pISSN 1598-2998, eISSN 2005-9256

https://doi.org/10.4143/crt.2020.424

Cancer Res Treat. 2021;53(1):148-161

Original Article

An Innovative Prognostic Model Based on Four Genes in Asian Patient with Gastric Cancer

Jiahui Chen[1], **Anqiang Wang**[1], **Jun Ji**[2,3], **Kai Zhou**[1], **Zhaode Bu**[1], **Guoqing Lyu**[4], **Jiafu Ji**[1]

[1]*Department of Gastrointestinal Surgery, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing,* [2]*Department of Gastrointestinal Surgery, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University, Jinan,* [3]*First Affiliated Hospital of Baotou Medical College, General Surgery, Baotou,* [4]*Department of Gastrointestinal Surgery, Peking University Shenzhen Hospital, Shenzhen, China*

Purpose Gastric cancer (GC) has substantial biological differences between Asian and non-Asian populations, which makes it difficult to have a unified predictive measure for all people. We aimed to identify novel prognostic biomarkers to help predict the prognosis of Asian GC patients.

Materials and Methods We investigated the differential gene expression between GC and normal tissues of GSE66229. Univariate, multivariate and Lasso Cox regression analyses were conducted to establish a four-gene-related prognostic model based on the risk score. The risk score was based on a linear combination of the expression levels of individual genes multiplied by their multivariate Cox regression coefficients. Validation of the prognostic model was conducted using The Cancer Genome Atlas (TCGA) database. A nomogram containing clinical characteristics and the prognostic model was established to predict the prognosis of Asian GC patients.

Results Four genes (*RBPMS2*, *RGN*, *PLEKHS1*, and *CT83*) were selected to establish the prognostic model, and it was validated in the TCGA Asian cohort. Receiver operating characteristic analysis confirmed the sensitivity and specificity of the prognostic model. Based on the prognostic model, a nomogram containing clinical characteristics and the prognostic model was established, and Harrell's concordance index of the nomogram for evaluating the overall survival significantly higher than the model only focuses on the pathologic stage (0.74 vs. 0.64, $p < 0.001$).

Conclusion The four-gene-related prognostic model and the nomogram based on it are reliable tools for predicting the overall survival of Asian GC patients.

Key words Stomach neoplasms, Overall survival, Prognostic model, Risk score, Nomograms

Introduction

Gastric cancer (GC) ranks as the third leading cause of cancer-related death worldwide. Every year, approximately one million people are diagnosed with GC, over half of which reside in Asia [1]. Surgical resection remains the most effective treatment for early and some advanced forms of GC, and over the last decade, chemotherapy has been considered a standard therapeutic regimen for patients with advanced or metastatic GC [2]. However, most patients with GC will eventually metastasize or relapse and the prognosis is still unsatisfactory [3]. Therefore, there is an urgent need to explore novel prognostic biomarkers to increase the accuracy of prognosis prediction and seize therapeutic opportunities for GC patients.

GC has substantial biological differences between Asian and non-Asian populations [4], which makes it difficult to have a unified predictive measure for all people. Due to the complex biological and molecular mechanisms underlying GC, traditional predict methods relying on clinical data such as serum examination, imaging examination, and pathologic information are limited.

In this genomic era, high-throughput platforms such as sequencing and microarrays play an increasingly important role in the field of oncology and make precision medicine possible. Using the hepatocellular carcinoma data set of The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), Long et al. [5] established a prognostic model for overall survival (OS) prediction, suggesting that these approaches and data have a wide range of clinical applications. Hou

Correspondence: Jiafu Ji
Department of Gastrointestinal Surgery Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, No. 52 Fucheng Road, Haidian District, Beijing 100142, China
Tel: 86-10-88196970 Fax: 86-10-88196970 E-mail: jijiafu@hsc.pku.edu.cn

Co-correspondence: Guoqing Lyu
Department of Gastrointestinal Surgery, Peking University Shenzhen Hospital, No. 1120 Lianhua Road, Shenzhen 518036, China
Tel: 86-0755-83923333 Fax: 86-0755-83061340 E-mail: 365973269@qq.com

*Jiahui Chen, Anqiang Wang, and Jun Ji contributed equally to this work.

Received May 8, 2020 **Accepted** August 28, 2020
Published Online August 31, 2020

et al. [6] analyzed the GC expression profile data in the GEO database through Lasso regression analysis, and obtained 11 genes related to prognosis, and thought that the selection of more than five genes can predict the prognosis. Cheong et al. [7] established a 4-gene chemotherapy prediction model based on the expression profile data of GC after D2 surgery, which can assess whether patients can benefit from postoperative adjuvant chemotherapy. The incidence of GC in Asian populations is much higher than that of other races. However, at present, there is no prognostic model based on comprehensive clinical data and gene expression data, which can make a more accurate judgment on the prognosis of Asian GC patients.

In our study, we used a GEO dataset from Korea to establish a four-gene prognostic model, including mRNA processing factor 2 (RBPMS2), regucalcin (RGN), pleckstrin homology domain containing S1 (PLEKHS1), and cancer/testis antigen 83 (CT83). All patients were classified into a low-risk group and a high-risk group. The prognostic model was validated in the Asian GC cohort in the TCGA database. Finally, a nomogram including clinical characteristics and prognostic model was established for OS prediction. As a whole, these predictive methods will help stratify Asian patients with GC more accurately and promote more precise treatment for all of them.

Materials and Methods

1. Microarray data

The gene expression profile matrix of GSE66229 was downloaded from the GEO database, and the clinical information was obtained from the corresponding published literature [8]. GSE66229 is composed of the GSE62254 and GSE66222 datasets, which were based on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array, Santa Clara, CA). The GSE62254 dataset contains 300 GC samples, and the GSE66222 dataset contains 100 adjacent nontumorous gastric tissues. Patients with GSE66222 are all included in GSE62254. We excluded two patients with multiple primary tumors. A total of 298 patients with 397 samples (99 adjacent nontumorous and 298 tumor tissues) were selected for further analyses.

2. RNA-sequencing data

The RNA-sequencing data and the corresponding clinical information used for validation were downloaded from the TCGA database. Among these GC patients, there were 348 patients with 373 samples (30 adjacent nontumorous and 343 tumor tissues). We excluded 35 patients with incomplete clinical information. Finally, 313 patients with 336 samples (28 adjacent nontumorous and 308 tumor tissues) were selected as the TCGA global cohort. According to the patients' race, the cohort was further divided into an Asian cohort and a non-Asian cohort. The Asian cohort contained 63 patients with 68 samples (6 nontumorous and 62 tumor tissues), and the non-Asian cohort contained 250 patients with 268 samples (22 nontumorous and 246 tumor tissues).

3. Identification of differentially expressed mRNA

Firstly, we obtained the original mRNA expression profiles of GC from the GSE66229 dataset and merged them into one file. The RMA algorithm is used to normalize and Log2 transform the expression data in R environment. The average value of gene expression was taken when duplicate data were found. Genes with an average expression value > 1 were retained, while the low abundance sequencing data were deleted. The mRNA microarray includes 20,486 mRNA expression profiles. Next, we calculated the identification of differentially expressed genes (DEGs) with the help of the Limma package [9], where genes with logFC > 2 or logFC < –2 and adjusted $p < 0.001$ were considered for subsequent analysis, the cutoff value of LogFC was also used in other study [10]. The Pheatmap package and Volcanomap package were applied to describe the DEGs via the R language.

4. Establishment and verification of the prognostic model

Univariate, Lasso-penalized, and multivariate Cox regression analyses were performed to reduce candidate genes and explore the correlation between the mRNA expression levels and patient OS.

In the univariate Cox regression, mRNA expression was considered to be significant when the $p < 0.001$. Then, we conducted Lasso-penalized Cox regression to further reduce the number of genes. For the selection operator of Lasso-penalized Cox regression, we sub-sampled the dataset that was replaced one thousand times and selected markers with a repetition frequency greater than 900. The tuning parameters were determined by the expected generalization error, which was estimated by 10-fold cross-validation and information-based Akaike Information Criteria/Bayesian Information Criteria. To ensure that the error was within one standard error from the minimal (MSE), the maximum value of the lambda was adopted, which was defined as "1-MSE" lambda. Then, the expression data of candidate genes was standardized through the scale method in R Software (R Foundation for Statistical Computing, Vienna, Austria). To evaluate individual genes as independent prognostic factors for OS, we performed a multivariate Cox regression analysis. The stepwise approach was applied to further select the best model. Ultimately, a prognostic model based on four genes was established. The risk score was based on a linear combi-

nation of the expression levels of individual genes multiplied by their multivariate Cox regression coefficients (β): risk score=(expression level of RBPMS2×β)+(expression level of RGN×β)+(expression level of PLEKHS1×β)+(expression level of CT83×β). The X-tile software was used to determine the optimal cutoff value [11]. Kaplan-Meier survival analyses were performed for the low-risk group and the high-risk group and the log-rank test was used to compare survival rates. A time-dependent receiver operating characteristic (ROC) curve was performed to assess the predictive power of the predictive model. Independence analysis of diagnostic models with other clinical characteristics was conducted by univariate and multivariate Cox regression analysis.

The relationship between the expression of the four genes and OS in the GSE62254 dataset was verified in a Kaplan-Meier plotter (KmPlot, https://kmplot.com/analysis/index.php?p=service&cancer=gastric). Validation of the prognostic model was performed using the TCGA database. The expression data of the four genes was standardized by the scale method as well in validation cohorts.

5. Validation of the expression levels of the four genes in TCGA cohorts

We extracted the expression levels of the four genes in the GEO cohort, TCGA global cohort, and TCGA Asian cohort to further explored the expression patterns of the four genes. The different expression patterns between the low-risk group and the high-risk group in the three cohorts were analyzed by Wilcoxon signed-rank test. And the analysis was performed with statistical software GraphPad Prism ver. 7.0 (GraphPad Software, Inc., San Diego, CA). The p-values are two-sided, and $p < 0.05$ was considered statistically significant.

6. Establishment and validation of the results in a nomogram

Nomogram is widely used as prognostic devices in oncology research. This approach can generate an individual probability of a specific clinical event via the integration of different determinant and prognostic variables [12-14]. In our study, we established a nomogram to evaluate the 1-year, 3-year, and 5-year OS probability of GC patients in the GEO cohort. We adopted Harrell's concordance index (C-index) to evaluate the predictive power of the nomogram (combined model). The C-index is calculated using a 1,000 resampled boot method, whose values range from 0.5 to 1.0. In general, the C-index is less accurate at 0.50-0.70, moderately accurate at 0.71-0.90, and highly accurate above 0.90. The calibration of the nomogram was accomplished by drawing a comparison between the prediction probabilities and the observation probabilities. The closer the probabilities were to the reference line, the higher the consistency of the model. Simultaneously, single prognostic factors were utilized to construct the nomograms, and the prediction accuracy with the nomogram was compared by using C-index, ROC analysis, and decision curve analysis (DCA). DCA is a novel method for evaluating prognostic strategies, with the ability to visualize the clinical effectiveness of the nomogram [15]. All of the above methods were implemented in the R language.

Results

1. Identification of DEGs between GC and nontumorous tissues

To better illustrate our research, we have drawn an analysis process flow chart (Fig. 1). In the mRNA expression profile of GC patients (n=298), a total of 252 DEGs (LogFC > 2 or LogFC < –2, adjusted $p < 0.001$) were found in comparison with normal tissues (n=99). Among these DEGs, 163 genes were downregulated and 89 genes were overexpressed. Differentially expressed mRNA heatmaps and volcano maps are shown in S1 and S2 Figs.

Then, univariate Cox regression analysis was applied to explore the DEGs related to OS, and 130 genes ($p < 0.001$) were identified after primary filtration (Fig. 1). To further reduce the candidate genes, we conducted a Lasso-penalized Cox analysis. As a result, five genes met the screening criteria, which appeared more than 900 times in 1,000 screenings (S3 Fig.). Ultimately, we performed a stepwise multivariate Cox regression analysis, and four candidate genes were finally selected to establish the prognostic model. Among them, RNA binding protein, *RBPMS2*, and *RGN* were downregulated in tumor tissues, and the other two genes, *PLEKHS1* and *CT83*, were overexpressed in tumor tissues.

2. The prognostic model shows good performance in risk stratification and ROC curve verification for GC patients

We built a predictive model using the four genes selected above and divided the GC patients into a low-risk group and a high-risk group according to the risk scores. The risk score=(expression level of RBPMS2×0.308)+(expression level of RGN×0.192)+(expression level of PLEKHS1×–0.197)+(expression level of CT83×–0.190). RBPMS2 (hazard ratio [HR], 1.360; 95% confidence interval [CI], 1.172 to 1.579) and RGN (HR, 1.211; 95% CI, 1.010 to 1.453) showed positive coefficients, indicating that these two genes are high-risk factors in GS, while PLEKHS1 (HR, 0.821; 95% CI, 0.680 to 0.991) and CT83 (HR, 0.827; 95% CI, 0.694 to 0.985) showed negative coefficients, suggesting that their overexpression signified a longer OS. Verification in KM Plot supports the assumption (Fig. 2A), the high expression of RBPMS2 and RGN was negatively correlated with the patient's prognosis,

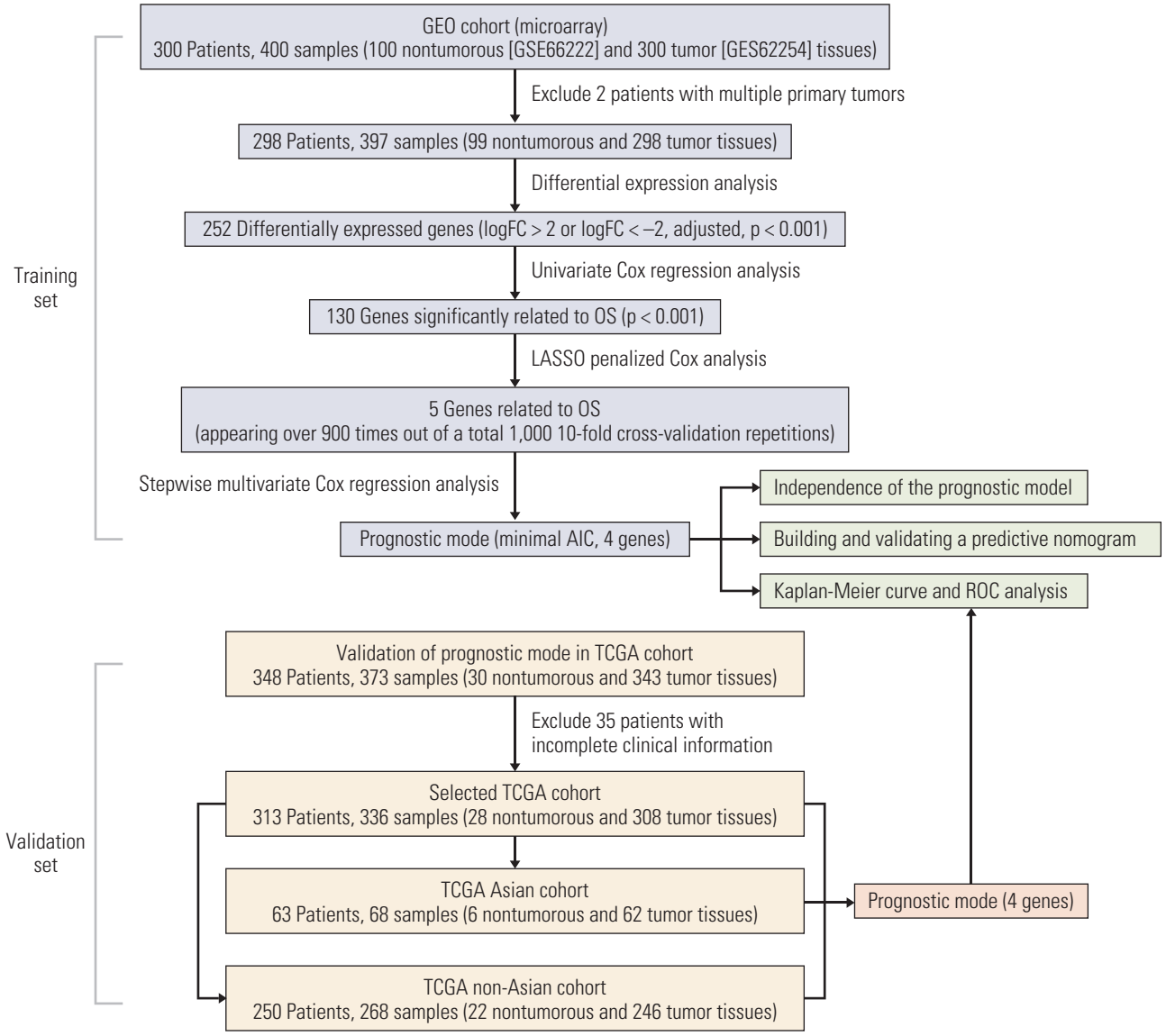

Fig. 1. The flowchart is used to describe the establishment and verification of the prognostic model. AIC, Akaike Information Criteria; GEO, Gene Expression Omnibus; OS, overall survival; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas.

while the expression of PLEKHS1 and CT83 was opposite, and it was positively correlated with the patient's prognosis. The optimal cutoff value of the risk score determined by X-tile software was 1.2 (S4 Fig.). As a result, 88 patients were classified into the high-risk group, and the other 210 patients were classified into the low-risk group. The Kaplan-Meier survival curves of the two risk groups showed significant differences (low-risk group vs. high-risk group: median OS, > 100 months vs. 21.9 months, respectively; $p < 0.001$) (Fig. 2B). ROC analysis confirmed the sensitivity and specificity of the prognostic model. The area under curves (AUCs) of the prognostic model were 0.684, 0.697, 0.699, 0.736, and 0.750 for 0.5-year, 1-year, 2-year, 3-year, and 5-year survival, respectively (Fig. 2C). In general, the prognostic model shows good performance in risk stratification for GC patients.

3. Univariate and multivariate Cox regression analysis proves that the prognostic model has good independence

We conducted univariate and multivariate Cox regression analyses on the diagnostic model to assess its independent predictive value with other conventional clinical factors in 298 GC patients from the GEO cohort. The univariate Cox regression showed that Lauren type (intestinal/diffuse/mixed), lymphatic and lymphovascular invasion (yes/no),

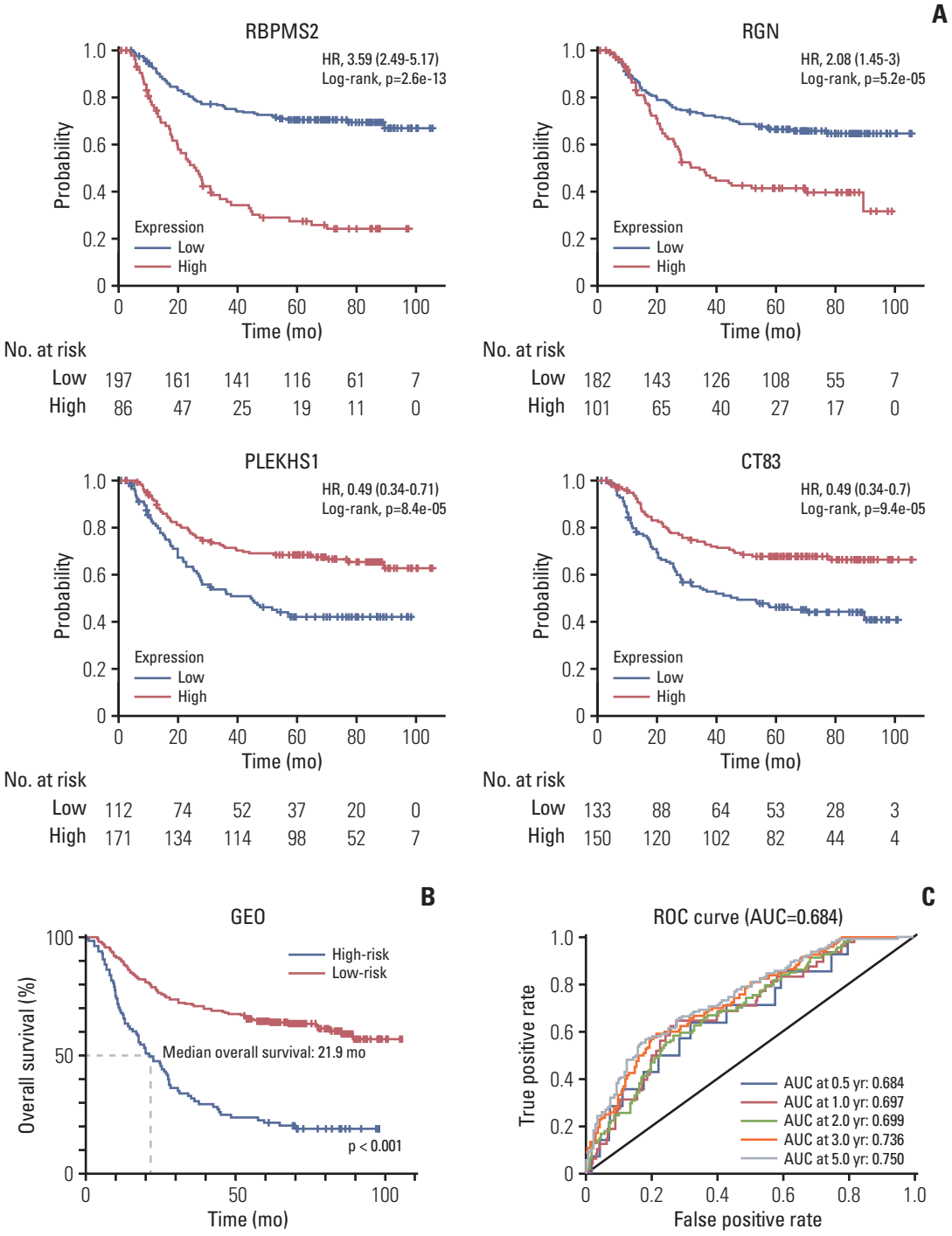

Fig. 2. Relationship between gene expression and prognosis of the four prognostic genes in KmPlot (A). Kaplan-Meier curve (B) and time-dependent receiver operating characteristic (ROC) curve (C) of the prognostic model in the Gene Expression Omnibus (GEO) cohort. The Kaplan-Meier curve shows the overall survival of patients in the high-risk group and the low-risk group distinguished by the optimal cutoff value. The ROC curve confirms the sensitivity and specificity of the prognostic model. *(Continued to the next page)*

D

Characterisric	HR (95% CI)	p-value
Univariate analysis		
Age	1.011 (0.996-1.026)	0.156
Sex (male/female)	0.917 (0.656-1.282)	0.612
EBV ISH (positive/negative)	1.045 (0.549-1.989)	0.893
Lauren type (intestinal/diffuse/mixed)	1.552 (1.178-2.044)	0.002
Lymphatic and lymphovascular invasion (yes/no)	2.230 (1.424-3.493)	< 0.001
Borrmann type (1/2/3/4)	1.931 (1.534-2.431)	< 0.001
Tumor location (cardia/body/antrum/whole)	0.887 (0.705-1.116)	0.307
pTNM stage (stage III+IV/stage I+II)	3.411 (2.343-4.964)	< 0.001
Progosis model (high/low)	3.283 (2.378-4.532)	< 0.001
Multivariate analysis		
Age	1.033 (1.015-1.051)	< 0.001
Lauren type (intestinal/diffuse/mixed)	1.176 (0.857-1.613)	0.316
Lymphatic and lymphovascular invation (yes/no)	1.518 (0.956-2.410)	0.077
Borrmann type (1/2/3/4)	1.302 (1.004-1.688)	0.047
pTNM stage (stage III+IV/stage I+II)	2.562 (1.671-3.928)	< 0.001
Progosis model (high/low)	2.708 (1.891-3.877)	< 0.001

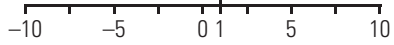

Fig. 2. *(Continued from the previous page)* Univariate and multivariate Cox regression analysis of prognostic model and other conventional clinical factors with overall survival (D). Red is not statistically significant and blue is statistically significant. AUC, area under curves; CI, confidence interval; EBV, Epstein-Barr virus; HR, hazard ratio; ISH, *in situ* hybridization.

Borrmann type (1/2/3/4), pTNM stage (stage III+IV/stage I+II), and prognostic model (high/low) were correlated with OS (Fig. 2D). However, age, sex, Epstein-Barr virus in situ hybridization status, and tumor location had no prognostic value in univariate Cox regression. Then, we performed multivariate Cox regression analysis on meaningful indicators in the results of univariate Cox regression analysis, and age was also included in the analysis. Finally, age, Borrmann type, pTNM stage, and prognostic model were shown to be independent predictors of OS (Fig. 2D).

4. The predictive model works well in the TCGA Asian population and is not effective in non-Asian populations

Validation of the prognostic model was conducted using TCGA database. In the TCGA global cohort, 110 patients were classified into the high-risk group, and the other 198 patients were classified into the low-risk group. Consistent with the result in the training set, patients in the low-risk group had a better prognosis than the higher-risk group (p=0.008) (Fig. 3A). However, the difference in OS is not as obvious as in the training set. Considering that GC has significant biological differences between different races, we further validated the prognostic model in TCGA Asian cohort and TCGA non-Asian cohort. In TCGA Asian cohort, a total of 23 patients were classified into the high-risk group, and 39 patients were classified into the low-risk group. The survival curves are highly consistent with the GEO cohort (p=0.002) (Fig. 3C). However, in the TCGA non-Asian cohort, the survival curve was similar to the TCGA global cohort, but it is not statistically significant (p=0.115) (Fig. 3E). Furthermore, ROC analysis confirmed the sensitivity and specificity of the prognostic model in the TCGA Asian cohort (AUC, 0.694), and AUCs for 0.5-year, 1-year, 2-year, 3-year, and 5-year survival were 0.694, 0.748, 0.741, 0.754, and 0.754 respectively (Fig. 3D). In the TCGA global cohort and TCGA non-Asian cohort, the AUCs are 0.593 and 0.567 (Fig. 3B and F), suggesting that the prognostic model has limited predictive power in these two cohorts.

5. Verification of the expression levels of the four genes in TCGA datasets

In the GEO cohort, RBPMS2 and RGN were overexpressed, while PLEKHS1 and CT83 were downregulated in the high-risk group (Fig. 4A). To further verify the accuracy of this result, we compared the expression of these four genes in the TCGA global cohort (Fig. 4B) and the TCGA Asian cohort (Fig. 4C). The results were similar to the GEO cohort.

6. Establishment and validation of the nomogram

To establish a more convenient and accurate survival pre-

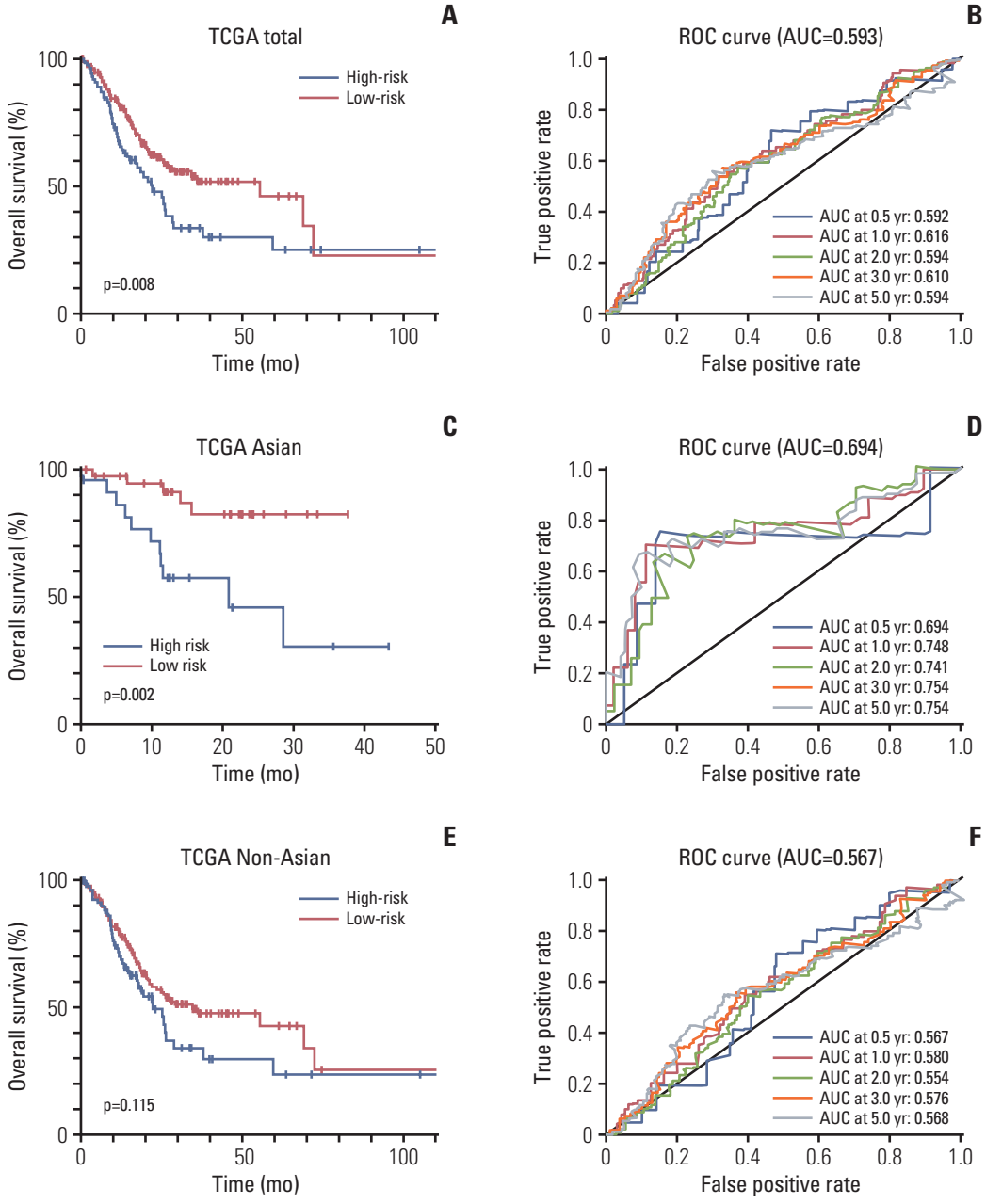

Fig. 3. Kaplan-Meier curve and time-dependent receiver operating characteristic (ROC) curve for the verification set in The Cancer Genome Atlas (TCGA) global cohort (A, B), TCGA Asian cohort (C, D), and the TCGA non-Asian cohort (E, F). The Kaplan-Meier curve shows the overall survival of patients in the high-risk group and the low-risk group distinguished by the same cutoff point as the prognostic model. AUC, area under curves.

diction method for patients with GC, we built a nomogram using the results of multivariate Cox regression analysis. The OS-related factors include age, pathologic stage (III+IV/I+II), Borrmann type (1/2/3/4), and prognostic model (high/low) (Fig. 5A). We used calibration plots to verify the nomogram (Fig. 5B). The C-index for the nomogram (combined model) was 0.74 (95% CI, 0.71 to 0.78), which was significantly higher than in other models. (Table 1). ROC analysis and DCA were performed to compare the prediction accuracy between single clinical factors and the nomogram. The DCA curve showed that the nomogram had the strongest predictive power and accuracy among several predictive models (Fig.

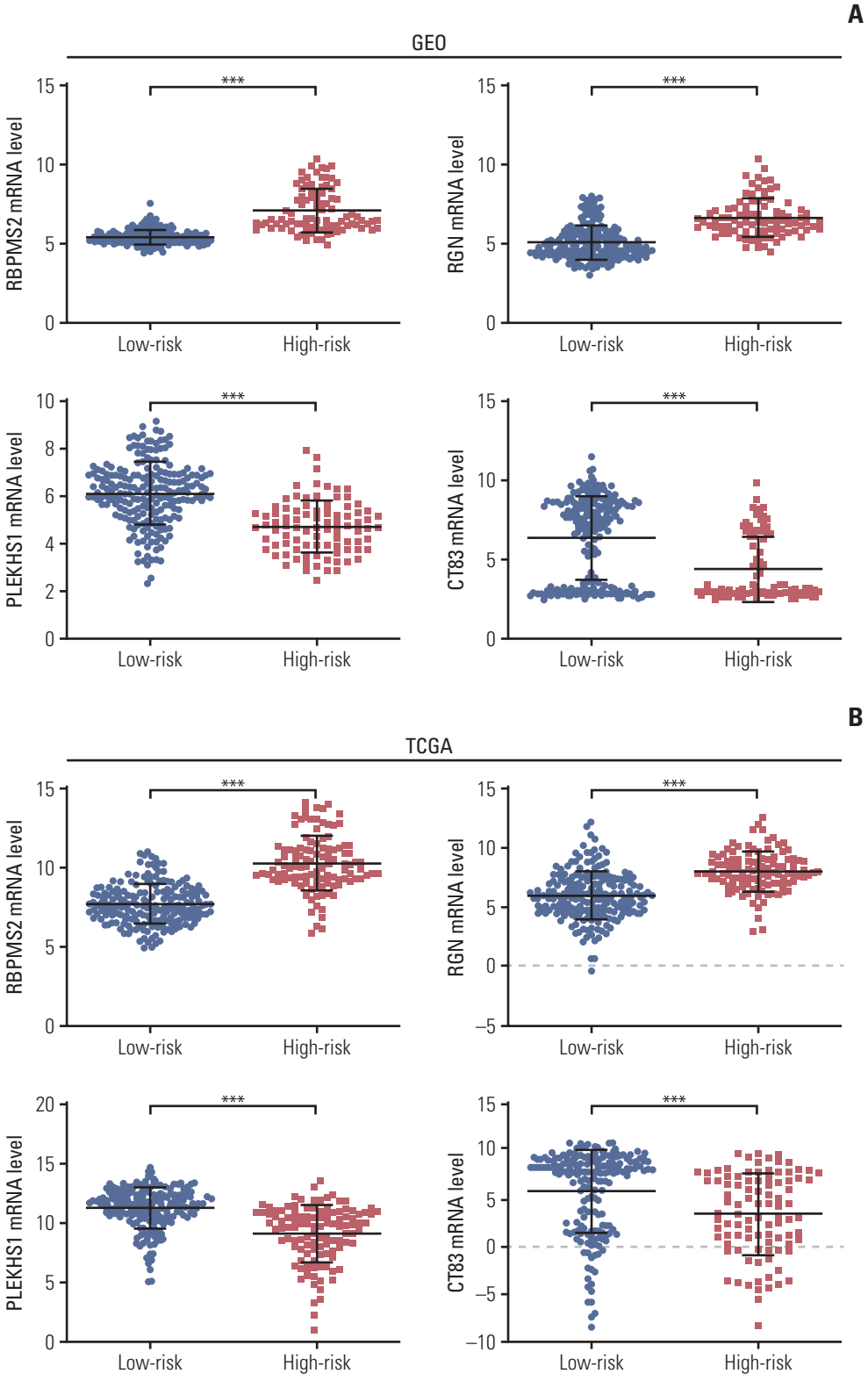

Fig. 4. The expression of the four prognostic genes in low-risk groups and high-risk groups of each cohort (*p < 0.01, ***p < 0.0001). The expression levels of the four genes in the Gene Expression Omnibus (GEO) cohort (A), in the The Cancer Genome Atlas (TCGA) global cohort (B), and in the TCGA Asian cohort (C). *(Continued to the next page)*

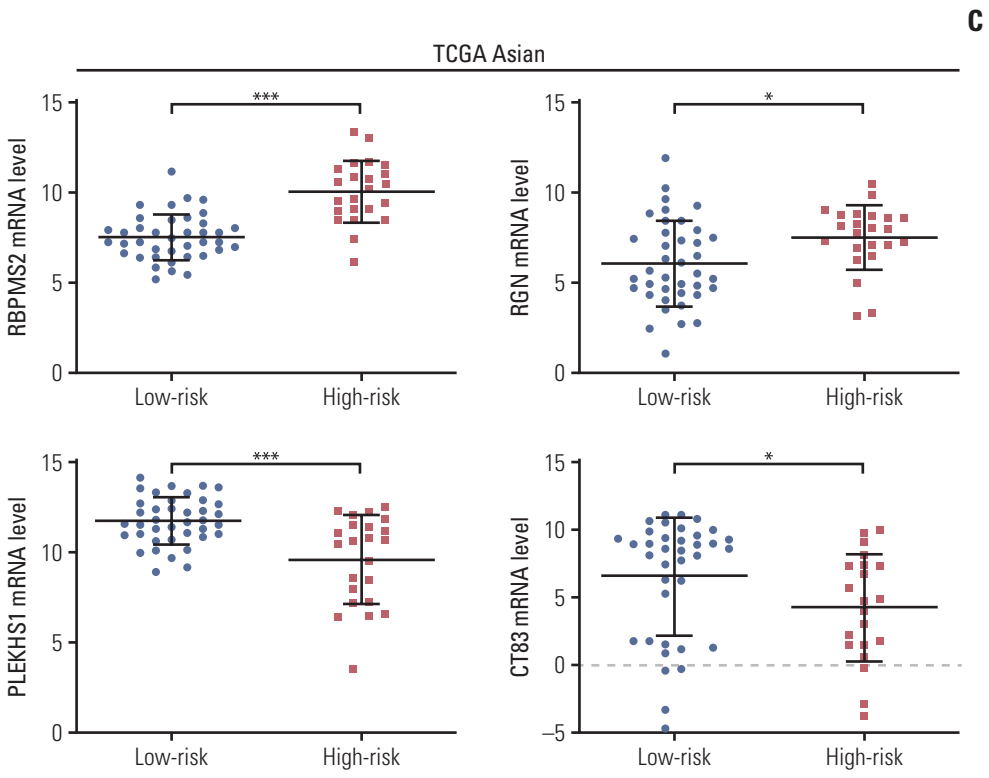

Fig. 4. *(Continued from the previous page)*

6A). The results of the ROC analysis show that the AUC of the nomogram was the largest (Fig. 6B).

Discussion

GC remains one of the most commonly diagnosed malignancies and leads to cancer death worldwide, especially in Asia [1]. There is tremendous pressure on developing techniques for the prevention and treatment of GC. Survival prediction affects the choices of further treatment options. Traditional survival prediction methods using pathological information and serum tumor marker levels as the main means have played an important role in clinical practice over the past few decades [16]. However, with the advent of the era of precision therapy, it is gradually becoming difficult for this predictive approach to meet clinical needs. GC has significant biological and epidemiological differences between Asian and non-Asian populations; it is difficult to find a one-size-fits-all approach for all patients. Therefore, in order to improve the efficacy of GC and reduce mortality, there is an urgent need for different survival prediction methods for patients of different races.

The GSE66229 dataset contains microarray profiles of gastric tumors from Asian patients. Previously, Cristescu et al. [8] used multi-omics data (including GSE66229) to classify GC into MSS/TP53+, MSS/TP53-, MSS/EMT, and MSI molecular types and described the molecular characteristics of GC at the genetic level. Moreover, Zhang et al. [17] used the GSE66229 dataset to investigate the significance of crosstalk between long non-coding RNA and mRNA in GC. In our study, we used the GEO dataset GSE66229 to select four genes, *RBPMS2*, *RGN*, *PLEKHS1*, and *CT83*, to establish a diagnostic model. The coefficient and hazard ratio showed that *RBPMS2* and *RGN* are high-risk factors in GC, while the overexpression of *PLEKHS1* and *CT83* signify a longer OS. The AUCs of the ROC curves were 0.684, 0.697, 0.699, 0.736, and 0.750 for 0.5-year, 1-year, 2-year, 3-year, and 5-year survival, respectively. This model performed well in survival prediction. In this model, patients can be assigned to low-risk or high-risk groups based on risk scores, providing a basis for further precision treatment. Diagnostic models can also be used to guide postoperative adjuvant chemotherapy. For patients in the high-risk group, a more potent combination chemotherapy regimen should be used to inhibit and destroy tumor cells. Moreover, in response to high-risk groups, a more detailed review strategy can be developed to enable early detection of recurrent cases. We also demon-

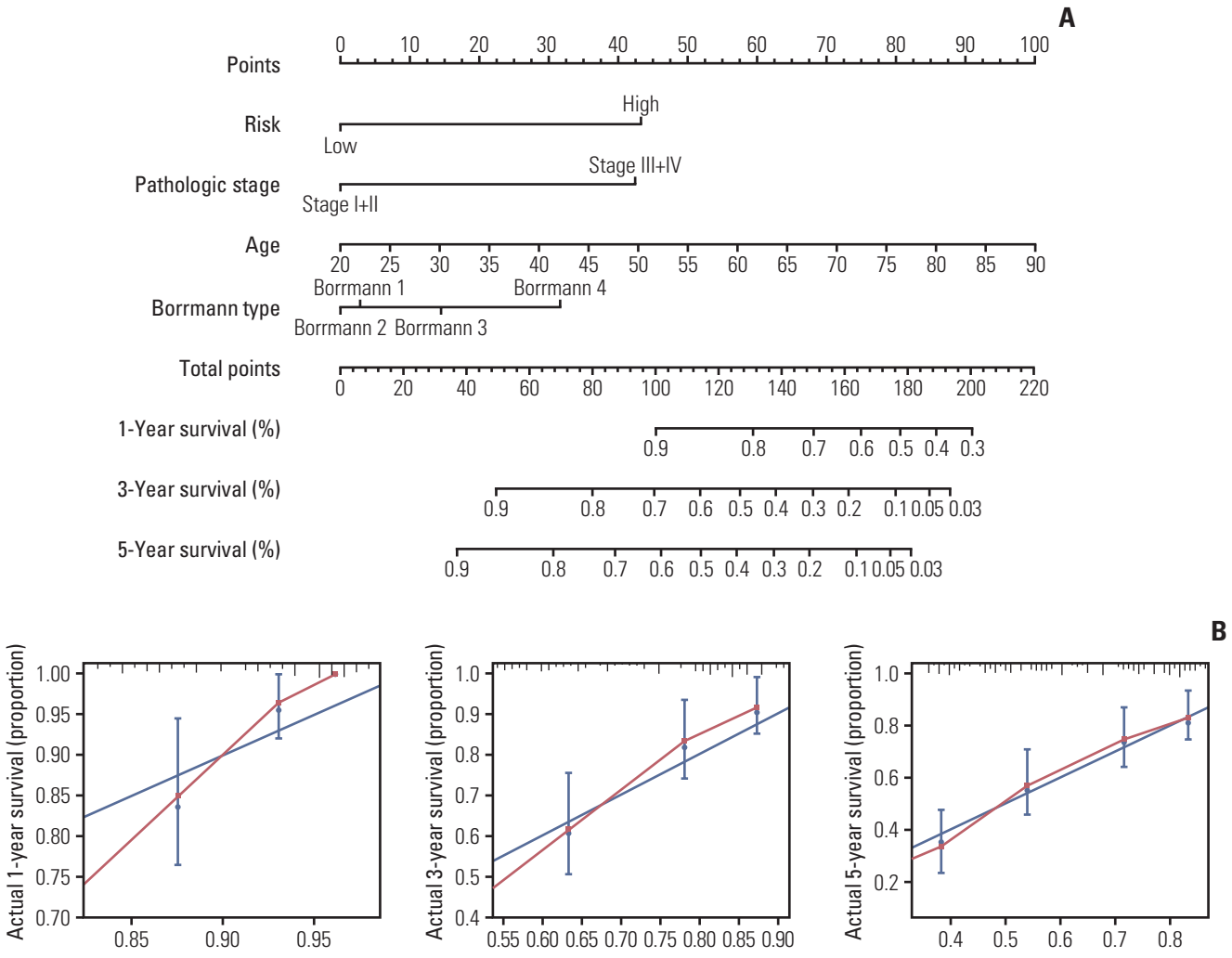

Fig. 5. The nomogram is used to predict 1-year, 3-year, and 5-year survival rates for Asian gastric cancer patients (A). The nomogram is applied by adding up the scores projected on the corresponding scale for each factor. The total number of scores project on the bottom scale represents the probability of 1-year, 3-year, and 5-year overall survival. (B) The calibration plots of the nomogram, the X-axis represents the survival rate predicted by the nomogram, and the Y-axis represents the actual survival rate calculated by Kaplan-Meier analysis. AUC, area under curves.

Table 1. C-index comparison between nomogram (combined model) and models constructed by single clinical characteristics

Model	C-index (95% CI)	p-value
Prognostic model	0.63 (0.59-0.67)	< 0.001
Pathologic stage model	0.64 (0.61-0.68)	< 0.001
Age model	0.55 (0.50-0.60)	0.534
Borrmann type model	0.63 (0.58-0.67)	< 0.001
Nomogram (Combined model)	0.74 (0.71-0.78)	< 0.001
Nomogram vs. Prognostic model	-	< 0.001
Nomogram vs. Pathologic stage model	-	< 0.001
Nomogram vs. Age model	-	< 0.001
Nomogram vs. Borrmann type model	-	0.002

CI, confidence interval.

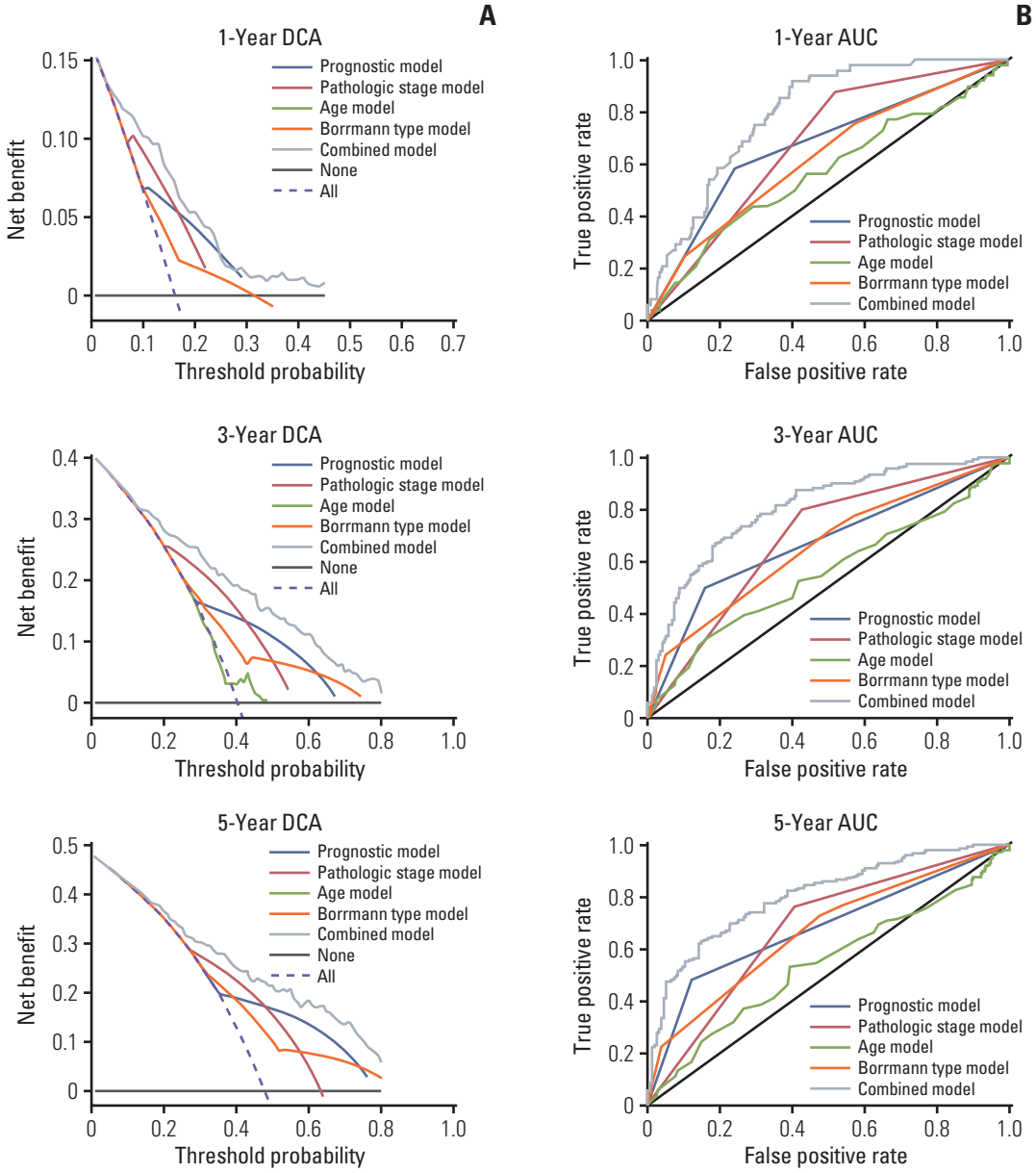

Fig. 6. Decision curve analysis (DCA) curves and the time-dependent receiver operating characteristic (ROC) curves for the nomogram. (A) DCA curve can visually evaluate the predictive power of the nomogram. The calculated net benefit (Y-axis) corresponds to the threshold probability of 1-year, 3-year, and 5-year survival rates on the X-axis. The solid gray line represents the probability that no patient will survive for 1 year, 3 years, or 5 years. The yellow dashed line represents the probability that all patients will live for 1 year, 3 years, and 5 years. The black, red, green, dark blue, light blue, and purple represent the nomograms. (B) The time-dependent ROC curve assesses the accuracy of the nomogram.

strated the independence of the prognostic model with other clinical data in GC.

Next, we validated the diagnostic model in the TCGA global cohort. Considering that GC has significant biological differences between different races [18], we further validated it in the TCGA Asian cohort and the TCGA non-Asian cohort. And the results were finally verified in the TCGA Asian cohort, demonstrating that the prognostic model based on the GEO dataset from Korea is capable of dividing Asian GC patients into a high-risk group or a low-risk group and predicting OS. However, the predictive power was limited in the non-Asian population. Similar ethnic differences

have also been found in epidemiological investigations in the United States. The incidence of GC in Japanese immigrants to Hawaii was higher than that of local residents in Hawaii [19]. Interethnic differences can, therefore, play an important role in the development and progression of GC.

The protein encoded by RBPMS2 is a member of the RNA recognition motif-containing protein family and plays an important role in regulating the development and dedifferentiation of digestive smooth muscle cells [20]. Previous research in chick embryos showed that RBPMS2 was expressed in the early stages of the visceral smooth muscle cell and gradually decreased with the maturity of smooth muscle cells. In differentiated primary cultured smooth muscle cells, ectopic expression of RBPMS2 upregulates cell proliferation rates and inhibits cell contractile function [21]. Aberrant elevated expression of RBPMS2 can be specifically observed in gastrointestinal mesenchymal neoplasm and digestive myopathy syndrome, demonstrating that the regulated expression of RBPMS2 is important for the proper development and differentiation of visceral smooth muscle cells [21,22] In our research, the expression of RBPMS2 in the high-risk group was significantly higher than that in the low-risk group. This finding suggests that cases in the high-risk group are more likely to invade the muscular layer and cause abnormal functioning of smooth muscle cells. RBPMS2 can be defined as a novel marker of visceral smooth muscle remodeling for characterizing smooth muscle layer invasion in gastrointestinal cancer.

RGN is a protein-coding gene. Previous studies on breast cancer and lung cancer have demonstrated that it may play a crucial role in maintaining intracellular Ca^{2+} homeostasis, suppressing cell proliferation, inhibiting oncogene expression, and increasing tumor suppressor gene expression [23,24]. Our research also found that RGN was downregulated in tumor tissues compared with adjacent nontumor tissues. However, in GC patients, RGN expression is higher in the high-risk group than in the low-risk group. To further validate the relationship between RGN expression levels and survival time, we validated the GEO data using the KmPlot website and found that the results regarding RGN expression in lung cancer and breast cancer are consistent with those reported in other works in the literature, but the results were reversed in GC. The RGN high-expression group has a shorter survival period, which coincides with our findings.

At present, there are few studies on PLEKHS1, and the prognostic value of PLEKHS1 in GC has not been verified in previous studies. Mutations in non-coding regions of PLEKHS1 were found in cancer patients according to a genome-wide analysis, which may be related to the degree of tumor malignancy [25]. Our study found that PLEKHS1 is a protective factor in patients with GC and that its expression is higher in the low-risk group, as in previous reports [26].

CT83, also known as KK-LC-1, is a gene that encodes a member of the cancer-testis antigenic protein family and is only expressed in malignant tumor tissue and testicular germ cells [27-29]. Based on this feature, Marcinkowski et al. [27] believe that CT83 may become an attractive target antigen for chimeric antigen receptor T-cell immunotherapy (CAR-T). In our results, the expression level of CT83 in the low-risk group was significantly higher than that in the high-risk group. We speculate that CT83 may be related to the body's anti-tumor response. In normal tissues, the expression of CT83 is at an extremely low level. When tumors occur, the expression level of CT83 may be related to the immune response of the body against tumors. The higher the expression levels of CT83, the stronger the body's ability to fight tumors, so the better the prognosis. Moreover, in the study of the early diagnosis of GC, Futawatari et al. [30] found that high CT83 expression rates can be frequently detected in the early stage of GC. Therefore, CT83 can be used as a potential marker for the early diagnosis and treatment of GC.

The current risk assessment for patients with GC is based primarily on the TNM staging system [14]. This anatomical-based approach has played an important role in clinical practice over the past few decades, but it is still difficult to explain why patients with the same staging receive different prognoses. With the advent of the era of precision therapy, doctors need more accurate methods for patient risk analysis. Our nomogram combines genetic information with GC clinical data, to make the prediction ability 15.6% higher than that only focus on the pathologic stage (C-index, 0.74 vs. 0.64; $p < 0.001$). Another advantage of our predictive model is that it only needs to detect the expression levels of the four genes rather than somatic mutations in patients, which can be achieved by some simple and economical methods, greatly reducing the cost of sequencing.

However, there are still some limitations to this research. Our nomograms were not externally validated in the TCGA database because of the lack of data on Bormann type. In the future, we will detect the expression levels of the four genes in more clinical samples to verify and improve this prediction model.

To sum up, our four-gene–related prognostic model and the nomogram are reliable tools for predicting the OS of Asian patients. However, the predictive power is limited in the non-Asian population. Our nomogram will help stratify Asian patients with GC more accurately and promote more precise treatment for all of them.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (https://www.e-crt.org).

Ethical Statement

All the obtained data were used according to the GEO and TCGA data access policies, as well as publication guidelines. Both mRNA profile data and clinical information are publicly available and open-access. This study does not involve animal studies. Therefore, the study does not need to be approved by the local ethics committee.

Author Contributions

Conceived and designed the analysis: Ji JF, Lyu G.
Collected the data: Zhou K, Chen J.
Contributed data or analysis tools: Bu Z, Wang A.
Performed the analysis: Ji J, Chen J.
Wrote the paper: Chen J, Wang A.

Conflicts of Interest

Conflicts of interest relevant to this article was not reported.

Acknowledgments

This work was supported by the National Key Technology Support Program (No. 2014BA109B02), the Beijing Municipal Science and Technology Project (No. D131100005313010), the National Science Foundation for Young Scientists of China (81802735), and the grants from 'San Ming' Project of Shenzhen city, China (No. SZSM201612051).

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394-424.
2. National Health Commission of The People's Republic of China. Chinese guidelines for diagnosis and treatment of gastric cancer 2018 (English version). Chin J Cancer Res. 2019;31:707-37.
3. Chen X, Liu H, Li G, Yu J. Implications of clinical research on adjuvant chemotherapy for gastric cancer: where to go next? Chin J Cancer Res. 2019;31:892-900.
4. Lin SJ, Gagnon-Bartsch JA, Tan IB, Earle S, Ruff L, Pettinger K, et al. Signatures of tumour immunity distinguish Asian and non-Asian gastric adenocarcinomas. Gut. 2015;64:1721-31.
5. Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, et al. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. J Cell Mol Med. 2018;22:5928-38.
6. Hou JY, Wang YG, Ma SJ, Yang BY, Li QP. Identification of a prognostic 5-Gene expression signature for gastric cancer. J Cancer Res Clin Oncol. 2017;143:619-29.
7. Cheong JH, Yang HK, Kim H, Kim WH, Kim YW, Kook MC, et al. Predictive test for chemotherapy response in resectable gastric cancer: a multi-cohort, retrospective analysis. Lancet Oncol. 2018;19:629-38.
8. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med. 2015;21:449-56.
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
10. Long J, Wang A, Bai Y, Lin J, Yang X, Wang D, et al. Development and validation of a TP53-associated immune prognostic model for hepatocellular carcinoma. EBioMedicine. 2019;42:363-74.
11. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res. 2004;10:7252-9.
12. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol. 2015;16:e173-80.
13. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26:1364-70.
14. Chen S, Chen X, Nie R, Ou Yang L, Liu A, Li Y, et al. A nomogram to predict prognosis for gastric cancer with peritoneal dissemination. Chin J Cancer Res. 2018;30:449-59.
15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74.
16. Sano T, Coit DG, Kim HH, Roviello F, Kassab P, Wittekind C, et al. Proposal of a new stage grouping of gastric cancer for TNM classification: International Gastric Cancer Association staging project. Gastric Cancer. 2017;20:217-25.
17. Zhang J, Yuan Y, Wei Z, Ren J, Hou X, Yang D, et al. Crosstalk between prognostic long noncoding RNAs and messenger RNAs as transcriptional hallmarks in gastric cancer. Epigenomics. 2018;10:433-43.
18. Jin H, Pinheiro PS, Callahan KE, Altekruse SF. Examining the gastric cancer survival gap between Asians and whites in the United States. Gastric Cancer. 2017;20:573-82.
19. Kolonel LN, Hankin JH, Nomura AM. Multiethnic studies of diet, nutrition, and cancer in Hawaii. Princess Takamatsu Symp. 1985;16:29-40.
20. Sagnol S, Yang Y, Bessin Y, Allemand F, Hapkova I, Notarnicola C, et al. Homodimerization of RBPMS2 through a new RRM-interaction motif is necessary to control smooth muscle plasticity. Nucleic Acids Res. 2014;42:10173-84.
21. Notarnicola C, Rouleau C, Le Guen L, Virsolvy A, Richard S,

Faure S, et al. The RNA-binding protein RBPMS2 regulates development of gastrointestinal smooth muscle. Gastroenterology. 2012;143:687-97.
22. Hapkova I, Skarda J, Rouleau C, Thys A, Notarnicola C, Janikova M, et al. High expression of the RNA-binding protein RBPMS2 in gastrointestinal stromal tumors. Exp Mol Pathol. 2013;94:314-21.
23. Yamaguchi M, Osuka S, Shoji M, Weitzmann MN, Murata T. Survival of lung cancer patients is prolonged with higher regucalcin gene expression: suppressed proliferation of lung adenocarcinoma A549 cells in vitro. Mol Cell Biochem. 2017;430:37-46.
24. Yamaguchi M, Osuka S, Weitzmann MN, Shoji M, Murata T. Increased regucalcin gene expression extends survival in breast cancer patients: overexpression of regucalcin suppresses the proliferation and metastatic bone activity in MDA-MB-231 human breast cancer cells in vitro. Int J Oncol. 2016;49:812-22.
25. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet. 2014;46:1160-5.
26. Liu X, Wu J, Zhang D, Bing Z, Tian J, Ni M, et al. Identification of potential key genes associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. Front Genet. 2018;9:265.
27. Marcinkowski B, Stevanovic S, Helman SR, Norberg SM, Serna C, Jin B, et al. Cancer targeting by TCR gene-engineered T cells directed against Kita-Kyushu Lung Cancer Antigen-1. J Immunother Cancer. 2019;7:229.
28. Paret C, Simon P, Vormbrock K, Bender C, Kolsch A, Breitkreuz A, et al. CXorf61 is a target for T cell based immunotherapy of triple-negative breast cancer. Oncotarget. 2015;6:25356-67.
29. Shigematsu Y, Hanagiri T, Shiota H, Kuroda K, Baba T, Mizukami M, et al. Clinical significance of cancer/testis antigens expression in patients with non-small cell lung cancer. Lung Cancer. 2010;68:105-10.
30. Futawatari N, Fukuyama T, Yamamura R, Shida A, Takahashi Y, Nishi Y, et al. Early gastric cancer frequently has high expression of KK-LC-1, a cancer-testis antigen. World J Gastroenterol. 2017;23:8200-6.