

Article

Three predictive models for fear of cancer recurrence among recovering breast cancer patients

Huiying Yang¹, Jing Zhang^{2,*}, Xiaolu Li³¹ Lishui Hospital of Traditional Chinese Medicine, No. 800, Zhongshan Street, Liandu District, Lishui 323000, Zhejiang, China² The Fourth Affiliated Hospital of Harbin Medical University, Harbin 151001, China³ Department of General Practice, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan 250021, Shandong, China

* Corresponding author: Jing Zhang, 13704815852@163.com

CITATION

Yang H, Zhang J, Li X. Three predictive models for fear of cancer recurrence among recovering breast cancer patients. *Psycho-Oncologie*. 2025; 19(3): 3642.
<https://doi.org/10.18282/po3642>

ARTICLE INFO

Received: 4 July 2024

Revised: 1 November 2024

Accepted: 14 November 2024

Available online: 15 September 2025

COPYRIGHT



Copyright © 2025 by author(s).

Psycho-Oncologie is published by PiscoMed Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.

<https://creativecommons.org/licenses/by/4.0/>

Abstract: Objectives: Fear of cancer recurrence (FCR) is a common psychological issue among breast cancer patients, yet predictive methods are lacking. This study aimed to construct and compare three models for predicting FCR among recovering breast cancer patients. **Methods:** A total of 490 breast cancer patients in rehabilitation were recruited from a university hospital in Harbin, China. Factors affecting FCR were collected using specific questionnaires. Multivariate analysis identified effective influencing factors, which were used as independent variables in a logistic regression model, an artificial neural network (ANN) model, and a random forest model. The dependent variable was high-level FCR. Models were validated using 10×10 cross-validation, and their performance was compared using AUC values, sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. **Results:** Six factors influenced FCR: age, occupational status, chemotherapy, self-image, cancer-caused fatigue, and disease perception. The logistic regression model showed an accuracy of 83.7%, sensitivity of 79.3%, specificity of 86.9%, positive predictive value of 85.1%, negative predictive value of 81.7%, and an AUC value of 0.746. The ANN model had an accuracy of 78.2%, sensitivity of 80.0%, specificity of 75.8%, positive predictive value of 81.9%, negative predictive value of 73.4%, and an AUC value of 0.746. The random forest model demonstrated an accuracy of 84.4%, sensitivity of 89.4%, specificity of 77.4%, positive predictive value of 84.4%, negative predictive value of 84.2%, and an AUC value of 0.892. **Conclusion:** The random forest model outperformed the ANN and logistic regression models in predicting FCR in recovering breast cancer patients, making it a valuable tool for screening high-risk FCR groups during rehabilitation.

Keywords: predictive models; oncology; breast cancer; rehabilitation

1. Introduction

Breast cancer is the most prevalent malignancy in China [1]. China ranks first in the world in terms of the number of new cases of breast cancer per year, accounting for approximately 18.4% of global breast cancer cases [2]. However, with improved early screening rates, medical technology, and increased awareness of health care, the 5-year survival rate for breast cancer patients in China is 72.7% [3]. However, during the 5-year survival period, it is important to pay attention not only to the changes in disease indicators but also to the psychological problems of patients, with fear of cancer recurrence being one of the most common and important psychological problems [4]. Fear of cancer recurrence (FCR) refers to patients' fear and concern about the possible metastasis returning and progressing in the future [5]. FCR, which is a common psychological problem, is prevalent in cancer patients to varying degrees

and can persist for long periods after the end of cancer treatment, with some patients consistently experiencing high levels of FCR [6]. Studies have shown that 72% of cancer patients have FCR during treatment and recovery, with 49% of them having high levels of fear [7]. Breast cancer patients have the highest FCR of all cancer patients [8]. Low levels of fear, which can be seen as normal psychological manifestations, can lead to positive health-seeking behavior and raise patients' awareness of recurrence, but high levels of fear can seriously affect patients' lives and treatment, leading to over screening or avoidance of screening [9]. Therefore, it is essential to screen patients with high FCR as early as possible and to identify the factors that influence them.

Previous studies have shown that demographic characteristics and disease-related factors, such as age, occupation, marital status, education level [10–13], time of illness, mode of surgery, radiotherapy, chemotherapy [14–17], and psychosocial factors, such as self-image [18], anxiety [19] depression [20], social support [21], cancer-related fatigue [22], hope [23], and disease perception [24], are associated with FCR, but the impact on recovering breast cancer patients is not known.

Accurately predicting the risk of FCR in cancer patients is an effective way to construct predictive models by incorporating specific predictors [25]. However, FCR predictive models for breast cancer have not been identified. Artificial neural networks combine several simple parallel processing units into a system that processes information by responding dynamically to external inputs [26]. The underlying neural network, which has a three-layer internal structure, including output, hidden and input layers, is widely used in medical research, such as test data analysis, bioactivity detection, medical expert system building and predictive model construction, due to its simultaneous processing of different information, high nonlinearity, good fault tolerance and computational accuracy, as well as low sample size requirements [27]. The random forest forms a new training set based on the original training set through bootstrap sampling. Each new training set is used to build a decision tree by means of random feature selection, and the random forest is a combination of decision trees [28]. Random forest not only improves the accuracy of the model but also estimates the generalization error and correlation coefficient using OOB data and shows good predictive performance for small samples and low-dimensional data [29]. Traditional logistic regression models are easy to use and can intuitively identify protective and risk factors for each influence.

Therefore, in this study, a cross-sectional survey was conducted to identify the effective influencing factors, and based on this, three models for predicting FCR in recovering breast cancer patients were constructed using the effective influencing factors as the independent variables and FCR with and without high levels as the dependent variables for logistic regression, artificial neural network and random forest.

2. Methods

In this study, a cross-sectional design was used to construct three FCR prediction models for breast cancer patients in recovery.

2.1. Participants

Study participants were surveyed from September 2021 to November 2022 in the departments of General Surgery, Oncology, and Radiotherapy at University Hospital in Harbin, China.

The eligibility criteria were as follows:

(i) pathological diagnosis of breast cancer; (ii) recovering patients who have undergone surgery and have completed radiotherapy and/or chemotherapy (during which endocrine therapy can be administered) from 1 week to 5 years later, an expected survival period of ≥ 1 year; (iii) over age 18; (iv) able to speak and read in Chinese; (v) awareness of the disease and written informed consent.

We excluded individuals who had (i) severe cognitive impairment; (iii) psychiatric or psychotic disorders; (iv) coma, delirium, lethargy and other disorders of consciousness; (v) diagnosed pathological anxiety or depression; and (vi) cancer that recurred or metastasized.

The questionnaire survey is preceded by a uniform training of nurses in the relevant departments, who then explained the purpose of the survey and distributed questionnaires to patients who met the criteria. Outpatient and inpatient follow-up visits were face-to-face. Those who were unable to participate independently were able to have a family member complete the survey for them.

2.2. Sample size

The sample size of this study for a logistic regression model is at least 10-20 times the number of independent variables. Considering that 15% of the questionnaires might not be useable due to omissions or other reasons, the independent variable in this study is 28, taken 15 times, for a sample size of $n = 28 \times 15 (1 + 15\%) = 483$.

The study was approved by local human research ethics committees in accordance with the Declaration of Helsinki, and all participants gave informed consent. The ethical review number is 2021-WZYSL SC-08.

2.3. Research measures

Demographic and clinical characteristics

Demographic and clinical characteristics were obtained with a self-report questionnaire, including age, sex, household income status, marital status, occupational status, education level, mode of surgery, chemotherapy and/or radiotherapy, cancer stage and duration of illness [30].

2.3.1. Fear of cancer recurrence

Gu et al. [31] translated the Simplified Disease Progression Scale (FoP-Q-SF) to measure FCR using the Chinese version. Responses were measure on a 5-point Likert scale ranging from 12–60, with higher scores indicating greater fear and a FoP-Q-SF score of no less than 34 indicating a high fear of relapse. Cronbach's alpha coefficient was 0.822 in this study.

2.3.2. Self-image

Hopwood et al. [32] developed and Zhang [33] translated the Chinese version of the self-image scale was used. The scale consists of 10 items on three dimensions

measured on a four-point scale with a total score of 0 to 30, with a BIS score of ≥ 10 indicating that the patient has a self-image problem. Cronbach's alpha coefficient was 0.624 in this study.

2.3.3. Hope

Zhao and Wang [34] translated the Chinese version of the Hope Scale, which consists of 12 items in three dimensions, was used. Cronbach's alpha coefficient was 0.872 in this study.

2.3.4. Social support

Zimet [35] adopted and Jiang et al. [36] developed the social support scale. The Chinese version of the scale consists of 12 items in 2 dimensions on a 7-point scale. The higher the score is, the more supported the individual feels. Cronbach's alpha coefficient was 0.950 in this study.

2.3.5. Cancer Fatigue Scale

Okuyama [37] adopted and Zhang et al. [38] developed, this scale is specific to fatigue in cancer patients and covers 15 items in three dimensions on a 5-point scale, with total scores ranging from 0 to 66, with higher scores indicating higher levels of fatigue. Cronbach's alpha coefficient was 0.891 in this study.

2.3.6. Depression

Adopted from Wang [39], the Chinese version of this scale has a 9-item questionnaire that assesses depression over the past two weeks on a 4-point scale. Cronbach's alpha coefficient was 0.756 in this study.

2.3.7. Anxiety

Adopted from 2013 Betty Wang's [39] Chinese version, this scale has 7 items assessing anxiety over the past two weeks and is scored on a 4-point scale. Cronbach's alpha coefficient was 0.711 in this study.

2.3.8. The brief illness perception questionnaire

Using a simplified version of the disease perception questionnaire developed by Broadbent and Mei et al. [40] translated this scale has three dimensions and a total of nine items. Higher scores represent a higher level of perceived disease threat to the patient. The ninth question is open-ended and is not surveyed for analysis in this study. Cronbach's alpha coefficient was 0.726 in this study.

2.4. Statistical analysis

Data were analyzed using SPSS version 26.0. Patients were categorized into two groups based on their Fear of Cancer Recurrence (FCR) scores: a low FCR group (< 34 points) and a high FCR group (≥ 34 points). Continuous variables were expressed as means with standard deviations, while categorical variables were presented as percentages and composition ratios. For comparison between groups, the chi-square (χ^2) test was employed for categorical variables, and independent t-tests were used for continuous variables. Significant factors from the univariate analysis ($P < 0.05$) were further evaluated using forward stepwise multivariate analysis to identify independent predictors of FCR. These significant variables were then used to build three predictive models: a logistic regression model, an artificial neural network (ANN) model, and a

random forest model. For each model, 10×10 cross-validation was performed to assess performance. The following evaluation metrics were calculated for each model: area under the curve (AUC), sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and F-measure. These metrics were used to compare the predictive power and overall effectiveness of each model in identifying breast cancer patients at high risk of FCR during recovery.

3. Results

Demographic and clinical characteristics of the low FCR group and the high FCR group.

A total of 490 breast cancer patients in recovery were included in the study. The differences in demographic and clinical characteristics between the low FCR group and the high FCR group were compared separately as shown in **Table 1**.

Table 1. Demographic and tumor characteristics ($N = 490$).

Factors	Overall	<34 points	≥34 points	χ^2	P
	490	208	282		
Sociodemographic factors					
Age (years)				15.511	0.000**
<45	30	21	9		
45~65	397	170	227		
>65	63	17	46		
Monthly per capita household income (yuan)				3.008	0.222
Less than \$1000	64	21	43		
1000~2500	144	61	83		
Above 2500	282	126	156		
Education level				0.876	0.831
Lower secondary and less	91	38	53		
High school or secondary school	173	76	97		
Tertiary	165	66	99		
Bachelor's degree or more	61	28	33		
Marital status				0.798	0.850
Unmarried	10	4	6		
Married	416	180	236		
Divorced	22	8	14		
Bereaved spouse	42	16	26		
Occupational status				27.839	0.000**
Employed	316	151	165		
No position	96	45	51		
Retirement	78	12	66		
Clinical features					
Duration of illness (months)				1.949	0.583

Table 1. (Continued).

Factors	Overall	<34 points	≥34 points	χ^2	<i>P</i>
<12	114	43	71		
12~23	111	52	59		
24~35	121	51	70		
≥36	144	62	82		
Cancer stage				2.949	0.229
I	49	18	31		
II	198	93	105		
III	243	97	146		
Surgical type				7.471	0.113
Modified radical surgery	65	34	31		
Classical radical surgery	76	35	41		
Simple mastectomy	196	84	112		
Partial mastectomy	81	25	56		
Breast conservation	72	30	42		
Chemotherapy				12.531	0.000**
Yes	283	101	182		
None	207	107	100		
Radiotherapy				1.995	0.158
Yes	242	95	147		
None	248	113	135		

The mean FCR score was 39.93 ± 7.78 , the mean physical health dimension score was 19.96 ± 4.07 , and the mean social family dimension score was 19.97 ± 4.59 . Among them, a recurrence fear scale score ≥ 34 was the cutoff, with 208 cases (42.45%) having low FCR and 282 cases (57.55%) having high FCR. The differences in patients' observed variables between the low FCR group and the high FCR group were compared separately as shown in **Table 2**.

Table 2. Comparison of patient variables between the two groups ($N = 490$).

	<34 points	≥34 marks	<i>F</i>	<i>P</i>
Self-image	20.89 ± 4.58	22.77 ± 2.43	24.121	0.000**
Cognitive dimensions	8.28 ± 2.24	9.22 ± 1.48		
Behavioral dimensions	4.30 ± 1.21	4.55 ± 1.00		
Emotional dimension	8.31 ± 1.94	9.00 ± 1.66		
Cancer-related fatigue	43.76 ± 14.91	49.93 ± 9.76	59.972	0.000**
Somatic fatigue	13.80 ± 7.11	16.25 ± 4.85		
Emotional exhaustion	8.81 ± 4.72	6.97 ± 3.52		
Cognitive fatigue	21.15 ± 12.81	26.71 ± 9.04		
Social support	47.46 ± 18.67	46.31 ± 18.32	0.777	0.378
External support	31.57 ± 12.85	30.72 ± 12.49		
Intrafamily support	15.89 ± 6.35	15.59 ± 6.33		

Table 2. (Continued).

	<34 points	≥34 marks	<i>F</i>	<i>P</i>
Anxiety	18.12 ± 4.55	18.43 ± 4.12	2.678	0.102
Depression	23.11 ± 5.63	23.15 ± 5.23	3.785	0.052
Hope	29.01 ± 8.14	28.74 ± 8.12	0.043	0.836
Maintaining intimate relationships with others	9.81 ± 3.15	9.41 ± 3.03		
Taking positive action	9.42 ± 3.11	9.44 ± 3.11		
A positive attitude toward the future and reality	9.78 ± 2.88	9.90 ± 3.02		
Disease Awareness	45.08 ± 12.47	50.67 ± 7.76	39.241	0.000**
Cognition	32.85 ± 9.29	36.62 ± 5.88		
Emotions	6.75 ± 2.69	7.70 ± 2.39		
Disease comprehension	5.47 ± 2.62	6.35 ± 2.54		

* Indicates $P < 0.05$, ** indicates $P < 0.01$.

The results of the logistic regression analysis showed that there were six final screening factors (see **Table 3**), namely, age, occupational status, chemotherapy, self-image, cancer-caused fatigue and disease perception, of which age and occupational status were protective factors; chemotherapy, self-image, cancer-caused fatigue and disease perception were the main risk factors for developing FCR.

Table 3. Multiple linear regression results of FCR dimensions.

Variable name	Regression coefficient	Standard error	<i>P</i>	<i>OR</i> value	95% CL	
					Lower limit	Upper limit
Constants	−4.834	1.13.	0.000	0.008		
Age			0.020			
<45	−1.539	0.551	0.005	0.215	0.073	0.632
45~65	−0.592	0.354	0.094	0.553	0.276	1.107
Occupational status			0.000			
Employed	−1.816	0.368	0.000	0.163	0.079	0.335
Separation from service	−1.585	0.408	0.000	0.205	0.092	0.456
Chemotherapy	0.708	0.220	0.001	2.030	1.319	3.126
Self-image	0.153	0.038	0.000	1.165	1.080	1.256
Cancer-related fatigue	0.031	0.011	0.003	1.032	1.011	1.053
Disease Awareness	0.041	0.014	0.003	1.042	1.041	1.070

A logistic regression model was established using the stepwise forward method with the six influencing factors screened by multifactor analysis as the independent variables and the occurrence of high FCR in patients as the dependent variable. The variables were assigned as shown in **Table 4**, and the resulting early warning model is shown in Equation (1). The predictive effect of the model was evaluated by the sample classification table in the logistic regression model (**Table 4**). The model accuracy was 83.7%, the sensitivity and specificity were 79.3% and 86.9% (**Table 5**),

the positive and negative predictive values were 85.1% and 81.7%, respectively, and the area under the ROC curve was 0.746.

Table 4. Logistic regression model variable assignment table.

Variable name	Meaning of variables	Assignment
X1	Age	
X1a	<45	0
X1b	45~65	1
X1c	>65	2
X2	Occupational status	
X2a	Employed	0
X2b	Separation from service	1
X2c	Retirement	2
X3	Chemotherapy	
X3a	Yes	0
X3b	None	1
X4	Self-image	Continuous variables
X5	Cancer-related fatigue	Continuous variables
X6	Disease Awareness	Continuous variables
Y	FCR	
	Low FCR	0
	High FCR	1

Equation (1):

$$\text{Logit}(P) = -4.834 - 1.539X1a - 0.592X1b - 1.816X2a - 1.585X2b + 0.708X3a + 0.153X4 + 0.031X5 + 0.041X6 \quad (1)$$

Table 5. Logistic regression model classification table.

	Projections		Accuracy
	Low FCR	High FCR	
Actual			
Low FCR	165	43	79.3
High FCR	37	245	86.9

Python statistical software was used to build artificial neural network models. Taking the 17 influencing factors of the cross-sectional survey as input variables and taking the occurrence of high recurrence fear as the output variable, a multilayer perceptron neural network model was established and 10×10 cross-validation was conducted (see Appendix for coding languages). It was verified that the accuracy of the model was 78.2%, the sensitivity and specificity were 80.0% and 75.8%, the positive and negative predictive values were 81.9% and 73.4%, respectively, and the area under the ROC curve was 0.818. According to the different ANNs of predictor variable values, the amount of change predicted by the network model was detected, and the importance of the predictor variable summarized. According to the ratio of the importance index to the maximum index value, the order of the importance of the

normalized predictors can be obtained. The importance order of the predictors determined by the model is shown in **Table 6** and **Figure 1**.

Table 6. Importance of independent variables.

Predictive variables	Importance	Normalized Importance
Cancer-related fatigue	0.082	100.0%
Occupational status	0.059	72.0%
Education level	0.054	65.9%
Disease awareness	0.049	59.8%
Radiotherapy	0.048	58.5%
Anxiety	0.043	52.4%
Per capita monthly household income	0.041	50.0%
Depression	0.035	42.7%
Sick time	0.033	40.2%
Hope	0.032	39.0%
Disease stage	0.029	35.3%
Social support	0.028	34.1%
Self-image	0.025	30.5%
Operation method	0.024	29.3%
Chemotherapy	0.014	17.1%
Age	0.011	13.4%
Marital status	0.010	12.2%

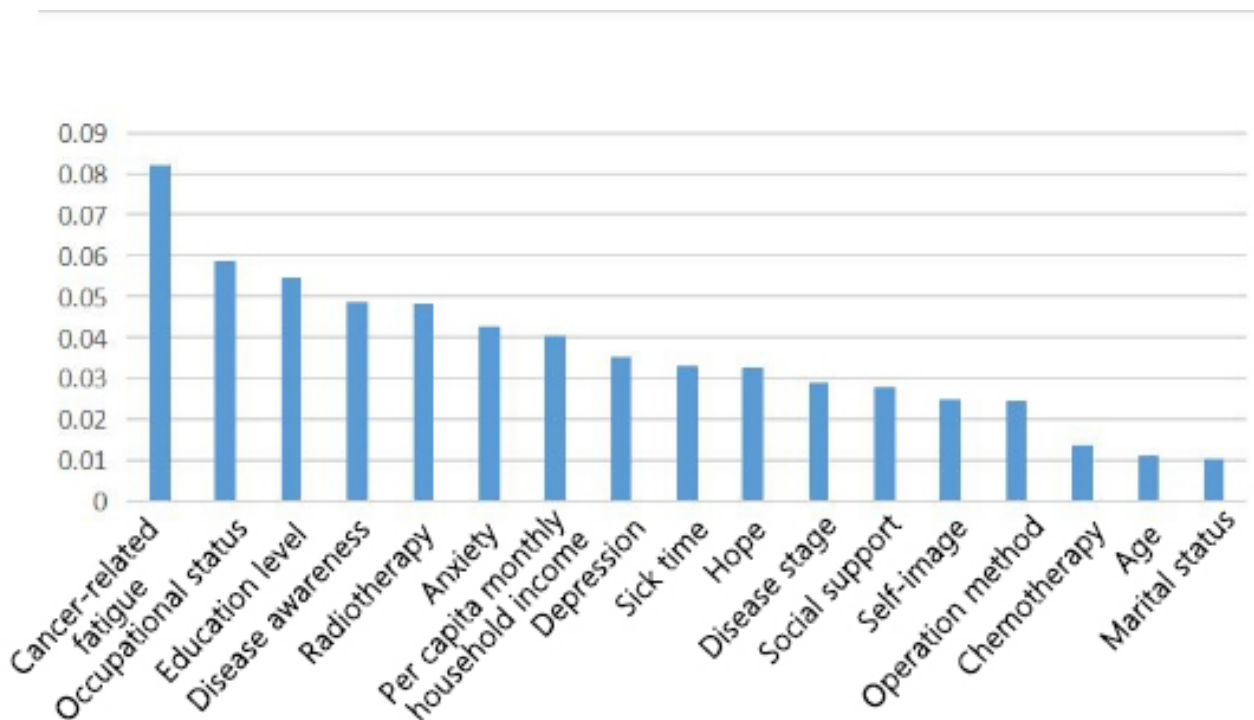


Figure 1. Importance of independent variables.

Python statistical software was used to build a random forest model. The 17 influencing factors in the cross-sectional survey were used as input variables, and the high fear of recurrence was used as an output variable. The Sklearn database was selected for modeling, Metrics is used as the result indicator output library of the model, and Panda and Numpy are used as the data processing library of the model, Matplotlib. Python is used as the independent variable importance processing tool, and Make-blods is used to input the influencing factors as model features. A 10×10 cross-validation was used (see Appendix for coding language). The accuracy, sensitivity and specificity of the model were verified to be 84.4%, 89.4% and 77.4%, respectively; the positive and negative predictive values were 84.4% and 84.2%, respectively; and the area under the ROC curve was 0.892. According to the different ANNs of predictor variable values, the amount of change predicted by the network model is detected, and the importance of the predictor variable is summarized. According to the ratio of the importance index to the maximum index value, the order of the importance of the normalized predictor variables can be obtained. The specific order is shown in **Table 7** and **Figure 2**.

Table 7. Importance of independent variables.

Predictive variables	Importance	Normalized importance
Cancer-related fatigue	0.176	100.0%
Disease awareness	0.096	54.5%
Self-image	0.096	54.5%
Sick time	0.085	48.3%
Social support	0.079	44.9%
Anxiety	0.068	38.6%
Hope	0.067	38.1%
Depression	0.062	35.2%
Occupational status	0.058	33.0%
Operation method	0.051	29.0%
Education level	0.038	21.6%
Age	0.036	20.5%
Disease stage	0.020	11.4%
Per capita monthly household income	0.019	10.8%
Radiotherapy	0.016	9.1%
Marital status	0.015	8.5%
Chemotherapy	0.015	8.5%

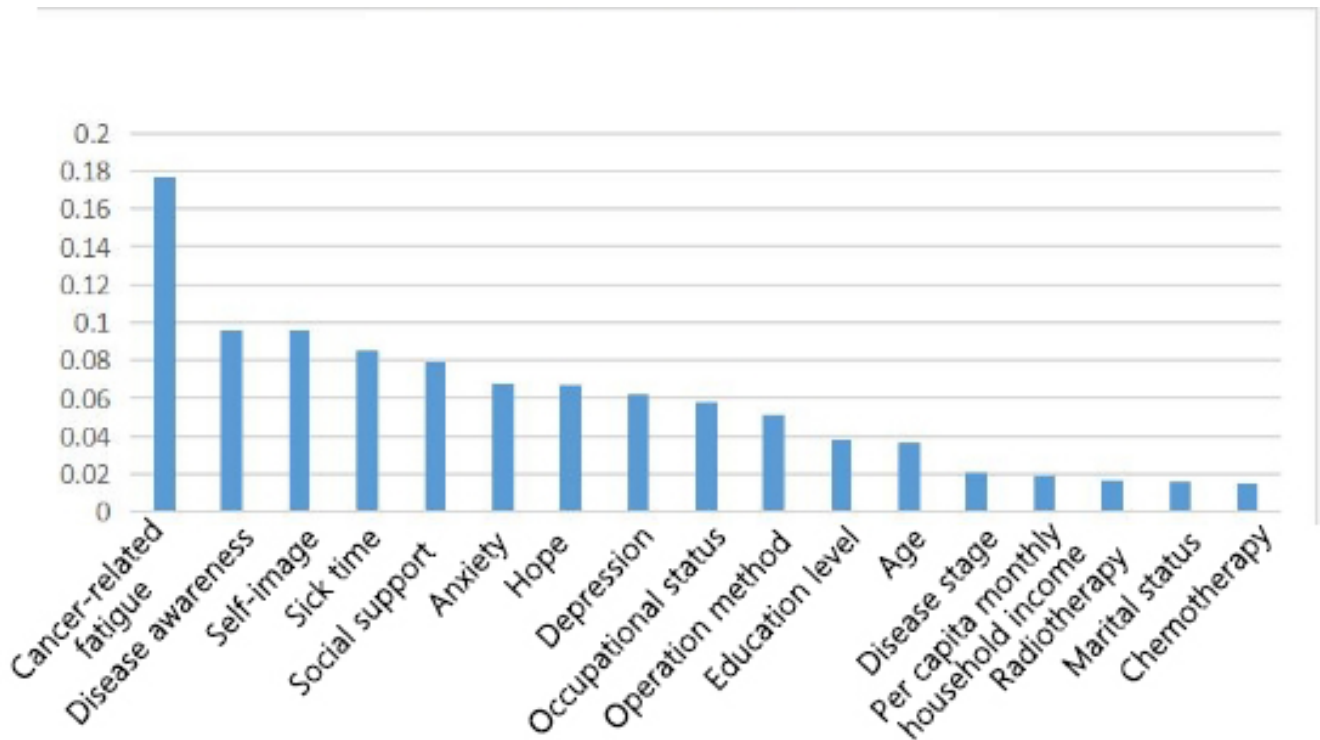


Figure 2. Importance of independent variables.

The results of the evaluation parameters of the three models are compared as shown in **Table 8** and **Figure 3** below and The accuracy of all three models was more than 70%, indicating high prediction accuracy; the AUC values of all three models were greater than 0.7, indicating that the models had better discrimination in predicting the fear of recurrence in recovering breast cancer patients, while the AUC values of the random forest model were better than those of the logistic regression model and the neural network model. According to the assessment criteria of AUC value, the closer the AUC value is to 1, the better the differentiation effect of the model, indicating that the random forest model has better prediction performance than the logistic regression model and the neural network model.

Table 8. Comparison of the evaluation of the prediction performance of the three models.

Modeling methods	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value	F-metric	AUC	AUC 95% CI	
								Lower limit	Upper limit
LR	0.837	0.793	0.869	0.851	0.817	0.860	0.746	0.694	0.795
ANN	0.782	0.800	0.758	0.819	0.734	0.810	0.818	0.500	0.890
RF	0.844	0.894	0.774	0.844	0.842	0.869	0.892	0.774	0.868

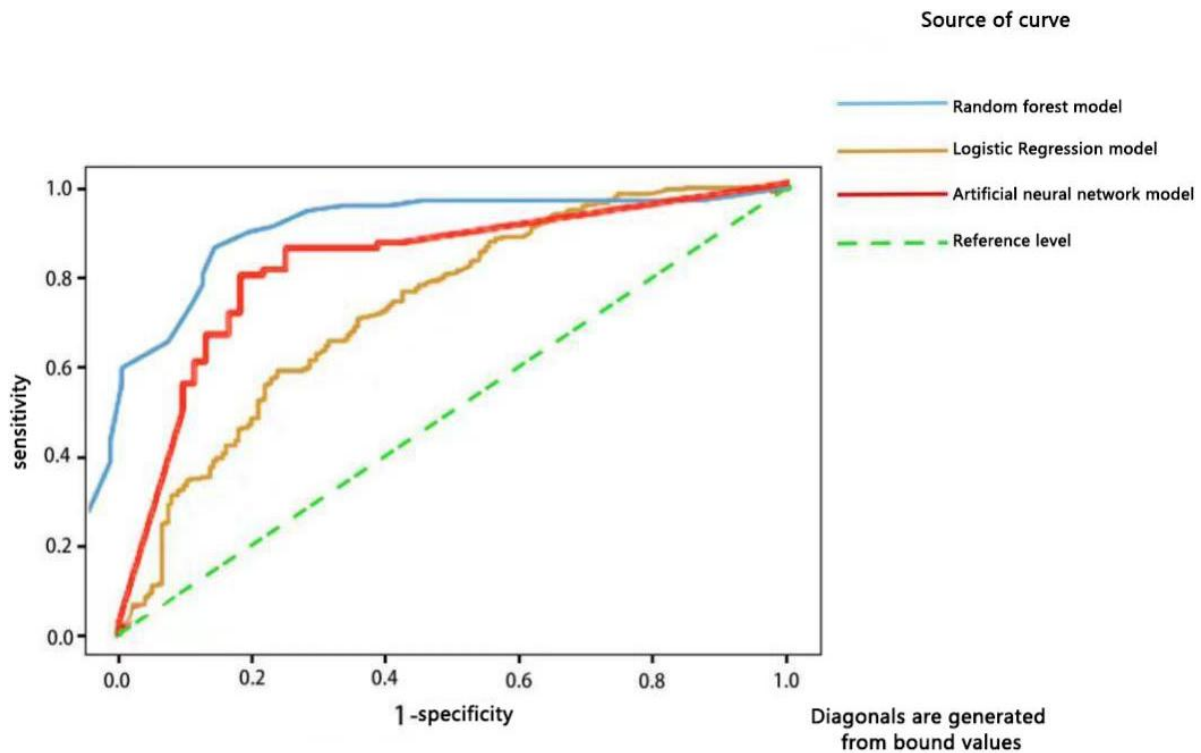


Figure 3. ROC curves for the three models.

4. Discussion

The results of this study showed that the mean score of FCR was 39.93 ± 7.78 , which is similar to the study of Uemoto et al. [41], with 57.56% of patients having a high FCR, a lower percentage than in the study of Luo et al. [42] for newly diagnosed cancer patients, but much higher than Reed [15]. This may be because newly diagnosed patients have not had enough time to deal with the threat of the disease and the physical and life changes; moreover, compared to other malignancies, breast cancer patients have a higher survival rate and better prognosis. The proportion of Chinese cancer survivors with FCR is still much higher than that in Western countries because FCR has been studied earlier in developed countries and more psychological support may already be in place. Therefore, health care professionals need to guide breast cancer patients to reduce their FCR and improve their quality of life after they have largely completed treatment and returned to daily life.

Age was an influencing factor in the FCR of recovering breast cancer patients, younger patients had a higher FCR, the results of Nahm et al. [43] and Starreveld et al. [44] are consistent. The lower acceptability of cancer in younger patients may be related to the fact that younger patients are less able to cope with death, and the disease and treatment may cause changes in fertility and debilitating side effects. Moreover, younger patients have fewer peers and family members who can understand and comfort them, and it is more difficult to make complex medical decisions calmly. Younger women with breast cancer also have the social responsibility of raising children and supporting elderly family members, which may also increase FCR [45]. Therefore, in the future, health workers should provide younger breast cancer patients with appropriate attention and psychological support. In addition, family members of

breast cancer patients should also be given health education and the opportunity to share their stress and find comfort in their daily lives.

Occupation was an independent predictor of FCR in recovering breast cancer patients, and the level of FCR in patients who were employed was higher than that in those who had retired, which was in line with Bae [46]. This may also be one of the reasons for the higher FCR among the patients in this study. This is because younger patients are the main source of income for their families. The patient is therefore more concerned that a relapse will make it more difficult to continue working. This suggests that healthcare professionals should focus on patients with jobs and provide more psychological support and rehabilitation advice.

Chemotherapy was found to be an influencing factor in FCR for recovering breast cancer patients. The results of Yang et al. [16] and Luo et al. [42] are consistent. A meta-analysis that included 30,176 cancer patients showed a weak but significant relationship between chemotherapy and FCR. Chemotherapy, as one of the main cancer treatments, may cause adverse effects, such as sleep problems, depression and anxiety, which may last months or even years after the completion of treatment; therefore, the physical symptoms and psychological difficulties associated with chemotherapy contribute to a higher fear of relapse. This suggests that healthcare workers should intervene early with patients undergoing chemotherapy or about to undergo chemotherapy, for example, by informing them in advance about the possible adverse effects of chemotherapy and preventive and palliative measures and advising them that certain symptoms during chemotherapy are normal physical reactions to chemotherapy and not signs of cancer recurrence.

In this study, the high FCR group had lower levels of self-image than the low FCR group, and the mean scores for both groups suggested that patients had a high level of self-image dissatisfaction, the self-image score in Wu et al. [47] study was lower than in this study. This may be because patients found themselves different from others and were overly concerned about changes in their image because only 14.7% received breast-conserving surgery. The study showed that patient self-image levels negatively predicted FCR, i.e., low self-image levels increased FCR in recovering breast cancer patients, which is similar to Graves' study [23]. Patients' perceptions of changes in their image and resulting behaviors, such as excessive body concerns and deliberate avoidance of people, lead to FCR of cancer-causing more harm to their bodies. At the same time, patients may not be concerned about their self-image during treatment, but they return to their daily lives as treatment ends. Changes in body image may be a constant reminder of the painful experiences they have had with the disease and treatment, causing them to fear that the same thing will happen again [18]. At the same time, some patients with breast cancer may have a low level of self-image due to the loss of their breasts and endocrine disorders due to medication and obesity [48]. Therefore, nursing staff should always pay attention to not only the physiological changes but also the corresponding psychological changes of patients after their body image has been changed due to treatment. For patients with hair loss due to chemotherapy should be taught about hair regrowth.

Fatigue in somatic symptoms was not only the most important independent variable factor in the random forest model but also showed an important predictive role in the ANN model, which is in line with previous findings. Patients often cite

physical symptoms as a sign of cancer recurrence [49]. Patients who frequently want to rest or experience frequent fatigue may reflect on the possibility of cancer recurrence due to their inability to identify the causes of fatigue, which may exacerbate their fears, while patients may also cite lack of concentration and energy as signs of disease recurrence. Based on the cancer threat interpretation model, patients are more likely to interpret physical symptoms such as fatigue as a physical threat rather than a state of physical awareness, which may lead to FCR and an unhealthy focus on physical changes, resulting in lower quality of life and overuse of health resources [50]. If survivors interpret their symptoms as potential signs of cancer recurrence, then fear and worry will increase, which in turn will further prompt survivors to interpret their symptoms as threatening. Cancer-caused fatigue may act as a trigger for patients to recall physical discomfort during treatment, exacerbating patients' fear of recurrence [51]. Therefore, it is important for health care professionals to pay attention to patients' somatic symptoms, especially cancer-related fatigue, and develop targeted interventions to alleviate or prevent the occurrence of adverse physical symptoms such as fatigue. Guiding patients to correctly identify their fatigue and develop a correct perception of somatic symptoms is also a priority for future work.

In this study, the disease perception score of the high FCR group was 50.67 ± 7.76 . In terms of disease perception, patients believed that the disease would last longer and seriously affect their futures and that the disease would cause many symptoms and the disease would be difficult to control. The results of this study suggest that illness perception predicts FCR. The results of this study are similar to those of Soriano and Mikrut [52]. A study of 110 breast cancer survivors in the UK demonstrated [53] that inadequate patient perception of disease and biased understanding can lead to high FCR. Individuals' perceptions of the disease vary, and patients' perceptions of the disease as a threat may lead to avoidance behaviors. According to the Common Sense Model of Disease Self-Regulation, disease perceptions are a key trigger, and individuals' perceptions of the disease as incurable, treated with lifelong medication and with severe consequences can lead to increased levels of FCR. The focus of the nursing staff's work is therefore on correcting the patient's poor perceptions and encouraging the patient to face the problem positively rather than avoiding it.

This study used a cross-sectional design to construct the model, which is in line with the thinking of several studies at home and abroad. However, FCR is a dynamic process, and the cross-sectional study may have some limitations in terms of causality. The cohort study may be able to compensate for this limitation, and the results can be further improved by adding cohort studies based on this study in the future. The selection of appropriate predictor variables into the model is an important part of constructing a model with good predictive performance. Each variable in this study can be measured directly or indirectly, facilitating the reproducibility of the model application. The final six variables screened by multifactor analysis were used as independent variables to build logistic regression models, artificial neural network models and random forest models. Age, occupational status, chemotherapy, self-image, cancer-related fatigue and disease perception, with age and occupational status as protective factors; chemotherapy, self-image level, cancer-caused fatigue and disease perception are the main risk factors for developing FCR. Literacy and monthly per

capita household income were statistically significant in the univariate analysis but did not enter the final model. However, in the Starreveld study, literacy and monthly per capita household income were predictors of FCR. The relationship between relevant predictor variables and FCR could therefore be further discussed in future longitudinal studies.

Three statistical methods were used in this study to develop different predictive models. The artificial neural network model has the advantage of the ability to model nonlinear relationships in high-dimensional datasets and to classify dependent variables more accurately while being able to predict complex relationships between variables. A limitation, however, is the inability to account for weights. The random forest model is based on a machine learning algorithm and has no restrictions on the number of datasets and predictor variables, no covariance and no overfitting problems. The logistic regression model is simple to use and can be used to visualize protective and risk factors but is prone to overfitting and covariance for multiple influences. This is consistent with a number of domestic and international studies and suggests that in future studies, appropriate modeling methods can be chosen.

The AUC value of the final logistic regression model was 0.746, the AUC value of the artificial neural network model was 0.818, and the AUC value of the random forest model was 0.892, all of which were between 0.7 and 0.9, indicating that all three models had good predictive ability, but the performance of the random forest model was better than that of the traditional model. In this study, the sensitivity, accuracy and specificity of the models were compared using a comparative approach, which is the same as the evaluation method of previous studies [54]. The results of this study showed that the random forest model outperformed the other two models, which is consistent with previous studies [55]. The random forest model and the logistic regression model were used as feature selection tools for the screening of patients with awake obstructive sleep apnea, and the performance of the classification methods was compared [54]. Retrospectively studied 5814 patients, 1673 of whom had pressure ulcer events, using random forest and artificial neural network models to construct separate pressure ulcer prediction models. A total of 19 variables were included, and the significance of the screening variables was assessed. The results showed that the random forest model had high accuracy in predicting pressure ulcers.

The logistic regression model, the artificial neural network model and the random forest model constructed in this study can all better distinguish between breast cancer patients with low FCR and those with high FCR in recovery. However, the random forest model has better performance, which indicates that machine learning methods have great potential in predicting FCR. This tool can be used in the future to screen out high-risk individuals prone to high FCR among breast cancer patients. In the future, this tool could be used to screen breast cancer patients for those at high risk of FCR and to intervene and study them more precisely.

5. Limitations

Due to time constraints, we conducted only a cross-sectional survey. A longitudinal study could not be conducted to determine the trajectory of change in FCR among breast cancer patients in recovery. Second, this study used questionnaires

to ensure the relative objectivity of the survey, and the selection of patients with an education level of primary school and above may have biased the findings. Third, due to human and environmental constraints, this study used a convenience sampling method to enroll participants. Thus, all samples were taken from one city, and there was a lack of research on recovering breast cancer patients from hospitals in other regions. We acknowledge the limitation of recruiting participants from a single hospital in a specific region, which may restrict the generalizability of our findings. Future studies should include a more diverse sample from multiple regions and different hospital types to enhance the applicability of the results. Additionally, important factors such as psychosocial support and economic status, which were not included in the current model, should be considered in future research. While the random forest model showed superior performance, its practical application may be limited by its computational complexity and cost, which should be taken into account when implementing it in clinical practice.

6. Conclusion

The predictor variables screened by the cross-sectional survey in this study were age, occupational status, chemotherapy, self-image, cancer-causing fatigue and disease perception, where age and occupational status were protective factors; chemotherapy, level of self-image, cancer-causing fatigue and disease perception were risk factors for developing FCR.

The constructed artificial neural network model, logistic regression model and random forest model all provided good discrimination, but the random forest model had better predictive results. Therefore, the random forest model can be used as a reference in screening, prevention and precision intervention for recovering breast cancer patients at high risk of FCR.

Author contributions: Conceptualization, HY and JZ; data curation, HY and XL; formal analysis, HY; writing—original draft preparation, HY and JZ. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: Thank you to every participant in this study.

Ethical approval: The study was conducted in accordance with the Declaration of Helsinki, and approved by Ethics Committee of the Fourth Affiliated Hospital of Harbin Medical University, and all participants gave informed consent. The ethical review number is 2021-WZYSLL SC-08.

Availability of data and materials: The data in this article will not be quoted in commentary articles. If you need the original data, please contact the first author, but please ensure that the original data may not be used without authorization.

Conflict of interest: The authors declare no conflict of interest.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021; 71(3): 209-249. doi: 10.3322/caac.21660

2. Chen M, Lv Q. Global analysis of breast cancer incidence and mortality based on population-based registry data, 2000-2020. *Chinese clinical journal of thoracic and cardiovascular surgery*. 2022; 29(04): 401-406.
3. Zhang M, Peng P, Wu C, et al. Analysis of breast cancer incidence and death among women in China's tumor registry regions from 2008-2012. *Chinese Journal of Oncology*. 2019; 04: 315-320.
4. Loughan AR, Lanoye A, Aslanzadeh FJ, et al. Fear of Cancer Recurrence and Death Anxiety: Unaddressed Concerns for Adult Neuro-oncology Patients. *Journal of Clinical Psychology in Medical Settings*. 2019; 28(1): 16-30. doi: 10.1007/s10880-019-09690-8
5. Dennett AM, Peiris CL, Shields N, et al. From Cancer Rehabilitation to Recreation: A Coordinated Approach to Increasing Physical Activity. *Physical Therapy*. 2020; 100(11): 2049-2059. doi: 10.1093/ptj/pzaa135
6. Rauff Søndergaard S, Ellekjær LB, Bechmann T, et al. Shared decision making with breast cancer patients: impact on patient engagement and fear of recurrence. Protocol for a Danish randomized trial in radiotherapy (DBCG RT SDM). *Acta Oncologica*. 2021; 60(8): 1032-1037. doi: 10.1080/0284186x.2021.1921261
7. Thewes B, Husson O, Poort H, et al. Fear of Cancer Recurrence in an Era of Personalized Medicine. *Journal of Clinical Oncology*. 2017; 35(29): 3275-3278. doi: 10.1200/jco.2017.72.8212
8. Taylor TR, Huntley ED, Sween J, et al. An Exploratory Analysis of Fear of Recurrence among African-American Breast Cancer Survivors. *International Journal of Behavioral Medicine*. 2011; 19(3): 280-287. doi: 10.1007/s12529-011-9183-4
9. Humphris G. Fear of recurrence and possible cases of anxiety and depression in orofacial cancer patients. *International Journal of Oral and Maxillofacial Surgery*. 2003; 32(5): 486-491. doi: 10.1054/ijom.2002.0399
10. Nahm SH, Blinman P, Butler S, et al. Factors associated with fear of cancer recurrence in breast and colorectal cancer survivors: A cross-sectional study of cancer survivors. *Asia-Pacific Journal of Clinical Oncology*. 2020; 17(3): 222-229. doi: 10.1111/ajco.13434
11. Skandarajah AR, Lisy K, Ward A, et al. Patient-reported outcomes in survivors of breast cancer one, three, and five years post-diagnosis: a cancer registry-based feasibility study. *Quality of Life Research*. 2020; 30(2): 385-394. doi: 10.1007/s11136-020-02652-w
12. Muldbücker P, Steinmann D, Christiansen H, et al. Are women more afraid than men? Fear of recurrence in couples with cancer – predictors and sex-role-specific differences. *Journal of Psychosocial Oncology*. 2020; 39(1): 89-104. doi: 10.1080/07347332.2020.1762823
13. Niu L, Liang Y, Niu M. Factors influencing fear of cancer recurrence in patients with breast cancer: Evidence from a survey in Yancheng, China. *Journal of Obstetrics and Gynaecology Research*. 2019; 45(7): 1319-1327. doi: 10.1111/jog.13978
14. Heidkamp P, Breidenbach C, Hiltrop K, et al. Individual courses and determinants of fear of cancer recurrence in long-term breast cancer survivors with and without recurrence. *Supportive Care in Cancer*. 2021; 29(12): 7647-7657. doi: 10.1007/s00520-021-06329-z
15. Reed SC, Bell JF, Miglioretti DL, et al. Fear of cancer recurrence and associations with mental health status and individual characteristics among cancer survivors: Findings from a nationally representative sample. *Journal of Psychosocial Oncology*. 2019; 38(2): 125-142. doi: 10.1080/07347332.2019.1649338
16. Yang Y, Wen Y, Bedi C, et al. The relationship between cancer patient's fear of recurrence and chemotherapy: A systematic review and meta-analysis. *Journal of Psychosomatic Research*. 2017; 98: 55-63. doi: 10.1016/j.jpsychores.2017.05.002
17. Yang Y, Cameron J, Humphris G. The relationship between cancer patient's fear of recurrence and radiotherapy: a systematic review and meta-analysis. *Psycho-Oncology*. 2016; 26(6): 738-746. doi: 10.1002/pon.4224
18. Lane BE, Garland SN, Chalifour K, et al. Prevalence and factors associated with fear of recurrence in a mixed sample of young adults with cancer. *Journal of Cancer Survivorship*. 2019; 13(6): 842-851. doi: 10.1007/s11764-019-00802-9
19. Chumdaeng S, Soivong P, Sethabouppha H, et al. Health problems among breast cancer survivors after completing conventional treatments: A cross-sectional study. *Nursing & Health Sciences*. 2020; 22(2): 436-444. doi: 10.1111/nhs.12678
20. Dinkel A, Marten-Mittag B, Krensreiter K. Association Between Daily Worry, Pathological Worry, and Fear of Progression in Patients with Cancer. *Frontiers in Psychology*. 2021; 12. doi: 10.3389/fpsyg.2021.648623
21. Gormley M, Ghazal L, Fu MR, et al. An Integrative Review on Factors Contributing to Fear of Cancer Recurrence Among Young Adult Breast Cancer Survivors. *Cancer Nursing*. 2020; 45(1): E10-E26. doi: 10.1097/ncc.0000000000000858
22. Cho D, Chu Q, Lu Q. Associations among physical symptoms, fear of cancer recurrence, and emotional well-being among Chinese American breast cancer survivors: a path model. *Supportive Care in Cancer*. 2017; 26(6): 1755-1761. doi: 10.1007/s00520-017-4010-3

23. De Graves S, Aranda S. Living With Hope and Fear-The Uncertainty of Childhood Cancer After Relapse. *Cancer Nursing*. 2008; 31(4): 292-301. doi: 10.1097/01.ncc.0000305745.41582.73
24. Cessna Palas JM, Hyland KA, Nelson AM, et al. An examination of the relationship of patient modifiable and non-modifiable characteristics with fear of cancer recurrence among colorectal cancer survivors. *Supportive Care in Cancer*. 2020; 29(2): 869-876. doi: 10.1007/s00520-020-05552-4
25. Kussainova D, Tursynbekova A, Aimbetova G, et al. Factors affecting the fear of recurrence in Breast cancer patients. *Research Journal of Pharmacy and Technology*. Published online January 19, 2024: 314-322. doi: 10.52711/0974-360x.2024.00049
26. Kangi AK, Bahrapour A. Predicting the Survival of Gastric Cancer Patients Using Artificial and Bayesian Neural Networks. *Asian Pacific Journal of Cancer Prevention*. 2018, 19(2): 487-490. doi: 10.22034/APJCP.2018.19.2.487
27. Salmanian K, Marashian FS. Prediction of Death Anxiety based on Body Image Concerns Mediated by Disease Perception in Patients with Breast Cancer. *Clinical social work journal*. 2022; 13(5): 61-66. doi:10.22359/cswj_13_5_12
28. Salmanian K, Marashian FS. Prediction of Death Anxiety Based on Body Image Concerns Mediated by Disease Perception in Patients With Breast Cancer. *Research Square*. 2021; 1. doi: 10.21203/rs.3.rs-718958/v1
29. Huang TC, Hsu TC, Hsieh YH, et al. Utilizing Graph Neural Networks for Breast Cancer Prognosis Prediction with High-dimensional Genomic Data. 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2023.
30. Mehnert A, Herschbach P, Berg P, et al. Progredienzangst bei Brustkrebspatientinnen - Validierung der Kurzform des Progredienzangstfragebogens PA-F-KF/ Fear of progression in breast cancer patients – validation of the short form of the Fear of Progression Questionnaire (FoP-Q-SF). *Zeitschrift für Psychosomatische Medizin und Psychotherapie*. 2006; 52(3): 274-288. doi: 10.13109/zptm.2006.52.3.274
31. Gu Z, Qiu T, Yang S, et al. A Study on the Psychological Factors Affecting the Quality of Life Among Ovarian Cancer Patients in China *Cancer Management and Research*. 2020; 12: 905-912. doi: 10.2147/cmar.s239975
32. Hopwood P, Fletcher I, Lee A, Ghazal SA. A body image scale for use with cancer patients. *European Journal of Cancer*. 2001; 37(2): 189-197. doi: 10.1016/S0959-8049(00)00353-1
33. Zhang Y. A study on self-image and quality of life of prostate cancer patients [D]. Fudan University, 2010.
34. Zhao H, Wang J. Social support and hope for hemodialysis patients. *Chinese Journal of Nursing*. 2000; 05: 49-51.
35. Blumenthal JA, Burg MM, Barefoot J, et al. Social support, type A behavior, and coronary artery disease. *Psychosomatic Medicine*. 1987; 49(4): 331-340. doi: 10.1097/00006842-198707000-00002
36. Jiang Q, Huang L, Ren W. A study on the correlation between coping styles, social support and psychosomatic symptoms in cancer patients. *Chinese Journal of Mental Health*. 1996; 04: 160-161.
37. Okuyama T, Akechi T, Kugaya A, et al. Development and validation of the cancer fatigue scale: A brief, development and validation of the cancer fatigue scale: A brief, three-dimensional, self-rating scale for assessment of fatigue in cancer patients. *Journal of Pain and Symptom Management*. 2000; 19(1): 5-14. doi: 10.1016/S0885-3924(99)00138-4
38. Zhang F, Ding Y, Han L. Reliability and validity of the Chinese version of the Cancer Fatigue Scale. *Chinese Journal of Mental Health*. 2011; 25(11): 810-813.
39. Wang BD. A Study on the Application of PHQ-9 and GAD-7 in Patients with Malignant Tumors [D]. Central South University. 2013.
40. Mei Y, Li H, Yang Y, et al. Reliability test of the simplified version of the Chinese version of the disease perception questionnaire in female breast cancer patients. *Journal of Nursing*. 2015; 22(24): 11-14.
41. Uemoto Y, Uchida M, Kondo N, et al. Predictive factors for patients who need treatment for chronic post-surgical pain (CPSP) after breast cancer surgery. *Breast Cancer*. 2021; 28(6): 1346-1357. doi: 10.1007/s12282-021-01275-4
42. Luo X, Li W, Yang Y, et al. High Fear of Cancer Recurrence in Chinese Newly Diagnosed Cancer Patients. *Frontiers in Psychology*. 2020; 11. doi: 10.3389/fpsyg.2020.01287
43. Xie Y. 102P Predictive models associated with the presence of pathological complete response following neoadjuvant chemotherapy for breast cancer. *Annals of Oncology*. 2020; 31: S51. doi: 10.1016/j.annonc.2020.03.041
44. Starreveld DEJ, Markovitz SE, van Breukelen G, et al. The course of fear of cancer recurrence: Different patterns by age in breast cancer survivors. *Psycho-Oncology*. 2017; 27(1): 295-301. doi: 10.1002/pon.4505
45. Guzik P, McKinney JA, Ulack C, et al. Outcomes That Matter Most to Young Adults Diagnosed with Cancer: A Qualitative Study. *Journal of Adolescent and Young Adult Oncology*. 2020; 10(5). doi: 10.1089/jayao.2020.0150

46. Bae KR, Cho J. Changes after cancer diagnosis and return to work: experience of Korean cancer patients. *BMC Cancer*. 2021; 21(1). doi: 10.1186/s12885-021-07812-w
47. Wu TY, Chang TW, Chang SM, et al. Dynamic Changes Of Body Image And Quality Of Life In Breast Cancer Patients. *Cancer Management and Research*. 2019; 11: 10563-10571. doi: 10.2147/cmar.s223314
48. Liu J, Mahendran R, Chua SM, et al. Validation of the English and Mandarin versions of the Fear of Cancer Recurrence Inventory in an Asian population. *Journal of Health Psychology*. 2017; 25(5): 617-628. doi: 10.1177/1359105317727819
49. Park S, Sato Y, Takita Y, et al. Mindfulness-Based Cognitive Therapy for Psychological Distress, Fear of Cancer Recurrence, Fatigue, Spiritual Well-Being, and Quality of Life in Patients With Breast Cancer—A Randomized Controlled Trial. *Journal of Pain and Symptom Management*. 2020; 60(2): 381-389. doi: 10.1016/j.jpainsymman.2020.02.017
50. Cunningham SJ, Patton M, Schulte F, et al. Worry about somatic symptoms as a sign of cancer recurrence: prevalence and associations with fear of recurrence and quality of life in survivors of childhood cancer. *Psycho-Oncology*. 2021; 30(7): 1077-1085. doi: 10.1002/pon.5647
51. Soriano EC, Otto AK, LoSavio ST, et al. Fear of Cancer Recurrence and Inhibited Disclosure: Testing the Social-Cognitive Processing Model in Couples Coping With Breast Cancer. *Annals of Behavioral Medicine*. 2020; 55(3): 192-202. doi: 10.1093/abm/kaaa043
52. Mikrut EE, Panjwani AA, Cipollina R, et al. Emotional adjustment among parents of adolescents and young adults with cancer: the influence of social constraints on cognitive processing and fear of recurrence. *Journal of Behavioral Medicine*. 2019; 43(2): 237-245. doi: 10.1007/s10865-019-00072-x
53. Millan ID, Cameron J, Yang Y, et al. An observational mixed-methods approach to investigate the fear of cancer recurrence cognitive and emotional model by Lee-Jones et al with women with breast cancer during radiotherapy treatment. *ecancermedicalscience*. 2019; 13. doi: 10.3332/ecancer.2019.984
54. Mishra S, Thakkar HK, Mallick PK, et al. A sustainable IoHT based computationally intelligent healthcare monitoring system for lung cancer risk detection. *Sustainable Cities and Society*. 2021; 72: 103079. doi: 10.1016/j.scs.2021.103079
55. He JL, Xu HQ, Yang J, et al. Fear of disease progression among breast cancer patients in China: a meta-analysis of studies using the fear of progression questionnaire short form. *Frontiers in Psychology*. 2023; 14. doi: 10.3389/fpsyg.2023.1222798

Appendix

```

import pandas as pd
import numpy as np
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import classification_report
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import RocCurveDisplay
from sklearn.neural_network import MLPClassifier
from sklearn.inspection import permutation_importance
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['font.sans-serif']=['FangSong'] # Chinese font
settings.blackbodySimHei
plt.rcParams['axes.unicode_minus']=False # Solve the problem that the saved image is a
negative sign
# Problems that appear as blocks
# Read the data
data=pd.read_excel('Total data (With scores for each
dimension) 1014.xlsx')
X,y=X['age classification', 'Career status', 'Chemotherapy', 'Self-image', 'Cancer-
related fatigue', 'Disease cognition']
y=data['Fear of cancer recurrence']-1
# Random forest
# pretreatment
X_rf=X.copy()
# Multi-categorical variables that are out of order (more than 2)
The following needs to be done
X_rf['Age classification', 'Career status']=X['Age classification', 'Career status'].astype(str)
X_rf=pd.get_dummies(X_rf)
# Parametric search
# The parameters that the random forest is searching for
parameters={'n_estimators':[50,100,150,300,500,1000],
'max_depth':[3,5,7,9,11,None],
'min_samples_leaf':[1,3,5,7,9,11]}
rf=RandomForestClassifier(random_state=10)
clf=GridSearchCV(rf, parameters, cv=5, n_jobs=-1, verbose=4)
clf.fit(X_rf, y)
print(clf.best_params_)
best_rf=clf.best_estimator_
# Get the best model
# Feature importance
part_feature_importances_bp=pd.DataFrame({'feature':X_rf.columns, 'importance':best_rf.feature_importances_})
for col in ['Age classification', 'Career status']:
# Multi-categorical variables that are out of order (Greater than 2)
The following needs to be done
col_X=Feature_importances_rf[col].map(lambda x:col in x)
col_s=sum(feature_importances_rf.loc[col_X]['importance'])
feature_importances_rf=feature_importances_rf.loc[~col_X]
feature_importances_rf=feature_importances_rf.sort_values(['importance'], ascending=False, inplace=True)
feature_importances_rf.index=range(feature_importances_rf.shape[0])
feature_importances_rf['Normalization importance']=feature_importances_rf['importance']/max(feature_importances_rf['importance'])

```

```
[ 'importances'])feature_importances_rf.to_csv('feature_importances_rf
.csv)#Save the results#Drawingf,ax=plt.subplots(figsize=(12,9)ax=
sns.barplot(x='Normalization importance'.y='feature'.data=feature_importances
_rf.ci=None,orient='h')ax.set_ylabel('variable'size=16)
#Set up y axis labelsax.set_xlabel('importance', size=16)
#Set up x axis labelsax.tick_params(l=1,labels=13)
#plt. savefig(feature_importances_rf.png'.dpi 600.bbox inches=
'tight')plt.show()#Cross-validation probability y_pred_rf=cross_val_predict(best rf.
X_rf.y.
cv=5.method='predict_proba',n_jobs=-1)[:1]#10Fold cross-validation, Return
probabilityprint(classification_report(y
np.where(y_pred_rf>0.5,1,0).digits=4.target_names=['feminine'masculine
D)#Confusion matrix cm=confusion_matrix(y,
np.where(y_pred_rf>-0.5,1,0))disp
ConfusionMatrixDisplay(confusion_matrix=cm
display_labels=['feminine':masculine'])disp.plot(values format='d'.cmap=plt.cm.Blues)
plt. savefig('rf cm.png'.dpi=600.bbox_ inches 'tight') plt.show()
#rocfpr, tpr, thresholds roc_curve(y,y_pred_rf)roc_auc=auc(fpr,
tpr)display=RocCurveDisplay(fpr=fpr,tpr=tpr,roc_auc=roc_auc,estimator_name='Random
forest')display.plot('T')plt.savefigplt.show()#Calculate the results and save
themtn,fp,fn,tp=cm.ravel()result_rf=pd.Series()result_rf.loc['tpr']=tp/(tp+fn)result_rf.loc['spc']=tn/(tp+fn)result_rf.
loc['ppv'] tp/(tp+
fp)result_rf.loc['npv']=tn/(tn+fn)result_rf.to_excel('result_Random forest.xlsx')
```