

# Impact of Lifestyle Patterns on Breast Cancer Prognosis: Evidence from a UK Biobank Survival Study

Yishan Yao

18289275015@163.com

Macao Polytechnic University <https://orcid.org/0000-0001-5403-0887>

Xiaobing Zhai

Macao Polytechnic University

Zhichao Liang

Macao Polytechnic University

Chi Kin Lam

Macao Polytechnic University

Hui Xie

Macao Polytechnic University

Henry H Y Tong

Macao Polytechnic University

Tong Tong

Fuzhou University

Yan Chen

University of Nottingham

Ritse M. Mann

Radboud University: Radboud Universiteit

Muzhen He

Fujian Provincial Hospital

Kefeng Li

Macao Polytechnic University

Tao Tan

Macao Polytechnic University

---

## Research Article

**Keywords:** Survival analysis, Kaplan-Meier, XGBoost, TimeROC, Lifestyle Patterns

**Posted Date:** January 20th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-8588282/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Breast Cancer on July 17th, 2026. See the published version at <https://doi.org/10.1007/s12282-026-01893-w>.

# Abstract

## Background:

Understanding how lifestyle habits influence breast cancer survival is crucial for improving long-term outcomes and guiding individualized care. The purpose of this study is to comprehensively understand how dietary habits, exercise frequency, sleep quality and the frequency of tobacco and alcohol use affect the survival of breast cancer patients by analyzing the synergistic effects of multiple lifestyle habits.

## Methods:

We analyzed data from 14,901 female breast cancer patients in the UK Biobank, a large-scale population-based cohort. Three feature selection methods were applied to identify key prognostic variables. Five survival risk models were compared to investigate the relationship between breast cancer survival and lifestyle habits. Kaplan-Meier survival analysis and SHapley Additive exPlanations (SHAP) interpretability analysis were performed on risk scores calculated from the best-performing model.

## Results:

Through feature selection, we identified that age, number of medications taken, Body Mass Index (BMI), alcohol intake, diastolic blood pressure and smoking status were pinpointed as critical characteristics influencing the survival of breast cancer patients. Among all models, the eXtreme Gradient Boosting (XGBoost) model demonstrated the highest predictive performance (3 years: AUC = 0.72, 6 years: AUC = 0.727, 9 years: AUC = 0.749). Kaplan-Meier analysis showed that patients classified as high-risk based on the median risk score had significantly worse survival outcomes ( $P < 0.0001$ ). SHAP analysis further confirmed the dominant influence of age and BMI on mortality risk.

## Conclusions:

This study highlights the prognostic value of lifestyle habits in breast cancer survival. By integrating routine health indicators into interpretable machine learning models, our findings provide a practical foundation for individualized risk assessment and lifestyle-based intervention strategies.

# Introduction

Breast cancer remains a significant global health concern, with its incidence continuing to increase among women worldwide. It is now the most commonly diagnosed malignant tumor in women, second only to skin cancer, accounting for about 30% of all female cancers<sup>[1, 2]</sup>. According to the latest report released in February 2024 by the International Agency for Research on Cancer (IARC) of the World Health Organization, there were an estimated 19.97 million new cancer cases globally in 2022, including more than 2.3 million new cases of breast cancer. Moreover, approximately 670,000 women died from breast cancer, which puts this disease as the most frequent oncological cause of death in women worldwide<sup>[3]</sup>. While substantial efforts have been devoted to improving early detection and treatment,

understanding the factors that influence long term survival remains critically important. With the growing number of breast cancer survivors and their increasing motivation to adopt healthier lifestyles post-diagnosis<sup>[4]</sup>, identifying modifiable behavioral factors that may improve survival outcomes is essential. Such insights could contribute to the development of evidence-based, personalized survivorship strategies aimed at enhancing both longevity and quality of life.

Breast cancer survival risk is influenced by a broad spectrum of modifiable lifestyle habits, many of which are increasingly prevalent in modern societies. Adverse habits such as smoking, excessive alcohol consumption, poor stress management, altered reproductive patterns, obesity, and physical inactivity have all significant impacts on the survival risk of breast cancer. Among these, alcohol consumption has been identified as a major contributor to both mortality and disability<sup>[5]</sup>, while smoking<sup>[6]</sup> facilitates cancer progression by promoting metastasis and impairing overall health. Chronic psychological stress<sup>[7]</sup> may indirectly increase the mortality risk by disrupting the endocrine and immune system functions. Additionally, the widespread issue of overweight and obesity<sup>[8]</sup> can lead to hormonal imbalances and chronic inflammation, both of which may serve as potential triggers for breast cancer, and a sedentary lifestyle<sup>[9]</sup> further compromises the body's immune defense. In addition to these factors, dietary patterns have received increasing attention. Previous studies have demonstrated that high glycemic index or insulin load diets after diagnosis are associated with increased risks of breast cancer (specific mortality and all-cause mortality), emphasizing the long-term impact of nutritional choices on survivorship<sup>[10]</sup>. Furthermore, disparities in breast cancer incidence, screening access, and survival outcomes across racial and ethnic groups highlight the importance of developing individualized and culturally sensitive interventions<sup>[11]</sup>. Although these modifiable lifestyle factors have been studied individually, they frequently co-occur and may interact in complex ways, potentially influencing outcomes in a synergistic manner. Therefore, an integrated assessment of multiple behaviors is essential for a more comprehensive understanding of their collective influence on breast cancer survival.

Although some studies have explored the relationship between a single lifestyle and the survival status of breast cancer patients, studies that systematically evaluate the combined effects of multiple behaviors remain limited<sup>[12, 13]</sup>. Research focusing solely on isolated factors may fail to capture the intricate interplay and cumulative impact of real-world lifestyle patterns. Moreover, prior studies are often constrained by limited sample sizes or population heterogeneity, which may restrict the reliability and generalizability of their findings. To fill this gap, this study utilizes data from the UK Biobank, a large and well-characterized prospective cohort, to investigate how multiple lifestyle habits jointly influence the survival outcomes of breast cancer patients. Specifically, we assess the synergistic effects of dietary habits, physical activity, sleep quality, tobacco and alcohol consumption using a comprehensive machine learning framework. By identifying key behavioral predictors and quantifying their relative contributions to patient prognosis, this study aims to provide actionable evidence to inform personalized rehabilitation strategies and promote healthier lifestyles among breast cancer survivors.

## Methods

# Study Population

The UK Biobank is a large-scale prospective, population-based cohort study that recruited approximately 500,000 participants (229,041 men and 273,293 women) aged 40–69 years from 22 study evaluation centers across the UK between 2006 and 2010<sup>[14]</sup>. All participants were followed through a link in the NHS's electronic health record. At baseline, participants completed a self-administered touch-screen questionnaire that collected information on sociodemographic characteristics, health and medical history, and lifestyle-related exposures<sup>[15]</sup>. A detailed description of the recruitment procedures and population characteristics has been described elsewhere<sup>[16]</sup>. In addition to the questionnaire, participants also received physical measurements and provided blood samples, including blood specimens. All participants provided written informed consent, and the study was approved by the Northwest Multicenter Research Ethics Committee. Individuals who had requested withdraw from the UK Biobank cohort study were excluded from our analysis. Data from biological samples from the United Kingdom (<http://www.ukbiobank.ac.uk/>) are available to all researchers upon request.

## Ascertainment of analytic population

We restricted the study population to self-reported female participants with a history of breast cancer ( $n = 14,901$ ). In this study, we set the following strict inclusion criteria for these women with breast cancer to ensure the accuracy of the study subjects and the validity of the study: (a) Breast cancer diagnosed as the primary malignancy. (b) Death not on the same-day as diagnosis (i.e., survival time > 0). (c) Eventual therapeutic drugs administered to the patient are not intended to treat breast cancer.

## Exposure

Breast cancer cases were identified using inpatient admission records sourced from the Hospital Episode Statistics (England), the Scottish Morbidity Record (Scotland), and the Patient Episode Database (Wales). These records were supplemented with death registration records provided by the National Health Service (NHS) system in England and Wales, as well as the information and statistics departments in Scotland. The diagnosis of breast cancer was based on the International Classification of Diseases, Tenth Revision (ICD-10) coding system, with the ICD codes used to identify breast cancer having been selected and validated by the UK Biobank Outcomes Adjudication Group. Cancer diagnosis data were obtained through the NHS number linkage for participants in England and Wales, and through the NHS Central Register for participants in Scotland. Participants were followed from the date of their baseline assessment until the earliest of the following: a diagnosis of breast cancer, diagnosis of a non-breast malignancy (excluding non-melanoma skin cancer), loss to follow-up, or the end of the study period. Participants who were diagnosed with any cancer other than breast cancer and non-melanoma skin cancer during follow-up were censored from the study on the date of their diagnosis of those other malignancies. Specifically, the ICD-10 codes for breast cancer patients encompass the following detailed

classifications: C50.1, C50.2, C50.3, C50.4, C50.5, C50.6, C50.8, C50.9, D05.7, D05.9, and D48.9. These codes encompass both invasive and in situ subtypes, ensuring comprehensive case ascertainment.

## Covariates

Based on our clinical knowledge and prior literature, we conducted a comprehensive analysis of prognostic factors for breast cancer and selected features for evaluation. Covariates include sociodemographic, lifestyle, and genetic factors previously associated with breast cancer, which are considered here as potential confounders in determining the association between exposure and outcome. It included patient information (age, sex, and year of diagnosis) and survival information (survival period and survival status at the cut-off date of follow-up), including 19 characteristics such as personal age, smoking, alcohol consumption, BMI, frequency of insomnia, maternal genetic disease, sibling genetic disease, stroke, race, fruit intake, diastolic blood pressure, age at menopause, and age at menarche. Baseline age (in years) was calculated from the date of birth and the date the cancer was diagnosed. For categorical variables, smoking status ("never", "before", "now"), alcohol intake ("low frequency", "median frequency", "high frequency"), sleeplessness ("never/rarely", "sometimes", "usually"), nervous ("No", "Yes"), siblings with or without breast cancer ("no breast cancer", "breast cancer"), mothers with or without breast cancer ("no breast cancer", "breast cancers"), Caucasian by race ("No", "Yes"), fruit intake ("not eating", "eating 1 to 2", "≥3"), irritability ("No", "Yes"), chest pain ("No", "Yes"), exercise time ("<30 min", "≥30 min") were obtained by touch screen questionnaire. The proportion of participants with missing data on these covariates was low (< 3% of the sample), and confounding factors with missing data were uniformly attributed to the category "No" of categorical variables. For menopausal status at the time of recruitment, women were defined as premenopausal based on whether they had had periods in the previous year, while those who reported that their periods had stopped for at least one year were classified as postmenopausal. To minimize confounding due to treatment effects, 756 breast cancer patients who reported taking breast cancer-specific medications (e.g., capecitabine, tamoxifen, goserelin, anastrozole, fluorouracil, letrozole, cyclophosphamide, toremifene and exemestane) were excluded. Survival time was calculated from the data of breast cancer diagnosis to either the date of death or the end of follow-up (August 1, 2022). For participants who died, survival time was defined as the interval between diagnosis and death; for those who remained alive, it was defined as the time between diagnosis and August 1, 2022. Survival status was coded as "0" for alive and "1" for deceased.

## Ascertainment of cancer and death

The death registration records were obtained from the National Health Service (NHS) system in England and Wales, as well as from the information and statistics departments in Scotland. These records provide the exact dates of death for the women who died during the study period. The follow-up continued until each woman passed away, was lost to follow-up (i.e., lost contact or withdrew from the study), or until the end of the cancer incidence follow-up period on August 1, 2022, whichever came first.

The accuracy of the death registry has been verified through contact with the death registration offices. According to reports, the completeness of case identification by the UK Cancer Registration Service is approximately 98–99%, based on a study that linked routine cancer registrations with information in hospital episode statistics databases<sup>[17]</sup>. In this study, all death registration records pertain to deaths caused by breast cancer.

## Feature Selection

To identify the most informative prognostic variables, we applied three feature selection methods: Stepwise Cox regression (Stepcox)<sup>[18]</sup>, Least absolute shrinkage and selection operator (Lasso)<sup>[19]</sup> and eXtreme Gradient Boosting (XGBoost)<sup>[20]</sup>. The intersection of features selected by these three methods was used to determine the final set of key variables. The rationale and implementation of each method are briefly summarized below.

XGBoost employs a greedy algorithm combined with regularization techniques to iteratively select the optimal feature and split point during each tree node split, aiming to simplify the model structure and enhance generalization performance. Lasso, on the other hand, applies pressure to the regression coefficients through an L1 regularization term, forcing some coefficients to zero, thereby achieving effective feature selection and dimensionality reduction. Stepcox selects the optimal variable combination by gradually introducing or excluding variables based on statistical criteria such as Akaike Information Criterion (AIC) to identify the best-fitting model.

## Prediction Models

In this study, we systematically evaluate the prediction performance based on five models: Cox Regression<sup>[21]</sup>, DeepSurv<sup>[22]</sup>, Enet<sup>[23]</sup>, GlmBoost<sup>[24]</sup> and XGBoost, aiming to identify the optimal prediction model. The specific mechanisms by which each model utilizes the characteristic variables of breast cancer patients to predict survival outcomes is introduced below.

The Cox proportional hazards regression model assesses the impact of different factors on survival risk by analyzing the relationship between patients' characteristic variables and survival time. During the prediction process, the Cox model calculates a risk score for each patient based on their characteristic variables and uses this risk score to predict the patient's survival time or survival probability.

DeepSurv is a survival analysis method that leverages feedforward neural networks to capture complex nonlinear relationships between input features and survival risk. During the prediction process, DeepSurv takes patients' characteristic variables as input, passes them through multiple layers of neural networks to perform nonlinear transformations, and ultimately outputs the predicted survival probability or risk score.

XGBoost is a powerful gradient boosting algorithm that iteratively trains multiple weak learners (typically decision trees) and combines them to achieve robust predictive capability. During the prediction process, XGBoost makes progressive predictions through a sequence of trained decision trees based on the input characteristic variables. Each decision tree makes decisions based on specific feature split points, and the final prediction value is obtained by weighted summation of the predictions from all decision trees.

Elastic Network (ENet) is a deep learning-based model that is particularly suitable for handling high-dimensional and sparse data. During the prediction process, ENet receives the input characteristic variables and processes them through multiple layers of neural networks for feature extraction and transformation.

Gradient Boosting with Component-wise Linear Models (GLMBoost) is a predictive model that combines Generalized Linear Models (GLMs) with gradient boosting techniques. It iteratively trains multiple GLMs as weak learners using a gradient boosting algorithm and combines them to achieve robust predictive capability. During the prediction process, GLMBoost makes progressive predictions through a sequence of trained GLMs based on the input characteristic variables. Each GLM calculates a prediction value based on a specific linear combination and link function, and the final prediction value is obtained by weighted summation of the predictions from all GLMs.

## Statistical Analysis

Participants were followed from the date of the baseline assessment center visit to the date of breast cancer registration (ICD-10 C50), date of death, date lost to follow-up, or end of cancer incidence follow-up (August 1, 2022). Women diagnosed with any cancer other than breast cancer (except nonmelanoma skin cancer) during follow-up were censored at the date of diagnosis<sup>[25]</sup>.

In this study, we first selected eligible female participants with complete information from the UK Biobank database, resulting in a cohort characterized by 20 comprehensive demographic, clinical, and lifestyle-related variables. Feature selection was then conducted using Stepcox, Lasso, and XGBoost, resulting in the identification of intersection features common to all three methods. By taking the intersection of three feature selection methods, we can reduce the biases that may arise from using a single method, allowing for a more comprehensive evaluation of the importance of features and enabling the accurate identification of key factors. Model training and evaluation are carried out on the features obtained by intersection. We performed a survival analysis using patient and tumor information to identify characteristics that significantly affect patient survival. These analyses were performed after excluding patients with incomplete information.

Subsequently, we systematically evaluate the prediction performance based on the five models (Cox regression, DeepSurv, ENet, GLMBoost, and XGBoost). By using the training set to train the model and the test set to verify the accuracy of the model, we ensure the rigor of the evaluation process. In order to comprehensively evaluate the prediction effect of each model, we construct a Time-dependent receiver

operating characteristic curve (TimeROC)<sup>[26]</sup> curve, and use area under the curve (AUC) as the core performance evaluation index, which has a range from 0.5 to 1.0, and is the key statistic to measure the performance of a binary classifier. AUC values, and 95% confidence intervals (CIs) for each model were calculated for accurate model performance comparisons across years of data.

A risk score was calculated for each patient after incorporating expression values for each feature, where weights reflected the relative contribution of each variable to overall survival risk.. Patients were then divided into low risk group and high risk group based on the median risk score. Kaplan–Meier survival analysis was used to compare survival differences between these groups in both the training cohort and four external validation cohorts.

Finally, model interpretability was assessed using SHAP value<sup>[27]</sup>, which quantify the contribution of each feature to the predicted risk score. This analysis enabled the identification of key predictors that influence breast cancer prognosis. The overall analysis workflow are presented in Fig. 1.

## Results

### Population Characteristics

Among 502,164 participants in the UK biobank, a total of 16,540 cases (3.29%) of breast cancer were diagnosed among 273,175 female participants. To ensure data integrity and maintain methodological rigor, we rigorously excluded cases with incomplete information (see Fig. 2 for the exclusion criteria and screening process). After this meticulous screening process, a final total of 14,901 patients were included in this study. To validate the effectiveness and generalization ability of the model, the study population was randomly divided into two subsets: a training set ( $n = 10431$ ; 70%) and a validation set ( $n = 4470$ ; 30%).

Subsequent data analysis revealed that participants with higher risk scores were more likely to exhibit a distinct set of risk factors. In particular, individuals in the high-risk group were more likely to have the following characteristics: older age, higher BMI, current smoking habits, frequent alcohol intake, usually sleeplessness problems, non-white ethnic background, little or no consumption of fresh fruit, history of irritability, experience chest tightness or chest pain, late age at menopause or late age at menarche, high age at first birth, relatively low Polygenic risk score (PRS). Baseline characteristics of study participants are presented in Table 1. Furthermore, we present the annual number of death among breast cancer patients in Fig. 3 providing the survival status and mortality trends across the follow-up period, These statistics were subsequently used as input for downstream survival modeling.

Table 1  
Baseline characteristics of the study population

| Characteristic          | High risk<br>(N = 7205) | Low risk<br>(N = 7696) | Overall<br>(N = 14901) |
|-------------------------|-------------------------|------------------------|------------------------|
| Age, mean(SD)           | 62.89(8.80)             | 53.93(7.96)            | 58.26(9.50)            |
| BMI, mean(SD)           | 27.41(5.07)             | 27.17(4.83)            | 27.28(4.95)            |
| Smoking status, No. (%) |                         |                        |                        |
| Never                   | 4240(58.8%)             | 4413(57.3%)            | 8653(58.0%)            |
| Previous                | 2317(32.2%)             | 2759(35.8%)            | 5076(34.6%)            |
| Current                 | 648(9.0%)               | 524(6.9%)              | 1172(7.4%)             |
| (Continued)             |                         |                        |                        |

- **Table.1** (Continued)

| <b>Characteristic<sup>a,b</sup></b>       | <b>High risk<br/>(N = 8217)</b> | <b>Low risk<br/>(N = 8315)</b> | <b>Overall<br/>(N = 16532)</b> |
|-------------------------------------------|---------------------------------|--------------------------------|--------------------------------|
| <b>Alcohol intake, No. (%)</b>            |                                 |                                |                                |
| Low frequency                             | 1726(23.9%)                     | 1829(23.8%)                    | 3555(23.9%)                    |
| Median frequency                          | 2661(36.9%)                     | 2961(38.5%)                    | 5622(37.7%)                    |
| High frequency                            | 2818(39.2%)                     | 2906(37.7%)                    | 5724(38.4%)                    |
| <b>Sleeplessness, No. (%)</b>             |                                 |                                |                                |
| Never/rarely                              | 1324(18.4%)                     | 1130(14.7%)                    | 2454(16.5%)                    |
| Sometimes                                 | 3525(48.9%)                     | 3751(48.7%)                    | 7276(48.8%)                    |
| Usually                                   | 2356(32.7%)                     | 2815(36.6%)                    | 5171(34.7%)                    |
| <b>Nervous, No. (%)</b>                   |                                 |                                |                                |
| No                                        | 5365(74.5%)                     | 5760(74.2%)                    | 11075(74.3%)                   |
| Yes                                       | 1840(25.5%)                     | 1986(25.8%)                    | 3826(25.7%)                    |
| <b>Siblings illnesses, No. (%)</b>        |                                 |                                |                                |
| No Breast Cancer                          | 6807(94.5%)                     | 7133(92.7%)                    | 13940(93.6%)                   |
| Breast Cancer                             | 398(5.5%)                       | 563(7.3%)                      | 961(6.4%)                      |
| <b>Mother illnesses, No. (%)</b>          |                                 |                                |                                |
| No Breast Cancer                          | 6383(88.6%)                     | 6729(87.4%)                    | 13112(88.0%)                   |
| Breast Cancer                             | 822(11.4%)                      | 967(12.6%)                     | 1789(12.0%)                    |
| <b>Ethnic White, No. (%)</b>              |                                 |                                |                                |
| No                                        | 329(4.6%)                       | 241(3.1%)                      | 570(3.8%)                      |
| Yes                                       | 6876(95.4%)                     | 7455(96.9%)                    | 14331(96.2%)                   |
| <b>Fruit intake, No. (%)</b>              |                                 |                                |                                |
| Don't eat                                 | 543(7.5%)                       | 426(5.5%)                      | 969(6.5%)                      |
| Eat 1–2                                   | 3859(53.6%)                     | 3866(50.2%)                    | 7725(51.8%)                    |
| Eat ≥ 3                                   | 2803(38.9%)                     | 3404(44.3%)                    | 6207(41.7%)                    |
| <b>Diastolic blood pressure, mean(SD)</b> | <b>81.20(10.22)</b>             | <b>81.38(9.97)</b>             | <b>81.28(10.09%)</b>           |
| <b>Irritability, No. (%)</b>              |                                 |                                |                                |
| No                                        | 5461(75.8%)                     | 5917(76.9%)                    | 11378(76.4%)                   |

| <b>Characteristic<sup>a,b</sup></b>  | <b>High risk<br/>(N = 8217)</b> | <b>Low risk<br/>(N = 8315)</b> | <b>Overall<br/>(N = 16532)</b> |
|--------------------------------------|---------------------------------|--------------------------------|--------------------------------|
| <b>Alcohol intake, No. (%)</b>       |                                 |                                |                                |
| Yes                                  | 1744(24.2%)                     | 1779(23.1%)                    | 3523(23.6%)                    |
| <b>Chest pain, No. (%)</b>           |                                 |                                |                                |
| No                                   | 6129(85.1%)                     | 6606(85.8%)                    | 12735(85.5%)                   |
| Yes                                  | 1076(14.9%)                     | 1090(14.2%)                    | 2166(14.5%)                    |
| <b>Exercises duration, No. (%)</b>   |                                 |                                |                                |
| < 30 min                             | 4591(63.7%)                     | 4980(64.7%)                    | 9571(64.2%)                    |
| ≥ 30 min                             | 2614(36.3%)                     | 2716(35.3%)                    | 5330(35.8%)                    |
| <b>Menopause age, mean(SD)</b>       | 50.09(3.82)                     | 49.55(4.21)                    | 49.81(4.04%)                   |
| <b>Periods started age, mean(SD)</b> | 12.90(1.58)                     | 12.89(1.58)                    | 12.90(1.58)                    |
| <b>First birth age, mean(SD)</b>     | 25.38(3.83)                     | 25.33(3.67)                    | 25.35(3.75)                    |
| <b>Treatment number, mean(SD)</b>    | 2.63(2.71)                      | 3.06(2.69)                     | 2.85(2.71)                     |
| <b>Standard PRS, mean(SD)</b>        | 0.40(0.98)                      | 0.47(0.97)                     | 0.44(0.97)                     |

<sup>a</sup> Characteristics of eligible female participants. BMI ¼ body mass index; PRS ¼ polygenic risk score; SD ¼ standard deviation.

<sup>b</sup> Mean (SD) for continuous variables and n (%) for categorical variables.

To improve the prediction performance of the model, we adopt a comprehensive feature selection strategy, which combines the advantages of Lasso regression, Stepcox regression and XGBoost. Specifically, by computing the intersection of the feature sets provided by each of the three methods individually, we obtained a rigorous model evaluation process.

In the Lasso regression analysis, we construct a penalized Cox proportional hazards model and determine the optimal penalty parameter (Lambda. Min) using cross-validation (CV) technique, which is finally determined to be 0.00087. To show the results of Lasso regression, we provide two charts: one is the Lasso regression coefficient path chart (Fig. 4A), which intuitively shows which features are selected by the model and the size of their coefficients by revealing the sparsity of the contribution of each feature to the model; and the cross-validation error plot (Fig. 4B), which shows how the model's performance varies across different penalty values and supports the selection of the optimal lambda. For the XGBoost method, we first calculate the importance score of each variable and select the top 10 most influential features based on their contribution to predictive accuracy. To visually demonstrate the

importance of these features, we create a feature importance ranking diagram (Fig. 4C), which intuitively displays the relative importance of each feature in the model through a clear distribution of weights, providing a solid foundation for subsequent model interpretation and optimization. Using the Stepcox method, we have identified features with significant impacts ( $P < 0.05$ ), enabling a more targeted understanding of prognostic variables. These features were further refined and incorporated into downstream modeling processes (Fig. 4D).

By integrating the results from three distinct feature selection methods (Lasso regression, stepwise Cox regression, and XGBoost) and applying advanced technical means such as cross-validation, stepwise regression and feature importance evaluation, we identified six overlapping features consistently selected across all methods. These key features include age, BMI, smoking status, number of medications, alcohol intake and diastolic blood pressure (Fig. 5).

- Through the visual display of a Venn diagram, we can observe that although three different feature screening strategies are used, these six features are consistently identified in all methods and constitute the core part of feature intersection. Based on this finding, we have reason to speculate that these six characteristics may play a particularly significant role in predicting or influencing the survival of breast cancer patients. This inference not only emphasizes the importance of these features in the prognosis of breast cancer, but also provides valuable clues and directions for further study of the pathogenesis of breast cancer and personalized treatment strategies.

## Model Evaluation

To thoroughly evaluate the time-varying characteristics of the prediction accuracy of the model, we adopted the TimeROC method to ensure the reliability, stability and time sensitivity of the evaluation results. Specifically, we randomly divided the dataset into a training set and a test set, and constructed a TimeROC curve for both subsets across different time intervals. This curve can intuitively show the trade-off relationship between sensitivity (true positive rate) and specificity (false positive rate) of the model at different time points and different thresholds, thereby enabling a more nuanced understanding of the model's predictive performance over time. Furthermore, we calculated the area AUC values for three key time points of 3 years, 6 years and 9 years respectively (Fig. 6). By comparing the AUC values across different time intervals, we aimed to better understand the performance differences of the model for different prediction periods, thus providing a scientific basis for the subsequent model application.

In the comparative analysis of survival predictions for breast cancer patients at 3-year, 6-year, and 9-year marks, the XGBoost model consistently demonstrated superior and stable predictive performance. Figure 6 presents the specific AUC values for five models, with the length of the error values represented by the length of the lines. Specifically, both in the training and validation sets, the AUC values achieved by the XGBoost model were significantly higher than those of the Cox model, DeepSurv model, GlmBoost model, and Enet model. This result demonstrates the accuracy and reliability of the XGBoost model in breast cancer survival prediction, with its advantages being evident in predicting 3-year, 6-year, and even

9-year survival rates. As the prediction time span extended, the XGBoost model not only maintained its efficient and accurate predictive ability but also exhibited stable performance in cross-validation, with AUC values consistently remaining at a high level. Therefore, within the scope of this study, the XGBoost is recognized as the most efficient, accurate, and suitable tool for long-term survival prediction in breast cancer patients. It is expected to offer sustained and reliable support for clinical decision-making and individualized patient management.

We further use the XGBoost model to predict the risk score for each sample. Based on the median risk score derived from the training set, patients were divided into high risk and low risk groups based on the median risk score. The Kaplan-Meier survival analysis showed that, in the test cohort, patients in the high risk group had significantly poorer survival outcomes compared to those in the low risk group ( $P < 0.0001$ , Fig. 7A).

To better understand the decision-making process of the model, we ranked the importance of the features by calculating and analyzing the SHAP value. This analysis enabled us to quantify the contribution of each feature to the predicted risk score, thereby identifying the variables with the greatest influence on survival prediction. Since age and menopausal age show a large number of positive and negative impact points (Fig. 7B), it shows that age and BMI have a significant and variable impact on the results predicted by the model under different observation values. For diastolic blood pressure and alcohol intake, although two color dots were also shown, the distribution of dots was more concentrated, indicating that the effect was relatively small. Then, we give the importance ranking of the 6 selected features (Fig. 7C). It indicates that, in this study, age and BMI may be the most pivotal factors influencing the survival duration of breast cancer patients.

## Discussion

In this study, we conducted survival analysis on data from breast cancer patients in the UK Biobank, which also provides extensive information on personal health factors such as lifestyle habits. We employed three feature selection methods to identify key factors and utilized five survival analysis algorithms to establish predictive models for 3-year, 6-year, and 9-year survival outcomes. Ultimately, the XGBoost model demonstrated the best performance. Furthermore, we adopted the Kaplan-Meier method to explore and visually present the impact of various factors on the survival duration of breast cancer patients. Additionally, we utilized SHAP analysis to ascertain which features had the most significant influence on the model's predictions.

In addition, we conducted a comprehensive evaluation of five different machine learning models in terms of their performance in predicting survival outcomes using complex medical data. Among them, DeepSurv, as a deep learning algorithm, demonstrated its unique advantages, particularly in modeling survival data that exhibit both linear and nonlinear risk characteristics, showing superiority compared to the traditional Cox Proportional Hazards algorithm. This finding is consistent with the viewpoint expressed in literature<sup>[23]</sup>, which states that the DeepSurv model can be seen as an extension of the Faraggi-Simon network and experimentally verifies that DeepSurv outperforms the Cox algorithm when

handling survival data containing both linear and nonlinear risk functions. However, it is worth noting that despite DeepSurv's breakthrough in performance, it did not surpass the XGBoost model in this study. This result aligns with the research conducted by Léo Grinsztajn et al.<sup>[28]</sup>, which reported that tree-based methods are easier to achieve good predictions on tabular data than deep learning. Therefore, deep learning algorithms have an advantage in capturing complex nonlinear relationships, while machine learning algorithms, especially ensemble learning algorithms, excel in processing large-scale data and improving prediction accuracy and stability. In practical applications, the choice of algorithm depends on the nature of the specific problem, the characteristics of the data, and the available computational resources.

We noted that the Kaplan-Meier curve results from both the training and validation cohorts showed that the survival probability of the high-risk group was significantly lower than that of the low-risk group, with  $P\text{-value} < 0.0001$ . This finding not only verifies the validity of the features screened in our study, but also further emphasizes the importance of these features in breast cancer survival analysis.

In this study, we utilized three different feature selection methods to consistently identify 6 key variables that potentially play critical roles in the prognosis of breast cancer patients. SHAP interpretability analysis further confirmed the prognostic importance of these features, as each was shown to have a substantial positive or negative impact on the model's predictions. Although features such as fruit intake and psychological stress (e.g., nervousness) were not selected as key features, they demonstrated moderate importance during the initial screening process, indicating that these factors also exert a certain influence on the prognosis of breast cancer patients.

Interestingly, our research findings indicate that the genetic situations of siblings and mothers do not have a significant impact on the survival prognosis of breast cancer patients. This may be attributed to the fact that the development of breast cancer involves multigene inheritance and complex interactions with environmental factors, making the contribution of single genetic factors relatively limited in overall prognosis assessment and thereby affecting their significance in prognostic analysis.

Beyond traditional prognostic factors, our analysis revealed a noteworthy association between the use of non-breast cancer medications and increased mortality risk. Specifically, our research data clearly shows that patients who take more non breast cancer drugs face a higher risk of mortality than those who take less of these drugs. Prior evidence suggests that certain non-breast cancer drugs may contribute to tumor progression<sup>[29, 30]</sup>, or may interact adversely with oncologic treatments<sup>[31–34]</sup>. These interactions can reduce the therapeutic effect, aggravate side effects, or even trigger new health problems, thus directly or indirectly affecting the survival of patients.

In our in-depth analysis of how age influences survival predictions in breast cancer patients, SHAP charts presented a key finding: elderly individuals exhibit a significant negative trend in the impact on the prediction model, while young individuals show a positive trend. This result clearly indicates that the prognosis of elderly breast cancer patients is significantly worse than that of young patients in our data.

set. This finding is consistent with previous studies, which have reported substantial age-related differences in survival, often attributed to factors such as reduced physiological reserve, comorbidities, and increased susceptibility to treatment-related complications in older adults<sup>[35]</sup>. Further supporting this observation, additional research has identified age  $\geq 60$  as an independent predictor of poor prognosis in female breast cancer patients<sup>[36]</sup>. Together, these findings reinforce the notion that increasing age is a critical factor associated with elevated mortality risk following breast cancer diagnosis.

Moreover, the adverse impact of high BMI on breast cancer prognosis was also demonstrated. This conclusion is consistent with previous studies, which also found a strong association between high levels of obesity and overall mortality, breast cancer-specific mortality, recurrence, and distant metastasis in breast cancer patients, which further supports the association between obesity and poor prognosis<sup>[37]</sup>. These findings further emphasize the crucial role of BMI in the assessment of breast cancer prognosis and provide new insights for future exploration of breast cancer prevention and treatment strategies.

Our study also explored the impact of other characteristics on individual breast cancer survival risk. Among these, high smoking frequency and early menopausal age have been confirmed to decrease survival for breast cancer patients. Frequent smoking behavior not only elevates the risk of developing cancer but also adversely affects the prognosis of cancer patients<sup>[5]</sup>. Although the effect of alcohol intake on survival was relatively modest in our analysis, we observed that lower levels of alcohol consumption were unexpectedly associated with worse survival outcomes. This may be due to studies suggesting that moderate alcohol consumption may have a certain protective effect on cardiovascular health, thereby indirectly influencing the survival of breast cancer patients<sup>[38]</sup>, this conclusion still requires further verification.

From a clinical perspective, our model provides an accessible and cost-effective approach to predict survival outcomes in breast cancer patients. By utilizing only comprehensive personal health factors as features, our prediction model achieved accuracy rates of 72%, 73%, and 75% respectively in the 3-year, 6-year, and 9-year assessments. This result fully demonstrates that, even under resource-limited conditions, by reasonably selecting and utilizing key health indicators, we can still provide extremely valuable references for the prognosis assessment of breast cancer patients.

- However, this study also has certain limitations. Firstly, the relatively small sample size may limit the generalizability and reliability of the research results. Secondly, we did not consider all factors that may affect the survival of breast cancer patients, such as genetics, lifestyle, and socioeconomic status, which are important prognostic factors for breast cancer. Finally, in terms of data processing, although combining "No" and "Unknown" into a single category can enhance data consistency and analytical robustness, it also has obvious drawbacks, including information loss, potential bias introduction, and reduced data transparency. Future research will focus on validating the results of this study across multiple cohorts, while expanding the sample size and incorporating

more factors, such as socioeconomic status and specific medical test data, into the analysis, as well as conducting prospective randomized clinical trials to verify our findings.

## Conclusion

This study conducted an in-depth analysis of the synergistic effects among individual health-related factors and systematically demonstrated how lifestyle habits collectively influence the survival outcomes of breast cancer patients. The survival analysis model developed herein demonstrates promising performance in predicting the survival duration of breast cancer patients. It not only provides robust scientific evidence for the prognosis assessment of breast cancer patients but also provides solid guidance for personalized habit and lifestyle adjustment, possessing significant clinical guidance value.

## Abbreviations

*AIC* (Akaike Information Criterion)

*AUC* (area under the curve)

*BMI* (Body Mass Index)

*CI*s (confidence intervals)

*CV* (cross-validation)

*ENet* (Elastic Network)

*GlmBoost* (Gradient Boosting with Component-wise Linear Models)

*GLMs* (Generalized Linear Models)

*Lasso* (Least absolute shrinkage and selection operator)

*NHS* (National Health Service)

*PRS* (Polygenic risk score)

*SHAP* (SHapley Additive exPlanations)

*Stepcox* (Stepwise Cox regression)

*TimeROC* (Time-dependent receiver operating curve)

*XGBoost* (eXtreme Gradient Boosting)

# Declarations

## Author Contributions

Yishan Yao (Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing-original draft; Writing-review & editing), Xiaobing Zhai (Formal analysis; Writing-review & editing), Zhichao Liang (Formal analysis; Writing-review & editing), Chi Kin Lam (Supervision; Writing-review & editing), Hui Xie (Supervision), Tong, H.H.Y. (Data Curation), Ritse Mann (Supervision), Tong Tong (Writing-review & editing), Yan Chen (Supervision), Muzhen He (Supervision), Kefeng Li (Data Curation), Tao Tan (Writing-review & editing; Supervision; Conceptualization)

## Funding

The funding sources played no role in the study design, data collection, data analysis, and interpretation of results or the decisions made in preparation and submission of the article.

UK Biobank has received core funding from the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency, the Welsh Government, British Heart Foundation, Cancer Research UK and Diabetes UK, and National Institute for Health and Care Research.

The details of UK Biobank core funding and additional funding are reported at <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/our-funding>.

## Conflicts of Interest

The authors declare no potential conflicts of interest.

## Availability of Data and Materials Statements

This research has been conducted using data from the UK Biobank under Application Number 99946. The UK Biobank data are not publicly available due to participant privacy and data sharing restrictions. Researchers can apply for access to the data through the UK Biobank resource at <https://www.ukbiobank.ac.uk>.

**Ethical approval:** The details of UK Biobank research ethics approval are elaborated at <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>.

# References

1. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast Cancer Stat 2022 CA: cancer J Clin. 2022;72(6):524–41.
2. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. Lancet (London England). 2021;397(10286):1750–69.

3. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I et al. Global cancer statistics 2022:GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin*. 2024.
4. Ellsworth RE, et al. Impact of lifestyle factors on prognosis among breast cancer survivors in the USA. *Expert Rev pharmacoeconomics outcomes Res* vol. 2012;12(4):451–64. 10.1586/erp.12.37.
5. GBD 2016 Alcohol Collaborators. *Lancet* (London England). 2018;392(10152):1015–35. [https://doi.org/10.1016/S0140-6736\(18\)31310-2](https://doi.org/10.1016/S0140-6736(18)31310-2). Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016.
6. Passarelli MN, Newcomb PA, Hampton JM, et al. Cigarette Smoking Before and After Breast Cancer Diagnosis: Mortality From Breast Cancer and Smoking-Related Diseases. *J Clin Oncol*. 2016;34(12):1315–22. 10.1200/JCO.2015.63.9328.
7. Printz C. Anxiety and depression may contribute to an increased risk of death in some cancers. *Cancer*. 2017;123(13):2389–2389.
8. Dehesh T, Fadaghi S, Seyedi M, et al. The relation between obesity and breast cancer risk in women by considering menstruation status and geographical variations: a systematic review and meta-analysis. *BMC Womens Health*. 2023;23:392. <https://doi.org/10.1186/s12905-023-02543-5>.
9. Soldato D, Michiels S, Havas J, Di Meglio A, Pagliuca M, Franzoi MA, Pistilli B, Iyengar NM, Cottu P, Lerebours F, Coutant C, Bertaut A, Tredan O, Vanlemmens L, Jouannaud C, Hrab I, Everhard S, Martin AL, André F, Vaz-Luis I, Jones LW. Dose/Exposure Relationship of Exercise and Distant Recurrence in Primary Breast Cancer. *J Clin Oncol*. 2024;42(25):3022–32. Epub 2024 Jun 5. PMID: 38838281; PMCID: PMC11361355.
10. Maryam S, Farvid RM, Tamimi EM, Poole, Wendy Y, Chen BA, Rosner WC, Willett MD, Holmes A. Heather Eliassen; Postdiagnostic Dietary Glycemic Index, Glycemic Load, Dietary Insulin Index, and Insulin Load and Breast Cancer Survival. *Cancer Epidemiol Biomarkers Prev* 1 Febr. 2021;30(2):335–43.
11. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J Clin*. 2017;67:439–48.
12. Kreklau A et al. An Observational Study on Breast Cancer Survival and Lifestyle Related Risk Factors. *In vivo* (Athens, Greece) vol. 35,2 (2021): 1007–1015. 10.21873/in vivo.12344
13. Parada H Jr, et al. Lifestyle Patterns and Survival Following Breast Cancer in the Carolina Breast Cancer Study. *Epidemiol* (Cambridge Mass) vol. 2019;30(1):83–92. 10.1097/EDE.0000000000000933.
14. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
15. <https://>
16. Foster HME, Celis-Morales CA, Nicholl BI, Petermann-Rocha F, Pell JP, Gill JMR, et al. The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy

- lifestyle factors and health outcomes: a prospective analysis of the UK Biobank cohort. *Lancet Public Health*. 2018;3(12):e576–85.
17. Moller H, Richards S, Hanchett N, Riaz SP, Luchtenborg M, Holmberg L, et al. Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates. *Br J Cancer*. 2011;105:170–6.
  18. Hastie TJ, Pregibon D. In: Chambers S, Hastie TJ, editors. *Generalized linear models*. Chapter 6 of *Statistical Models*. Wadsworth & Brooks/Cole; 1992.
  19. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J royal Stat Soc Ser b-methodological*. 1996;58:267–88.
  20. Chen T, Guestrin C, Data Mining. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and (2016)*: n. pag.
  21. Cox DR. Regression models and life-tables[J]. *J Roy Stat Soc: Ser B (Methodol)*. 1972;34(2):187–202.
  22. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24. 10.1186/s12874-018-0482-1. PMID: 29482517; PMCID: PMC5828433.
  23. Paszke A, Chaurasia A, Kim S et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation[J]. 2016.10.48550/arXiv.1606.02147
  24. Peter Buehlmann and Bin Yu. Boosting with the L2 loss: regression and classification. *J Am Stat Assoc*. 2003;98:324–39.
  25. Guo W, Key TJ, Reeves GK. Adiposity and breast cancer risk in postmenopausal women: results from the UK Biobank prospective cohort. *Int J Cancer*. 2018;143:1037–46.
  26. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol*. 2017;17:53. <https://doi.org/10.1186/s12874-017-0332-6>.
  27. Lundberg S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:170507874*.
  28. Grinsztajn L, Oyallon E. Why do tree-based models still outperform deep learning on tabular data? 2022, *arXiv:2207.08815*.
  29. Collins JA, Blake JM, Crosignani PG. Breast cancer risk with postmenopausal hormonal treatment. *Hum Reprod Update*. 2005;11:545–60.
  30. Støer NC, et al. Drug use and cancer risk: a drug-wide association study (DWAS) in Norway. *Cancer Epidemiol Biomark Prev*. 2021;30:682–9.
  31. Baker AF, Dorr RT. Drug interactions with the taxanes: clinical implications. *Cancer Treat Rev*. 2001;27:221–33.
  32. Hamy A-S, et al. Celecoxib with neoadjuvant chemotherapy for breast cancer might worsen outcomes differentially by cox-2 expression and er status: exploratory analysis of the REMAGUS02

- trial. JCO. 2019;37:624–35.
33. Tyler T. Drug interactions in metastatic breast cancer. J Oncol Pharm Pr. 2011;17:236–45.
  34. Bibi R, et al. Prevalence of potential drug-drug interactions in breast cancer patients and determination of their risk factors. J Oncol Pharm Pr. 2021;27:1616–22.
  35. Zhou Y, Wen Y, Xiang Z, Ma J, Lin Y, Huang Y, Chen C. Cancer Survival Trends in Southeastern China, 2011–2021: A Population-Based Study. Clin Epidemiol. 2024;16:45–56. PMID: 38318284; PMCID: PMC10840559.
  36. Chen HL, Zhou MQ, Tian W, Meng KX, He HF. Effect of Age on Breast Cancer Patient Prognoses: A Population-Based Study Using the SEER 18 Database. PLoS ONE. 2016;11(10):e0165409. 10.1371/journal.pone.0165409. PMID: 27798652; PMCID: PMC5087840.
  37. Pang Y, Wei Y, Kartsonaki C. Associations of adiposity and weight change with recurrence and survival in breast cancer patients: a systematic review and meta-analysis. Breast Cancer. 2022;29:575–88. <https://doi.org/10.1007/s12282-022-01355-z>.
  38. Newcomb PA, Kampman E, Trentham-Dietz A, et al. Alcohol consumption before and after breast cancer diagnosis: associations with survival from breast cancer, cardiovascular disease, and other causes. J Clin Oncol. 2013;31(16):1939–46. 10.1200/JCO.2012.46.5765.

## Figures

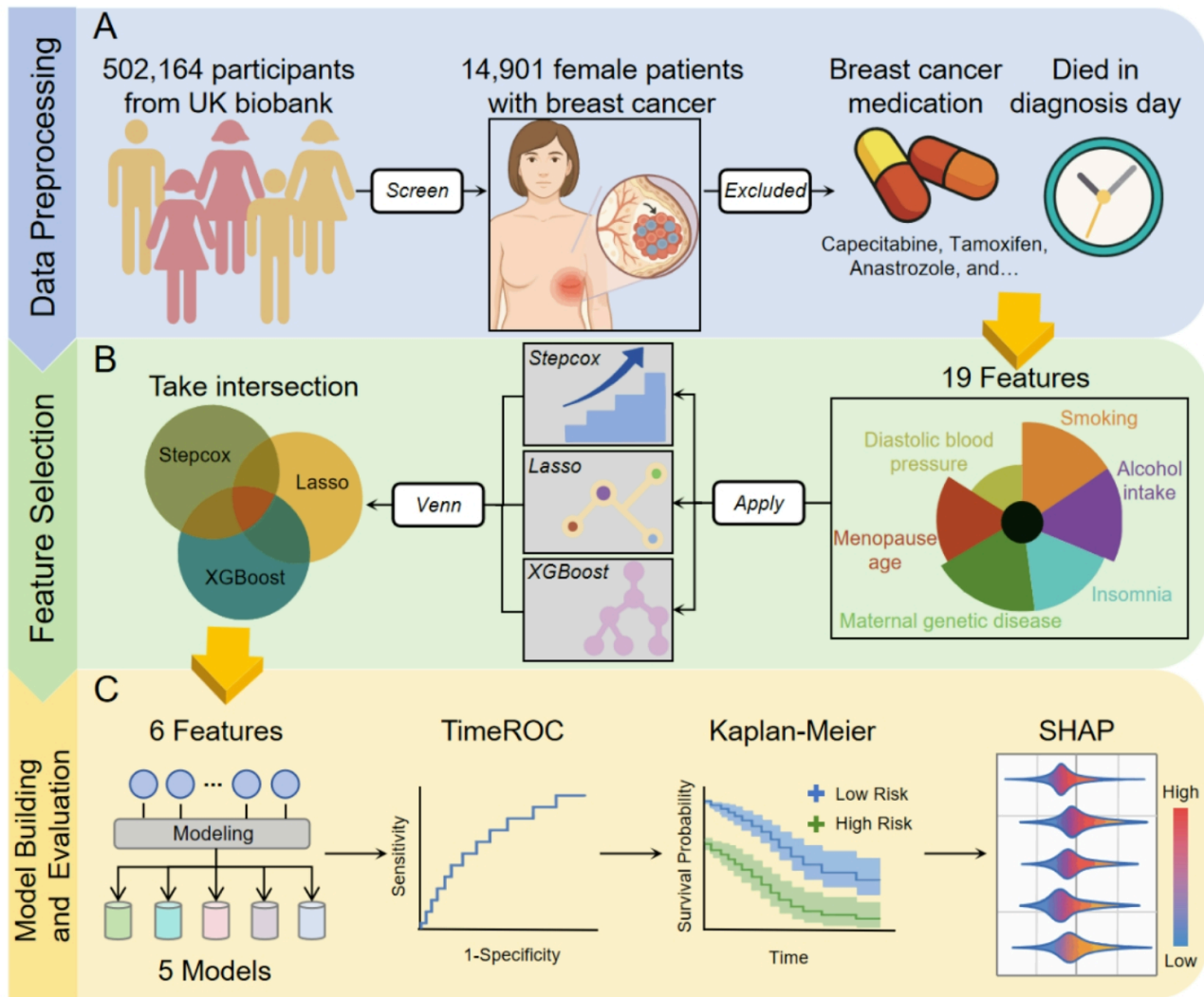


**Figure 1**

Workflow overview with data preprocessing, feature selection, model building, and model evaluation. **A** Data preprocessing. All patients are screened according to the inclusion criteria. **B** Conduct feature selection based on individual health factors. Three feature selection methods are applied to screen 19 features, and the features obtained from their intersection are taken as key features. **C** Model Building and Evaluation. Five survival models were constructed with key features, and the model evaluation was carried out by using TimeROC and KM curves, and the interpretability analysis was carried out by using SHAP diagrams.

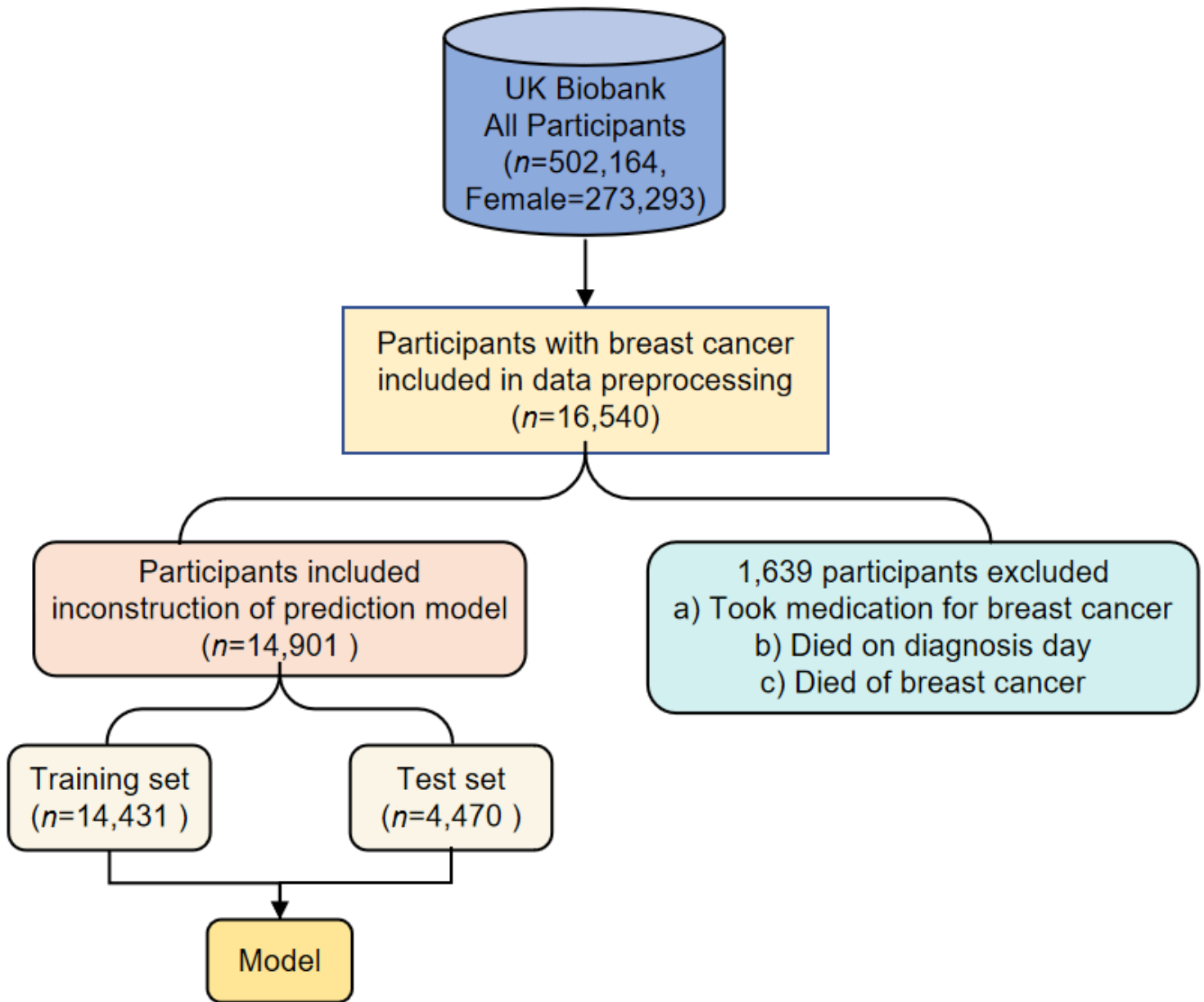


**Figure 2**

Regarding the process of data preprocessing and dataset partitioning.

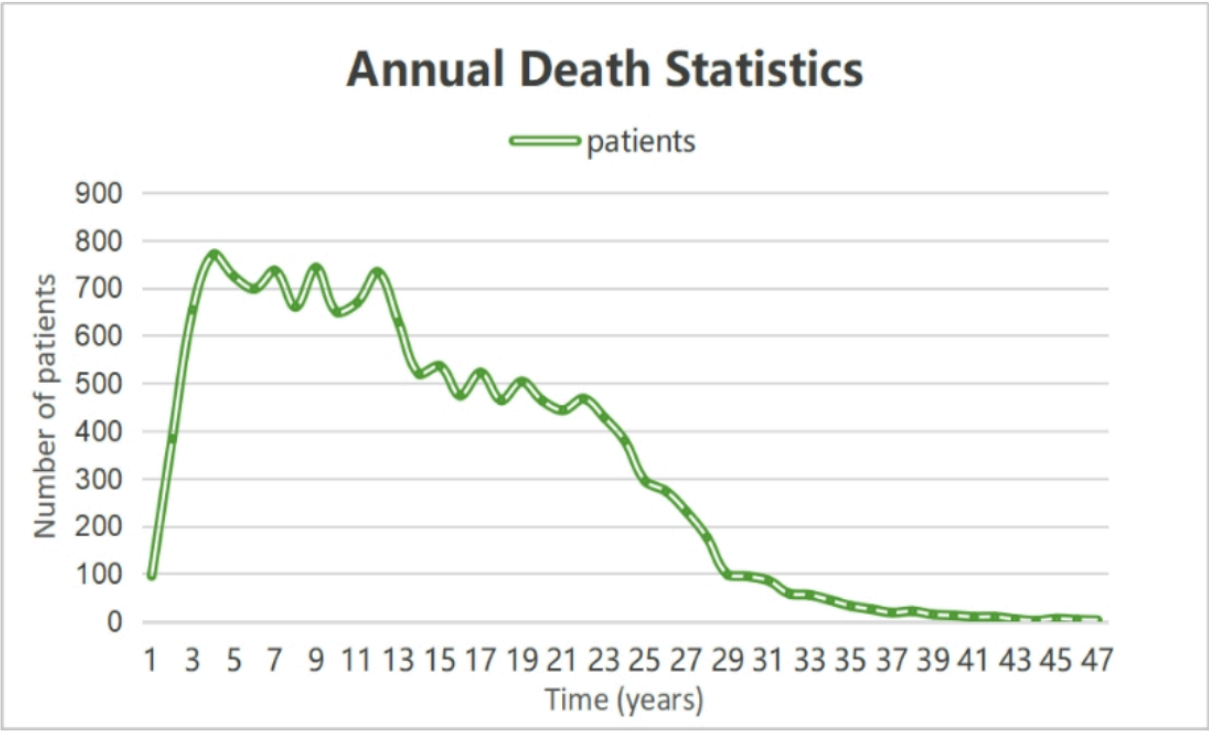


**Figure 3**

The aggregated statistical data on annual deaths.

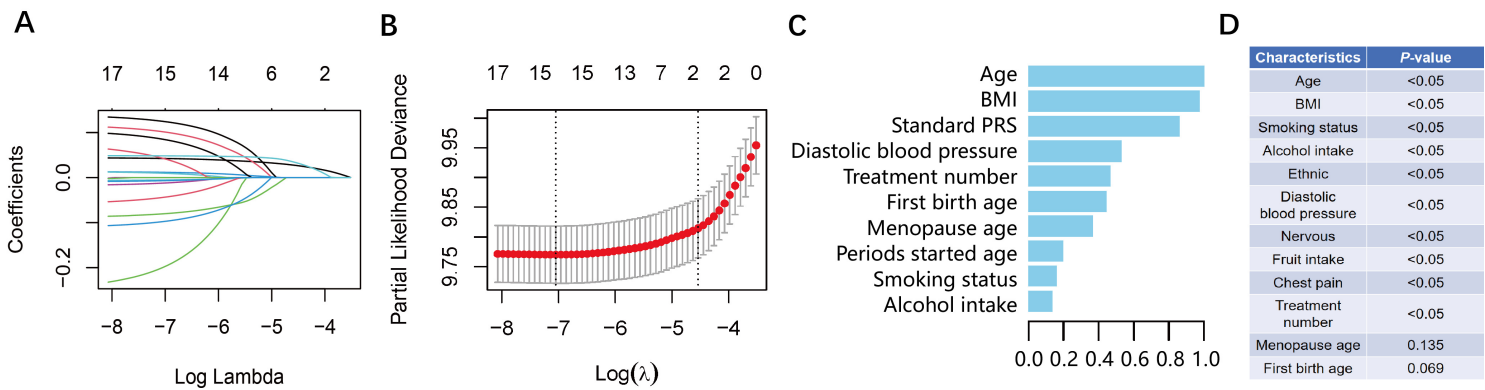


**Figure 4**

Feature selection. **A** Pathway plot of the Lasso regression coefficients **B** Impact plot of penalty parameter of the Lasso regression **C** Characteristics importance ranking plot of XGBoost **D** Characteristics P-value table of Stepcox

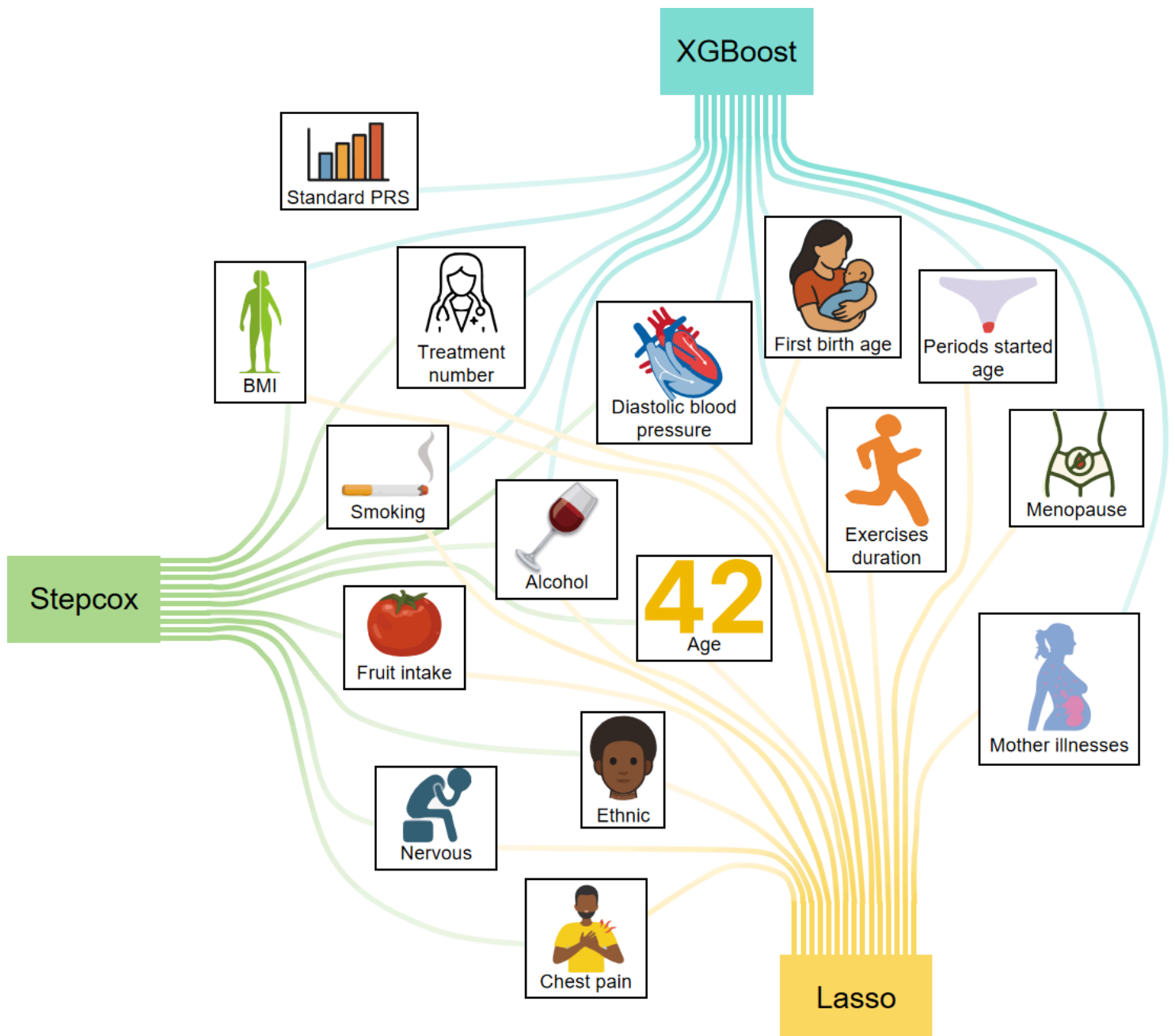


**Figure 5**

Network Venn diagram for three feature selection methods

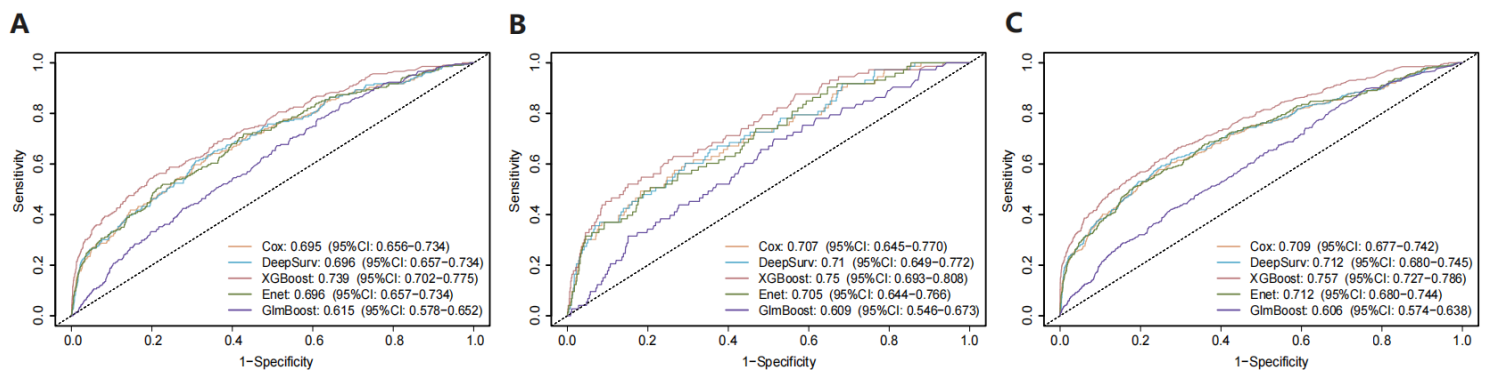


Figure 6

The ROC curves for the test cohort at three, six, and nine years.

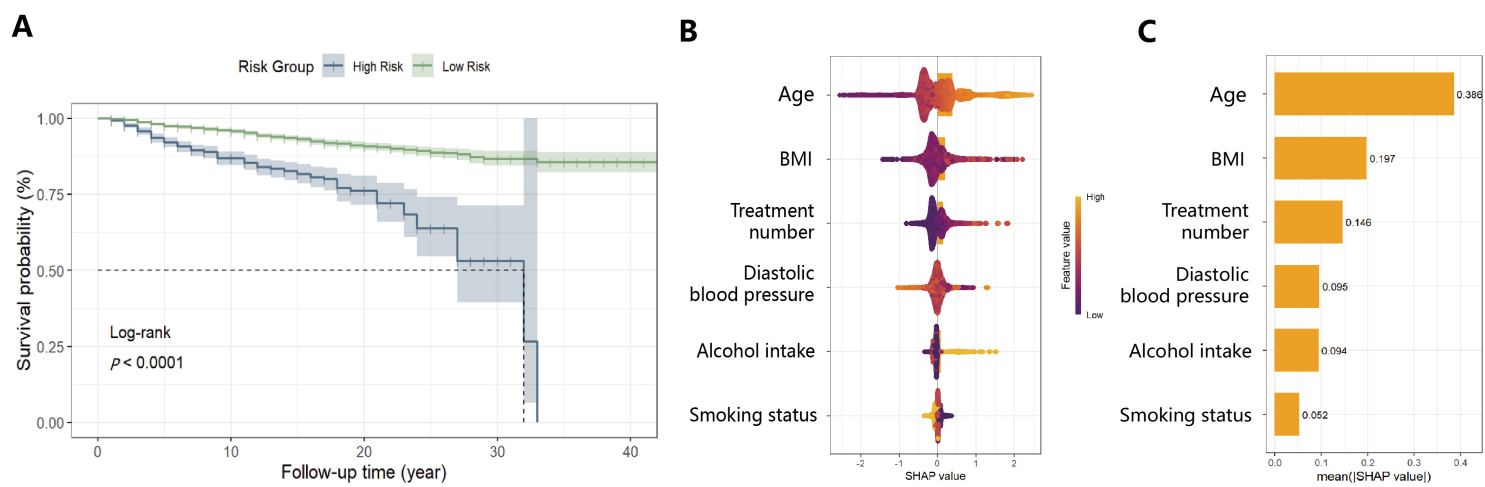


Figure 7

**A** The Kaplan-Meier curves for the high risk and low risk groups of the test cohort

**B** Feature ranking-distinguishing feature value of SHAP interpretability analysis

**C** Feature ranking-not distinguishing feature value of SHAP interpretability analysis