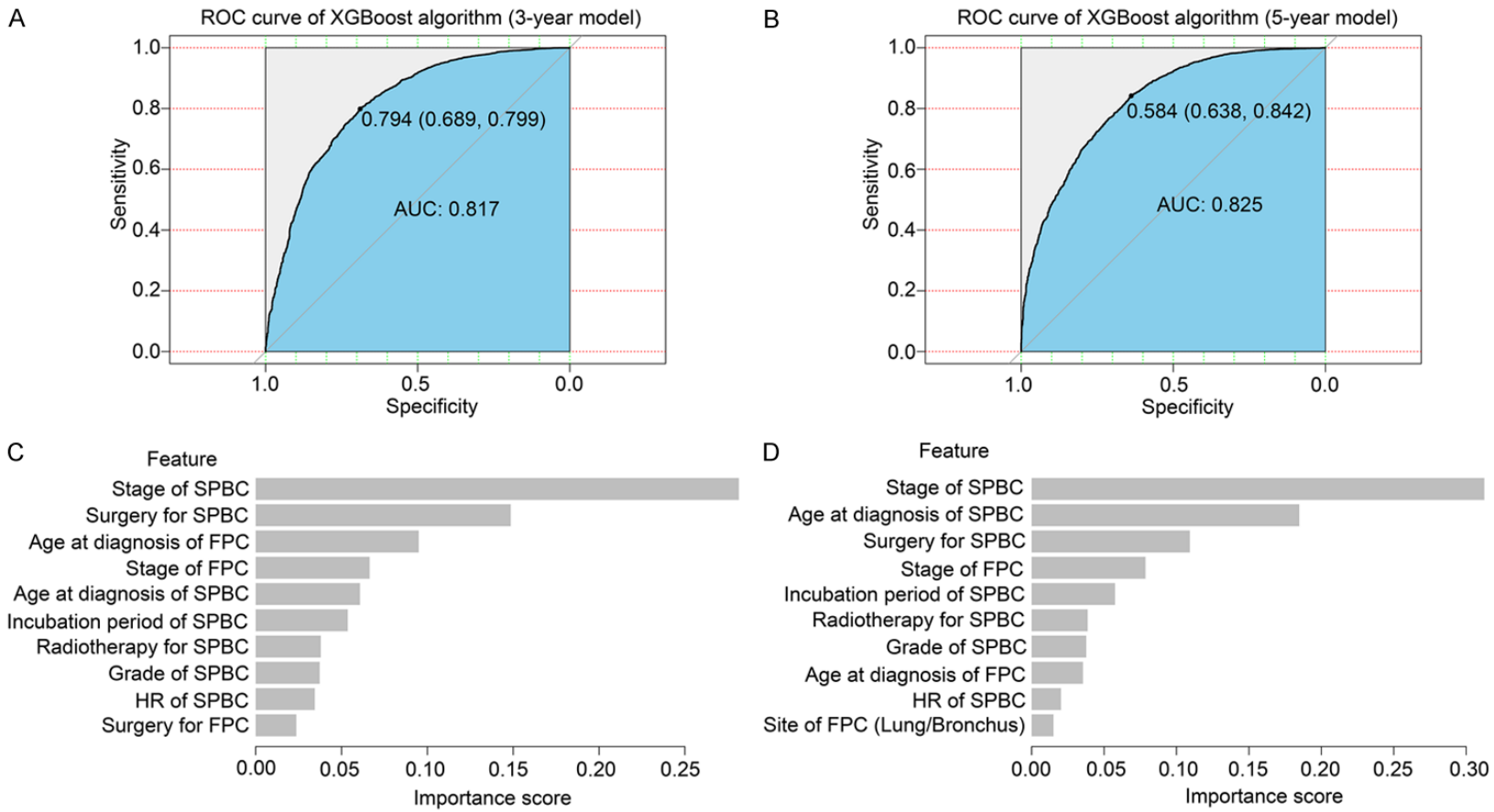


Figure 5. Evaluation of XGBoost prognostic models. ROC curve: receiver operating characteristic curve; AUC: area under curve; SPBC: second primary breast cancer; FPC: first primary cancer; HR: hormone receptor.

Table 6. Performance of prognostic models built by machine learning algorithms on test sets

	AUC	
	3-year survival (test data)	5-year survival (test data)
XGBoost	0.817	0.825
RF	0.800	0.799
LR	0.810	0.789
ANN	0.815	0.787
ID3	0.643	0.684
SVM	0.623	0.711
KNN	0.601	0.615

AUC: area under curve; XGBoost: extreme gradient boosting; ID3: iterative dichotomiser 3; SVM: support vector machines; KNN: k-nearest neighbor; RF: random forest; LR: logistic regression; ANN: artificial neural network.

BC [18]. In terms of treatment, a lower proportion of patients with SPBC received surgery, radiotherapy and chemotherapy than those with PBC. Nevertheless, clinical features, pathological features and therapy variables all have an impact on the prognosis of SPBC. It would take a lot of effort and might be statistically incorrect to compare the prognosis of patients with SPBC and PBC directly.

When we performed PSM to further balance clinical and pathological information and differences in treatment patterns, we found that mortality rates in SPBC patients remained greater than in PBC patients. More critically, the prognosis of SPBC patients was greatly overestimated due to variations in median follow-up times. After excluding differences in the clinical characteristics of BC patients, the top 10 FPC were assessed in SPBC patients, who accounted for approximately 95% of all patients. As a result, 51.18% of SPBC patients had first primary breast cancer, resulting in the highest SIR and poor prognosis. This suggests that there is a direct connection between first primary breast cancer and SPBC. Genetic mutations and genetic factors play a key role in this process. Studies show that BRCA1 mutations are common in triple-negative breast cancer [19, 20]. Another factor is that approximately 75% of BRCA1-associated breast cancers are hormone receptor negative [21-23]. Based on studies of Japanese breast cancer patients, there is a clear association between BRCA1 mutations and familial inheritance in young BC patients [24, 25]. Because of genetic factors, we believe that patients with BRCA1 mutations are prone to develop the disease at a younger age. Furthermore, we believe that the proportion of HR+ BC is significantly lower in younger patients, which adversely affects the prognosis of patients. Therefore, BC patients with a BRCA1/2 mutation and a family history of the disease may require prophylactic contralateral mastectomy. Consistent findings were also observed in the first primary ovarian cancer group, again associated with genetic mutations [26-28]. Unexpectedly, no significant increase was found in SIR in SPBC among ovarian cancer survivors, which has not been previously reported in the literature. A number of factors were hypothesized for this finding, including the advanced stage and poor prognosis of some ovarian cancer patients and the protective effect of oophorectomy on SPBC development [29, 30]. The proportion of ovarian cancer survivors with the HR-/HER2- subtype of SPBC is much higher than that of patients with PBC, and even higher than that of survivors of first primary breast cancer. In addition to the genetic factors mentioned above, studies have shown that ovarian cancer surgery, chemotherapy and suppression of ovarian function may all lead to a significant decrease in hormone levels and reduce the risk of HR+ BC [31, 32].

Thyroid cancer survivors who developed SPBC have the second highest SIR, but their prognosis is even better than that of patients with PBC who have only one tumor. This has been previously reported in the literature [33], but no reasonable epidemiological explanation has been proposed. Our hypothesis is based on several facts. Firstly, the median age of onset of the two primary cancers is much younger, and regular tumor surveillance facilitates early tumor discovery and prompt treatment. Secondly, we found a higher proportion of HR+ subtypes in SPBC compared with PBC, which is associated with a better prognosis. In addition, patients with primary thyroid cancer were treated with radioactive iodine-131 in 48% of cases, while there was no evidence that this group of patients had a higher chance of developing SPBC [34-36]. For the first time in our study, a statistically significant decrease of 43.84% in the proportion of HR-/HER2- subtypes was found in SPBC patients with first pri-

mary thyroid cancer treated with iodine-131. Furthermore, radioisotope treatment reduced the SMR of SPBC in thyroid cancer survivors. Is it possible that iodine-131 treatment influences the prognosis of SPBC by affecting the molecular subtypes of the disease? This finding could have significant clinical and molecular implications and may also serve as a starting point for future research.

Compared with survivors of other cancers, survivors with lung/bronchus cancer had a much poorer overall prognosis and a much higher SMR of SPBC. Although this had been reported before, no available results can explain the reasons behind it [37]. Based on our research findings, on the one hand, these patients had an older median age of onset and a higher proportion of HR-/HER2- subtype which suggested a poorer prognosis for these patients. On the other hand, because of the short interval time of the two primary cancers and the high doses of radiotherapy and chemotherapy that lung cancer patients will get [38], patients will be in poor underlying physical conditions and will find it challenging to tolerate treatment of SPBC. Therefore, it is crucial to pay special attention to the occurrence of SPBC in lung cancer survivors to achieve early identification and treatment, promote nutritional support therapy, minimize drug toxicity and side effects, and ultimately improve the survival rate of these patients.

The proportion of HR+ SPBC in our study increased with age, which may be related to the body's estrogen levels [32]. However, there was no evidence of a bimodal distribution [39], which was comparable to that reported by Armitage and Doll [40, 41]. The estrogen-dependent proliferation of tumor cells is more pronounced in younger individuals [42, 43]. The proportion of HR+ breast cancers in SPBC was reduced following treatment of primary breast or ovarian cancer with endocrine therapy or surgical debulking [31, 32]. In addition, in the presence of genetic factors, these individuals have a tendency to develop early and are more likely to develop HR-/HER2- subtype disease [21, 22, 24]. Thus, we think that the percentage of HR+ subtypes is lower in younger SPBC patients and increases with age as the genetic and hormone-dependent aspects of the disease diminish. Furthermore, the percentage of HR+ subtypes stabilizes after the age of 65 years, which may be due to a decrease in the number of tumor-inducing cell divisions after menopause [44].

For non-BC survivors, we found a positive correlation between SIR and prognosis and the percentage of HR+ subtypes, which has not been reported in SPBC patients. For example, patients with primary thyroid cancer had higher SIR, better prognosis, and a greater percentage of HR+ subtypes in SPBC. Whereas the opposite was true for SIR, prognosis, and percentage of HR+ subtypes in SPBC in primary lung/bronchus cancer survivors. Tumor-specific mortality was mainly determined by the first and SPC, especially in patients with poorer stage and higher malignancy. In the case of primary lung cancer, for example, patients with later age of onset, shorter latency and higher tumor malignancy had a significantly higher risk of tumor-specific death and a lower SIR of SPBC. In contrast, SPBC patients with first primary thyroid cancer, have a lower risk of cancer-specific death and a higher SIR of SPBC.

However, it was difficult to estimate the 3- and 5-year survival rates of SPBC patients based on these results alone. Nomograms have been established in previous studies to predict the survival rates of BC patients whereas its accuracy is low (only about 0.7) for patients with multiple cancers [13]. The study by Luo et al. used only two machine learning algorithms to predict patient prognosis, which is clearly an insufficient number of algorithms and not accurate enough (AUC=0.737) [45]. To our knowledge, there is no literature estimating the prognosis of SPBC patients. Therefore, we developed an XGBoost model based on the clinicopathological characteristics of SPBC patients. In addition, our XGBoost model is trained with 10-fold cross-validation to improve the robustness of the model. As we evaluated, our model outperforms traditional machine learning algorithms and artificial neural network in terms of predicting the survival rate of SPBC patients, suggesting that XGBoost is superior in tackling this type of tasks. The most critical features impacting the survival rates of SPBC patients included surgery for SPBC, stage and age at diagnosis of two primary cancers, and latency period of SPBC, suggesting that SPBC may be a major cause of patient' death. But FPC still has a significant detrimental effect on the OS of SPBC patients.

This is the first comprehensive study to examine the relationship among FPC, treatment modality, epidemiological factors, molecular subtypes and SIR and SMR of SPBC, and the impact of the interaction of these factors on prognosis. Most of the previous studies conducted have been limited to epidemiological studies and interpretation of SIR, with little emphasis on systematic analysis and interpretation. Although our study is based on a large and accurate cancer database, gene-level association analysis is lacking. Genetic factors are important endogenous factors in the development of SPC, while the genotype-phenotype correlation of SPC remains unclear. Current studies on SPBC genotypes are limited and more research are needed to explain the relationship between SPBC and FPC at the genetic level. Despite these limitations, our study has important implications for patients with SPBC.

In conclusion, we investigated the relationship between female cancer patients and SPBC, compared SIR, SMR, clinical characteristics and prognosis of SPBC patients with different FPC, and constructed an XGBoost model to assess the survival of SPBC patients. Our findings revealed an important role of FPC in the development of SPBC and provided additional theoretical support for clinical follow-up and identification of SPBC.

Data analyzed during the study are openly available via the National Cancer Institute's Surveillance, Epidemiology, and End Results program (SEER) (https://SEER.cancer.gov/), or the Supplementary Raw Data.

Acknowledgements

We would like to thank all the developers of the R programming package for selflessly sharing their code. This work was supported by the National Natural Science Foundation of China (82174164; 81901886), Shaanxi Administration of Traditional Chinese Medicine (2021-ZZ-JC019) and the Fundamental Research Funds for the Central Universities (xzy012020040).

Disclosure of conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Address correspondence to: Shuqun Zhang and Yinbin Zhang, Department of Oncology, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157 West Fifth Street, Xi'an, Shaanxi, P. R. China. E-mail: shuqun_zhang1971@163.com (SQZ); 23227119@qq.com (YBZ)

References

[1] Sung H, Hyun N, Leach CR, Yabroff KR and Jemal A. Association of first primary cancer with risk of subsequent primary cancer among survivors of adult-onset cancers in the United States. JAMA 2020; 324: 2521-2535.

[2] Wright FC, Look Hong NJ, Quan ML, Beyfuss K, Temple S, Covelli A, Baxter N and Gagliardi AR. Indications for contralateral prophylactic mastectomy: a consensus statement using modified Delphi methodology. Ann Surg 2018; 267: 271-279.

[3] Bhatia S and Sklar C. Second cancers in survivors of childhood cancer. Nat Rev Cancer 2002; 2: 124-132.

[4] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022; 72: 7-33.

[5] Maxwell KN, Wenz BM, Kulkarni A, Wubbenhorst B, D'Andrea K, Weathers B, Goodman N, Vijai J, Lilyquist J, Hart SN, Slavin TP, Schrader KA, Ravichandran V, Thomas T, Hu C, Robson ME, Peterlongo P, Bonanni B, Ford JM, Garber JE, Neuhausen SL, Shah PD, Bradbury AR, DeMichele AM, Offit K, Weitzel JN, Couch FJ, Domchek SM and Nathanson KL. Mutation rates in cancer susceptibility genes in patients with breast cancer with multiple primary cancers. JCO Precis Oncol 2020; 4: PO.19.00301.

[6] US Preventive Services Task Force; Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, Caughey AB, Doubeni CA, Epling JW Jr, Kubik M, Landefeld CS, Mangione CM, Pbert L, Silverstein M, Simon MA, Tseng CW and Wong JB. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer: US preventive services task force recommendation statement. JAMA 2019; 322: 652-665.

[7] Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, Jervis S, van Leeuwen FE, Milne RL, Andrieu N, Goldgar DE, Terry MB, Rookus MA, Easton DF, Antoniou AC; BRCA1 and BRCA2 Cohort Consortium; McGuffog L, Evans DG, Barrowdale D, Frost D, Adlard J, Ong KR, Izatt L, Tischkowitz M, Eeles R, Davidson R, Hodgson S, Ellis S, Nogues C, Lasset C, Stoppa-Lyonnet D, Fricker JP, Faivre L, Berthet P, Hooning MJ, van der Kolk LE, Kets CM, Adank MA, John EM, Chung WK, Andrulis IL, Southey M, Daly MB, Buys SS, Osorio A, En-

gel C, Kast K, Schmutzler RK, Caldes T, Jakubowska A, Simard J, Friedlander ML, McLachlan SA, Machackova E, Foretova L, Tan YY, Singer CF, Olah E, Gerdes AM, Arver B and Olsson H. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA 2017; 317: 2402-2416.

[8] Domchek SM, Jhaveri K, Patil S, Stopfer JE, Hudis C, Powers J, Stadler Z, Goldstein L, Kauff N, Khasraw M, Offit K, Nathanson KL and Robson M. Risk of metachronous breast cancer after BRCA mutation-associated ovarian cancer. Cancer 2013; 119: 1344-1348.

[9] Bernstein JL; WECARE Study Collaborative Group; Concannon P. ATM, radiation, and the risk of second primary breast cancer. Int J Radiat Biol 2017; 93: 1121-1127.

[10] Shin SJ, Dodd-Eaton EB, Gao F, Bojadzieva J, Chen J, Kong X, Amos CI, Ning J, Strong LC and Wang W. Penetrance estimates over time to first and second primary cancer diagnosis in families with Li-Fraumeni syndrome: a single institution perspective. Cancer Res 2020; 80: 347-353.

[11] Breast Cancer Association Consortium; Dorling L, Carvalho S, Allen J, González-Neira A, Luccarini C, Wahlström C, Pooley KA, Parsons MT, Fortuno C, Wang Q, Bolla MK, Dennis J, Keeman R, Alonso MR, Álvarez N, Herraez B, Fernandez V, Núñez-Torres R, Osorio A, Valcich J, Li M, Törngren T, Harrington PA, Baynes C, Conroy DM, Decker B, Fachal L, Mavaddat N, Ahearn T, Aittomäki K, Antonenkova NN, Arnold N, Arveux P, Ausems MGEM, Auvinen P, Becher H, Beckmann MW, Behrens S, Bermisheva M, Białkowska K, Blomqvist C, Bogdanova NV, Bogdanova-Markov N, Bojesen SE, Bonanni B, Børresen-Dale AL, Brauch H, Bremer M, Briceno I, Brüning T, Burwinkel B, Cameron DA, Camp NJ, Campbell A, Carracedo A, Castelao JE, Cessna MH, Chanock SJ, Christiansen H, Collée JM, Cordina-Duverger E, Cornelissen S, Czene K, Dörk T, Ekici AB, Engel C, Eriksson M, Fasching PA, Figueroa J, Flyger H, Försti A, Gabrielson M, Gago-Dominguez M, Georgoulias V, Gil F, Giles GG, Glendon G, Garcia EBG, Alnæs GIG, Guénel P, Hadjisavvas A, Haeberle L, Hahnen E, Hall P, Hamann U, Harkness EF, Hartikainen JM, Hartman M, He W, Heemskerk-Gerritsen BAM, Hillemanns P, Hogervorst FBL, Hollestelle A, Ho WK, Hooning MJ, Howell A, Humphreys K, Idris F, Jakubowska A, Jung A, Kapoor PM, Kerin MJ, Khusnutdinova E, Kim SW, Ko YD, Kosma VM, Kristensen VN, Kyriacou K, Lakeman IMM, Lee JW, Lee MH, Li J, Lindblom A, Lo WY, Loizidou MA, Lophatananon A, Lubiński J, MacInnis RJ, Madsen MJ, Mannermaa A, Manoochehri M, Manoukian S, Margolin S, Martinez ME, Maurer T, Mavroudis D, McLean C, Meindl A, Mensenkamp AR, Michailidou K, Miller N, Mohd Taib NA, Muir K, Mulligan AM, Nevanlinna H, Newman WG, Nordestgaard BG, Ng PS, Oosterwijk JC, Park SK, Park-Simon TW, Perez JIA, Peterlongo P, Porteous DJ, Prajzendanc K, Prokofyeva D, Radice P, Rashid MU, Rhenius V, Rookus MA, Rüdiger T, Saloustros E, Sawyer EJ, Schmutzler RK, Schneeweiss A, Schürmann P, Shah M, Sohn C, Southey MC, Surowy H, Suvanto M, Thanasitthichai S, Tomlinson I, Torres D, Truong T, Tzardi M, Valova Y, van Asperen CJ, Van Dam RM, van den Ouweland AMW, van der Kolk LE, van Veen EM, Wendt C, Williams JA, Yang XR, Yoon SY, Zamora MP, Evans DG, de la Hoya M, Simard J, Antoniou AC, Borg Å, Andrulis IL, Chang-Claude J, García-Closas M, Chenevix-Trench G, Milne RL, Pharoah PDP, Schmidt MK, Spurdle AB, Vreeswijk MPG, Benitez J, Dunning AM, Kvist A, Teo SH, Devilee P and Easton DF. Breast cancer risk genes - association analysis in more than 113,000 women. N Engl J Med 2021; 384: 428-439.

[12] Cheng Y, Huang Z, Liao Q, Yu X, Jiang H, He Y, Yao S, Nie S and Liu L. Risk of second primary breast cancer among cancer survivors: implications for prevention and screening practice. PLoS One 2020; 15: e0232800.

[13] Bao S, Jiang M, Wang X, Hua Y, Zeng T, Yang Y, Yang F, Yan X, Sun C, Yang M, Fu Z, Huang X, Li J, Wu H, Li W, Tang J and Yin Y. Nonmetastatic breast cancer patients subsequently developing second primary malignancy: a population-based study. Cancer Med 2021; 10: 8662-8672.

[14] Hong J, Wei R, Nie C, Leonteva A, Han X, Du X, Wang J, Zhu L, Tian W and Zhou H. The risk and prognosis of secondary primary malignancy in lung cancer: a population-based study. Future Oncol 2021; 17: 4497-4509.

[15] Kong J, Yu G, Si W, Li G, Chai J, Liu Y and Liu J. Second primary malignancies in patients with hepatocellular carcinoma: a population-based analysis. Front Oncol 2021; 11: 713637.

[16] El-Tamer MB, Homel P and Wait RB. Is race a poor prognostic factor in breast cancer? J Am Coll Surg 1999; 189: 41-45.

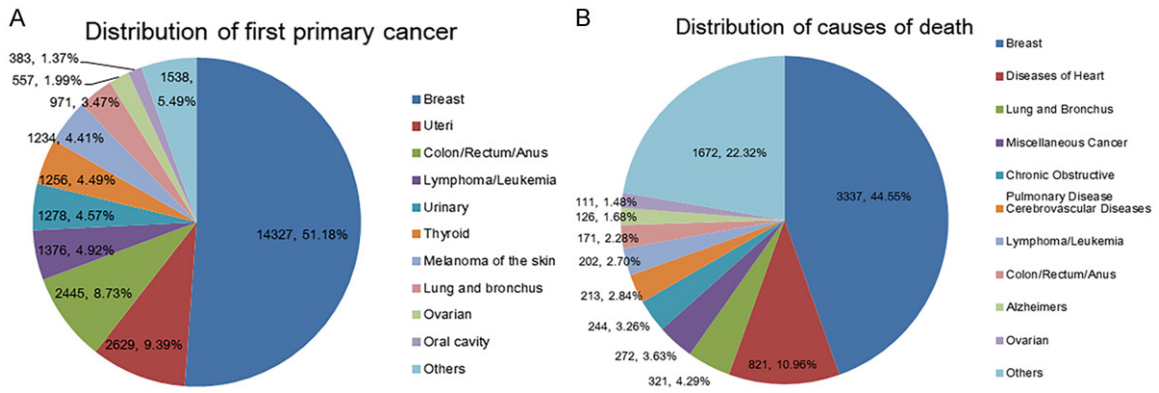
[17] Martínez ME, Unkart JT, Tao L, Kroenke CH, Schwab R, Komenaka I and Gomez SL. Prognostic significance of marital status in breast cancer survival: a population-based study. PLoS One 2017; 12: e0175515.

[18] Hwang KT, Kim J, Jung J, Kim BH, Park JH, Jeon SY, Hwang KR, Roh EY, Park JH and Kim SJ. Long-term prognostic effect of hormone receptor subtype on breast cancer. Breast Cancer Res Treat 2020; 179: 139-151.

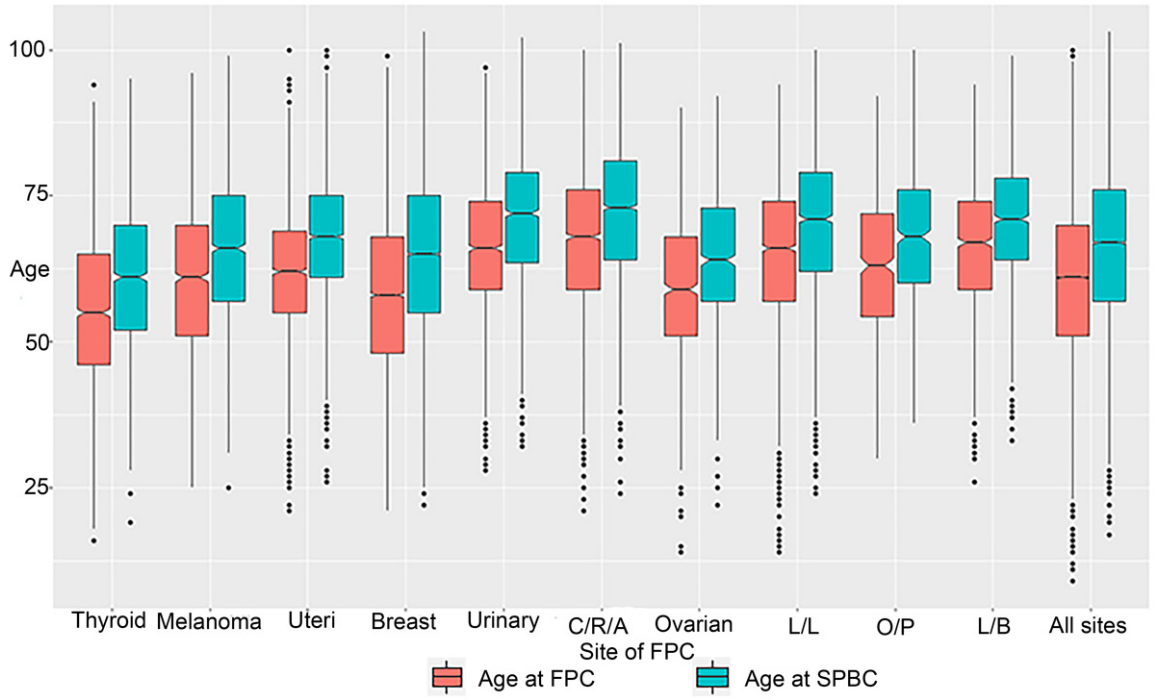
[19] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490: 61-70.

[20] Prat A, Adamo B, Cheang MC, Anders CK, Carey LA and Perou CM. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. Oncologist 2013; 18: 123-133.
[21] Karp SE, Tonin PN, Bégin LR, Martinez JJ, Zhang JC, Pollak MN and Foulkes WD. Influence of BRCA1 mutations on nuclear grade and estrogen receptor status of breast carcinoma in Ashkenazi Jewish women. Cancer 1997; 80: 435-441.
[22] Loman N, Johannsson O, Bendahl PO, Borg A, Fernö M and Olsson H. Steroid receptors in hereditary breast carcinomas associated with BRCA1 or BRCA2 mutations or unknown susceptibility genes. Cancer 1998; 83: 310-319.
[23] Foulkes WD, Metcalfe K, Sun P, Hanna WM, Lynch HT, Ghadirian P, Tung N, Olopade OI, Weber BL, McLennan J, Olivotto IA, Bégin LR and Narod SA. Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type. Clin Cancer Res 2004; 10: 2029-2034.
[24] Kataoka A, Tokunaga E, Masuda N, Shien T, Kawabata K and Miyashita M. Clinicopathological features of young patients (<35 years of age) with breast cancer in a Japanese Breast Cancer Society supported study. Breast Cancer 2014; 21: 643-650.
[25] Sugano K, Nakamura S, Ando J, Takayama S, Kamata H, Sekiguchi I, Ubukata M, Kodama T, Arai M, Kasumi F, Hirai Y, Ikeda T, Jinno H, Kitajima M, Aoki D, Hirasawa A, Takeda Y, Yazaki K, Fukutomi T, Kinoshita T, Tsunematsu R, Yoshida T, Izumi M, Umezawa S, Yagata H, Komatsu H, Arimori N, Matoba N, Gondo N, Yokoyama S and Miki Y. Cross-sectional analysis of germline BRCA1 and BRCA2 mutations in Japanese patients suspected to have hereditary breast/ovarian cancer. Cancer Sci 2008; 99: 1967-1976.
[26] King MC, Marks JH and Mandell JB; New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 2003; 302: 643-646.
[27] Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Brody LC and Tucker MA. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. N Engl J Med 1997; 336: 1401-1408.
[28] Chen S and Parmigiani G. Meta-analysis of BRCA1 and BRCA2 penetrance. J Clin Oncol 2007; 25: 1329-1333.
[29] Kauff ND, Satagopan JM, Robson ME, Scheuer L, Hensley M, Hudis CA, Ellis NA, Boyd J, Borgen PI, Barakat RR, Norton L, Castiel M, Nafa K and Offit K. Risk-reducing salpingo-oophorectomy in women with a BRCA1 or BRCA2 mutation. N Engl J Med 2002; 346: 1609-1615.
[30] Parazzini F, Braga C, La Vecchia C, Negri E, Acerboni S and Franceschi S. Hysterectomy, oophorectomy in premenopause, and risk of breast cancer. Obstet Gynecol 1997; 90: 453-456.
[31] Robinson WR, Nichols HB, Tse CK, Olshan AF and Troester MA. Associations of premenopausal hysterectomy and oophorectomy with breast cancer among black and white women: the Carolina breast cancer study, 1993-2001. Am J Epidemiol 2016; 184: 388-399.
[32] Dutra MC, Rezende MA, de Andrade VP, Soares FA, Ribeiro MV, de Paula EC and Gobbi H. Immunophenotype and evolution of breast carcinomas: a comparison between very young and postmenopausal women. Rev Bras Ginecol Obstet 2009; 31: 54-60.
[33] Cheng W, Shen X and Xing M. Decreased breast cancer-specific mortality risk in patients with a history of thyroid cancer. PLoS One 2019; 14: e0221093.
[34] Sawka AM, Thabane L, Parlea L, Ibrahim-Zada I, Tsang RW, Brierley JD, Straus S, Ezzat S and Goldstein DP. Second primary malignancy risk after radioactive iodine treatment for thyroid cancer: a systematic review and meta-analysis. Thyroid 2009; 19: 451-457.
[35] Zhang Y, Liang J, Li H, Cong H and Lin Y. Risk of second primary breast cancer after radioactive iodine treatment in thyroid cancer: a systematic review and meta-analysis. Nucl Med Commun 2016; 37: 110-115.
[36] Berrington de Gonzalez A, Curtis RE, Kry SF, Gilbert E, Lamart S, Berg CD, Stovall M and Ron E. Proportion of second cancers attributable to radiotherapy treatment in adults: a cohort study in the US SEER cancer registries. Lancet Oncol 2011; 12: 353-360.
[37] Wang C, Hu K, Deng L, He W, Fang F, Tamimi RM and Lu D. Increased risk of breast cancer-specific mortality among cancer survivors who developed breast cancer as a second malignancy. BMC Cancer 2021; 21: 491.
[38] Kaplan HG, Malmgren JA and Atwood MK. Increased incidence of myelodysplastic syndrome and acute myeloid leukemia following breast cancer treatment with radiation alone or combined with chemotherapy: a registry cohort analysis 1990-2005. BMC Cancer 2011; 11: 260.
[39] Anderson WF, Rosenberg PS, Prat A, Perou CM and Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst 2014; 106: dju165.
[40] Armitage P and Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954; 8: 1-12.

[41] Armitage P and Doll R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Br J Cancer 1957; 11: 161-169.
[42] Erić I, Petek Erić A, Kristek J, Koprivčić I and Babić M. Breast cancer in young women: pathologic and immunohistochemical features. Acta Clin Croat 2018; 57: 497-502.
[43] Walker RA, Lees E, Webb MB and Dearing SJ. Breast carcinomas occurring in young women (<35 years) are different. Br J Cancer 1996; 74: 1796-1800.
[44] Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer 2001; 1: 157-162.
[45] Luo L, Lin H, Huang J, Lin B, Huang F and Luo H. Risk factors and prognostic nomogram for patients with second primary cancers after lung cancer using classical statistics and machine learning. Clin Exp Med 2022; [Epub ahead of print].



Supplementary Figure 1. The distribution of first primary cancer and the causes of death of second primary breast cancer (SPBC) patients. A: Distribution of first primary cancer; B: Distribution of causes of death of second primary breast cancer (SPBC) patients.



Supplementary Figure 2. Median age of onset of FPC and SPBC. FPC: first primary cancer; SPBC: second primary breast cancer; C/R/A: colon/rectum/anus; L/L: lymphoma/leukemia; O/P: oral/pharynx; L/B: lung/bronchus.

Supplementary Table 1. Univariate COX analysis of clinical features extracted from SEER database

	Hazard Ratio (HR)	95% CI	P
Primary Site of SPBC			
Low-inner	Reference		
Low-outer	1.02	0.90-1.15	0.807
Upper-inner	0.80	0.71-0.90	***
Upper-outer	0.93	0.84-1.03	0.146
Central portion	1.15	1.01-1.31	*
Other site	1.17	1.06-1.29	**
Historical Type of SPBC			
IDC	Reference		
ILC	1.16	1.07-1.25	***
Mixed	0.94	0.85-1.03	0.177
Other	1.10	1.03-1.18	**
Race			
White	Reference		
Black	1.44	1.35-1.54	***
Other	0.83	0.75-0.91	***
Marital Status			
Married	Reference		
Single	1.34	1.25-1.44	***
Other	1.74	1.66-1.83	***
Site of FPC			
Breast	Reference		
Oral/Pharynx	1.20	1.00-1.44	*
Colon/Rectum/Anus	1.07	0.99-1.16	0.086
Lymphoma/Leukemia	1.16	1.05-1.28	**
Lung/Bronchus	1.97	1.78-2.17	***
Melanoma of skin	0.62	0.54-0.71	***
Ovarian	1.13	0.97-1.32	0.124
Thyroid	0.45	0.39-0.53	***
Urinary	1.04	0.94-1.16	0.437
Uteri	0.72	0.66-0.79	***
Other sites	1.40	1.28-1.53	***
HR of SPBC			
Negative	Reference		
Positive	0.67	0.64-0.71	***
HER2 of SPBC			
Negative	Reference		
Positive	1.16	1.05-1.28	**
Grade of SPBC			
Well	Reference		
Moderately	1.36	1.27-1.45	***
Poorly	1.87	1.76-2.00	***
Stage of SPBC			
I	Reference		
II	1.62	1.54-1.71	***
III	3.31	3.09-3.53	***
IV	8.85	8.19-9.55	***

Radiotherapy for SPBC			
Without	Reference		
With	0.60	0.57-0.63	***
Chemotherapy for SPBC			
Without	Reference		
With	0.94	0.89-0.99	*
Surgery for SPBC			
Without	Reference		
With	0.19	0.18-0.20	***
Age at diagnosis of SPBC	1.03	1.03-1.04	***
Incubation Period	0.99	0.99-1.00	***
Stage of FPC			
I	Reference		
II	1.44	1.36-1.53	***
III	2.27	2.12-2.44	***
IV	2.59	2.38-2.81	***
Grade of FPC			
Well	Reference		
Moderately	1.30	1.21-1.39	***
Poorly	1.67	1.55-1.80	***
Age at diagnosis of FPC	1.04	1.03-1.04	***
Radiotherapy for FPC			
Without	Reference		
With	0.89	0.84-0.93	***
Chemotherapy for FPC			
Without	Reference		
With	1.22	1.17-1.28	***
Surgery for FPC			
Without	Reference		
With	0.53	0.50-0.56	***

Notes: *P<0.05; **P<0.01; ***P<0.001. SPBC: second primary breast cancer; FPC: first primary cancer; IDC: infiltrating ductal carcinoma; ILC: infiltrating lobular carcinoma.

Supplementary R Code. R code for Propensity score matching (PSM) and XGBoost model.

```
# PSM R Code


Setwd(dir="c:/Users/a/Desktop/")

library(tidyverse)

library(openxlsx)

library(rJava)

library(xlsxjars)

library(dplyr)

library(readxl)

all <- read.csv("allbreastforPSM.csv",header = TRUE)

all[1:4,1:4]

all2 <- all %>% filter(age1 != "15") %>%

filter(stage.1 == "I"|stage.1 == "II"|stage.1 == "III"|stage.1 == "IV") %>%

filter(race == "Black"|race == "Other"|race == "White")

all22 <- all2 %>%

select(ID,age1,stage.1,grade,historical_type,HR,primary.site.labeled,Laterality,Marital.status,surgery.
recode,chemotherapy.recode,radiation.recode,race)

all222 <- all22[!all22$ID %in% names(which(table(all22$ID)>1)),]

summary (all222)

names(all222)  <-  c("ID","age","stage","grade","historical_type","HR","primary.site.labeled","Laterality",
"Marital.status","surgery.recode","chemotherapy.recode","radiation.recode","Race")

Save(all222,file = "all222.RData")

second <- read.csv("secondbreastforPSM.csv",header = TRUE)

second[1:4,1:4]

second2 <- second %>% filter(age1 != "15") %>%

filter(stage.1 == "I"|stage.1 == "II"|stage.1 == "III"|stage.1 == "IV") %>%

filter(race == "Black"|race == "Other"|race == "White")

second22 <- second2 %>%

select(ID,age1,stage.1,grade,historical_type,HR,primary.site.labeled,Laterality,Marital.status,surgery.
recode,chemotherapy.recode,radiation.recode,race)


Summary(second22)

Names(second22)  <-  c("ID","age","stage","grade","historical_type","HR","primary.site.labeled","Latera-
lity","Marital.status","surgery.recode","chemotherapy.recode","radiation.recode","Race")
```

```
#second2$Type <- “second”


first <- all222 %>%

filter(! ID %in% second22$ID)

first[1:4,1:4]

first$Type <- “first”


first <- first[!duplicated(first$ID),]

second22 <- second22[!duplicated(second22$ID),]

second22$Type <- “second”


merge_1_2 <- rbind(first,second22)

merge_1_2$radiation.recode <- ifelse (merge_1_2$radiation.recode == “without”,“without”,“with”)


save(all222,second22,merge_1_2,first,file = “Step0_data_0917.RData”)


load(“Step0_data_0917.RData”)


set.seed(1234)

library(MatchIt)

merge_1_2.new <- na.omit(merge_1_2)

merge_1_2.new$radiation.recode <- as.factor (merge_1_2.new$radiation.recode)

merge_1_2.new2 <- merge_1_2.new

merge_1_2.new2$Type <- as.logical(merge_1_2.new2$Type == ‘second’)

summary(merge_1_2.new2)


#merge_1_2.new22 <- merge_1_2.new2[sample(nrow(merge_1_2.new2),10000),]

m.out <- matchit(data = merge_1_2.new2,formula = Type~age + grade + stage+ historical_type
+HR+primary.site.labeled+Laterality+Marital.status+surgery.recode+chemotherapy.recode+radiation.
recode+Race,method = “nearest”,distance = “logit”,replace = FALSE,ratio = 2,caliper = 0.05)

save(m.out,file = “Step1_m.out.RData”)

load(“Step1_m.out.RData”)

dta_m <- match.data(m.out)

write.csv(dta_m, file = “secondbreast_afterPSM_2X_0917.csv”)

d1 <- m.out$match.matrix
```

```
d2 <- rownames(d1)
rownames(d1) <- NULL
pairs <- cbind(d2, d1)
colnames(pairs)[1:2] <- c("group1","group2")
write.csv(pairs, file = "secondbreast_2X_0917.csv")
head(dta_m)
dta_m_2 <- subset(dta_m,Type == "TRUE")
second_raw <- read.xlsx("secondbreast_raw_data.xlsx")
second_raw[1:4,1:4]
add <- second[,c("ID","month.since.index")]
secondbreast_selected <- second_raw %>% filter(ID %in% dta_m_2$ID) %>%
inner_join(add,by = "ID")
secondbreast_selected$breast_related_OS <- secondbreast_selected$month.since.index.y
write.xlsx(secondbreast_selected,"secondbreast_selected.xlsx")

#XGBOOST R code
Setwd(dir="c:/Users/a/Desktop/")
Library("caret")
Library("xgboost")
Library("stringr")
Library("Matrix")
Library("pROC")
Library("discretization")
Library("DiagrammeR")
see<-read.csv("secondbreast5.csv",header = TRUE)
seer1 <- see[18:21]
chi1<-chiM(seer1,alpha=0.05)
head(chi1$Disc.data,10)
chi1$cutp
a1 <- as.matrix(chi1$Disc.data)
a2 <- model.matrix(~see$Marital.status-1,see)
a3 <- model.matrix(~see$primary.site.of.SPBC-1,see)
a4 <- model.matrix(~see$historical_type.of.SPBC-1,see)
a5 <- model.matrix(~see$Site.of.FPC-1,see)
```

```
a6 <- model.matrix(~see$Race-1,see)

yinzi <- cbind(a2,a3,a4,a5,a6)

seer <- cbind(see[6:17],yinzi,a1)

set.seed(70)

train_sub <- sample(nrow(seer),7/10*nrow(seer))

train_data <- seer[train_sub,]

test_data <- seer[-train_sub,]

traindata1 <- data.matrix(train_data[c(1:42)])

traindata2 <- Matrix(traindata1,sparse = TRUE)

traindata3 <- train_data[43]

traindata4 <- list(data=traindata2,label=traindata3)

dtrain <- xgb.DMatrix(data = traindata4$data,label=traindata4$label)

testset1 <- data.matrix(test_data[c(1:42)])

testset2 <- Matrix(testset1,sparse = TRUE)

testset3 <- test_data[43]

testset4 <- list(data=testset2,label=testset3)

dtest <- xgb.DMatrix(data = testset4$data,label=testset4$label)

xgb<-xgboost(data=dtrain,max_depth=4,eta=0.17,subsample=0.7,objective='binary:logistic',nrou
nd=25)


importance <- xgb.importance(traindata2@Dimnames[[2]],model = xgb)

xgb.plot.importance(importance_matrix = importance[1:10])

pred1 <- data.frame(predict(xgb,testset2,type='response'))

xgb_lr.train.modelroc <- roc(test_data$group, pred1$predict.xgb..testset2..type....response..)

plot(xgb_lr.train.modelroc,print.auc=TRUE,auc.polygon=TRUE,grid=c(0.1,0.2),grid.col=c("green",
"red"),max.auc.polygon=TRUE,auc.polygon.col="skyblue",print.thres=TRUE,main='ROC curve of Xgboo-
st algorithm')


pred2<- data.frame(predict(xgb,traindata2,type='response'))

xgb_lr.train.modelroc <- roc(train_data$group, pred2$predict.xgb..traindata2..type....response..)

plot(xgb_lr.train.modelroc,print.auc=TRUE,auc.polygon=TRUE,grid=c(0.1,0.2),grid.col=c("green",
"red"),max.auc.polygon=TRUE,auc.polygon.col="skyblue",print.thres=TRUE,main='ROC curve of Xgboo-
st algorithm')


save(xgb, file = "xgboost.RData")

load("xgboost.Rdata")
```