

Methods and Applications

An Improved Training Algorithm Based on Ensemble Penalized Cox Regression for Predicting Absolute Cancer Risk

Liyuan Liu^{1,2,&}; Fu Yang^{3,&}; Yeye Fan²; Chunyu Kao³; Fei Wang^{1,4}; Lixiang Yu^{1,4};
Yong He^{2,3}; Jiadong Ji^{3,#}; Zhigang Yu^{1,4,#}

ABSTRACT

Introduction: Biases in cancer incidence characteristics have led to significant imbalances in databases constructed by prospective cohort studies. Since they use imbalanced databases, many traditional algorithms for training cancer risk prediction models perform poorly.

Methods: To improve prediction performance, we introduced a Bagging ensemble framework to an absolute risk model based on ensemble penalized Cox regression (EPCR). We then tested whether the EPCR model outperformed other traditional regression models by varying the censoring rate of the simulated data.

Results: Six different simulation studies were performed with 100 replicates. To assess model performance, we calculated mean false discovery rate, false omission rate, true positive rate, true negative rate, and the areas under the receiver operating characteristic curve (AUC) values. We found that the EPCR procedure could reduce the false discovery rate (FDR) for important variables at the same true positive rate (TPR), thereby achieving more accurate variable screening. In addition, we used the EPCR procedure to build a breast cancer risk prediction model based on the Breast Cancer Cohort Study in Chinese Women database. AUCs for 3- and 5-year predictions were 0.691 and 0.642, representing improvements of 0.189 and 0.117 over the classical Gail model, respectively.

Discussion: We conclude that the EPCR procedure can overcome challenges posed by imbalanced data and improve the performance of cancer risk assessment tools.

Most cancer predictions involve imbalanced binary classification datasets, i.e., the number of instances of cases is far smaller than the number of instances of controls. We are more concerned about predicting

cases because misclassification of cases can be more costly (1). However, traditional supervised learning algorithms do not possess high predictive accuracy for minority classes. The “ensemble learning” approach for statistical modeling is a powerful method for generating highly accurate predictive models, in which Bagging (2–3), a simple yet effective ensemble method, has been employed in many practical applications (4). This paper proposes building an ensemble penalized Cox regression (EPCR) model for disease risk prediction and validates the accuracy of the method through numerical simulations and an empirical study on a Breast Cancer Chinese Women database.

METHODS

Ensemble Penalized Cox Regression Model

We propose an ensemble penalized Cox regression (EPCR) model based on penalized Cox regression (PCR) models (5–8) (Supplementary Figure S1, available in <https://weekly.chinacdc.cn/>). For the original dataset $D = \{\tilde{T}_i, \Delta_i, Z_i\} (i = 1, \dots, n)$, we first use a repeated sampling technique to generate B bootstrap data sets from the original data set by $D^{(k)} = \{\tilde{T}_i^{(k)}, \Delta_i^{(k)}, Z_i^{(k)}\}_{i=1}^n (k = 1, \dots, B)$. Next, a set of base learners $\hat{P}^{(k)}(a, \tau, Z) (k = 1, \dots, B)$ are trained by the PCR algorithm independently on $\tilde{D}^{(k)} (k = 1, \dots, B)$. More details about PCR algorithm are provided in Supplementary Materials. For each sample in the test set, EPCR achieves prediction by averaging the probability prediction values given by each of these B base learners.

Simulation Study

To assess the predictive accuracy of the proposed EPCR procedure and to compare its performance to alternative methods — i.e., Cox regression based on a stepwise procedure [using Akaike Information Criterion (AIC) or Bayesian Information Criterion

(BIC)] or single PCR — we conducted simulation studies across a range of conditions by varying the censoring rate or dimensionality of predictors. We were particularly interested in assessing the ability of the proposed EPCR procedure to correctly identify important predictors associated with cancer as well as the accuracy of the EPCR procedure in predicting cancer risk.

Each p -dimensional predictor is assumed to be a continuous variable generated from a multivariate normal distribution with a mean (μ) of zero and a covariance matrix $\Sigma = (0.8^{|j-i|})$, $i, j = 1, \dots, p$. The first 15 of the p -dimensional predictors were assumed to be genuinely associated with the onset of cancer. For simplicity, we specified the regression coefficients of the Cox model as 1.5 for the first five predictors, 1 for predictors 6–10, 0.5 for predictors 11–15, and 0 for the rest.

By specifying different baseline hazard functions $h_0(t)$, we can generate different survival times T that obey different distributions (6). To obtain this value, the survival function $S(t)$ was first generated through a uniform distribution $U(0, 1)$, and T is then generated using the following equation:

$$T = H_0^{-1}[-\log(S(t)) \exp(-\beta'Z)] \quad (1)$$

Here, H_0^{-1} denotes the inverse function of the cumulative hazard function $H_0(t) = \int_0^t h_0(u) du$. For simplicity, we specify $h_0(t)$ as 1, at which point the survival time T follows an exponential distribution. Furthermore, we generated the censoring metric Δ from a Bernoulli $b(0, 1 - r)$ distribution, where r is the censoring rate.

Varying the dimensionality of predictors p and censoring rate r , our simulation study considered the following six main settings:

Setting 1: $n = 1,000, p = 100, r = 30\%$;

Setting 2: $n = 1,000, p = 100, r = 50\%$;

Setting 3: $n = 1,000, p = 100, r = 70\%$;

Setting 4: $n = 1,000, p = 50, r = 30\%$;

Setting 5: $n = 1,000, p = 50, r = 50\%$;

Setting 6: $n = 1,000, p = 50, r = 70\%$.

For each setting, bootstrap times B is specified as 200 and simulated data are split into two parts: 70% to train the models and 30% as a test dataset for comparing model performance. The simulation study was repeated 100 times for each setting. Mean values of the four evaluation metrics [“false discovery rate (FDR),” “false omission rate (FOR),” “true positive rate (TPR),” and “true negative rate (TNR)” for variable screening] were calculated to test whether the

important predictors could be correctly identified by the models. Finally, the area under the receiver operating characteristic curve (AUC) was calculated for each model to test how well each model could be used for prediction of the onset of cancer.

Empirical Study: Application to a Real-World Cancer Cohort

To validate the disease risk prediction validity of the proposed EPCR model, we applied it to the Shandong sub-database from Breast Cancer Cohort Study in Chinese Women (BCCS-CW) (9) to develop a candidate breast cancer incidence risk predictor. The workflow of this part of the study is presented in [Supplementary Figure S2](#) (available in <https://weekly.chinacdc.cn/>).

The onset of breast cancer was treated as the outcome event and individuals who had not yet developed breast cancer were censoring data. We considered the age of individuals with breast cancer to be the age at which the patient received the first cancer diagnosis, and the age of individuals who had not yet developed breast cancer as the age registered at baseline. We randomly selected 70% individuals from the case and control groups respectively to form a training set for model development; the remaining 30% of the control group was used as a test dataset. The EPCR procedure was performed on the training set to generate an absolute risk prediction model for breast cancer, and this was then used to estimate the probability of onset in the test group over the next three or five years. Similarly, the bootstrap times B is specified as 200. Based on actual three- and five-year follow-up results, receiver operating characteristic (ROC) curves were plotted to assess model performance, where a single PCR model and a classical Gail model (10) were used for comparison.

All the analyses were performed in the R software (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria). Packages “glmnet” and “gbm” were used to construct the EPCR model, “pROC” was used to plot the ROC curve, and “Table 1” was used to create a demographic characteristics table. $P < 0.05$ was considered statistically significant ($\alpha = 0.05$).

RESULTS

[Table 1](#) summarizes the mean values of the 5 evaluation metrics for 100 replications of each simulation setting. These results clearly show that the

TABLE 1. The mean values of 5 metrics for the 6 models over 100 replicate experiments for each simulation setting.

Method	FDR	FOR	TPR	TNR	AUC
Setting 1: 30% censoring					
Traditional approach					
Stepwise-AIC*	0.766	0.127	0.344	0.795	0.721
Stepwise-BIC	0.173	0.125	0.201	0.99	0.733
PCR-LASSO [†]	0.275	0.009	0.952	0.922	0.863
PCR-EN ($\alpha = 0.5$)	0.375	0.005	0.973	0.878	0.873
Ensemble approach					
EPCR-LASSO [§]	0.111	0.011	0.936	0.977	0.878
EPCR-EN ($\alpha = 0.5$)	0.202	0.007	0.963	0.952	0.878
Setting 2: 50% censoring					
Traditional approach					
Stepwise-AIC	0.795	0.134	0.317	0.779	0.704
Stepwise-BIC	0.239	0.130	0.168	0.986	0.704
PCR-LASSO	0.321	0.011	0.939	0.907	0.858
PCR-EN ($\alpha = 0.5$)	0.407	0.007	0.965	0.864	0.869
Ensemble approach					
EPCR-LASSO	0.169	0.017	0.903	0.964	0.865
EPCR-EN ($\alpha = 0.5$)	0.255	0.012	0.937	0.936	0.874
Setting 3: 70% censoring					
Traditional approach					
Stepwise-AIC	0.809	0.140	0.299	0.76	0.690
Stepwise-BIC	0.301	0.136	0.124	0.984	0.678
PCR-LASSO	0.368	0.018	0.905	0.892	0.842
PCR-EN ($\alpha = 0.5$)	0.46	0.011	0.945	0.842	0.855
Ensemble approach					
EPCR-LASSO	0.242	0.028	0.843	0.945	0.864
EPCR-EN ($\alpha = 0.5$)	0.348	0.018	0.903	0.904	0.872
Setting 4: 30% censoring					
Traditional approach					
Stepwise-AIC	0.555	0.260	0.337	0.809	0.733
Stepwise-BIC	0.098	0.258	0.199	0.987	0.732
PCR-LASSO	0.191	0.020	0.955	0.888	0.858
PCR-EN ($\alpha = 0.5$)	0.257	0.010	0.979	0.834	0.882
Ensemble approach					
EPCR-LASSO	0.093	0.028	0.935	0.954	0.883
EPCR-EN ($\alpha = 0.5$)	0.163	0.017	0.963	0.909	0.894
Setting 5: 50% censoring					
Traditional approach					
Stepwise-AIC	0.609	0.272	0.315	0.784	0.713
Stepwise-BIC	0.121	0.267	0.161	0.985	0.705
PCR-LASSO	0.207	0.027	0.941	0.878	0.853
PCR-EN ($\alpha = 0.5$)	0.277	0.016	0.969	0.818	0.867

TABLE 1. (Continued)

Method	FDR	FOR	TPR	TNR	AUC
Ensemble approach					
EPCR-LASSO	0.115	0.040	0.905	0.943	0.877
EPCR-EN ($\alpha = 0.5$)	0.179	0.026	0.941	0.901	0.877
Setting 6: 70% censoring					
Traditional approach					
Stepwise-AIC	0.617	0.281	0.281	0.793	0.716
Stepwise-BIC	0.127	0.275	0.128	0.987	0.696
PCR-LASSO	0.271	0.047	0.903	0.835	0.836
PCR-EN ($\alpha = 0.5$)	0.322	0.032	0.940	0.783	0.851
Ensemble approach					
EPCR-LASSO	0.149	0.066	0.845	0.927	0.862
EPCR-EN ($\alpha = 0.5$)	0.217	0.047	0.898	0.875	0.870

Abbreviation: EPCR=Ensemble penalized Cox regression; PCR=Penalized Cox regression; AUC=Areas under the receiver operating characteristic curve; EN=Elastic net; FDR=False discovery rate; FOR=False omission rate; TPR=True positive rate; TNR=True negative rate; AIC=Akaike Information Criterion; BIC=Bayesian Information Criterion; LASSO=Least absolute shrinkage and selection operator.

* The method "Stepwise-AIC (BIC)" refers to fitting a Cox model using stepwise procedures based on AIC (BIC) criterion.

† The method "PCR-LASSO [EN ($\alpha = 0.5$)]" refers to a Cox model with a LASSO-Type [EN-Type ($\alpha = 0.5$)] penalty.

§ The method "EPCR-LASSO [EN ($\alpha = 0.5$)]" refers to an Ensemble Penalized Cox Regression model whose base models were trained by Cox Regression algorithm with a LASSO-Type [EN-Type ($\alpha = 0.5$)] penalty.

EPCR-least absolute shrinkage and selection operator (LASSO) model has the lowest FDR, which indicates that the model has the lowest probability of incorrectly screening out unimportant variables, while its TPR is also at a high level among all models. So EPCR-LASSO model is better able to correctly screen out important models compared to other models. Furthermore, a comparison of the EPCR-elastic net (EN) and PCR-EN models showed that the introduction of the ensemble framework was able to reduce the FDR of variable screening while maintaining a similar FOR. As shown in Figure 1, the AUCs based on the risk scores estimated by the EPCR procedure were higher than those from the other models at all six settings.

The censoring rate reflects the level of imbalance in the database. The higher the censoring rate, the more imbalanced the database is and the lower the percentage of cases in the database. As seen in Table 1, we observed increases in mean FDR, decreases in mean TPR and AUC for all models as the censoring rate increased; however, the EPCR-LASSO and EPCR-EN models (i.e., those that used the ensemble framework) consistently performed better than their competitors. For example, at $n = 1,000$ and $p = 100$, the PCR-LASSO model's FDR increased to 0.368 when the censoring rate was increased to 70%, meaning that more than a third of the important variables identified by the model were incorrect. In contrast, the EPCR-

LASSO model was able to reduce this error by 0.146. Finally, it is worth noting that among the ensemble methods, the EPCR model with the LASSO penalty performed better overall during variable screening than the model with the elastic net ($\alpha = 0.5$) penalty. Both were used for prediction with comparable accuracy, and here the elastic net penalty model performed slightly better.

For the empirical study, Supplementary Table S1 (available in <https://weekly.chinacdc.cn/>) shows the baseline population characteristics of risk factors in the Shandong sub-dataset across overall, cases and controls. The proportion of cases present in this dataset was only 0.3%, which is a serious imbalance. For the EPCR model, the AUC for 3- and 5-year predictions were 0.691 and 0.642, respectively, while those are 0.502 and 0.525, respectively, for the Gail model (see Figure 2). Supplementary Figure S3 (available in <https://weekly.chinacdc.cn/>) shows factor importance scores based on the EPCR model. See Supplementary Materials (available in <https://weekly.chinacdc.cn/>) for details of the factor importance measures for the EPCR model. Here the red line indicates the importance score threshold that distinguished important from unimportant variables. This analysis revealed that life satisfaction, dysmenorrhea, number of miscarriages, and breastfeeding were all predicted to be influential variables, a finding that is consistent with empirical data.

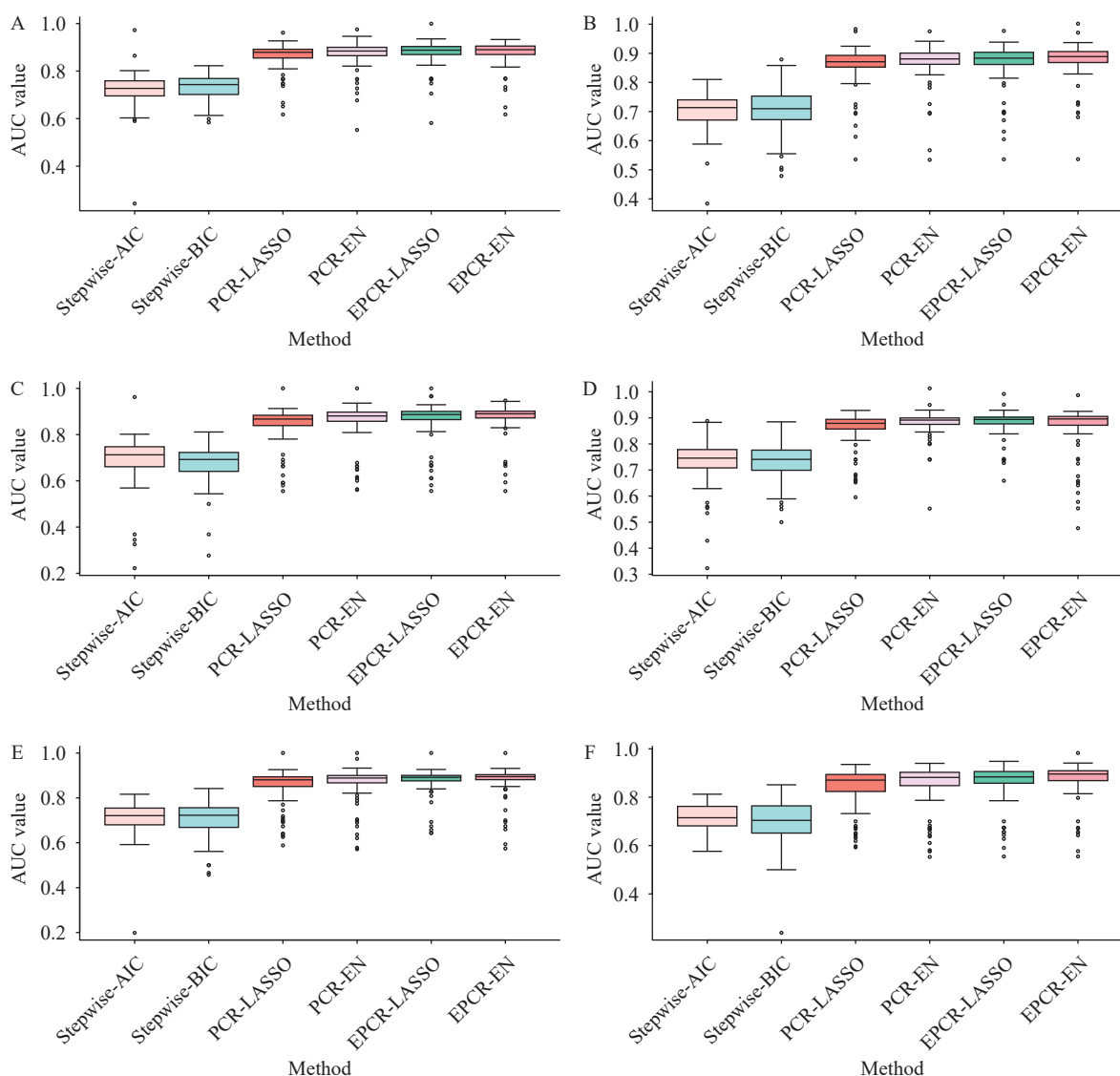


FIGURE 1. Box plots of AUC values for each modeling method. Data show boxplots of 100 replicates of settings 1–6. (A) $n=1,000$, $p=100$, 30% censoring; (B) $n=1,000$, $p=100$, 50% censoring; (C) $n=1,000$, $p=100$, 70% censoring; (D) $n=1,000$, $p=50$, 30% censoring; (E) $n=1,000$, $p=50$, 50% censoring; (F) $n=1,000$, $p=50$, 70% censoring.

Abbreviation: EPCR=Ensemble penalized Cox regression; PCR=Penalized Cox regression; AUC=Areas under the receiver operating characteristic curve; EN=Elastic net; AIC=Akaike Information Criterion; BIC=Bayesian Information Criterion; LASSO=Least absolute shrinkage and selection operator.

DISCUSSION

Most existing cancer prediction models can be divided into absolute risk models and relative risk models. The latter, however, is actually a single classifier that can only predict whether an individual is at high risk or not, but not an individual's risk of developing cancer over time in the future. The widely-used Gail model (10), a breast cancer risk assessment tool, is an absolute risk model based on five breast cancer risk factors and their interactions.

In recent years, ML has been used to improve the

predictive performance of cancer prediction models. Most current studies have focused on ML methods using classifiers such as k-nearest neighbor (KNN) (11), random forest (12) (i.e., for the identification of high-risk individuals), or Support Vector Machine (SVM) (13) or logistic regression models (i.e., for the prediction of relative risk). Moreover, most of these models only utilize the label of cancer or not in the sample, and the follow-up information of the data is not fully utilized.

At the same time, given that the databases used to develop tumorigenesis risk prediction models are

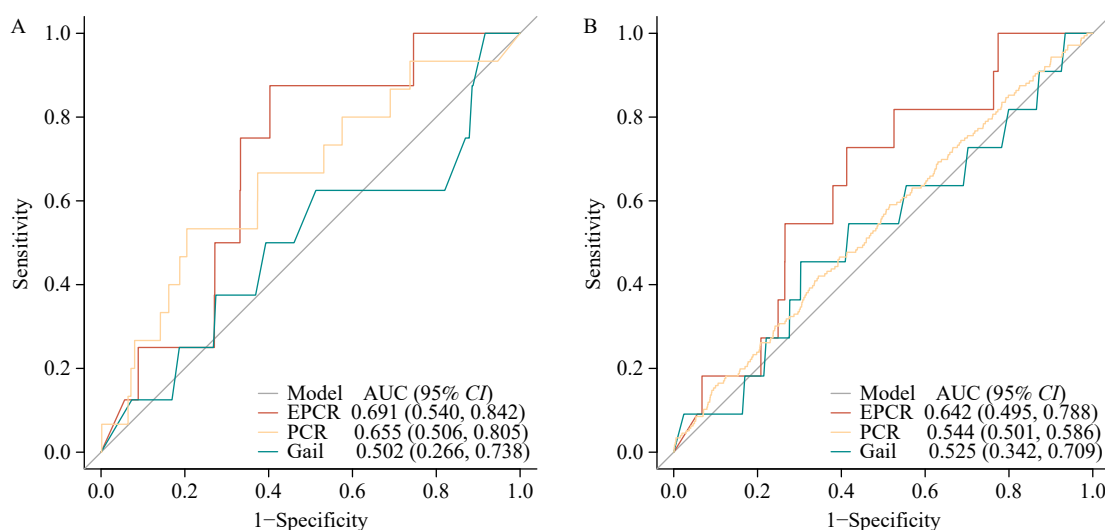


FIGURE 2. The ROC curve for 3- and 5-year model predictions of disease onset. (A) 3-year ROC; (B) 5-year ROC.

Note: Red indicates the ROC curve of the EPCR model, orange indicates the ROC curve of the PCR model, and lime green indicates the ROC curve of the Gail model.

Abbreviation: ROC=receiver operating characteristic; EPCR=ensemble penalized Cox regression; PCR=penalized Cox regression; AUC=the areas under the receiver operating characteristic curve.

mostly imbalanced, we propose applying ensemble learning methods to improve prediction performance. Specifically, the Bagging ensemble framework can be used to be able to better handle imbalanced data. Here, a PCR model was used as the base predictor, since it can make full use of follow-up information while also being able to adapt to high-dimensional data. Several simulation studies were carried out to verify the effectiveness of this method under different censoring rates settings. As shown in Table 1, the AUC based on the risk scores estimated by the EPCR model was consistently higher than that of a single PCR model or a traditional stepwise regression model under all settings. This suggests that the introduction of the Bagging ensemble framework can improve the predictive performance of PCR models, and this advantage becomes more apparent as the censoring rate increases. For example, compared to penalized logistics regression (PLR)-LASSO, the AUC of ensemble penalized logistics regression (EPLR)-LASSO increased by 1.5% and 2.2% for 30% and 70% deletion rates when $n = 1,000$, $p = 100$, respectively.

In addition, the EPCR model allows for a more robust data-driven identification of risk factors. Under all simulation settings, we calculated FDR, FOR, TPR, and TNR values for variable screening. These results showed that EPCR-LASSO had the lowest FDR while maintaining a very high TPR. For example, for Setting 1, EPCR-LASSO had the lowest FDR (0.164 lower than PCR-LASSO) as well as a TPR greater than 0.93.

This means that the variables identified by the EPCR-LASSO approach contain the fewest insignificant variables and the most significant variables compared to the other five models; that is, EPCR-LASSO is a more accurate approach for the identification of significant variables. Moreover, EPCR-LASSO continued to perform the best as the censoring rate increased. In addition, we also found that EPCR-EN can also significantly reduce FDR while maintaining the same level of TPR as PCR-EN. Taken together, these results suggest that the EPCR procedure is the best choice to use to identify important risk factors. For cancers whose etiology is unknown, the number of cases that can be used to train a prediction model is extremely small. Therefore, the exclusion or inclusion of a case can have a significant impact on the selection of risk factors. The EPCR model benefits from the Bagging ensemble framework to more robustly identify risk factors (14), which in turn can provide a more meaningful reference for studies of disease etiology.

Next, we developed and validated a breast cancer risk prediction model by analyzing the large BCCS-CW database using the EPCR procedure. Compared to the classical Gail model, our model achieved a higher degree of discrimination with higher accuracy (Figure 2). The AUC for 3- and 5-year predictions of the EPCR model were 0.691 and 0.642, which represented improvements of 0.189 and 0.117 over the classic Gail model, respectively. The other published absolute risk prediction model for the Chinese

population showed a maximum AUC of only 0.634 (15). The difference between our results and this model further demonstrates that cancer prediction models developed by the EPCR procedure are more accurate in identifying high-risk populations and may be more useful for rationally allocating healthcare resources under medical constraints.

However, it is also important to be aware that the EPCR model developed here has limitations. The application of the EPCR procedure to develop disease prediction models is only applicable where the corresponding risk factors satisfy the proportional hazards assumption. This is because the EPCR model is actually an average of multiple COX regression models. However, in most cases, especially those containing high-dimensional data, the proportional hazards assumption does not hold. Therefore, the EPCR model is more suitable for short-term disease risk prediction. As shown in Figure 2, the 5-year AUC based on the risk score estimated by the EPCR model is lower than the 3-year AUC. Therefore, the actual effectiveness of the EPCR model in predicting risk may be lower when applied to a longer (e.g., 10-year) timeframes.

Conflicts of interest: No conflicts of interest reported.

Funding: Supported by the China Postdoctoral Science Foundation (grants 2021M691911 and 2021M701997); National Key Research and Development Program of China (2016YF0901301); and the General programs of Natural Science Foundation of Shandong Province (ZR2021MH243).

doi: 10.46234/ccdcw2023.037

Corresponding authors: Jiadong Ji, jiadong@sdu.edu.cn; Zhigang Yu, yuzhigang@sdu.edu.cn.

¹ Department of Breast Surgery, The Second Hospital, Cheeloo College of Medicine, Shandong University, Jinan City, Shandong Province, China; ² School of Mathematics, Shandong University, Jinan City, Shandong Province, China; ³ Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan City, Shandong Province, China; ⁴ Institute of Translational Medicine of Breast Disease Prevention and Treatment, Shandong University, Jinan City, Shandong Province, China.

[‡] Joint first authors.

Submitted: February 07, 2023; Accepted: February 28, 2023

REFERENCES

1. Maloof MA. Learning when data sets are imbalanced and when costs are unequal and unknown. In: ICML-2003 workshop on learning from imbalanced data sets II. Washington: ICLM. 2003. <https://www.site.uottawa.ca/~nat/Workshop2003/maloof-icml03-wids.pdf>.
2. Breiman L. Bagging predictors. *Mach Learn* 1996;24(2):123 – 40. <http://dx.doi.org/10.1023/A:1018054314350>.
3. Liang G, Zhang C. Empirical study of bagging predictors on medical data. In: Conferences in research and practice in information technology series. Ballarat, Australia: OPUS. 2010:31.
4. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 2003;19(9):1090 – 9. <http://dx.doi.org/10.1093/bioinformatics/btg038>.
5. Cox DR. Regression models and life-tables. *J Roy Stat Soc B Methodol* 1972;34(2):187 – 220. <http://dx.doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
6. Cox DR. Partial likelihood. *Biometrika* 1975;62(2):269 – 76. <http://dx.doi.org/10.1093/biomet/62.2.269>.
7. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B StatMethodol* 2005;67(2):301 – 20. <http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x>.
8. Gui J, Li HZ. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005;21(13):3001 – 8. <http://dx.doi.org/10.1093/bioinformatics/bti422>.
9. Bao HL, Liu LY, Fang LW, Cong S, Fu ZT, Tang JL, et al. The Breast Cancer Cohort Study in Chinese Women: the methodology of population-based cohort and baseline characteristics. *Chin J Epidemiol* 2020;41(12):2040 – 5. <http://dx.doi.org/10.3760/cma.j.cn112338-20200507-00695>. (In Chinese).
10. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81(24):1879 – 86. <http://dx.doi.org/10.1093/jnci/81.24.1879>.
11. Chen HL, Huang CC, Yu XG, Xu X, Sun X, Wang G, et al. An efficient diagnosis system for detection of Parkinson's disease using fuzzy *k*-nearest neighbor approach. *Expert Syst Appl* 2013;40(1):263 – 71. <http://dx.doi.org/10.1016/j.eswa.2012.07.014>.
12. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 2019;7:81542 – 54. <http://dx.doi.org/10.1109/ACCESS.2019.2923707>.
13. Yu W, Liu TB, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inf Decis Making* 2010;10(1):16. <http://dx.doi.org/10.1186/1472-6947-10-16>.
14. Alelyani S. Stable bagging feature selection on medical data. *J Big Data* 2021;8(1):11. <http://dx.doi.org/10.1186/S40537-020-00385-8>.
15. Han YT, Lv J, Yu CQ, Guo Y, Bian Z, Hu YZ, et al. Development and external validation of a breast cancer absolute risk prediction model in Chinese population. *Breast Cancer Res* 2021;23(1):62. <http://dx.doi.org/10.1186/s13058-021-01439-2>.

SUPPLEMENTARY MATERIAL

Penalized Cox Regression Model

In this study, the outcome event of interest is the onset of cancer, and the initial event is the birth of the individual, so survival time is defined as the time span from birth to diagnosis of cancer — i.e., the age of the individual when cancer is first diagnosed. Similarly, we defined the censoring time as the age of the individual at the cutoff of the cancer time observation process. We first give a concise description of the EPCR model. Denote the data set as $D = \{T_i, \Delta_i, Z_i\} (i = 1, \dots, n)$, where $T_i = \min(T_i, C_i)$, T_i is the survival time of the i th individual, C_i is the censoring time of the i th individual, $\Delta_i = I(T_i \leq C_i)$ indicates that the i th individual was diagnosed with cancer ($\Delta_i = 1$) or was censored ($\Delta_i = 0$), $I(\cdot)$ is the indicator function, and Z_i is the p -dimensional predictor associated with the onset of cancer.

Given a p -dimensional predictor vector Z , the Cox (1) proportional hazard model specifies that an individual's hazard function for the onset of cancer at age t takes the form

$$\lambda(tZ) = \lambda_0(t) e^{\beta^T Z} \quad (1)$$

Here, $\beta = (\beta_1, \dots, \beta_p)$ represents a p -dimensional vector of unknown regression parameters, and $\lambda_0(t)$ is an arbitrary baseline hazard function. Let $R_i = \{j : T_j \geq T_i\}$ be the set of individuals who are still at risk at age T_i , Cox (1–2) proposed that the maximum partial likelihood estimator $\hat{\beta}$ is the maximizer of partial likelihood:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} Pl(\beta) = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^n \left\{ \frac{e^{\beta^T Z_i}}{\sum_{j \in R_i} e^{\beta^T Z_j}} \right\}^{\Delta_i} \quad (2)$$

or equivalently the maximizer of log partial likelihood:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} Pl(\beta) = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^n \Delta_i \left\{ \beta^T Z_i - \ln \sum_{j \in R_i} e^{\beta^T Z_j} \right\} \quad (3)$$

Due to the numerous predictive indicators of the onset of cancer and the low incidence of cancer, cancer-related datasets often have high dimensionality, strong correlational structure, and a small sample size. It is particularly important to efficiently screen out those predictors related to the onset of cancer from a large number of predictors. This paper screens out these predictors by imposing a penalty on the regression parameter β in the traditional Cox model. The PCR model estimates β as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} -Pl(\beta) + P(\beta) \quad (4)$$

where $P(\beta)$ is an elastic net (3) penalty function for β :

$$P(\beta) = \lambda [\alpha \beta_1 + (1 - \alpha) \beta_2^2] \quad (5)$$

The baseline hazard function $\lambda_0(t)$ from in formula (1) is also unknown. After estimating regression parameter $\hat{\beta}$, we can estimate the cumulative baseline hazard function using a Breslow estimator (4):

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^n \frac{I(\widetilde{T}_i \leq t) \Delta_i}{\sum_{j \in R_i} e^{\widehat{\beta}^T Z_j(\widetilde{T}_i)}} \quad (6)$$

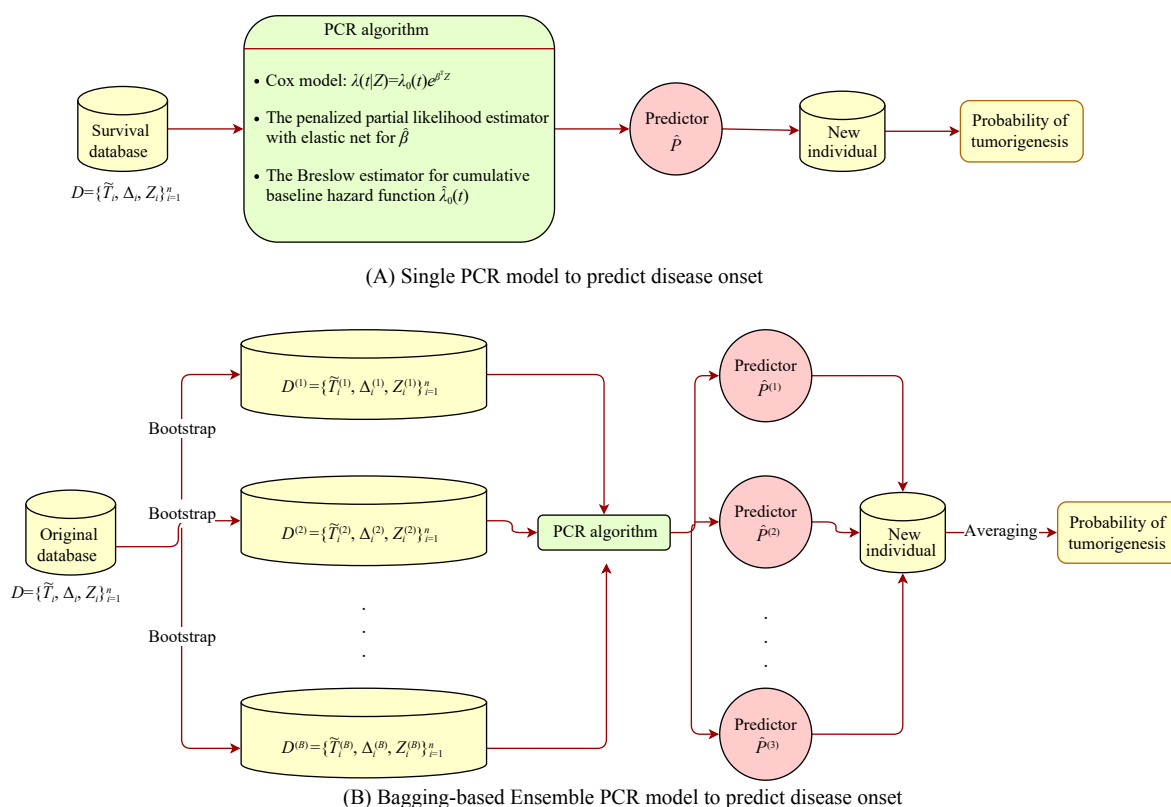
Therefore, if it is known that an individual whose predictor Z is Z^* does not have cancer at age a , then the PCR model (as shown in [Supplementary Figure S1A](#)) predicts the probability of developing cancer within τ years as

$$\widehat{P}(a, \tau, Z^*) = \int_a^{a+\tau} -\widehat{\lambda}_0(t) \exp(\widehat{\beta}^T Z^*) \frac{\widehat{S}(t)}{\widehat{S}(a)} dt \quad (7)$$

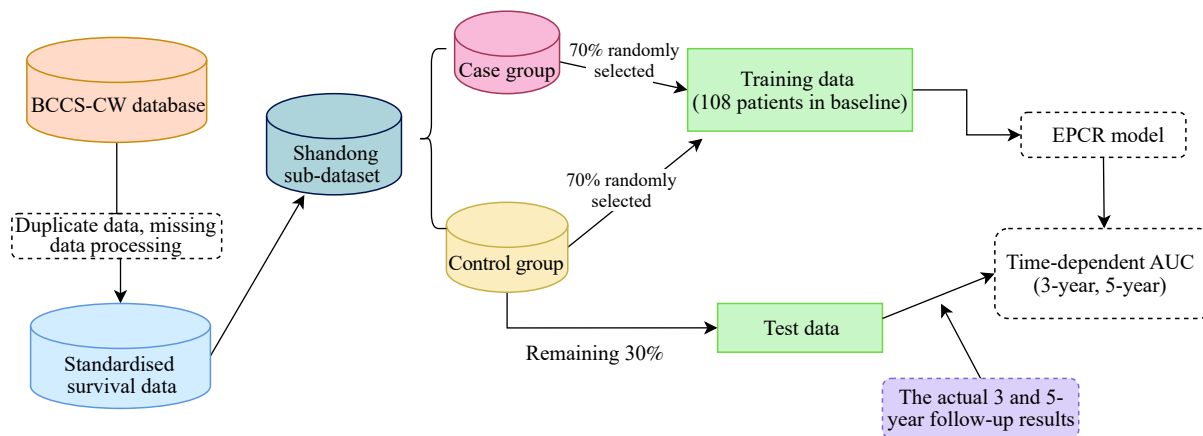
where $\widehat{S}(t) = \int_0^t -\widehat{\lambda}_0(u) \exp(\widehat{\beta}^T Z^*) du$ is the estimator of the survival function.

Predictor Importance Measures

As shown in [Supplementary Figure S1B](#), the EPCR model establishes B different and mutually independent penalized COX models. Thus, the EPCR model creates a $B \times p$ importance assessment matrix for the p -dimensional predictors, denoted E . Let $E(b, j)$ denote the (b, j) th entry of E . We then have



SUPPLEMENTARY FIGURE S1. Flowcharts for the single PCR model and the Bagging-based ensemble PCR model. Abbreviation: PCR=Penalized Cox regression.



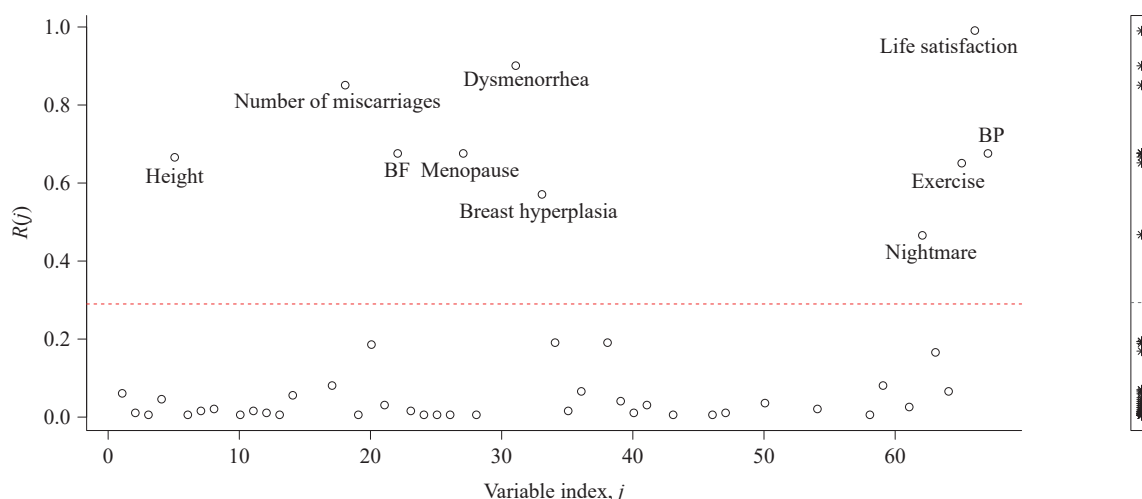
SUPPLEMENTARY FIGURE S2. Workflow of the empirical study.

Abbreviation: BCCS-CW=Breast Cancer Cohort Study in Chinese Women; EPCR=Ensemble penalized Cox regression; AUC=Areas under the receiver operating characteristic curve.

$$E(b, j) = \begin{cases} 1, & \text{If the coefficient of the } j\text{th predictor in the } b\text{th PCR model is nonzero} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Using a majority-vote summary (5), the EPCR model quantifies the importance of the j th predictor for disease occurrence as $R(j) = \frac{1}{B} \sum_{b=1}^B E(b, j), j = 1, \dots, p$.

Based on the definition of the importance assessment matrix E , it is easy to see that the importance assessment indicator $R(j)$ for the j th predictor is actually the frequency by which the j th predictor is selected by all EPCR base models. Thus, the larger $R(j)$ is, the more important predictor j is. Some studies (5) have sorted predictors



SUPPLEMENTARY FIGURE S3. Factors influencing the development of breast cancer identified by the EPCR-based assessment.

Note: Left (wide) panel shows $R(j)$ against j . Right (narrow) panel shows $R(j)$ sorted in order to identify the largest gap; for more details please see Methods. Variable j with an $R(j)$ value above the red line indicates a significant variable identified by the EPCR model. Supplementary Table S1 lists the meaning of the variable indices for explanation.

Abbreviation: EPCR=Ensemble penalized Cox regression; BP=Bean product; BF=Breast feeding.

according to the values $R(1), R(2), \dots, R(p)$, then searched for the maximum gap between any consecutive entries, ultimately choosing predictor j if $R(j)$ is above this gap.

In our study, we simply specify a certain cutoff and select predictors where $R(j)$ is greater than the cutoff. The cutoff can be $\frac{1}{p} \sum_{j=1}^p R(j)$, or 0.5 (indicating that half of the base models of EPCR have selected this variable).

Four Evaluation Metrics in Simulation Study

Denote the set of variables that are truly associated with the response variable as M , and the set of important variables comprising the model screening is \widehat{M} . In simulation studies, the following indicators are mainly used to measure the ability of the model to correctly screen variables:

- 1) False Discovery Rate (FDR) = $\frac{|\widehat{M} \cap M^c|}{|\widehat{M}|}$;
- 2) False Omission Rate (FOR) = $\frac{|\widehat{M}^c \cap M|}{|\widehat{M}^c|}$;
- 3) True Positive Rate (TPR) = $\frac{|\widehat{M} \cap M|}{|M|}$;
- 4) True Negative Rate (TNR) = $\frac{|\widehat{M}^c \cap M^c|}{|M^c|}$.

SUPPLEMENTARY TABLE S1. Population characteristics for Shandong subset of the BCCS-CW database, overall and by diagnosis of BC in the baseline.

Risk factor	Overall (N=60,397)	Case (N=154)	Control (N=60,243)	P value*
Age [†] , mean (SD)	43.56 (11.58)	46.82 (8.55)	43.55 (11.58)	<0.001
Location, n (%)				
Rural	7,903 (13.1)	27 (17.5)	7,876 (13.1)	0.129
Urban	52,494 (86.9)	127 (82.5)	52,367 (86.9)	
Occupation, n (%)				
Farmer	43,158 (71.5)	102 (66.2)	43,056 (71.5)	0.272
Worker	8,838 (14.6)	19 (12.3)	8,819 (14.6)	

Continued

Risk factor	Overall (N=60,397)	Case (N=154)	Control (N=60,243)	P value*
Teacher	336 (0.6)	1 (0.6)	335 (0.6)	
Civil service	164 (0.3)	1 (0.6)	163 (0.3)	
Individual traders	989 (1.6)	4 (2.6)	985 (1.6)	
Driver	32 (0.1)	0 (0.0)	32 (0.1)	
Services	469 (0.8)	0 (0.0)	469 (0.8)	
Staff	853 (1.4)	2 (1.3)	851 (1.4)	
Housewife	4,527 (7.5)	19 (12.3)	4,508 (7.5)	
Health care	858 (1.4)	5 (3.2)	853 (1.4)	
Student	2 (0.0)	0 (0.0)	2 (0.0)	
Others	171 (0.3)	1 (0.6)	170 (0.3)	
Education year, mean (SD)	6.14 (3.99)	5.74 (4.08)	6.14 (3.98)	0.211
Education, n (%)				
Primary or below	29,772 (49.3)	81 (52.6)	29,691 (49.3)	0.133
Junior high school	22,177 (36.7)	49 (31.8)	22,128 (36.7)	
Senior middle or vocational high school	6,814 (11.3)	23 (14.9)	6,791 (11.3)	
University	1,634 (2.7)	1 (0.6)	1,633 (2.7)	
Height, mean (SD)	1.59 (0.05)	1.59 (0.05)	1.59 (0.05)	0.201
Weight, mean (SD)	59.65 (9.10)	63.85 (10.48)	59.64 (9.09)	<0.001
BMI, mean (SD)	23.66 (3.40)	25.12 (3.60)	23.66 (3.40)	<0.001
Waistline (Chi ²), mean (SD)	2.39 (0.25)	2.52 (0.28)	2.39 (0.25)	<0.001
Hip circumference (Chi ²), mean (SD)	3.00 (0.24)	3.10 (0.25)	3.00 (0.24)	<0.001
WHR, mean (SD)	0.80 (0.05)	0.81 (0.06)	0.80 (0.05)	0.001
Number of family members, mean (SD)	3.47 (1.12)	3.37 (1.31)	3.47 (1.12)	0.29
Family annual income, mean (SD)	14,928.10 (15,807.75)	15,853.40 (13,636.80)	14,925.74 (15,812.94)	0.467
Economic status, n (%)				
Very good	337 (0.6)	1 (0.6)	336 (0.6)	<0.001
Good	8,172 (13.5)	19 (12.3)	8,153 (13.5)	
Common	50,459 (83.5)	125 (81.2)	50,334 (83.6)	
Poor	1,368 (2.3)	7 (4.5)	1,361 (2.3)	
Very poor	61 (0.1)	2 (1.3)	59 (0.1)	
Social status, n (%)				
Very good	339 (0.6)	1 (0.6)	338 (0.6)	0.007
Good	7,880 (13.0)	20 (13.0)	7,860 (13.0)	
Common	51,529 (85.3)	129 (83.8)	51,400 (85.3)	
Poor	622 (1.0)	3 (1.9)	619 (1.0)	
Very poor	27 (0.0)	1 (0.6)	26 (0.0)	
Number of miscarriages, mean (SD)	0.44 (0.81)	0.60 (0.92)	0.44 (0.81)	0.014
Breast feeding, n (%)				
Yes	57,698 (95.5)	145 (94.2)	57,553 (95.5)	0.527
No	2699 (4.5)	9 (5.8)	2690 (4.5)	
Menopause, n (%)				
Yes	17,833 (29.5)	96 (62.3)	17,737 (29.4)	<0.001
No	42,564 (70.5)	58 (37.7)	42,506 (70.6)	

Continued

Risk factor	Overall (N=60,397)	Case (N=154)	Control (N=60,243)	P value*
Dysmenorrhea, n (%)				
Yes	47,855 (79.2)	123 (79.9)	47,732 (79.2)	0.924
No	12,542 (20.8)	31 (20.1)	12,511 (20.8)	
Breast hyperplasia, n (%)				
No	56,179 (93.0)	148 (96.1)	56,031 (93.0)	0.178
Yes	4,218 (7.0)	6 (3.9)	4,212 (7.0)	
Bean product, n (%)				
Almost every day	2,494 (4.1)	9 (5.8)	2,485 (4.1)	0.085
3–4 days a week	12,266 (20.3)	32 (20.8)	12,234 (20.3)	
1–2 days a week	26,792 (44.4)	54 (35.1)	26,738 (44.4)	
Almost never	18,845 (31.2)	59 (38.3)	18,786 (31.2)	
Sleep satisfaction, n (%)				
Very satisfied	11,345 (18.8)	22 (14.3)	11,323 (18.8)	0.04
Satisfied	42,765 (70.8)	106 (68.8)	42,659 (70.8)	
Not satisfied	6,133 (10.2)	25 (16.2)	6,108 (10.1)	
Very dissatisfied	154 (0.3)	1 (0.6)	153 (0.3)	
Nightmare, n (%)				
No	57,681 (95.5)	141 (91.6)	57,540 (95.5)	0.03
Yes	2,716 (4.5)	13 (8.4)	2,703 (4.5)	
Exercise, n (%)				
Yes	3,421 (5.7)	10 (6.5)	3,411 (5.7)	0.786
No	56,976 (94.3)	144 (93.5)	56,832 (94.3)	
Life satisfaction degree, n (%)				
<25	41,350 (68.5)	75 (48.7)	41,275 (68.5)	<0.001
≥25	19,047 (31.5)	79 (51.3)	18,968 (31.5)	
Behavioral prevention degree, n (%)				
<1	45,018 (74.5)	90 (58.4)	44,928 (74.6)	<0.001
≥1	15,379 (25.5)	64 (41.6)	15,315 (25.4)	
Awareness of BC degree, n (%)				
<8	49,180 (81.4)	107 (69.5)	49,073 (81.5)	<0.001
≥8	11,217 (18.6)	47 (30.5)	11,170 (18.5)	

Note: This table shows only the important variables in Supplementary Figure S3 as well as some essential variables; other insignificant variables are omitted. The omitted variables include marital status, number of birth, number of pregnancies, number of term pregnancies, age at 1st pregnancy at term, full-term birth, breast feeding duration (month), history of contraceptive use, age of menarche, menstruation regular, family history of breast, menstrual period, menstrual cycle, benign breast disease history, nipple discharge, mamma accessoria, nipple retraction, cervical cancer history, ovarian cancer history, ovarian cyst history, diabetes mellitus, hypertension, coronary heart disease, nephritis, fresh beans, red meat, dairy products, corn, carrot, fried foods, colored vegetables or fruit, garlic, ham, pickles, sleep duration, smoking, drinking, passive smoking, tea, insomnia, waking up early, and sleeping late. For numerical risk factors, the numbers and numbers within parentheses indicate the mean and SD, respectively; for categorical risk factors, the numbers and numbers within parentheses indicate the numbers of people and their percentages of the cohort size respectively.

Abbreviation: BMI=Body mass index; WHR=Waist-to-hip ratio; BC=Breast cancer; SD=Standard deviation.

* The P value of Comparing groups for statistical differences. T-Test was used for numerical risk factors and Chi-squared test for Categorical risk factors, respectively.

† Age of the case group is their age at diagnosis of breast cancer; age of the control group is the initial age at the start of follow-up.

§ A Chinese unit of length, which is equal to one third of a meter.

REFERENCES

- Cox DR. Regression models and life-tables. *J Roy Stat Soc B Methodol* 1972;34(2):187 – 220. <http://dx.doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- Cox DR. Partial likelihood. *Biometrika* 1975;62(2):269 – 76. <http://dx.doi.org/10.1093/biomet/62.2.269>.

3. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B Stat Methodol* 2005;67(2):301 – 20. <http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x>.
4. Breslow NE. Discussion of professor Cox's paper. *J Royal Stat Soc B* 1972;34:216-7. <https://doi.org/10.1111/j.2517-6161.1972.tb00900.x>.
5. Zhu M, Fan GZ. Variable selection by ensembles for the Cox model. *J Stat Comput Simul* 2011;81(12):1983 – 92. <http://dx.doi.org/10.1080/00949655.2010.511622>.
6. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005;24(11):1713 – 23. <http://dx.doi.org/10.1002/sim.2059>.