Immune-related Gene Data-based Molecular Subtyping Related to the Prognosis for Breast Cancer Patients

Guoyu Mu
the First Affiliated Hospital Of Dalian Medical University
Hong Ji
the Second Affiliated Hospital Of Dalian Medical University
Hui He
the First Affiliated Hospital Of Dalian Medical University
Hongjiang Wang (wanghongjiang_dmu@163.com)
the First Affiliated Hospital Of Dalian Medical University https://orcid.org/0000-0002-2096-6297

Research article

Keywords: Breast cancer, Immunotherapy, TCGA database

Posted Date: August 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-46061/v1

Version of Record: A version of this preprint was published at Breast Cancer on November 27th, 2020. See the published version at https://doi.org/10.1007/s12282-020-01191-z.

Abstract

Background Breast cancer (BC), the most frequently seen malignant tumor in female, is associated with increasing morbidity and mortality year by year. Generally, the available treatments for BC include surgery, chemotherapy, radiotherapy, endocrinotherapy and molecular targeted therapy. Typically, as molecular biology, immunology and pharmacogenomics develop, a growing amount of evidence has suggested that immunocyte infiltration into tumor microenvironment, together with the immunophenotype of tumor cells, would remarkably influence the development and malignant transformation of tumor; as a result, immunotherapy has become a promising therapy for treating BC, which would also affect patient prognosis.

Methods In this study, samples collected from TCGA and ImmPort database would be analyzed to search for specific immune-related genes affecting BC patient prognosis. A total of 64 immune-related genes with significant correlation with patient prognosis had been screened and performed shrinkage estimate, among which, 29 most representative ones with significant correlation with patient prognosis had been selected and utilized to establish the prognosis prediction model for BC patients (as referred to as the RiskScore equation). Thereafter, samples in both training set and test set would be substituted into the model, respectively; meanwhile, BC patients would also be divided based on the median RiskScore to assess the efficiency, accuracy and stability of the model in predicting and classifying patient prognosis. Subsequently, functional annotations, GO and KEGG signaling pathway enrichment analysis would be carried out among the 29 as-screened immune-related genes.

Results The results found that, these 29 genes could be mainly enriched to numerous BC- and immune microenvironment-related pathways. Eventually, the relationship between RiskScore and the sample clinical features as well as the signaling pathways would be analyzed.

Conclusions Our findings indicate that, the prognosis prediction model RiskScore established on the basis of the expression profiles of 29 immune-related genes has displayed high prediction accuracy and stability in identifying the immune features, which can guide the clinicians to diagnose and predict the prognosis for different immunophenotypes, in the meantime of offering numerous therapeutic targets for precisely treating BC in clinic using the identified subtype-specific immune molecules.

Background

Breast cancer (BC), the most common malignancy, is the leading cause of cancer-related death in female in the less developed countries, which has affected 882,900 cases and resulted in 324,300 deaths in 2012, accounting for 25% and 15% of cancer cases and cancer deaths among females, respectively[1]. Typically, BC is associated with the reproductive as well as endocrine risk factors, including the application of oral contraceptive, never having a child, and a long menstrual history[2]. On the other hand, the potentially modifiable risk factors are drinking, obesity, physical inactivity, and menopausal hormone therapy[3]. Some large-scale clinical data indicate that, systemic adjuvant chemotherapy is generally not

recommended for most early BC patients following surgery or radiotherapy, since chemotherapy would result in far greater toxic reactions than survival benefit to the patient[4–6]. However, BC will rapidly recur in high survival risk patients not undergoing chemotherapy, which may also progress into invasion and distant metastasis of para-carcinoma tissues[7]. Consequently, it is of crucial necessity to identify the related survival risk in patients through subgroup classification and early diagnosis, and to offer additional systemic adjuvant chemotherapy for high-risk patients.

According to recent studies, BC can be classified into four subtypes, including Luminal A (ER+/PR+/HER2−, grade 1 or grade 2), Luminal B (ER+/PR+/HER2+, or ER+/PR+/HER2 − grade 3), HER2 overexpression (ER−/PR−/HER2+), and triple negative breast cancer (TNBC, ER−/PR−/HER2−)[8]. Among then, the Luminal A subtype is associated with favorable prognosis, which is also sensitive to endocrine therapy; in this way, endocrine therapy alone can serve as the general treatment[9]. On the other hand, the Luminal B subtype is linked with high tumor proliferation rate, among which, the HER2 negative Luminal B subtype can be usually treated with endocrine therapy + chemotherapy, while the HER2-positive Luminal B subtype would be generally treated with chemotherapy + anti-HER2 treatment + endocrine therapy[10]. Moreover, the HER2 over-expression subtype is characterized by poor prognosis and rapid progression, which is mainly treated with chemotherapy + anti-HER2 therapy[11]. Specifically, the negative expression of ER, PR and HER2 in TNBC is associated with unique biological characteristics and potent heterogeneity, and no standard treatment is recommended for this subtype except for chemotherapy[12]. Recently, progresses have been made in the early diagnosis and treatment for BC, which render BC a treatable disease; however, multidrug resistance (MDR) remains a major challenging issue in treating metastatic BC, typically, the survival for patients with metastatic BC is only 2–3 years[13]. Unfortunately, such general classification method cannot accurately manifest the individual differences [14]. Notably, the existing large-scale databases containing gene expression data, including TCGA and ImmPort, allow us to search for the potentially reliable BC biomarkers to predict and classify patient prognosis[15].

Increasing evidence has supported that, immunocytes in tumor microenvironment can remarkably promote or inhibit tumor growth, which can thereby serve as the indicator for BC prognosis. Moreover, immune escape has been verified as the novel cancer marker[16]. Recently, great achievements have been attained in treating BC patients through immunotherapies, such as BC vaccines, monoclonal antibodies (MAbs), antibody-drug conjugates (ADCs), checkpoint inhibitors, and stimulatory molecule agonist antibodies[17–20]. Besides, tumor-infiltrating lymphocytes (TILs) and tumor-related macrophages in BC tissues have also been identified to exert crucial parts in the immune escape mechanism of tumor cells, and they are found to be remarkably related to patient prognosis[21, 22]. Nonetheless, the molecular events regarding tumor cell-immunocyte interaction in the BC microenvironment should be further examined and summarized, which contribute to determining their potential roles in predicting the prognosis for BC patients[23].

In this study, a prognosis prediction model for BC was developed and verified based on immune-related genes retrieved according to the clinical features of patients collected from TCGA and ImmPort

databases, which was promising to help clinicians to evaluate the therapeutic effect, prognosis, and therapeutic option for BC patients.

Materials And Methods

Pre-processing of preliminary sample data and initial screening of BC immune-related genes

The latest clinical information on follow-up had been downloaded on December 14th, 2018, through TCGA GDC API. A total of 1222 RNA-Seq data samples were involved, as presented in Table S1. 1109 of these 1222 data samples were tumor tissues, while the remaining 113 were normal tissues. Additionally, the immune-related gene set had also been downloaded from the ImmPort database on October 8th, 2018, as displayed in Table S2, and 1811 genes were covered.

At first, the retrieved 1109 RNA-seq data samples would be pre-processed according to the steps shown below: 1) 39 samples with no clinical data and 21 with 0 OS would be removed; 2) the normal tissue sample data would be removed; 3) genes of FPKM < 1 were also removed from all samples; and 4) only the expression profiles of immune-related genes would be preserved. Altogether 1376 genes had been employed for subsequent analysis of the model. The pre-processed data are shown in Table S3, while the sample statistics of clinical information are displayed in Table 1.

Secondly, 1068 samples had been classified into training set and test set, respectively, and random grouping with replacement would be performed for all samples for 500 times ahead of time to eliminate the impact of random allocation bias on model stability. Grouping would be conducted based on the training set: test set ratio of 0.7:0.3 since the BC sample size was over 1000. Specifically, the most suitable training set and test set would be selected based on the following criteria: 1) similar age distributions, clinical stages, follow-up times and death proportions between the two groups; and 2) close binary sample sizes in the two randomly divided datasets after clustering the gene expression profiles. The final training set data (n = 533) are displayed in Table S4, and the test set data are shown in Table S5 (n = 535). Meanwhile, the clinical information statistics of both test set and training set samples are presented in Table 1. The final information of both training set and test set samples has been displayed in Table 1. There was no significant difference between training set and test set data, as verified by the P-value, which had indicated reasonable sample grouping.

Table 1
sample statistics of training set and test set

Clinical Features	Overall	Train	Testing	Pvalue
OS	1068	533	535	0.862408
T	1068	533	532	0.356377
T1	279	155	124	
T2	616	291	325	
T3	132	70	62	
T4	38	17	21	
TX	3	0	3	
N	1068	526	525	0.613292
N0	502	256	246	
N1	357	182	175	
N2	119	58	61	
N3	73	30	43	
NX	17	7	10	
M	1068	461	444	0.688259
M0	883	451	432	
M1	22	10	12	
MX	163	72	91	
Stage	1068	521	525	0.424994
I	181	106	75	
II	606	297	309	
III	239	109	130	
IV	20	9	11	
X	22	12	10	
Age	1068	533	535	0.515704
0 ~ 40	75	42	33	
40 ~ 50	219	118	101	
50 ~ 60	283	143	140	

Clinical Features	Overall	Train	Testing	Pvalue
60 ~ 70	277	124	153	
70 ~ 100	214	106	108	
IHC_Her2	1068	306	296	0.701374
0	59	31	28	
1+	263	134	129	
2+	194	101	93	
3+	86	40	46	

Single-factor survival analysis of immune-related genes in the training set

All immune-related genes were analyzed using the univariate Cox proportional hazards regression model; at the same time, survival data were evaluated by the survival coxph function of R package[24], and $p < 0.05$ served as the significance threshold.

Screening of specific immune-related genes for BC prognosis, and construction of the prognosis prediction model

First of all, the Least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996) algorithm was employed to further narrow the range of prognosis-specific immune-related genes under the condition of maintaining high accuracy. Besides, the glmnet of R software packagewas used for lasso cox regression analysis. Next, to further compress the number of immun`e-related genes, the R package MASS was employed for stepwise regression analysis using the Akaike information criterion, which had taken into consideration about the model statistical fitting degree, as well as the number of parameters used in fitting. Typically, the StepAIC method in MASS package had originated from the most complex model, which had deleted one variable each time to reduce AIC, and a smaller value was indicative of a superior model, demonstrating sufficient fitting degree and fewer parameters of the model. The risk model of 17 genes (Table S7) had finally been obtained using this algorithm. The results of stepwise regression are presented in Table S8. The formula was as follows:

RiskScore = PIK3CA*0.025861691 + CCR7*0.014541227 + SEMA7A*0.158263093 + ACVR2A*-0.437173332 + CBL*0.231921725 + PLXNB2*0.014940811 + PLXND1*0.033074364 + APOBEC3F*-0.314321194 + NFATC2*-0.257156537 + NFKBIZ*-0.046977178 + TNFSF4*0.16976996 + DAXX*-0.034395422 + TLR2*0.023037905 + SEMA3B*-0.044973358 + HSPA2*-0.023131493 + TPT1*-0.001623522 + CCL22*-0.077745415.

Afterwards, expression profiles of related genes would be collected from both the training set and test set, respectively; subsequently, they would be incorporated into the model, so as to calculate the RiskScore of all samples. Then, the median RiskScore would serve as the threshold to classify the samples into high

risk group (Risk-H) and low risk group (Risk-L), respectively; afterwards, ROC analysis, KM analysis and gene clustering analysis would be performed to comprehensively assess the efficiency, accuracy and stability of the model in predicting and classifying the prognosis for BC patients, respectively.

Functional annotations and signaling pathway enrichment of immune-related genes specific to prognosis

The gene families of the 17 eventually screened genes would be annotated following the human gene classification in HGNC database[25]. Specifically, the clusterProfile of R software package would be employed in KEGG and GO enrichment analyses for the above-mentioned 17 immune-related genes specific to prognosis.

Correlation between Riskscore and the signaling pathways as well as the clinical features of samples

Firstly, the KEGG functional enrichment scores of all samples would be analyzed using the ssGSEA function of R software package GSVA[26]. Meanwhile, the correlation with RiskScore would be calculated, and clustering analysis would be carried out according to the enrichment score of each sample in each pathway.

Subsequently, the correlations of related factors (including T, N, M, Stage, Age and Her2 expression) with RiskScore would be evaluated, respectively. Afterwards, the nomogram model and forest plot would be established using the clinical features (such as T, N, M, Stage, Age and Her2 expression) as well as the RiskScore, and the correlations of RiskScore as well as various clinical features with patient survival would thereby be assessed.

Results

Retrieval of immune-related genes based on the survival and prognosis results of BC patients

Firstly, related data were downloaded from the TCGA and ImmPort databases, which were then pre-processed (see Materials and methods). Subsequently, all the immune-related genes and survival data would be analyzed by the univariate Cox proportional hazards regression model using the survival coxph function of R package, and the significance level was set at $P < 0.05$, as shown in Table S6. Eventually, a total of 62 significantly different immune-related genes in terms of prognosis had been discovered. The relationships of the p-values of these 62 genes with the HR and expression quantities were shown in Fig. 1.

Screening of prognosis-specific immune-related genes and construction of the prognosis prediction model for BC

Currently, 62 immune-related genes have been recognized, but such a large number of these genes will go against clinical detection. Consequently, the scope of immune-related genes should be further narrowed in the meantime of guaranteeing the high accuracy. Thus,the R software package glmnet was used for lasso cox regression to refine the prognostic genes identified above, leading to reduced gene numbers

from 62 to 29.Moreover,the R package MASS was employed for stepwise regression analysis using the Akaike information criterion, which had taken into consideration about the model statistical fitting degree and the number of parameters used for fitting. On the other hand, the StepAIC method in MASS package had originated from the most complex model, which had deleted one variable each time to reduce AIC, and a smaller value had suggested a superior model, indicating sufficient fitting degree and fewer parameters of the model. Finally, the risk model of 17 genes was obtained using this algorithm (Table S7). The formula is shown in materials and methods.

Subsequently, training set samples would be incorporated into the formula to calculate the RiskScore of all samples, and the median RiskScore would serve as the threshold to divide the samples into high risk (Risk-H) and low risk (Risk-L) groups. Furthermore, ROC analysis of the prognosis classification for RiskScore would be carried out using the survivalROC of R software package. The OS distribution of sample was around > 2 years (Fig.S1); as a result, the model predicting effect for 3-, 5- and 10-year survival had been evaluated in this study, respectively, with the average AUC of about 0.789, as presented in Fig. 2A. Besides, the sample distribution in Risk-H and Risk-L groups under different OS is presented in Fig. 2B. As could be observed, no obvious difference in sample size was detected between the 0- and 1-year Risk-H and Risk-L groups; besides, the sample size in Risk-H group after the 5th year would be dramatically smaller than that in Risk-L group, which had become markedly significant as the OS extended (Fig. 2C). The clustering results of training set samples are presented in Fig. 2D. Obviously, the above-mentioned 17 genes could be markedly clustered into the high and low expression groups, respectively, while samples in the training set could also be assigned into two groups, and the RiskScore values of the two subclasses would also be compared (Fig. 2E).

Additionally, to further confirm the stability and reliability of the prognosis prediction model, the expression profiles of these 17 genes had been collected from the test set, which were subsequently incorporated into the model for model verification; at the same time, the RiskScore of samples would also be calculated. Afterwards, data in test set would be employed to evaluate the effects of the model on predicting the 3-, 5- and 10-year survivals, with the average 3-10-year AUC of 0.726, as displayed in Fig. 3A. Besides, the sample distribution in both Risk-H and Risk-L groups under different OS is also displayed in Fig. 3B. As could be observed, no distinct difference was detected in sample size between 0- and 1-year Risk-H and Risk-L groups; in addition, the sample size in Risk-H group after the 3rd year had been notably reduced compared with that in Risk-L group, which became more obvious as the OS extended (Fig. 3C). The clustering results for samples in the test set, and the difference in the RiskScore value between two groups, have been displayed in Fig. 3D and E, respectively.

On the other hand, to further validate the stability as well as reliability for the prognosis prediction model, the expression profile data among the above-mentioned 17 genes were extracted from a total of 1068 samples, followed by substitution into the model, so as to calculate the RiskScore values to validate the model as previously described. The series of results have been displayed in Fig. 4.Taken together, verification results based on the test set data suggested that, the prognosis model established on the

basis of the expression profiles of these 17 immune genes had displayed excellent prediction accuracy and stability in identifying the immune-related features.

Finally, the KM survival curves of the risk model constructed based on 17 genes in predicting the Risk-H and Risk-L groups at training set, test set and all-sample level are shown in Fig. 5. Figure 5A shows the KM survival curve of the training set ($p < 0.0001$), Fig. 5B displays the KM survival curve of the test set ($p < 0.01$), and Fig. 5C is the KM survival curve of all samples ($p < 0.0001$).

Functional annotations of immune-related genes and signaling pathway enrichment specific to prognosis

First of all, the gene families of the 17 obtained genes had been annotated in accordance with the human gene classification in HGNC database. As presented in Table 2, two genes had been enriched into the Plexins family, and two genes were also significantly enriched in the Semaphorins family ($p < 0.01$). Besides, the clusterProfile of R software package had also been employed for enrichment analyses of the 17 above-mentioned immune-related genes specific to prognosis. Results of GO enrichment are displayed in Fig. 6A, results of KEGG pathway enrichment are presented in Fig. 6B, and the related data are exhibited in Table S9 and Table S10, respectively. It could be observed from these results that, most of the above-mentioned genes could be enriched to multiple immunity- and cancer-related biological processes and signaling pathways.

Table 2
17 gene function annotation results

GeneFamily	Genes	pvalue	padj
Plexins	PLXNB2/PLXND1	2.77E-05	0.000471545
Semaphorins	SEMA7A/SEMA3B	0.000115948	0.001971113
Type 2 receptor serine/threonine kinases	ACVR2A	0.004392947	0.074680102
Nuclear factors of activated T-cells	NFATC2	0.004392947	0.074680102
Phosphatidylinositol 3-kinase subunits	PIK3CA	0.006582604	0.11190426
Toll like receptors	TLR2	0.008039856	0.136677545
Apolipoprotein B mRNA editing enzyme catalytic subunits	APOBEC3F	0.009495095	0.161416623
Heat shock 70 kDa proteins	HSPA2	0.013124409	0.223114961
Tumor necrosis factor superfamily	TNFSF4	0.013848769	0.235429069
Endogenous ligands	CCL22	0.155987949	1
Ankyrin repeat domain containing	NFKBIZ	0.164081906	1
Ring finger proteins	CBL	0.201688899	1
CD molecules	CCR7	0.253455841	1
unknown	DAXX/TPT1	1	1
Correlation of RiskScore with the signaling pathways and clinical features of samples			

First of all, the KEGG functional enrichment scores of samples in training set, test set and all samples would be analyzed, respectively, using the ssGSEA function of R software package GSVA. Moreover, the correlations with RiskScore would also be calculated according to the enrichment scores of all pathways in all samples. A total of 45 related KEGG pathways had been obtained, as displayed in Table S11-S13. Among them, the top 50% pathways had been selected for clustering analysis according to their enrichment scores, as shown in Fig. 7. It could be observed that, JAK/STAT signaling pathway, Insulin signaling pathway and Pathways in cancer had the best correlation with a correlation coefficient of about 0.36. Thereafter, the correlations of various factors (including T, N, M, Stage, Age and Her2 expression) with RiskScore would also be analyzed, as displayed in Fig. 8. Clearly, there were obvious associations of other features with RiskScore ($p < 0.05$), revealing that the RiskScore model was dependent on these clinical features. On the other hand, the nomogram model would be constructed using RiskScore along with the clinical features. Nomogram is a method to intuitively and effectively demonstrate the results of the risk model, which can conveniently predict outcomes. In the nomogram, the straight line length would be employed to examine the impacts of different variables (and their values) on the outcome. In this

study, the nomogram model had been established using the clinical features (including T, N, M, Stage, Age and Her2 expression), together with RiskScore, as presented in Fig. 9. According to the model results, the RiskScore features would remarkably affect the prediction results of survival rate, which had indicated that, the risk model constructed based on 17 genes could well predict the prognosis. Finally, the forest plot had been established using both RiskScore and clinical features. Notably, forest plot allows to simply and intuitively illustrate the pooled statistical results of different research factors, which generally treats an ineffective line vertical to the X-axis (generally at the coordinate of X = 1 or 0) as the center, while several segments parallel to the X-axis would be employed to represent the effect size and 95% confidence interval (CI) of each study. In this study, the forest plot had been constructed using the clinical features, such as T, N, Stage, Grade, Age, Alcohol and Smoking, and the RiskScore was also calculated by the risk model, as displayed in Fig. 10. As could be figured out, the HR of RiskScore had been evidently increased compared with those of other clinical features ($p < 0.05$). The multivariate cox-regression analyses of various clinical features and RiskScore are presented in Table S14.

Conclusions

BC is a highly complex and heterogeneous malignancy, which is associated with heterogeneous molecular profiles, clinical responses to therapeutics and prognoses [27]. Tumor heterogeneity is responsible for the various BC subtypes, while different subtypes will have different sensitivities to chemotherapy and prognosis [28]. In addition, no consistent therapeutic benefits can be achieved among different patients from clinical medication, which can be ascribed to their potential toxic and side effects. As a result, postoperative systemic adjuvant chemotherapy remains a source of controversy in clinical practice. Therefore, it is crucial to retrieve the potential BC biomarkers predicting patient prognosis and recurrence, as well as to carry out and benefit from the early adjuvant chemotherapy for high risk patients[29].

BC has been recognized to be immunogenic, which has involved multiple putative tumor-associated antigens (TAAs), such as HER-2 and Mucin 1 (MUC1) [30, 31]. Note worthily, these TAAs have been treated as the targets for developing new cancer vaccine and bispecific antibody (bsAbs) over the past decade, among which, some have been translated into tumor-specific immune responses and have been verified to be clinically beneficial[32]. Immunocytes in BC tissue are mainly composed of T-lymphocytes (70–80%), while the remaining components are derived from B-lymphocytes macrophages, natural killer cells and antigen-presenting cells (APCs)[33, 34]. Of them, T-cells can be activated through recognizing the APCs-presented tumor antigens; typically, the intensity and quality of T-cell activation signals are found to be related to a variety of interactions between receptor and ligand[35].

Plenty of evidence have supported that, immunocytes in tumor microenvironment can effectively enhance or suppress tumor growth, which can thereby serve as an indicator for the prognosis of BC. The interactions between the immune system and incipient cancer cells, which is also referred to as immunoediting, can be divided into 3 phases, namely, elimination, equilibrium, and escape[36]. Of them, elimination process suggests that, the innate and adaptive arms of the immune system will recognize the

incipient cancer cells presented by the new antigens (derived from mutations or translocations) on their surface, which is associated with MHC-I; alternatively, the distress signals can be expressed by the transformed cells with chromosomal changes (such as aneuploidy or hyperploidy); finally, the immune system will eliminate these abnormal cells[37]. The equilibrium status will be reached when the immune system fails to eliminate the transformed cells but can stop them from further progression, and such process has been deemed as the dormancy phase during the development of primary cancer. It is mediated by the equilibrium between cells and cytokines for promoting elimination (such as IL-12, IFN-γ, TNF-α, CD4 TH1, CD8 + T cells, NK cells and γδT cells), as well as those promoting the persistence of nascent tumor (including IL-23, IL-6, IL-10, TGF-β, NKT cells, CD4 Th2, Foxp3 + regulatory T [Treg] cells, and MDSCs)[38]. On the other hand, monocytes have exerted crucial part during this process, which may differentiate into proinflammatory M1 or anti-inflammatory M2 types under the effect of tumor microenvironment[39]. Immune escape of cancer cells may take place under various mechanisms. In HR-positive BC, the absence of strong tumor antigens and low MHC-I expression allow for tumor progression that is unnoticed by the immune system [40]. Estrogen has immunosuppressive effect on the tumor microenvironment, which can boost the tolerance of the weak immunogenic cancer; besides, estrogen receptor (ER) can be expressed in most immunocytes, including macrophages, T and B lymphocytes, as well as NK cells [41]. The immune response can be polarized to the Th2 – rather than the Th1-effector immune response in the presence of estrogen [42, 43]. In HER2-positive cancer cells, MHC-I presentation shows negative correlation with HER2 expression[44]. Typically, the triple negative breast cancer (TNBC) has exhibited a spectrum of MHC-I presentations and high antigen expression in tumor, but immune escape in TNBC is found to be primarily related to the development of the immunosuppressive tumor microenvironment (including Tregs, MDSCs and PD-1/PD-L1)[45]. As a result, in the era of immunotherapy, it is particularly important to retrieve the molecular events in tumor immune-microenvironment, so as to search for biomarkers related to survival prediction for BC patients of all subtypes.

In this study, 17 prognosis-specific immune-related genes have been discovered through mining, statistics and sorting of the big data such as TCGA and ImmPort; besides, the prognosis prediction model has also been constructed; the RiskScore of patients is calculated; finally, prediction and verification are also carried out. Our findings suggest that, the prognosis prediction model constructed based on the expression profiles of specific immune genes can further classify patients with definite clinical stage into different subgroups based on the predicted survival results. Moreover, the RiskScore is calculated according to the expression profiles of specific immune-related genes, which should be used in combination with the clinical features of patients, thus more precisely predicting BC patient survival. Taken together, this model contributes to identifying new markers for BC in clinic, which can provide multiple targets for the precise medical treatment for BC in the meantime of more accurately classifying patients at molecular subtype level. Furthermore, this model is promising to guide the clinicians in determining the prognosis, clinical diagnosis and medication for BC patients with different immunophenotypes.

List Of Abbreviations

Breast cancer (BC)

multidrug resistance (MDR)

The Cancer Genome Atlas (TCGA)

monoclonal antibodies (MAbs)

antibody-drug conjugates (ADCs)

tumor-infiltrating lymphocytes (TILs)

bispecific antibody (bsAbs)

triple negative breast cancer (TNBC)

Declarations

- **Ethics approval and consent to participate**

Not applicable

- **Consent for publication**

Not applicable

- **Availability of data and material**

The supplementary data used and generated during the current study are available from the corresponding authors on reasonable request.

- **Competing interests**

The authors declare that they have no competing interests.

- **Funding**

Not applicable

- **Authors' contributions**

GYM and HJW designed the study. GYM, HH collected and analyzed the data. Guoyu Mu, and HH wrote the manuscript. All authors discussed the results and contributed to the final draft of the manuscript. All

authors read and approved the manuscript and agree to be accountable for all aspects of the research in ensuring that the accuracy or integrity of the work are appropriately investigated and resolved.

- **Acknowledgements**

Not applicable.

References

1. Waks AG, Winer EP: **Breast Cancer Treatment**. *Jama* 2019, **321**(3):316.
2. Bernstein L: **Epidemiology of endocrine-related risk factors for breast cancer**. *Journal of mammary gland biology and neoplasia* 2002, **7**(1):3-15.
3. Seiler A, Chen MA, Brown RL, Fagundes CP: **Obesity, Dietary Factors, Nutrition, and Breast Cancer Risk**. *Current breast cancer reports* 2018, **10**(1):14-27.
4. Laas E, Hamy AS, Michel AS, Panchbhaya N, Faron M, Lam T, Carrez S, Pierga JY, Rouzier R, Lerebours F *et al*: **Impact of time to local recurrence on the occurrence of metastasis in breast cancer patients treated with neoadjuvant chemotherapy: A random forest survival approach**. *PloS one* 2019, **14**(1):e0208807.
5. Chaudhary LN, Wilkinson KH, Kong A: **Triple-Negative Breast Cancer: Who Should Receive Neoadjuvant Chemotherapy?** *Surgical oncology clinics of North America* 2018, **27**(1):141-153.
6. Charalampoudis P, Karakatsanis A: **Neoadjuvant chemotherapy for early breast cancer**. *The Lancet Oncology* 2018, **19**(3):e128.
7. Cheng Y, Wu Y, Wu L: **Gene Expression-Guided Adjuvant Chemotherapy in Breast Cancer**. *The New England journal of medicine* 2018, **379**(17):1680-1681.
8. Xiao W, Zheng S, Yang A, Zhang X, Zou Y, Tang H, Xie X: **Breast cancer subtypes and the risk of distant metastasis at initial diagnosis: a population-based study**. *Cancer management and research* 2018, **10**:5329-5338.
9. Park S, Lee SK, Paik HJ, Ryu JM, Kim I, Bae SY, Yu J, Kim SW, Lee JE, Nam SJ: **Adjuvant endocrine therapy alone in patients with node-positive, luminal A type breast cancer**. *Medicine* 2017, **96**(22):e6777.
10. Alfarsi L, Johnston S, Liu DX, Rakha E, Green AR: **Current issues with luminal subtype classification in terms of prediction of benefit from endocrine therapy in early breast cancer**. *Histopathology* 2018, **73**(4):545-558.
11. Veitch Z, Khan OF, Tilley D, Ribnikar D, Kostaras X, King K, Tang P, Lupichuk S: **Real-World Outcomes of Adjuvant Chemotherapy for Node-Negative and Node-Positive HER2-Positive Breast Cancer**. *Journal of the National Comprehensive Cancer Network : JNCCN* 2019, **17**(1):47-56.

12. De Laurentiis M, Cianniello D, Caputo R, Stanzione B, Arpino G, Cinieri S, Lorusso V, De Placido S: **Treatment of triple negative breast cancer (TNBC): current options and future perspectives**. *Cancer treatment reviews* 2010, **36 Suppl 3**:S80-86.
13. Li Y, Gao X, Yu Z, Liu B, Pan W, Li N, Tang B: **Reversing Multidrug Resistance by Multiplexed Gene Silencing for Enhanced Breast Cancer Chemotherapy**. *ACS applied materials & interfaces* 2018, **10**(18):15461-15466.
14. Lee G, Bang L, Kim SY, Kim D, Sohn KA: **Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer**. *BMC medical genomics* 2017, **10**(Suppl 1):28.
15. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, Hu Z, Zalocusky KA, Shankar RD, Shen-Orr SS *et al*: **ImmPort, toward repurposing of open access immunological assay data for translational and clinical research**. *Scientific data* 2018, **5**:180015.
16. Steven A, Seliger B: **The Role of Immune Escape and Immune Cell Infiltration in Breast Cancer**. *Breast care* 2018, **13**(1):16-21.
17. Allahverdiyev A, Tari G, Bagirova M, Abamor ES: **Current Approaches in Development of Immunotherapeutic Vaccines for Breast Cancer**. *Journal of breast cancer* 2018, **21**(4):343-353.
18. Cortes J, Curigliano G, Dieras V: **Expert perspectives on biosimilar monoclonal antibodies in breast cancer**. *Breast cancer research and treatment* 2014, **144**(2):233-239.
19. Bardia A: **Antibody-drug conjugates in breast cancer**. *Clinical advances in hematology & oncology : H&O* 2017, **15**(4):251-254.
20. Bischoff J: **Checkpoint Inhibitors in Breast Cancer - Current Status and Future Directions**. *Breast care* 2018, **13**(1):27-31.
21. Zabotina TN, Korotkova OV, Chertkova AI, Zakharova EN, Tabakov DV, Dzhgamadze NT, Savostikova MV, Artamonova EV, Khailenko VA, Kovalenko EI *et al*: **Tumor-Infiltrating Lymphocytes in Breast Cancer. Association with Clinical and Pathological Parameters**. *Bulletin of experimental biology and medicine* 2018, **166**(2):241-244.
22. Wang J, Chen H, Chen X, Lin H: **Expression of Tumor-Related Macrophages and Cytokines After Surgery of Triple-Negative Breast Cancer Patients and its Implications**. *Medical science monitor : international medical journal of experimental and clinical research* 2016, **22**:115-120.
23. Eltoukhy HS, Sinha G, Moore CA, Sandiford OA, Rameshwar P: **Immune modulation by a cellular network of mesenchymal stem cells and breast cancer cell subsets: Implication for cancer therapy**. *Cellular immunology* 2018, **326**:33-41.
24. Zhang Y, Li H, Zhang W, Che Y, Bai W, Huang G: **LASSObased CoxPH model identifies an 11lncRNA signature for prognosis prediction in gastric cancer**. *Molecular medicine reports* 2018, **18**(6):5579-5593.
25. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, Yates B, Bruford E: **Genenames.org: the HGNC and VGNC resources in 2019**. *Nucleic acids research* 2019, **47**(D1):D786-D792.

26. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data**. *BMC bioinformatics* 2013, **14**:7.
27. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.
28. Tazaki E, Shishido-Hara Y, Mizutani N, Nomura S, Isaka H, Ito H, Imi K, Imoto S, Kamma H: **Histopathologcial and clonal study of combined lobular and ductal carcinoma of the breast**. *Pathology international* 2013, **63**(6):297-304.
29. Shuai Y, Ma L: **Prognostic value of pathologic complete response and the alteration of breast cancer immunohistochemical biomarkers after neoadjuvant chemotherapy**. *Pathology, research and practice* 2019, **215**(1):29-33.
30. Fremd C, Stefanovic S, Beckhove P, Pritsch M, Lim H, Wallwiener M, Heil J, Golatta M, Rom J, Sohn C *et al*: **Mucin 1-specific B cell immune responses and their impact on overall survival in breast cancer patients**. *Oncoimmunology* 2016, **5**(1):e1057387.
31. Conley SJ, Bosco EE, Tice DA, Hollingsworth RE, Herbst R, Xiao Z: **HER2 drives Mucin-like 1 to control proliferation in breast cancer cells**. *Oncogene* 2016, **35**(32):4225-4234.
32. Ye H, Sun C, Ren P, Dai L, Peng B, Wang K, Qian W, Zhang J: **Mini-array of multiple tumor-associated antigens (TAAs) in the immunodiagnosis of breast cancer**. *Oncology letters* 2013, **5**(2):663-668.
33. Coventry BJ, Weightman MJ, Bradley J, Skinner JM: **Immune profiling in human breast cancer using high-sensitivity detection and analysis techniques**. *JRSM open* 2015, **6**(9):2054270415603909.
34. Pusztai L, Karn T, Safonov A, Abu-Khalaf MM, Bianchini G: **New Strategies in Breast Cancer: Immunotherapy**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2016, **22**(9):2105-2110.
35. Pardoll DM: **The blockade of immune checkpoints in cancer immunotherapy**. *Nature reviews Cancer* 2012, **12**(4):252-264.
36. Mittal D, Gubin MM, Schreiber RD, Smyth MJ: **New insights into cancer immunoediting and its three component phases–elimination, equilibrium and escape**. *Current opinion in immunology* 2014, **27**:16-25.
37. Croxford JL, Tang ML, Pan MF, Huang CW, Kamran N, Phua CM, Chng WJ, Ng SB, Raulet DH, Gasser S: **ATM-dependent spontaneous regression of early Emu-myc-induced murine B-cell leukemia depends on natural killer and T cells**. *Blood* 2013, **121**(13):2512-2521.
38. Wu X, Peng M, Huang B, Zhang H, Wang H, Huang B, Xue Z, Zhang L, Da Y, Yang D *et al*: **Immune microenvironment profiles of tumor immune equilibrium and immune escape states of mouse sarcoma**. *Cancer letters* 2013, **340**(1):124-133.
39. Jinushi M, Komohara Y: **Tumor-associated macrophages as an emerging target against tumors: Creating a new path from bench to bedside**. *Biochimica et biophysica acta* 2015, **1855**(2):123-130.
40. Lee HJ, Song IH, Park IA, Heo SH, Kim YA, Ahn JH, Gong G: **Differential expression of major histocompatibility complex class I in subtypes of breast cancer is associated with estrogen receptor and interferon signaling**. *Oncotarget* 2016, **7**(21):30119-30132.

41. Pierdominici M, Maselli A, Colasanti T, Giammarioli AM, Delunardo F, Vacirca D, Sanchez M, Giovannetti A, Malorni W, Ortona E: **Estrogen receptor profiles in human peripheral blood lymphocytes**. *Immunology letters* 2010, **132**(1-2):79-85.
42. Hu ZY, Xiao H, Xiao M, Tang Y, Sun J, Xie ZM, Ouyang Q: **Inducing or Preventing Subsequent Malignancies for Breast Cancer Survivors? Double-edged Sword of Estrogen Receptor and Progesterone Receptor**. *Clinical breast cancer* 2018, **18**(5):e1149-e1163.
43. Salem ML: **Estrogen, a double-edged sword: modulation of TH1- and TH2-mediated inflammations by differential regulation of TH1/TH2 cytokine production**. *Current drug targets Inflammation and allergy* 2004, **3**(1):97-104.
44. Inoue M, Mimura K, Izawa S, Shiraishi K, Inoue A, Shiba S, Watanabe M, Maruyama T, Kawaguchi Y, Inoue S *et al*: **Expression of MHC Class I on breast cancer cells correlates inversely with HER2 expression**. *Oncoimmunology* 2012, **1**(7):1104-1110.
45. Engel JB, Honig A, Kapp M, Hahne JC, Meyer SR, Dietl J, Segerer SE: **Mechanisms of tumor immune escape in triple-negative breast cancers (TNBC) with and without mutated BRCA 1**. *Archives of gynecology and obstetrics* 2014, **289**(1):141-147.

Figures

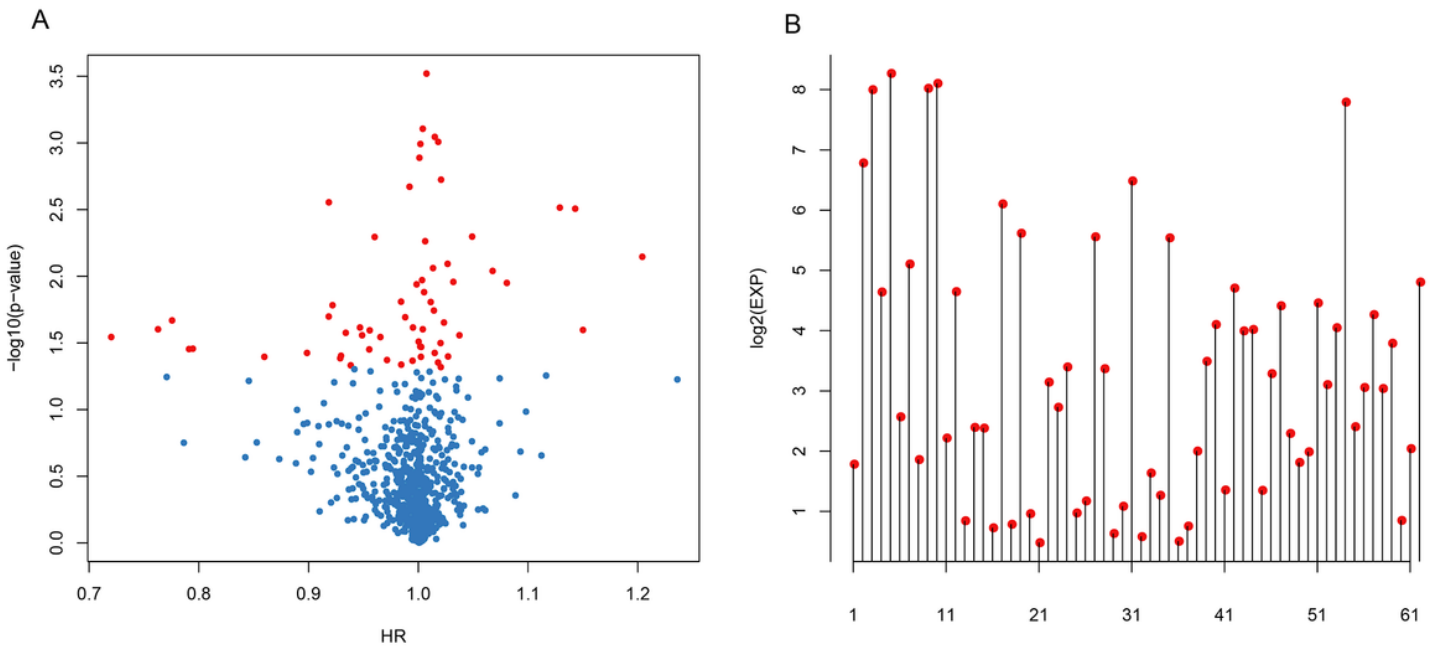

Figure 1

The relationships of the p-values of 62 genes and the HR, expression quantities. (A) The relationships of the p-values of 62 genes and the HR is displayed. (B) The relationships of the p-values of 62 genes and the expression levels. Red dots repre-sents significantly different immune-related genes (p□0.05) regarding prognosis.

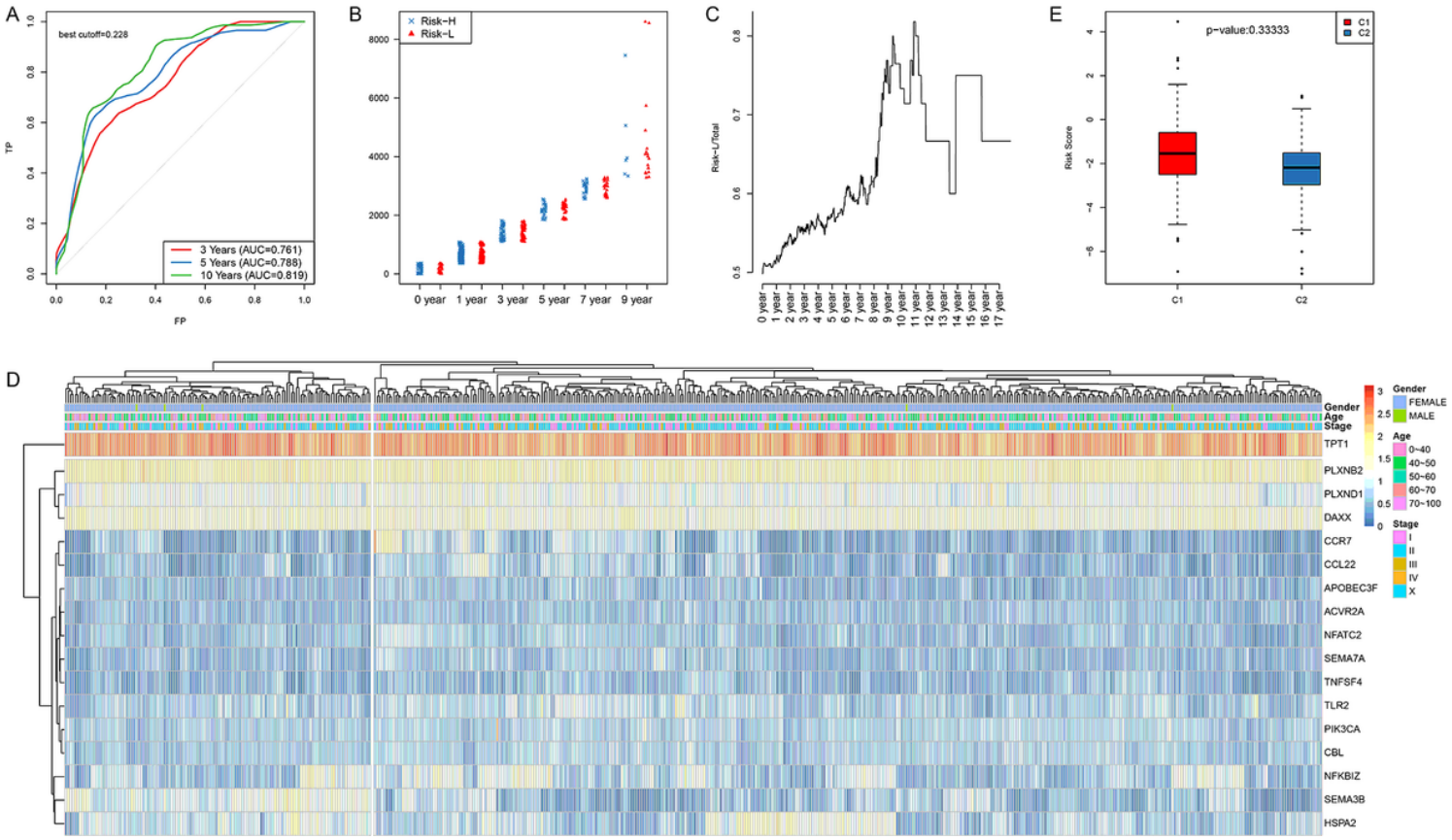

Figure 2

Verify the stability of the prognosis prediction model included 17 immune-related genes for BC patients in training set. (A) The survival predicted ROC curves of 17-gene risk model in training set. (B) The distribution of samples in Risk-H and Risk-L groups of training set divided through 17-gene risk model under different OS. (C) The level of Risk-L group/Total sample size with the extension in OS in the traning set. (D) The clustering results of training set samples. (E) Difference in the RiskScore between the two groups which had been clustered by the expression of 17 genes of training set samples.

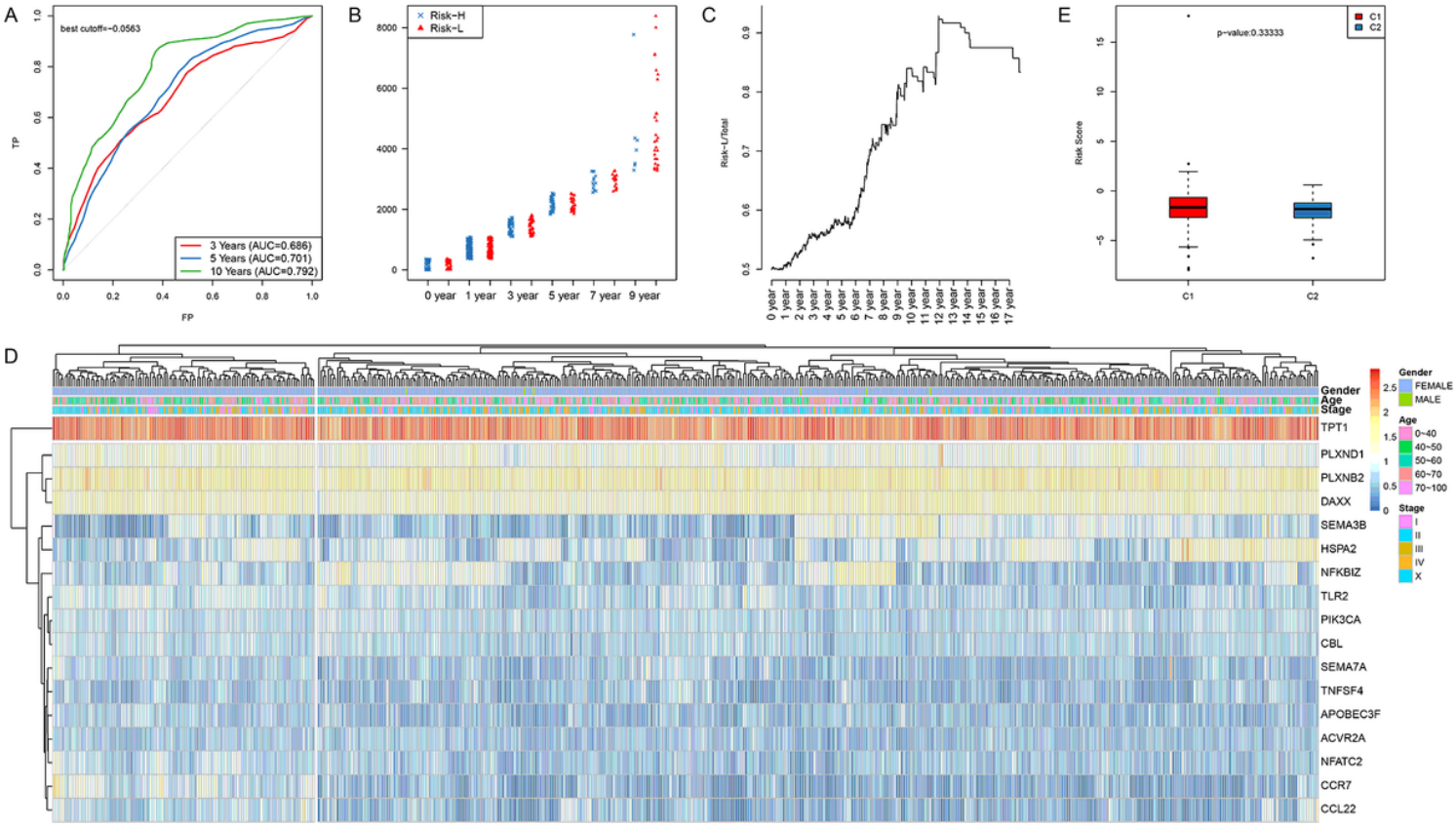

Figure 3

Verify the reliability of the prognosis prediction model included 17 im-mune-related genes for BC patients in test set. (A) The survival predicted ROC curves of 17-gene risk model in test set. (B) The distribution of samples in Risk-H and Risk-L groups of test set divided through 17-gene risk model under different OS. (C) The level of Risk-L group/Total sample size with the extension in OS in the test set. (D) The clustering results of test set samples. (E) Difference in the RiskScore between the two groups which had been clustered by the expression of 17 genes of test set sam-ples.

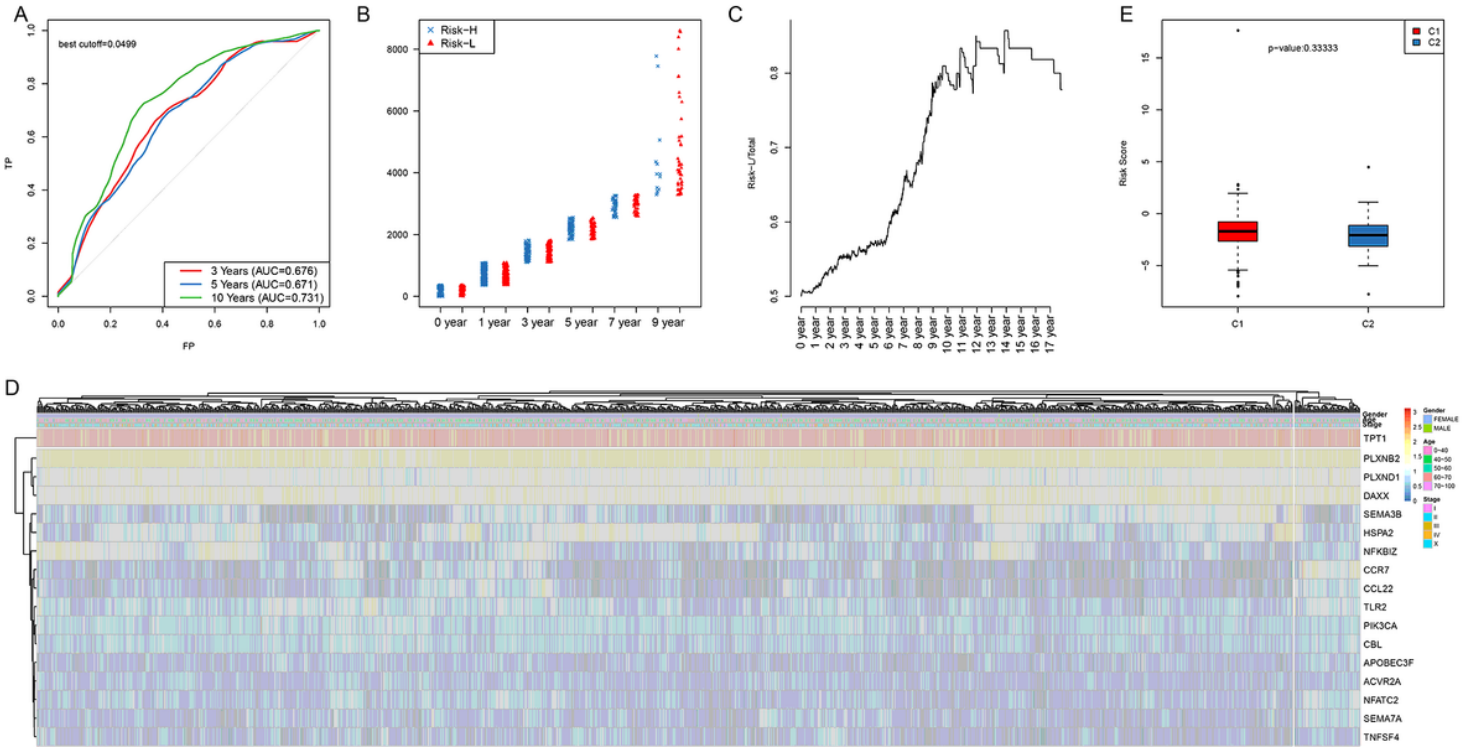

Figure 4

Verify the reliability of the prognosis prediction model included 17 im-mune-related genes for all the BC patients in both sets. (A) The survival predicted ROC curves of 17-gene risk model. (B) The distribution of all the samples in Risk-H and Risk-L groupsdivided through 17-gene risk model under different OS. (C) The level of Risk-L group/Total sample size with the extension in OS. (D) The clustering results of all the samples. (E) Difference in the RiskScore between the two groups which had been clustered by the expression of 17 genes.

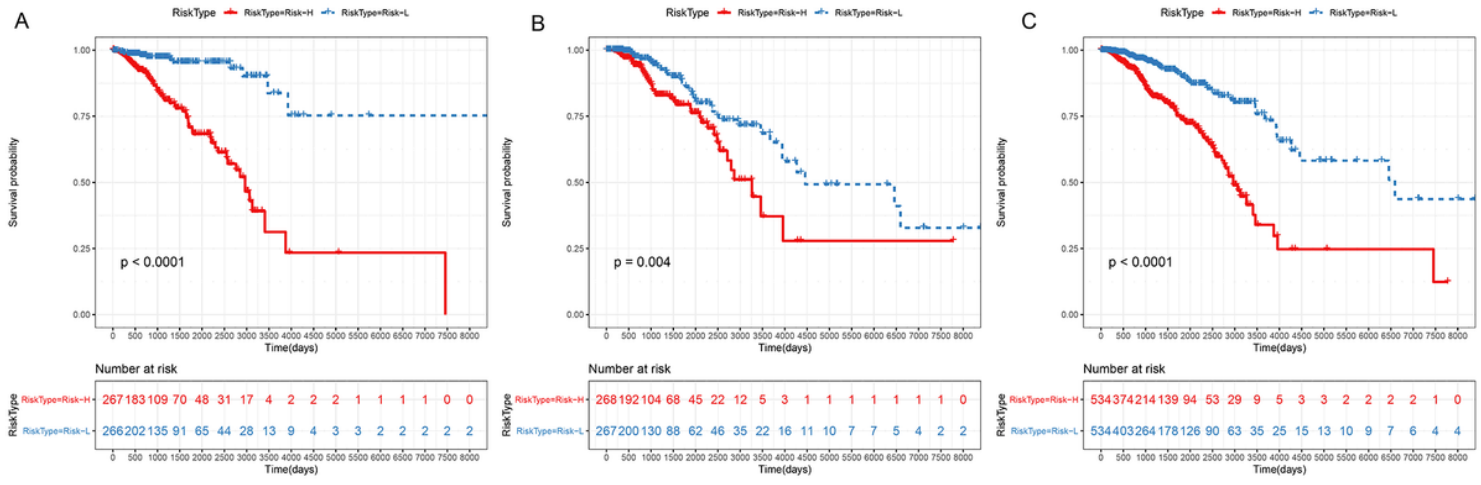

Figure 5

The KM survival curve of the 17-gene-based risk model in predicting the Risk-H and Risk-L groups on the training set (A), test set (B) and all samples (C).

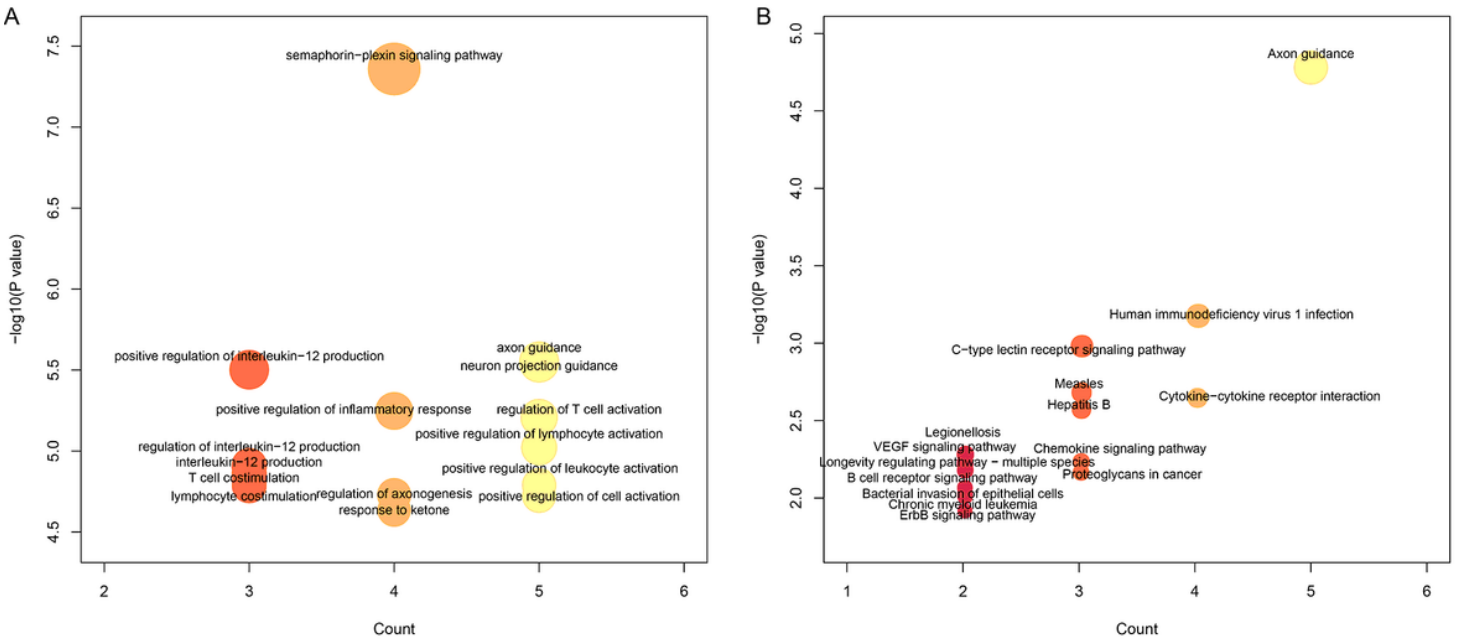

Figure 6

The GO (A) and KEGG pathway (B) enrichment analysis of the 17 specific im-mune-related genes.

Figure 7

Correlation of RiskScore with signaling pathways. KEGG functional enrichment score of each sample was analyzed, the correlation with RiskScore was calculated, re-spectively,based on the enrichment score of each pathway in each sample, and top 30 pathways related KEGG pathways were shown. Clustering analysis had to be carried out according to the enrichment scoreintraining set.

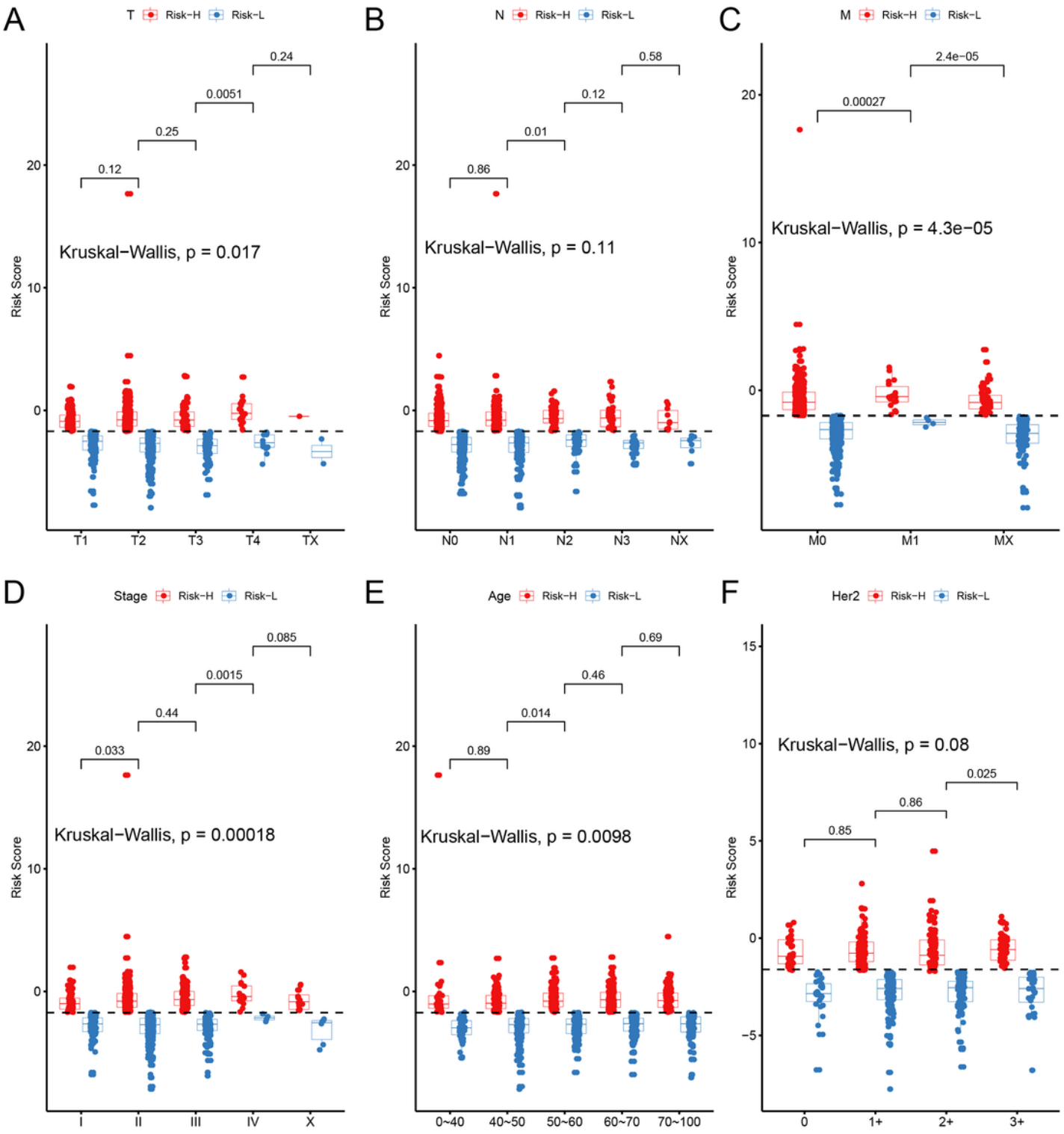

Figure 8

The relationships of different clinical factorswith RiskScorefor BC pa-tients.Comparison of RiskScore among different T (A),N (B), M (C), stage (D), age (E) and Her2 expression (F). The horizontal axis represents the different clinical fac-tors, and the vertical axis represents RiskScores.

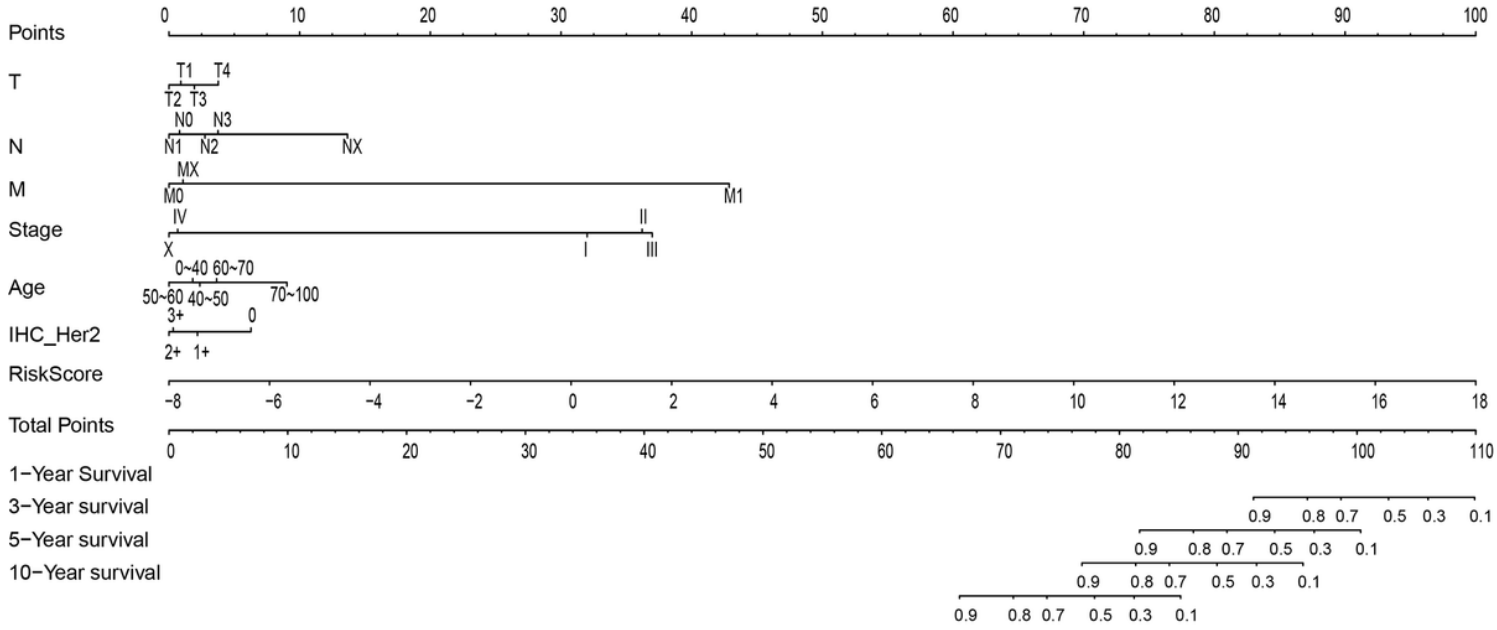

Figure 9

The nomogram model constructed by combining the clinical features (T, N, M, stage, age and Her2 expression) with RiskScorefor BC patients

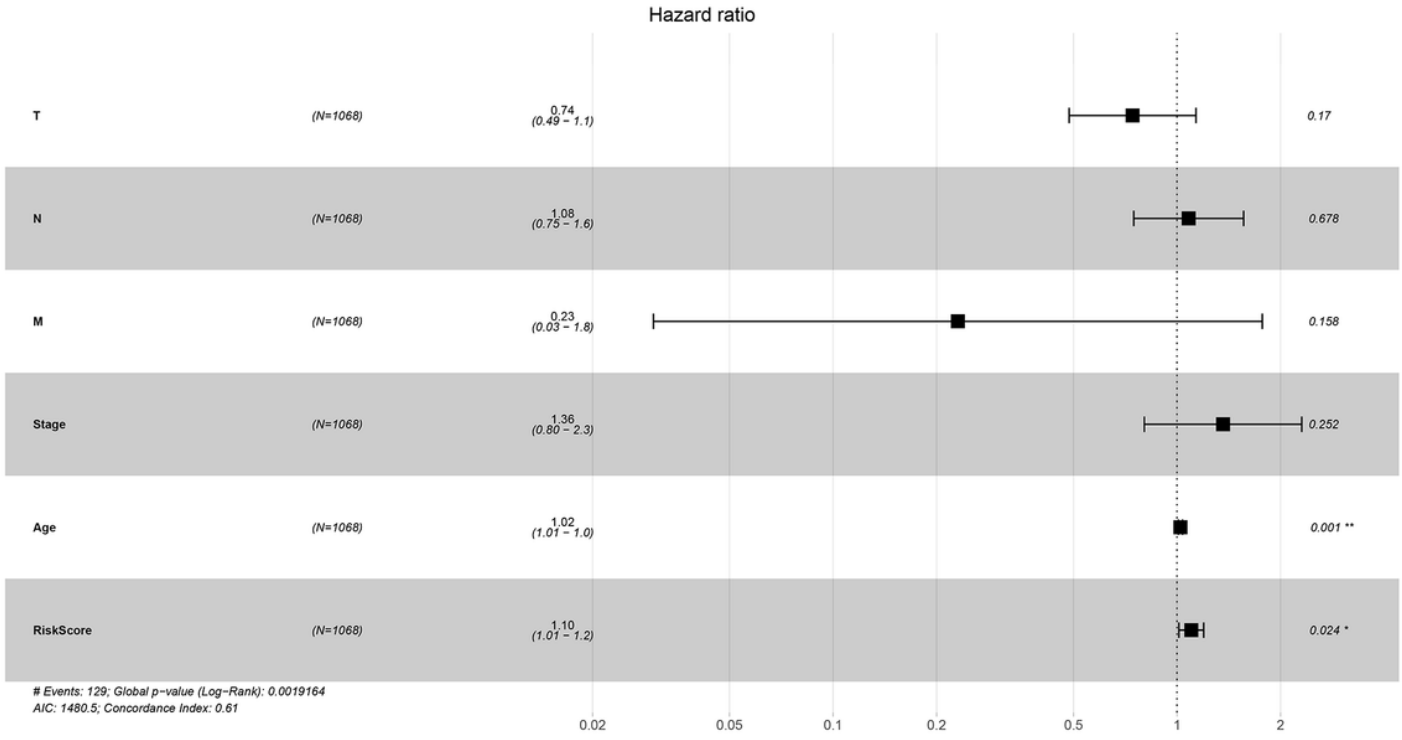

Figure 10

The forest plot constructed by combining the clinical features with RiskScorefor BC patients.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymanu.docx
- S6S14Table.xlsx
- S6Fig.tif
- S5Table.txt
- S5Fig.tif
- S4Table.txt
- S4Fig.tif
- S3Table.txt
- S3Fig.tif
- S2Table.txt
- S2Fig.tif
- S1Table.txt
- S1Fig.tif