



## OPEN ACCESS

## EDITED BY

Dharmendra Kumar Yadav,  
Gachon University, Republic of Korea

## REVIEWED BY

Yingying Zhu,  
Sun Yat-sen Memorial Hospital, China  
Shuheng Bai,  
The First Affiliated Hospital of Xi'an Jiaotong  
University, China

## \*CORRESPONDENCE

Lei Wang

✉ wlei81@126.com

Gang Sun

✉ sung853219@126.com

RECEIVED 03 October 2023

ACCEPTED 29 February 2024

PUBLISHED 20 March 2024

## CITATION

Wu M, Zhang T, Gao C, Zhao T, Wang L and  
Sun G (2024) Assessing of case-cohort  
design: a case study for breast cancer  
patients in Xinjiang, China.  
*Front. Oncol.* 14:1306255.  
doi: 10.3389/fonc.2024.1306255

## COPYRIGHT

© 2024 Wu, Zhang, Gao, Zhao, Wang and Sun.  
This is an open-access article distributed under  
the terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Assessing of case-cohort design: a case study for breast cancer patients in Xinjiang, China

Mengjuan Wu<sup>1</sup>, Tao Zhang<sup>1</sup>, Chunjie Gao<sup>1</sup>, Ting Zhao<sup>2</sup>,  
Lei Wang<sup>3\*</sup> and Gang Sun<sup>4,5\*</sup>

<sup>1</sup>Country College of Public Health, Xinjiang Medical University, Urumqi, China, <sup>2</sup>Department of Medical Record Management, The Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi Xinjiang, China, <sup>3</sup>Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi Xinjiang, China, <sup>4</sup>Xinjiang Cancer Center/ Key Laboratory of Oncology of Xinjiang Uyghur Autonomous Region, Urumqi, Xinjiang, China, <sup>5</sup>Department of Breast and Thyroid Surgery, The Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi, Xinjiang, China

**Objective:** To assess the effectiveness and clinical value of case-cohort design and determine prognostic factors of breast cancer patients in Xinjiang on the basis of case-cohort design.

**Methods:** The survival data with different sample characteristics were simulated by using Cox proportional risk models. To evaluate the effectiveness for the case-cohort, entire cohort, and simple random sampling design by comparing the mean, coefficient of variation, etc., of covariate parameters. Furthermore, the prognostic factors of breast cancer patients in Xinjiang were determined based on case-cohort sampling designs. The models were comprehensively evaluated by likelihood ratio test, the area under the receiver operating characteristic curve (AUC), and Akaike Information Criterion (AIC).

**Results:** In a simulations study, the case-cohort design shows better stability and improves the estimation efficiency when the censored rate is high. In the breast cancer data, molecular subtypes, T-stage, N-stage, M-stage, types of surgery, and postoperative chemotherapy were identified as the prognostic factors of patients in Xinjiang. These models based on the different sampling designs both passed the likelihood ratio test ( $p < 0.05$ ). Moreover, the model constructed under the case-cohort design had better fitting effect (AIC=3,999.96) and better discrimination (AUC=0.807).

**Conclusion:** Simulations study confirmed the effectiveness of case-cohort design and further determined the prognostic factors of breast cancer patients in Xinjiang based on this design, which presented the practicality of case-cohort design in actual data.

## KEYWORDS

case-cohort design, breast cancer, survival prognosis, Cox proportional hazards model, simulations study

## 1 Introduction

Breast cancer with a high mortality rate is one of the most widespread malignant tumors, which seriously threatens women's health and safety. Global Cancer Statistics 2020 pointed out that there were 2.27 million new cases of breast cancer worldwide, and approximately one in eight patients died of breast cancer in 2020 (1). Since the twenty-first century, the morbidity and mortality of female breast cancer in China have been continuously increasing (2), which would cause tremendous burden of breast cancer. Furthermore, breast cancer is highly heterogeneous, with the variety in molecular subtype, clinical stage, and other pathological features (3). The differences in tumor cell growth rate, invasion ability, and potential metastasis are strongly correlated with patients' survival prognosis (4). Survival analysis is widely applied to investigate the relationship among survival time, survival state, and important influencing factors of breast cancer patients. For instance, Ma et al. (5) studied the serum lipid changes in breast cancer patients during neoadjuvant chemotherapy and the impact of dyslipidemia on their prognosis. Zhou et al. (6) identified the potential prognostic factors of patients with triple-negative breast cancer and built the corresponding prediction model.

In China, there are endemic variety in the morbidity and mortality of breast cancer (7). Relevant studies (8–10) showed that the current situation of breast cancer in Xinjiang is different from that in other regions, with such features as lower incidence rate, luminal breast cancer appearing more frequently, and women aged 45–55 having a higher risk of developing this disease. At present, there have been many studies evaluating the prognostic risk factors of patients with breast cancer in Xinjiang (11–14); for instance, Shan et al. (11) investigated the clinicopathological features and prognostic characteristics of patients with triple-negative breast cancer in Xinjiang, based on clinical information for 319 patients. Fu et al. (13) focused on the difference in survival and prognosis of breast cancer patients with different molecular subtypes in Xinjiang. Cao et al. (14) evaluated the association of hypoxia-inducible factor-1 $\alpha$  and survivin with breast cancer prognosis in breast cancer patients. However, the sample size of some studies was relatively small (11, 12), and those studies were mainly focus on exploring the impact of molecular subtypes or gene expression on the prognosis of breast cancer patients (11, 13, 14). On the other hand, it is necessary to follow up a large number of research subjects over the long-term in survival analysis, which may inevitably cause certain omissions in the process of data collection. Realistically, the breast cancer patients followed up by the hospitals or cancer centers are equivalent to random sample from the overall population. Therefore, it could not totally represent the basic characteristics of the overall population to a certain extent. In particular, a previous study showed that the mortality rate of breast cancer in Xinjiang Cancer Registration Area was only approximately 8.72% (15). When the incidence of interested event in the follow-up subjects is lower, directly using the data of random samples would cause the insufficient power of statistical analysis (16). To decrease the sampling error produced by simple random sampling, Prentice (17) proposed the case-cohort design in 1986. On the basis of simple random sampling, the case-cohort design analyzes those

patients who have experienced outcome events in the full cohort, which is suitable for these studies with lower incidence of disease outcomes or higher costs of covariate collections (18–20), and compared with the simple random sampling, the case-cohort design may decrease the sample error (21, 22). Yu et al. (18) separately investigated the relationship between demographic characteristics, tumor histology, and time of onset and recurrence of nephroblastoma patients, under a case-cohort design. Cai et al. (19) employed a case-cohort design to identify the influencing factors of fungal infection in patients with hematopoietic cell transplant. Particularly, the case-cohort design is widely used to analyze the factors influencing morbidity or mortality of breast cancer (20–22). For example, based on the case-cohort design, Yang et al. (20) used additive risk model to explore the major prognostic factors of patients with breast cancer. The case-cohort design was employed to evaluate the prospective associations between perfluoroalkyl substances and breast cancer risk in (21). Yao et al. (22) used case-cohort design to investigate the association of serum biomarker of vitamin D status, 25-hydroxyvitamin D values with breast cancer recurrence, and survival prognosis. It was indicated that the results based on the case-cohort design with fewer samples were similar to those based on the full cohort. The case-cohort design could be not only suitable for large cohort studies with low incidence but also available reduce the cost and improve the efficiency. Furthermore, there may be a lack of repeatability in the analysis of actual clinical data; thus, using a case-cohort design could partly decrease the bias generated by random sampling. Therefore, it is significant to further determine the prognostic factors of breast cancer patients in Xinjiang by using a case-cohort design, which could contribute to explore patients' clinical treatments and improve their survival probability.

Inspired by the aforementioned discussion, in this paper, we first explored the effectiveness of the case-cohort sampling design by using simulated data. To do this, we employed the Cox proportional hazards model to fit the parameters of covariates in these models under full cohort, case-cohort with different sampling proportions, and simple random sampling designs, respectively, and then, we compared these estimated values of parameters for those models (such as the mean, standard deviation, coefficient of variation, and bias). Second, due to the fact that the mortality for the Xinjiang breast cancer patients was relatively lower, we further discussed the applicability of the case-cohort design in identifying the prognostic factors of breast cancer patients in Xinjiang, by comparing the comprehensive performance of these models established under the case-cohort and full cohort sampling designs. These results could offer scientific basis for evaluating the prognosis of breast cancer patients in Xinjiang.

## 2 Methods

### 2.1 Case-cohort design

In the case-cohort design, the random subcohort (denoted as  $S$ ) was selected by simple random sampling from the full cohort. We denoted  $S_i$  and  $\delta_i$  as indicator variables, respectively, whether the  $i$

th patient was included in the random subcohort and whether the  $i$  th patient experienced outcome events. That is, if the  $i$  th patient was included in the random subcohort, then  $S_i = 1$ , and if the  $i$  th patient experienced the outcome event (i.e., case), then  $\delta_i = 1$ . The case-cohort samples included the random subcohort and all cases outside the random subcohort (20) (see Figure 1). Denote  $CC_i$  as an indicator variable, the explicit expression is as follows,

$$CC_i = \begin{cases} 1, & \delta_i = 1 \text{ or } S_i = 1, \\ 0, & \delta_i = 0 \text{ and } S_i = 0. \end{cases}$$

In this paper, it is assumed that there are  $N$  independent individuals in total. For the survival data with censored, the Cox proportional hazards model is used for analysis. Let  $X$  be the covariable for the  $i$  th individual and  $\beta$  be the partial regression coefficient, then the basic form of the Cox proportional hazards model is,

$$h_i(t|X) = h_0(t) \exp(X\beta), \quad i = 1, \dots, N,$$

where  $h_0(t)$  denotes the baseline risk function. Since the case-cohort design is a biased sampling, the cases and non-cases in the case-cohort design are equally weighted. The pseudolikelihood is used to infer the partial regression coefficient  $\beta$ , then an estimator for  $\beta$  may be obtained by maximizing the pseudolikelihood function

$$L_p(\beta) = \prod_{i=1}^N \left[ \frac{\exp\{X_i^T \beta\}}{\sum_{j \in \mathcal{R}(T_i)} \exp\{X_j^T \beta\}} \right]^{\delta_i},$$

where  $T_i$  represents the observed true event time for  $i$  th patient, the risk set at time  $t$  denoted by  $\mathcal{R}(t) = \{j : T_j \geq t, j \in \mathcal{D}(t) \cup S\}$ , and  $\mathcal{D}(t)$  is the collection of cases at time  $t$ . Then, the maximum pseudolikelihood estimator for  $\beta$  be solved as

$$U_p(\beta) = \sum_{i=1}^N \delta_i \left[ X_i - \frac{\sum_{j \in \mathcal{R}(T_i)} \exp\{X_j^T \beta\} X_j}{\sum_{j \in \mathcal{R}(T_i)} \exp\{X_j^T \beta\}} \right] = 0.$$

## 2.2 Simulation study

Let  $T_i^*$  and  $C_i$  be the time that the interested event occurs or fails and the time that the  $i$  th patient was followed up or censored ( $i = 1, \dots, N$ ), respectively. If  $T_i^* \leq C_i$ , then the  $i$  th patient experienced the outcome event before the end of the observation period. Otherwise, if  $T_i^* > C_i$ , then the  $i$  th patient is censored. Thus, the observed true event time is defined as  $T_i = \min(T_i^*, C_i)$ . Whether or not each patient experienced the outcome event is given by the right censored indicative variable  $\delta_i = I(T_i^* \leq C_i)$ , where  $I(\cdot)$  is an indicator function.

The time that the interested event occurs or fails for an individual is usually described by using exponential, Weibull, lognormal, and Gamma distributions, etc. The censored time usually follows uniform, exponential distribution and so on (23). In this paper, the survival data were simulated based on the total number of the full cohort sample  $N = 5,000$ ,  $T_i^* \sim \text{Weibull}(\alpha, \lambda)$ , and  $C_i \sim \text{uniform}(0, \theta)$ , where the scale parameter  $\lambda = 4$ , the shape parameter  $\alpha = 2$ , and  $\theta$  denotes as censored rate. Given that this paper mainly focuses on categorical variables, Bernoulli distributions with different probabilities of occurrence were chosen to fit covariates during the simulation process. Therefore, assuming that there are three covariates for each individual, namely,  $X_1, X_2$ , and  $X_3$ , generated from Bernoulli distributions with success rates of 0.1, 0.5, and 0.9, respectively, i.e.,  $X_1 \sim B(0.1)$ ,  $X_2 \sim B(0.5)$ , and  $X_3 \sim B(0.9)$  (Table 1). Then, the Cox proportional hazards model is considered as follows:

$$h_i(t) = h_0(t) \exp(X_1\beta_1 + X_2\beta_2 + X_3\beta_3).$$

The simulated data with six different sample characteristics was simulated based on different censored rates and regression coefficients (Table 1).

In the following, we compared the parameter estimations of these sampling designs:

FC: parameter estimations based on full cohort ( $\hat{\beta}_{FC}$ ).

CCI: parameter estimations based on a case-cohort design with one-third proportion sample ( $\hat{\beta}_{CCI}$ ).

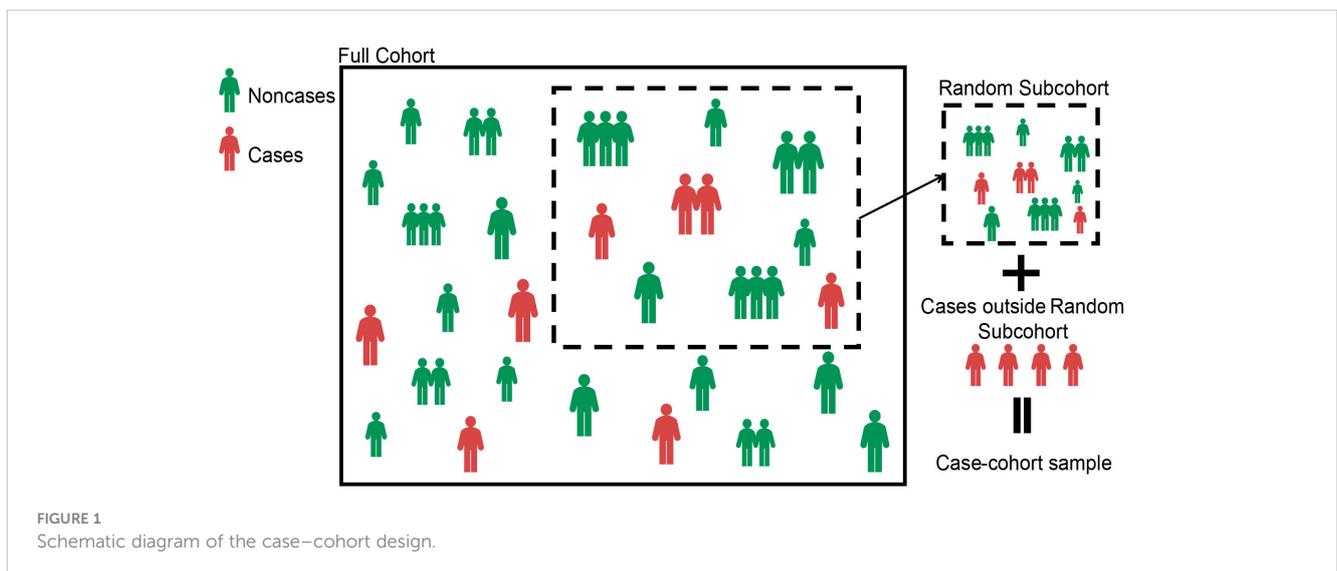


FIGURE 1 Schematic diagram of the case-cohort design.

TABLE 1 Different sample characteristics of simulated data.

Number	$\theta$	Covariates distribution	$\beta$
1	50%	$X_1 \sim B(0.1), X_2 \sim B(0.5), X_3 \sim B(0.9)$	$(1.5, 1.5, 1.5)^T$
2			$(-1.5, -1.5, -1.5)^T$
3	80%	$X_1 \sim B(0.1), X_2 \sim B(0.5), X_3 \sim B(0.9)$	$(1.5, 1.5, 1.5)^T$
4			$(-1.5, -1.5, -1.5)^T$
5	90%	$X_1 \sim B(0.1), X_2 \sim B(0.5), X_3 \sim B(0.9)$	$(1.5, 1.5, 1.5)^T$
6			$(-1.5, -1.5, -1.5)^T$

$\theta$  denotes as censored rate.  $X_1, X_2,$  and  $X_3$  indicate covariates.  $\beta$  denotes the corresponding estimated parameters.

CCII: parameter estimations based on a case-cohort design with one-sixth proportion sample ( $\hat{\beta}_{CCII}$ ).

RS: parameter estimations based on random subcohort with one-third proportion sample ( $\hat{\beta}_{RS}$ ).

The simulated data were sampled 1,000 times for parameter estimations. The mean, standard error of the mean (SE.mean), standard deviation (SD), coefficient of variation (CV), range, and bias of these parameters were compared to assess the performance of different sampling designs.

### 2.3 Analysis of breast cancer data

The breast cancer patients collected in this paper was sourced from the Affiliated Cancer Hospital of Xinjiang Medical University. Based on full cohort and case-cohort sampling designs, the survival data of these patients were analyzed to identify the independent prognostic factors of breast cancer patients in Xinjiang, by using Kaplan-Meier analysis, Cox proportional hazards model, and stepwise regression. Meanwhile, the parameter estimations of those models were compared to evaluate the comprehensive performance of these models based on case-cohort and full cohort sampling designs and then assess the effectiveness and clinical value of the case-cohort design.

Potential influencing factors such as survival status (life or death), survival time, basic demographic, and clinicopathological of patients were gathered. The patients' histological grades of tumors are divided into low, medium, and high. According to immunohistochemical technique, there are luminal A, luminal B, HER2 overexpression, and triple-negative breast cancer. The TNM staging system is divided into T stage (primary tumor), N stage (regional lymph nodes), and M stage (distant organ metastases). T stage was divided into T1 (tumor size,  $\leq 2$  cm), T2 (tumor size, 2–5 cm), T3 (tumor size,  $>5$  cm), and T4 (tumors of any size with direct extension to the chest wall and/or to the skin, that is ulceration or skin nodules, macroscopic nodules); N stage included N0 (no regional lymph node metastases), N1 (micrometastases, or metastases in one to three axillary lymph nodes), N2 (metastases in four to nine axillary lymph nodes), and N3 (metastases in 10 or more axillary lymph nodes); and M stage was split into M0 (no clinical or radiographic evidence of distant metastases) and M1

(distant metastases) (24). The types of surgery that patients underwent included no surgery, radical surgery, and breast-conserving surgery. In addition, the age [classified into three categories: younger group ( $\leq 45$  years), middle-aged group (46–69 years), and the elderly group ( $\geq 70$  years)] and postoperative chemotherapy of patient were also included.

The inclusion criteria for patients were 1) the age of patient was above 18, 2) tumor of primary site was only identified as breast cancer, and 3) the information of clinicopathological and follow-up were complete. Patients were excluded if 1) medical documents were unsigned, such as informed consent and patient instructions, at the time of admission, and 2) the information about the molecular subtypes, clinical stage, types of surgery, etc., were partial. A total of 8,226 breast cancer patients were followed up in this paper, and the end of the follow-up period was 31 December 2021. Among them, 7,948 patients were effectively followed up, with a follow-up rate of 96.62%. According to the inclusion and exclusion criteria, a total of 3,641 patients were ultimately included, of which 326 patients died (i.e., the censored rate more than 90%).

In this paper, all statistical analysis and visualization were conducted using R 4.1.3 software. A  $p < 0.05$  based on a two-tailed test was considered statistically significant.

### 2.4 Model evaluation

#### 2.4.1 Likelihood ratio test

The likelihood ratio test was used to evaluate Cox regression models in general and reflect the fitting effect of the models (25), based on the following formula,

$$\chi^2_v = -2\text{Log}L_i - (-2\text{Log}L_j),$$

where  $\chi^2_v \sim \chi^2(v)$ ,  $-2\text{Log}L_i$  represents the log-likelihood function value of a regression model with  $i$  parameters. The smaller the value of  $\chi^2_v$ , the better the fitting effect of the model.

#### 2.4.2 Akaike Information Criterion

Akaike Information Criterion (AIC) (26) is applicable to select the most effective model from various models and evaluate the validity of the modeling results. The general form of this is as follows

$$\text{AIC} = -2\ln(L) + 2K,$$

where  $L$  and  $K$  is the maximum likelihood function and the number of independent parameters, respectively. The smaller the AIC value is, which indicates a minimum discrepancy between the probability and the true distribution, the better the model is.

#### 2.4.3 Discrimination

The accuracy of the model predictions is evaluated on the basis of the discrimination. A model showed good discrimination if this model can distinguish whether the patient has reached the endpoint. The area under the receiver operating characteristic

(ROC) curve (AUC), which has a value of 0.5–1.0 and the discrimination is better with the higher value of AUC, was used to assess the discrimination of models (27).

### 3 Results

#### 3.1 Results of simulation

In the simulation data of six different sample characteristics, the parameters  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  of these models constructed in FC, RS, CCI, and CCII sampling designs were estimated, where full cohort  $N = N_{FC} = 5,000$  and subcohort  $N_{RS} = 1,666$  (see Table 2, Supplementary Tables S1, S2, respectively).

The estimated results of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  showed that its mean values were relatively close under different parameter settings of

four sampling designs. Its SE, mean, SD, CV, range, and bias were small, which demonstrated that Cox proportional hazards model presents the better ability in the analysis of the simulated data. Moreover, the findings showed that the fitting results of parameters in the RS and CCI sampling designs approached to the same with a large bias in the results of CCII sampling design when  $\theta = 50\%$ . For instance, in the scenario of  $\theta = 50\%$  and  $\beta_1 = \beta_2 = \beta_3 = -1.5$  (Supplementary Table S2), the bias value of  $\beta_3$  in CCII is approximately 0.02, and its SD, CV, and range are also larger than those of other sampling designs.

On the other hand, it was found that when the censored rate increases, the efficiency of simple random sampling design decreases, the range, SD, and CV of parameter estimations under this sampling design become larger, and then the possibility of outlier is increased. In actual application, there may be a large bias in the results of simple random sampling design without repeated

TABLE 2 The simulation results of  $\beta_1$  under different censored rate and sampling design.

$\theta$	$(\beta_1, \beta_2, \beta_3)^T$	Sampling design	$\beta_1$					
			Mean	SE.mean	SD	CV	Range	Bias
50%	$(1.5, 1.5, 1.5)^T$	FC	1.4998	0.0019	0.0594	0.0396	0.4566	0.0002
		RS	1.5064	0.0032	0.1006	0.0668	0.7138	-0.0064
		CCI	1.5043	0.0039	0.1218	0.0809	0.8582	-0.0043
		CCII	1.5198	0.0053	0.1689	0.1112	1.0474	-0.0198
	$(-1.5, -1.5, -1.5)^T$	FC	-1.5037	0.0029	0.0933	-0.0620	0.5794	0.0037
		RS	-1.5011	0.0050	0.1590	-0.1060	1.1000	0.0011
		CCI	-1.4970	0.0048	0.1533	-0.1024	0.9272	-0.0030
		CCII	-1.4922	0.0066	0.2074	-0.1390	1.3636	-0.0078
80%	$(1.5, 1.5, 1.5)^T$	FC	1.4989	0.0025	0.0785	0.0524	0.6556	0.0011
		RS	1.5049	0.0043	0.1363	0.0906	1.0314	-0.0049
		CCI	1.5062	0.0045	0.1420	0.0943	1.0461	-0.0062
		CCII	1.5185	0.0062	0.1963	0.1293	1.1437	-0.0185
	$(-1.5, -1.5, -1.5)^T$	FC	-1.5094	0.0058	0.1847	-0.1223	1.3911	0.0094
		RS	-1.5247	0.0105	0.3306	-0.2168	2.4861	0.0247
		CCI	-1.5041	0.0073	0.2311	-0.1537	1.7562	0.0041
		CCII	-1.4966	0.0087	0.2749	-0.1837	1.6536	-0.0034
90%	$(1.5, 1.5, 1.5)^T$	FC	1.4977	0.0032	0.1002	0.0669	0.7329	0.0023
		RS	1.5003	0.0058	0.1847	0.1231	1.2513	-0.0003
		CCI	1.5030	0.0050	0.1589	0.1057	1.2128	-0.0030
		CCII	1.5134	0.0068	0.2137	0.1412	1.2151	-0.0134
	$(-1.5, -1.5, -1.5)^T$	FC	-1.5420	0.0091	0.2888	-0.1873	1.8267	0.0420
		RS	-1.7807	0.0570	1.8040	-1.0131	17.9073	0.2807
		CCI	-1.5379	0.0102	0.3225	-0.2097	1.9279	0.0379
		CCII	-1.5257	0.0111	0.3510	-0.2301	2.5577	0.0257

$\theta$  denotes as censored rate.  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  indicate the estimated parameters. SE, mean standard error of the mean; SD, standard deviation; CV, coefficient of variation; FC, full cohort; RS, random subcohort; CCI, case-cohort design with one-third proportion sample; CCII, case-cohort design with one-sixth proportion sample.

sampling. For instance, under the RS sampling design, when  $\theta = 90\%$ ,  $\beta_1 = \beta_2 = \beta_3 = -1.5$  (Figure 2C) or 1.5 (Figure 2D), respectively, there are many outliers with the ranges of approximately 18 in the fitted values of  $\beta_3$  and  $\beta_1$ , which greatly exceeds the ranges of the estimated values under other sampling designs.

Moreover, when the censored rate is high (i.e.,  $\theta = 80\%$  or  $\theta = 90\%$ ), CCI and CCII sampling designs have good stability, with smaller dispersion degree and variation index of the parameters, especially CCI. CCI sampling design improves the estimation efficiency because only partial samples (approximately 40%) of the full cohort samples were used by this sampling design to reach the fitting result of FC sampling design, as shown in Table 2, Figure 2, and Supplementary Figures S1, S2. Therefore, when the sample censored rate was 90%, the sample error of the case-cohort design is smaller than that of simple random sampling.

### 3.2 Results of breast cancer data

In this paper, there were 3,641 breast cancer patients in Xinjiang with a censored rate of more than 90% as full cohort samples, of which only 326 patients experienced the outcome event (i.e., death). Hence, based on the results of the simulation in Section 3.1, the case-cohort design with a one-third sample proportion was selected to analyze these data. First, one-third of the patients were randomly selected as a random subcohort (1,214 patients) combining with all cases outside the subcohort, and then, a case-cohort sample with 1,418 patients was formed. The basic information about clinicopathological characteristics of patients is shown in Table 3. Furthermore, Kaplan–Meier analysis was performed to analyze the clinical data of patients based on the full cohort and case-cohort

sampling designs, as shown in Figures 3 and 4, respectively. Then, the statistically significant factors ( $p < 0.05$ ) in Kaplan–Meier analysis and factors with clinical practice value were added to the Cox regression model, and the significant prognostic factors were selected by bidirectional stepwise regression.

The fitting results of Cox regression model showed that the parameter estimations under the two sampling designs were very close (see Figure 5). It was finally determined that molecular subtypes, T stage, N stage, and M stage were the risk factors for prognosis of Xinjiang breast cancer patients ( $p < 0.05$  and  $HR > 1$ ). In detail, patients with clinicopathological features of triple-negative breast cancer, T3, N3, and M1 substages had the highest risk of death. Simultaneously, types of surgery and postoperative chemotherapy were protective factors for independent prognosis ( $p < 0.05$  and  $HR < 1$ ). Patients who underwent breast-conserving surgery, radical surgery, and postoperative chemotherapy had a lower risk of death than others who did not have surgery. Thus, a model that can effectively predict prognosis of patients has been established as follows:

$$h(t|X) = h_0(t) \exp(\beta_1 X_{\text{Molecular subtypes}} + \beta_2 X_{\text{T-stage}} + \beta_3 X_{\text{N-stage}} + \beta_4 X_{\text{M-stage}} + \beta_5 X_{\text{Types of surgery}} + \beta_6 X_{\text{Postoperative chemotherapy}})$$

Finally, the performances of these models established on the basis of the case-cohort (CCI) and full cohort (FC) sampling designs were comprehensively evaluated, as shown in Table 4. Both Cox proportional hazards models established under the two sampling designs passed the likelihood ratio test ( $p < 0.05$ ), where  $\chi^2_{\text{CCI}} = 490.05 < \chi^2_{\text{FC}} = 518.80$ . In addition, the AIC value (3,999.96) obtained by the CCI was also smaller compared with

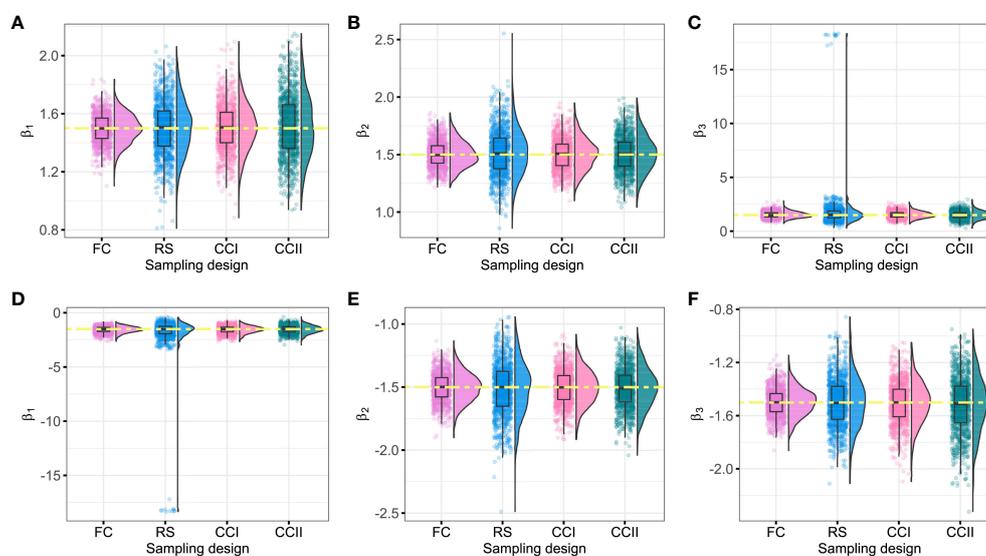


FIGURE 2

Fitting values of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  under different sampling designs ( $\theta = 90\%$ ). The yellow dashed line represents the initial value of the regression coefficients. (A–C) The fitting values of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  when the initial regression coefficients are 1.5; (D–F) the fitting values of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  when the initial regression coefficients are  $-1.5$ . FC, full cohort; RS, random subcohort; CCI, case-cohort design with one-third proportion sample; CCII, case-cohort design with one-sixth proportion sample.

TABLE 3 Basic information about clinicopathological characteristics of breast cancer patients in Xinjiang.

	Full cohort (N=3641)	Case-cohort (N=1418)
<b>Survival time</b>		
$\bar{x} \pm sd$	1,630 $\pm$ 578	1,520 $\pm$ 616
<b>Survival state</b>		
Survival	3,315(91.0%)	1,092(77.0%)
Death	326(9.0%)	326(23.0%)
<b>Age</b>		
$\leq 45$	691(19.0%)	273(19.3%)
46-69	2,609(71.7%)	988(69.7%)
$\geq 70$	341(9.4%)	157(11.1%)
<b>Histological grade</b>		
Low	138 (3.8%)	60 (4.2%)
Medium	2,606 (71.6%)	1,007 (71.0%)
High	897 (24.6%)	351 (24.8%)
<b>Molecular subtyping</b>		
Luminal A	473 (13.0%)	182 (12.8%)
Luminal B	2,218 (60.9%)	824 (58.1%)
HER2 overexpression	397 (10.9%)	175 (12.3%)
Triple-negative	553 (15.2%)	237 (16.7%)
<b>T-stage</b>		
T1	1,740 (47.8%)	632 (44.6%)
T2	1,620 (44.5%)	633 (44.6%)
T3	171 (4.7%)	90 (6.3%)
T4	110 (3.0%)	63 (4.4%)
<b>N-stage</b>		
N0	1,814 (49.8%)	636 (44.9%)
N1	1,140 (31.3%)	463 (32.7%)
N2	389 (10.7%)	156 (11.0%)
N3	298 (8.2%)	163 (11.5%)
<b>M-stage</b>		
M0	3,518 (96.6%)	1,335 (94.1%)
M1	123 (3.4%)	83 (5.9%)
<b>Types of surgery</b>		
No surgery	357 (9.8%)	179 (12.6%)
Breast-conserving surgery	630 (17.3%)	230 (16.2%)
Radical surgery	2,654 (72.9%)	1,009 (71.2%)
<b>Postoperative chemotherapy</b>		
No	635 (17.4%)	275 (19.4%)
Yes	3,006 (82.6%)	1,143 (80.6%)

$\bar{x} \pm sd$  denotes mean  $\pm$  standard deviation.

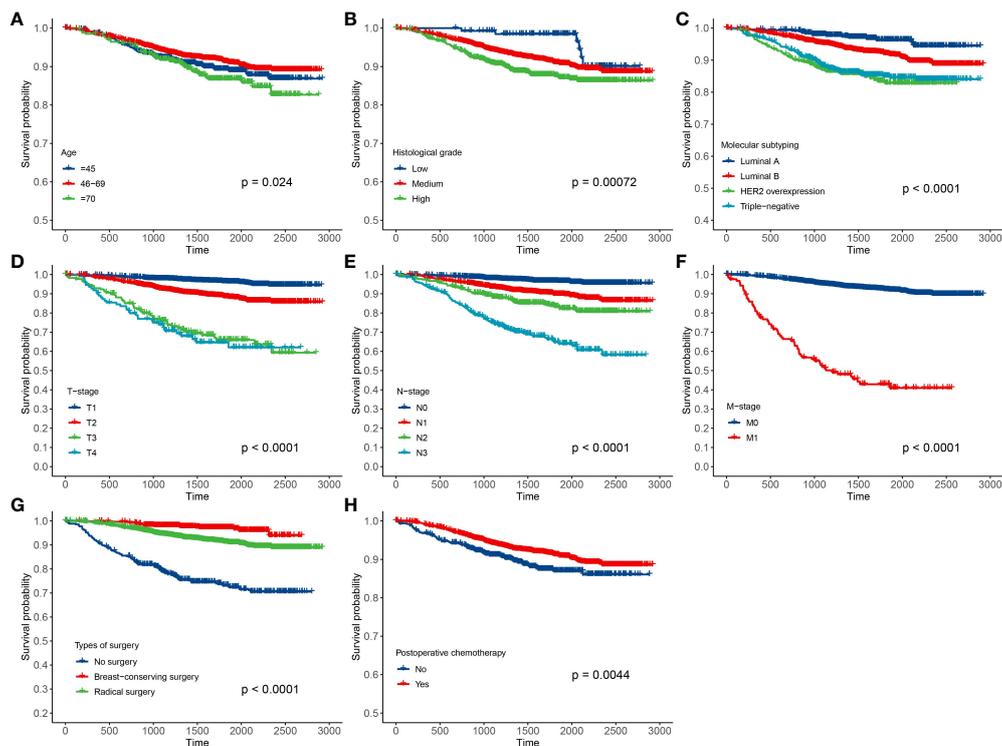
the FC, which indicated that the fitting effect of the case-cohort sampling design was better.

Moreover, ROC curves of Cox regression models under FC and CCI sampling designs were separately drawn to compare the discrimination of these models (Figure 6). It was shown that both AUC values were  $>0.8$ , and they were very close, which confirmed a good discrimination for the prognostic model constructed in this paper and also further verified that the case-cohort sampling design reached a better fitting effect only using approximately 38.9% of the full cohort samples.

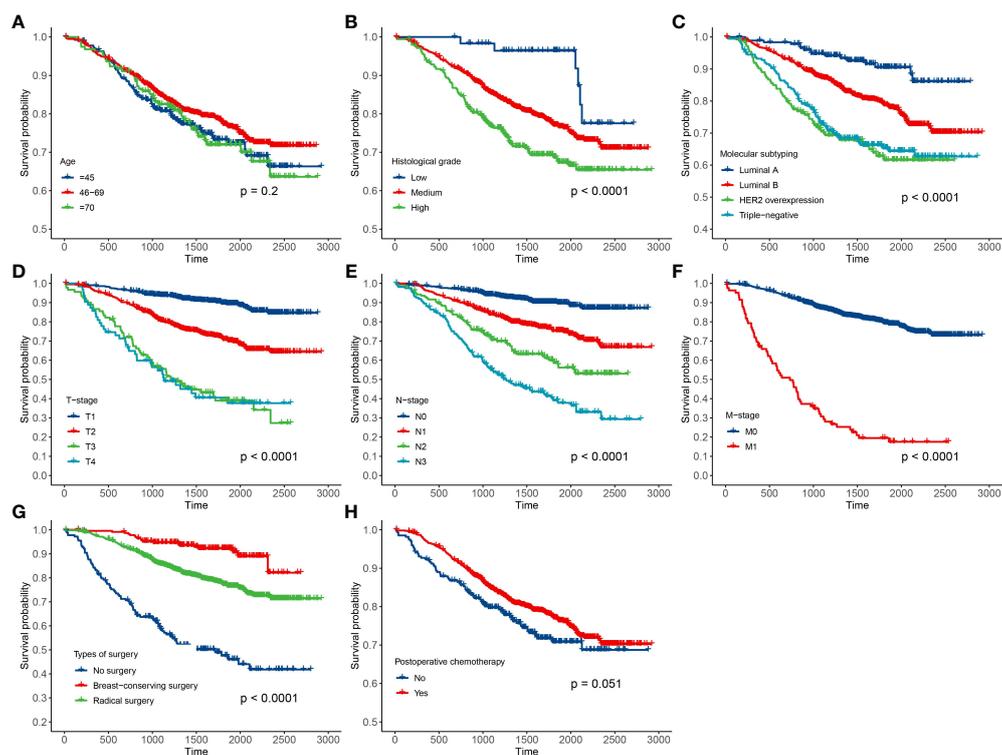
## 4 Discussion

Case-cohort design is suitable for cancer research with large cohort and low incidence, which could improve efficiency and reduce the cost of collecting redundant non-case data (18). One of the highlights of this paper is that the effectiveness of the case-cohort design was verified based on the Cox proportional risk model, and the different censored rates (50%, 80%, and 90%) and different sampling ratios (1/3, 1/6) were conducted in the simulation study. By simulating the survival data with different sample characteristics, this study estimated the coefficients of Cox regression models in FC, CCI, CCII, and RS sampling designs to assess the performance of the models and sampling designs, respectively. Our findings showed that the case-cohort design could improve the estimation efficiency, especially the higher censored rate. Since the morbidity of breast cancer has been an increasing tendency year by year in Xinjiang (10), and the mortality for the followed up Xinjiang breast cancer patients was relatively lower, using the case-cohort design could reduce the bias caused by random sampling, more effectively identify prognostic factors, and further promote the improvement of clinical prognostic methods. Therefore, based on the case-cohort design, this study analyzed the actual clinical data of breast cancer patients in Xinjiang to identify independent prognostic factors (molecular subtypes, T stage, N stage, M stage, types of surgery, and postoperative chemotherapy). Another innovation of this paper is that the performance of the model established under the full cohort and case-cohort in the actual data were comprehensively evaluated in breast cancer patients in Xinjiang by likelihood ratio test, AIC criterion, and discrimination. This further confirmed that the prognosis model constructed under the case-cohort sampling design had better fitting effect than that based on the full cohort sampling design, and the case-cohort sampling design showed certain applicability in the actual data.

The results of simulations in this paper displayed that the estimated mean values of regression coefficients were close to the given initial values in the survival data with different scenarios, indicating that Cox proportional hazards model could achieve the better fitting effect. In addition, when the censored rate was lower, the fitting results of the regression coefficients under the RS and CCI sampling designs were nearly the same, while there was a larger bias of the parameter estimations under CCII sampling designs. It demonstrated that not only the suitable sampling designs should



**FIGURE 3** Results of Kaplan–Meier analysis for the clinical data of breast cancer patients based on the full cohort sampling designs. (A) Age; (B) histological grade; (C) molecular subtyping; (D) T stage; (E) N stage; (F) M stage; (G) types of surgery; and (H) postoperative chemotherapy.



**FIGURE 4** Results of Kaplan–Meier analysis for the clinical data of breast cancer patients based on the case–cohort sampling designs. (A) Age; (B) histological grade; (C) molecular subtyping; (D) T stage; (E) N stage; (F) M stage; (G) types of surgery; and (H) postoperative chemotherapy.

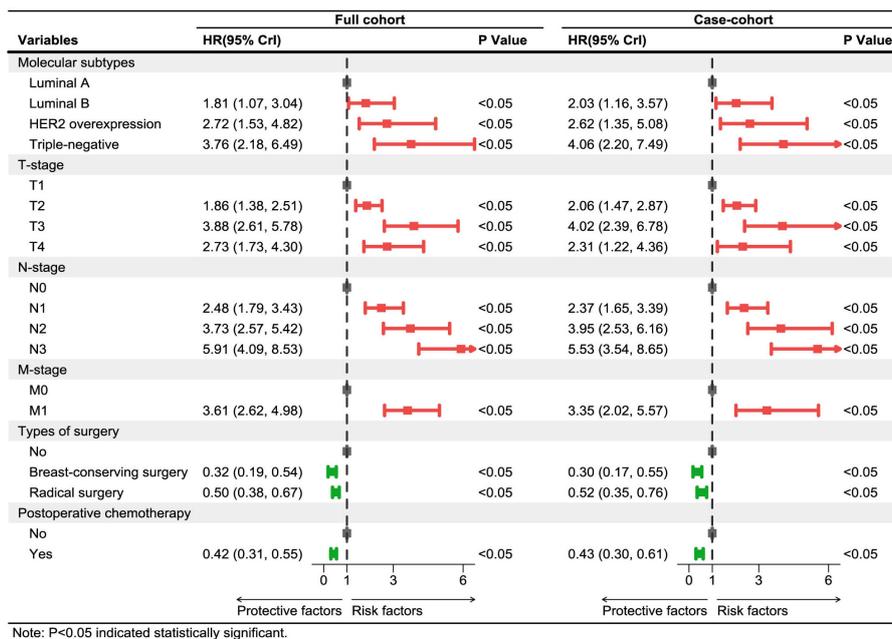


FIGURE 5

Multivariate Cox regression models of breast cancer patients in Xinjiang with full cohort and case-cohort sampling designs. The red lines and squares reflect the HR and 95%CrI for risk factors, while green reflects the HR and 95%CrI for protective factors. HR, hazard ratio; CrI, credibility interval.

be selected but also the sampling proportion should not be too small in the analysis; otherwise, it would also reduce the statistical efficiency. On the other hand, when the censored rate gradually increased, the parameter estimations under the single simple random sampling design would be more likely to generate outliers, which could result in the gradual decrease in efficiency under this sampling design. However, in actual applications, it is often difficult to conduct multiply repeated sampling, which may lead to a significant deviation in the obtained results. Meanwhile, our findings revealed that when the censored rate was higher, the CCI and CCII sampling designs had superior stability (i.e., there are fewer outliers and smaller deviations), especially CCI sampling design. Both estimated mean values under CCI and CCII sampling designs had smaller dispersion degree and variation index, and the CCI design results that only used 38.9% samples of the full cohort samples were close to FC design results. Moreover, using different types of covariates may have a certain influence on the simulation results, but this impact is relatively small, as demonstrated in the paper by Yang et al. (20), where there were slight differences between simulation results of the normal and

uniform distribution. To sum up, the simulation results of this paper confirmed that the case-cohort design is a cost-effective sampling design compared with simple random sampling design, which could improve the efficiency of estimation. In particular, the case-cohort design was more effective and stable when the interested events had a relatively lower incidence, which was consistent with the results in these references (18, 20, 28).

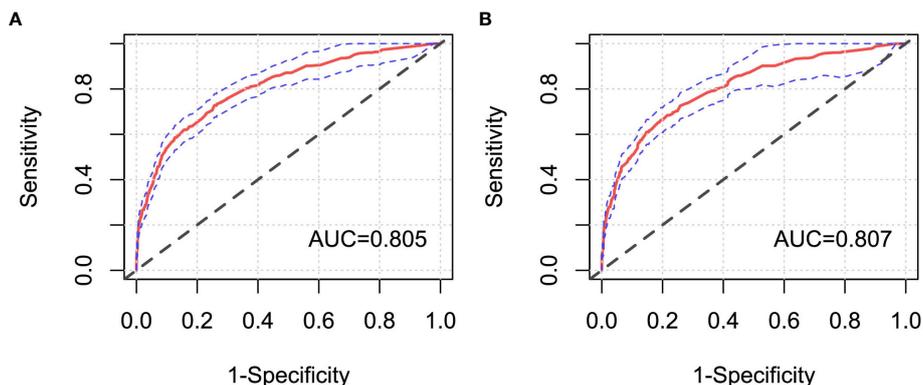
In this paper, the breast cancer patients followed up were registered in the Affiliated Cancer Hospital of Xinjiang Medical University, which could be regarded as a random sample from the overall population, with more than 90% censored rate. Thus, a one-third proportion of case-cohort sampling design was used to analyze these data, and the same Cox regression model was also simultaneously implemented in the full cohort sampling design to compare the difference between the two designs' results. The results showed that the prognosis of patients with triple-negative breast cancer was the worst, which may be the cause of the tumor cells of those patients being more aggressive and more prone to recurrence and metastasis (29). Luminal breast cancer patients had better prognosis and higher survival rate than other non-luminal

TABLE 4 Evaluation indexes of Cox regression models under different sampling designs.

Sampling designs	Likelihood ratio test		AIC	AUC (95%CrI)
	$\chi^2_v$	p-value		
Full cohort (FC)	518.80	<0.05 <sup>a</sup>	4690.78	0.805 (0.779, 0.832)
Case-cohort (CCI)	490.05	<0.05 <sup>a</sup>	3999.96	0.807 (0.780, 0.835)

AIC, Akaike Information Criterion; AUC, the area under the receiver operating characteristic curve. CrI, credible interval.

<sup>a</sup>p<0.05 indicates statistically significant.  $\chi^2_v$  denotes the value of chi-square test.



**FIGURE 6**  
ROC curves of Cox regression models under different sampling designs. (A) Full cohort design; (B) case-cohort design. AUC, the area under the receiver operating characteristic curve.

patients. Moreover, T, N, and M stages were independent risk prognostic factors of breast cancer patients. Patients with advanced T stage had larger tumors, more tumor cells, and the longer time for the tumor formation, so these patients would be more likely to develop into distant metastasis breast cancer ones. The later stage of N stage indicated greater probability, more numbers of lymph node metastases, and higher risk of death, which are typical clinical features of breast cancer progression (30). Because distant metastasis of breast cancer (i.e., M stage) means that the tumors of breast cancer could spread to the lung, liver, brain, and other parts of the body, the occurrence of distant metastasis (i.e., advanced breast cancer) would result in more difficult clinical treatment (31). Therefore, regular breast self-examination and clinical screening for women were recommended to achieve the purpose of early detection, early diagnosis, and early treatment, and then reduce the mortality and improve the prognosis of breast cancer patients. At the same time, it was also shown that the breast-conserving surgery [HR=0.30, 95%CrI: (0.17, 0.55)], radical surgery [HR=0.52, 95%CrI: (0.35, 0.76)], and postoperative chemotherapy [HR=0.43, 95%CrI: (0.30, 0.61)] were protective factors for breast cancer patients in Xinjiang. These surgeries could effectively reduce the size of the tumor, reduce the number of tumors, and control the spread of the disease, thereby greatly improving the survival probability for breast cancer patients. Initially, the radical surgery, as a common treatment, occupied a very important position. But now, breast-conserving surgery is more widely used to treat patients with early disease progression, with the characteristics of shorter operation time and lower incidence of postoperative complications (32). Standard postoperative adjuvant chemotherapy for patients could prevent the recurrence and control the metastasis of cancer to a certain extent, and it could reduce the pain, improve the quality of life, and then extend their life cycle for some patients with advanced stage (33). Finally, the likelihood ratio test, ROC curve, and AIC criteria were used to compare the superiority of model prediction in the full cohort and the case-cohort sampling designs. The comparison findings showed that both models under FC and CCI sampling designs passed the likelihood ratio test ( $p < 0.05$ ), and the model constructed under the CCI design had better fitting effect

(AIC=3,999.96) and better discrimination [AUC=0.807, 95%CrI: (0.780, 0.835)], which demonstrated that the case-cohort design was suitable to analyze the prognosis of breast cancer patients in Xinjiang.

There are some limitations in this study. On the one hand, we only employed Cox proportional hazards model with Prentice's weight method to investigate the effectivity and stability of the case-cohort design. However, different weighted estimation methods (such as Barlow and Self-Prentice method) or different statistical models (such as additive risk model) could also be applied to make statistical inference to be more accurate and effective under the case-cohort design when the weights of case-cohort samples are not mutually independent or the actual data do not follow the proportional hazards assumption. On the other hand, only the clinical data of breast cancer patients in Xinjiang were analyzed in this paper, but the applicability of the case-cohort design in the other regions or other cancers deserves to be further explored. Last but not least, the main purpose of our paper is to explore the factors affecting the prognosis of breast cancer patients in Xinjiang, based on the case-cohort design and Cox proportional risk model. Hence, we focused on the influence degree of different factors on the occurrence time of the event. It was needed to consider the impact of covariates on survival time and the chronological order of events; therefore, we only reported HR values in the outcome in this paper. In our future work, we will consider different methods such as logistic regression or propensity score to calculate different statistical indicators (such as OR and RR values) (34–36), in order to find the best reporting indicator for actual data with different sample characteristics.

## 5 Conclusion

In summary, this study demonstrated the effectivity and stability of the case-cohort design through simulating data and confirmed that this design could maintain a better estimation efficiency in cancers with high censored rate. Furthermore, independent prognostic factors of breast cancer patients in

Xinjiang were determined under the case-cohort design, and the practical fitting effect and useful application of the case-cohort design were demonstrated by comparing with the results based on full cohort design.

## Data availability statement

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 TiZ, zhaoting0557@163.com.

## Ethics statement

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). The participants provided their written informed consent to participate in this study.

## Author contributions

MW: Conceptualization, Software, Writing – original draft, Writing – review & editing. GS: Data curation, Funding acquisition, Writing – review & editing. TaZ: Software, Writing – original draft. CG: Methodology, Writing – original draft. TiZ: Data curation, Resources, Writing – review & editing. QZ: Data curation, Resources, Writing – review & editing. LW: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660
- Pu XY, Ma Y, Zhong ZG. Trend analysis of female breast cancer deaths in China between 2006 and 2020—Based on an Age-Period-Cohort Model. *Health Econ Res.* (2023) 40:28–33. doi: 10.14055/j.cnki.33-1056/f.2023.02.002
- Kang B, Lee J, Jung JH, Kim WW, Keum H, Park HY. Differences in clinical outcomes between HER2-negative and HER2-positive luminal B breast cancer. *Med (Baltimore).* (2023) 102:e34772. doi: 10.1097/MD.00000000000034772
- Tao ZQ, Shi AM, Lu CT, Song T, Zhang ZG, Zhao J. Breast cancer: Epidemiology and etiology. *Cell Biochem Biophys.* (2015) 72:333–8. doi: 10.1007/s12013-014-0459-6
- Ma Y, Lv M, Yuan P, Chen X, Liu Z. Dyslipidemia is associated with a poor prognosis of breast cancer in patients receiving neoadjuvant chemotherapy. *BMC Cancer.* (2023) 23:208. doi: 10.1186/s12885-023-10683-y
- Zhou HL, Chen DD. Prognosis of patients with triple-negative breast cancer: a population-based study from SEER database. *Clin Breast Cancer.* (2023) 23:e85–94. doi: 10.1016/j.clbc.2023.01.002
- He R, Zhu B, Liu J, Zhang N, Zhang WH, Mao Y. Women's cancers in China: A spatio-temporal epidemiology analysis. *BMC Womens Health.* (2021) 21:116. doi: 10.1186/s12905-021-01260-1
- Qiu XJ, Liu Y, Wu E, Meng T, Cheng F. Clinicopathological characteristics and prognosis analysis among 1006 cases of different molecular subtypes of breast cancer in Xinjiang. *Chin Clin Oncol.* (2015) 20:525–30.
- Wang C, Wang YH, Shen YY, Han W. Clinicopathological characteristics among different molecular subtypes of breast cancer in Xinjiang. *J Mod Oncol.* (2017) 25:1921–4. doi: 10.3969/j.issn.1672-4992.2017.12.017
- Li HF, Guo CM, Wang HY, Dilimurati A. Epidemiological analysis of 1701 cases of breast cancer in a Third Grade Hospital of Urumqi of Xinjiang province. *Pract J Cancer.* (2022) 37:975–9. doi: 10.3969/j.issn.10015930.2022.06.030
- Shan MH, Li HT, Luo L, Yao XM, Ma BL, Ma J. Analysis on clinico-pathological features and prognosis of TNBC patients in Xinjiang Region. *J Xinjiang Med Univ.* (2020) 43:63–68+73. doi: 10.3969/j.issn.1009-5551.2020.01.016
- Nie Y, Ying B, Lu Z, Sun T, Sun G. Predicting survival and prognosis of postoperative breast cancer brain metastasis: a population-based retrospective analysis. *Chin J Cancer Prev Treat.* (2023) 136:1699–707. doi: 10.1097/CM9.0000000000002674
- Fu AL, Liu YJ, Abudushalimu YMT, Zhao T. Correlation of molecular subtypes with survival and prognosis among females with breast cancer: a single-center analysis. *Chin J Cancer Prev Treat.* (2023) 30:587–92. doi: 10.16073/j.cnki.cjcp.2023.10.03
- Cao Q, Mushajiang M, Tang CQ, Ai XQ. Role of hypoxia-inducible factor-1 $\alpha$  and survivin in breast cancer recurrence and prognosis. *Heliyon.* (2023) 9:e14132. doi: 10.1016/j.heliyon.2023.e14132

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant Nos. 12061079 and 82060520), the Project of Top-notch Talents of Technological Youth of Xinjiang (Grant No. 2022TSYCCX0108), the Natural Science Foundation of Xinjiang (Grant No. 2022D01C287), and the Tianshan Cedar Talent Training Project of Science and Technology Department of Xinjiang Uygur Autonomous Region (Grant No. 2020XS14).

## 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.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2024.1306255/full#supplementary-material>

15. Zhao R, Liu LX, Ni MJ, Li SG, Yan YZ, Zhang XF, et al. Disease burden of premature death of Malignant neoplasms in Xinjiang cancer registries. *Modern Prev Med.* (2017) 44:1703–1707+1713. doi: CNKI:SUN:XDYF.0.2017-09-040
16. Onland-Moret NC, van der A DL, van der Schouw YT, Buschers W, Elias SG, van Gils CH, et al. Analysis of case-cohort data: a comparison of different methods. *J Clin Epidemiol.* (2007) 60:350–5. doi: 10.1016/j.jclinepi.2006.06.022
17. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika.* (1986) 73:1–11. doi: 10.2307/2336266
18. Yu JC, Cao YX. Diagnostics for the proportional hazards model with case-cohort data. *Acta Mathematica Sin (Chinese Series).* (2020) 63:137–48. doi: 10.3969/J.ISSN.0583-1431.2020.02.004
19. Cai J, Kim S. Correction: Case-cohort design in hematopoietic cell transplant studies. *Bone Marrow Transplant.* (2022) 57:145. doi: 10.1038/s41409-021-01522-4
20. Yang J, Ding JL. Additive risk regression analysis in case-cohort design and its application to breast cancer data. *J Math.* (2021) 41:270–82. doi: 10.13548/j.sxzz.2021.03.009
21. Feng Y, Bai Y, Lu Y, Chen M, Fu M, Guan X, et al. Plasma perfluoroalkyl substance exposure and incidence risk of breast cancer: A case-cohort study in the Dongfeng-Tongji cohort. *Environ pollut.* (2022) 306:119345. doi: 10.1016/j.envpol.2022.119345
22. Yao S, Kwan ML, Ergas IJ, Roh JM, Cheng TD, Hong CC, et al. Association of serum level of vitamin d at diagnosis with breast cancer survival: A case-cohort analysis in the pathways study. *JAMA Oncol.* (2017) 3:351–7. doi: 10.1001/jamaoncol.2016.4188
23. Wan F. Simulating survival data with predefined censoring rates for proportional hazards models. *Stat Med.* (2017) 36:838–54. doi: 10.1002/sim.7178
24. Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast cancer-major changes in the American joint committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin.* (2017) 67:290–303. doi: 10.3322/caac.21393
25. Mbona SV, Ndlovu P, Mwambi H, Ramroop S. Multiple imputation using chained equations for missing data in survival models: applied to multidrug-resistant tuberculosis and HIV data. *J Public Health Afr.* (2023) 14:2388. doi: 10.4081/jphia.2023.2388
26. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods.* (2012) 17:228–43. doi: 10.1037/a0027127
27. Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol.* (2018) 63:07TR01. doi: 10.1088/1361-6560/aab4b1
28. Tuo JY, Bi JH, Li ZY, Shen QM, Tan YT, Li HL, et al. [Statistical methods for relative risk estimation and applications in case-cohort study]. *Zhonghua Liu Xing Bing Xue Za Zhi.* (2022) 43:392–6. doi: 10.3760/cma.j.cn112338-20210812-00638
29. Metzger-Filho O, Tutt A, de Azambuja E, Saini KS, Viale G, Loi S, et al. Dissecting the heterogeneity of triple-negative breast cancer. *J Clin Oncol.* (2012) 30:1879–87. doi: 10.1200/JCO.2011.38.2010
30. Pereira ER, Jones D, Jung K, Padera TP. The lymph node microenvironment and its role in the progression of metastatic cancer. *Semin Cell Dev Biol.* (2015) 38:98–105. doi: 10.1016/j.semcdb.2015.01.008
31. Huo X, Li J, Zhao F, Ren D, Ahmad R, Yuan X, et al. The role of capecitabine-based neoadjuvant and adjuvant chemotherapy in early-stage triple-negative breast cancer: a systematic review and meta-analysis. *BMC Cancer.* (2021) 21:78. doi: 10.1186/s12885-021-07791-y
32. Qiu H, Xu WH, Kong J, Ding XJ, Chen DF. Effect of breast-conserving surgery and modified radical mastectomy on operation index, symptom checklist-90 score and prognosis in patients with early breast cancer. *Med (Baltimore).* (2022) 99:e19279. doi: 10.1097/MD.00000000000019279
33. Liu M, Yang J, Xu B, Zhang X. Tumor metastasis: Mechanistic insights and therapeutic interventions. *MedComm.* (2021) 2:587–617. doi: 10.1002/mco2.100
34. Noma H, Misumi M, Tanaka S. Risk ratio and risk difference estimation in case-cohort studies. *J Epidemiol.* (2023) 33:508–13. doi: 10.2188/jea.JE20210509
35. Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol.* (2007) 166:332–9. doi: 10.1093/aje/kwm069
36. Ek M, Roth B, Nilsson PM, Ohlsson B. Characteristics of endometriosis: A case-cohort study showing elevated IgG titers against the TSH receptor (TRAb) and mental comorbidity. *Eur J Obstet Gynecol Reprod Biol.* (2018) 231:8–14. doi: 10.1016/j.ejogrb.2018.09.034