

Original Article



OPEN ACCESS

Received: Jan 29, 2022

Revised: Mar 17, 2022

Accepted: Mar 17, 2022

Published online: Apr 26, 2022

Correspondence to

Hui Xu

Department of Oncology, the First Affiliated Hospital of Anhui Medical University; Anhui Provincial Cancer Institute/Anhui Provincial Office for Cancer Prevention and Control, No. 218 Jixi Road, Hefei 230022, Anhui Province, People's Republic of China.
Email: xuhui52088@163.com

Tai Ma

Department of Oncology, the First Affiliated Hospital of Anhui Medical University, No. 218 Jixi Road, Hefei 230022, Anhui Province, People's Republic of China.
Email: matai@ahmu.edu.cn

*Cheng Zhang and Minmin Xie contributed equally to this study.

Copyright © 2022. Korean Gastric Cancer Association

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ORCID iDs

Cheng Zhang 
<https://orcid.org/0000-0001-7665-8675>

Determination of Survival of Gastric Cancer Patients With Distant Lymph Node Metastasis Using Prealbumin Level and Prothrombin Time: Contour Plots Based on Random Survival Forest Algorithm on High-Dimensionality Clinical and Laboratory Datasets

Cheng Zhang ^{1,2,*}, Minmin Xie ^{1,*}, Yi Zhang ¹, Xiaopeng Zhang ³,
Chong Feng ³, Zhijun Wu ⁴, Ying Feng ¹, Yahui Yang ¹, Hui Xu ^{1,2}, Tai Ma ¹

¹Department of Oncology, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui Province, People's Republic of China

²Anhui Provincial Cancer Institute/Anhui Provincial Office for Cancer Prevention and Control, Hefei, People's Republic of China

³Department of Noncommunicable Diseases and Health Education, Hefei Center for Disease Prevention and Control, Hefei, People's Republic of China

⁴Department of Oncology, Ma'anshan Municipal People's Hospital, Ma'anshan, People's Republic of China





ABSTRACT

Purpose: This study aimed to identify prognostic factors for patients with distant lymph node-involved gastric cancer (GC) using a machine learning algorithm, a method that offers considerable advantages and new prospects for high-dimensional biomedical data exploration.

Materials and Methods: This study employed 79 features of clinical pathology, laboratory tests, and therapeutic details from 289 GC patients whose distant lymphadenopathy was presented as the first episode of recurrence or metastasis. Outcomes were measured as any-cause death events and survival months after distant lymph node metastasis. A prediction model was built based on possible outcome predictors using a random survival forest algorithm and confirmed by 5×5 nested cross-validation. The effects of single variables were interpreted using partial dependence plots. A contour plot was used to visually represent survival prediction based on 2 predictive features.

Results: The median survival time of patients with GC with distant nodal metastasis was 9.2 months. The optimal model incorporated the prealbumin level and the prothrombin time (PT), and yielded a prediction error of 0.353. The inclusion of other variables resulted in poorer model performance. Patients with higher serum prealbumin levels or shorter PTs had a significantly better prognosis. The predicted one-year survival rate was stratified and illustrated as a contour plot based on the combined effect the prealbumin level and the PT.

Conclusions: Machine learning is useful for identifying the important determinants of cancer survival using high-dimensional datasets. The prealbumin level and the PT on distant lymph node metastasis are the 2 most crucial factors in predicting the subsequent survival time of advanced GC.

Minmin Xie <https://orcid.org/0000-0002-4542-1541>Yi Zhang <https://orcid.org/0000-0003-1940-7081>Xiaopeng Zhang <https://orcid.org/0000-0002-0841-3956>Chong Feng <https://orcid.org/0000-0002-6645-7703>Zhijun Wu <https://orcid.org/0000-0003-3086-3889>Ying Feng <https://orcid.org/0000-0002-8238-5826>Yahui Yang <https://orcid.org/0000-0002-1878-5032>Hui Xu <https://orcid.org/0000-0002-0421-8125>Tai Ma <https://orcid.org/0000-0003-0911-850X>

Trial Registration

ChiCTR Identifier: [ChiCTR1800019978](https://www.chictr.org/ChiCTR1800019978)

Funding

This study was supported by the Anhui Provincial Key Research and Development Program (grant number 1804b06020351) and the First Affiliated Hospital of the Anhui Medical University Clinical Research Project (grant number LCYJ2021YB015).

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Author Contributions

Conceptualization: X.H., M.T.; Data curation: X.M., Z.Y., Z.X., F.C., W.Z., F.Y., Y.Y.; Formal analysis: Z.C., X.M.; Funding acquisition: X.H., M.T.; Investigation: X.M., Z.Y., Z.X., F.C., W.Z., F.Y., Y.Y.; Methodology: X.H., M.T.; Project administration: X.H., M.T.; Resources: X.M., Z.Y., Z.X., F.C., W.Z., F.Y., Y.Y.; Supervision: X.H., M.T.; Validation: X.M., Z.Y., Z.X., F.C., W.Z., F.Y., Y.Y.; Writing - original draft: Z.C., X.M.; Writing - review & editing: X.H., M.T., X.M., Z.Y., Z.X., F.C., W.Z., F.Y., Y.Y.

Trial Registration: ChiCTR Identifier: [ChiCTR1800019978](https://www.chictr.org/ChiCTR1800019978)**Keywords:** Stomach neoplasms; Lymphatic metastasis; Survival analysis; Supervised machine learning

INTRODUCTION

Patients with recurrent or metastatic gastric cancer (GC) have short life expectancies [1]. The tumor, node and metastasis staging system classifies patients with distant metastasis into the M1 subgroup without considering the impact of different metastatic sites [2]. In fact, different patterns of recurrence or metastasis typically imply different survival outcomes [3-5]. Distant lymph nodes are common sites of metastasis in advanced GC [6-9]; however, GC with nodal recurrence has a short survival time similar to a hematogenous relapse [5], which is different from the nature of other late-stage cancers (e.g., breast cancer with distant nodal metastasis has a better prognosis than hematogenous metastasis). More importantly, the risk factors contributing to the short survival time of this subgroup have not been identified.

Although use of gene signatures and molecular profiles has been suggested for survival prediction of GC [10-12], in view of their application in clinical practice, it is easier and more applicable to use universal laboratory analytes and clinical features as potential predictive markers. It has been demonstrated that peripheral blood parameters could be independent factors in predicting the survival of stage IV GC [13]; however, the selection of the candidate predictors is arbitrary or empirical and the inner links between the features are omitted. Therefore, it is uncertain whether the chosen independent variables can achieve the optimal power of prediction. Furthermore, conventional statistical analytical methods, such as traditional logistic regression models and Cox proportional hazard (PH) models, perform poorly when dealing with datasets that contain numerous noisy variables [14-16]. To overcome these problems, machine learning techniques have been developed and introduced in GC research, yielding satisfactory results compared to traditional methods [17-19].

To utilize the high volume of medical records for survival prediction of GC with distant node recurrence or metastasis, we collected demographic, pathological, therapeutic, and laboratory variables to create a high-dimensional dataset of advanced GC. The laboratory information contained 58 common analytes, including routine blood, liver function, kidney function, nutrition status, electrolyte, coagulation function, and tumor biomarker tests. The random survival forest (RSF) algorithm, a well-known machine learning technique, was employed to select important features and train the prediction model on patients with GC with synchronous or metachronous distant node metastasis.

MATERIALS AND METHODS

Patient enrollment and follow-up

The cohort was derived from a registered hospital-based ambispective cohort study on consecutive patients with gastric and esophagogastric junction carcinoma admitted at the First Affiliated Hospital of Anhui Medical University between January 2010 and December 2019 (ChiCTR1800019978, <http://www.chictr.org.cn/>). Individuals with primary metastatic GC (stage IV disease) and initial nonstaged IV disease who developed distant recurrence

after gastrectomy were eligible. To build a machine learning model, we selected patients whose metastatic sites included distant lymph nodes, because the number of patients in this subgroup was the largest in our cohort. **Supplementary Table 1** presents the mean post-metastasis survival of the patients in our cohort at each metastatic site. The diagnosis of distant lymphatic metastasis was confirmed by cytopathological, histopathological, and radiological examination. Patients with multiple primary malignant tumors were excluded from the study. All procedures performed in this study involving human participants were in accordance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards. This study was approved by the Ethics Committee of the First Affiliated Hospital of Anhui Medical University (reference number: Quick-PJ 2021-05-19). The requirement of informed consent was waived owing to the retrospective nature of this study. The workflow of this study is illustrated in **Fig. 1**.

The date of identification of distant lymph node metastasis was the starting point of the observation. The endpoint was death for any reason. The death date was acquired from the provincial population death register system or during a telephonic follow-up of the patients or their relatives conducted every 3 months. Survival time was defined as the interval (in months) between the start point and the death event or the last follow-up.

Data collection

To build a high-dimensional dataset of advanced GC with distant lymph node metastasis, we extracted demographic, clinical, and laboratory information from electronic medical records and combined these variables (features) with survival outcomes. Briefly, 20 demographic features, 28 clinical features, 76 laboratory features, the survival status, and the survival time were recorded. The laboratory variables were measured within one month before and after the identification of distant lymph node metastasis. Owing to patient heterogeneity and missing values at random, not all features contained sufficient valid values. To deal with this problem, we excluded variables in which over 30% observations were randomly missed and we recoded some variables by summarizing and integrating information. Variables that were clearly unrelated to survival after metastasis, based on existing knowledge, were also excluded. Finally, 81 variables were included to establish the model: survival status, survival time, 10 demographic features, 11 clinical features, and 58 laboratory features (**Supplementary Table 2**).

Model training and optimization

The model was built based on the RSF algorithm, a random forest method for the analysis of right-censored survival data [20]. Briefly, in this technique, bootstrap samples were drawn from the original data and a survival tree was grown for each bootstrap sample. Each bootstrap sample excluded on average 37% data, which are called out-of-bag (OOB) data. The cumulative hazard function (CHF) was calculated for each tree. A tree was grown starting at the root node and split at a node that maximized the survival according to the log-rank splitting rule. As the number of nodes increased and dissimilar cases became separated, each node in the tree became homogeneous and was populated by cases with similar survival times. The ensemble CHF was constructed by averaging all trees. The prediction error for the ensemble CHF was obtained using the OOB data [20].

The model was optimized and evaluated by nested cross-validation (nCV). Briefly, the data were split into five outer-loop folds, in which a single fold was selected as the test set, whereas the remaining four folds were merged and split into five inner-loop folds comprising

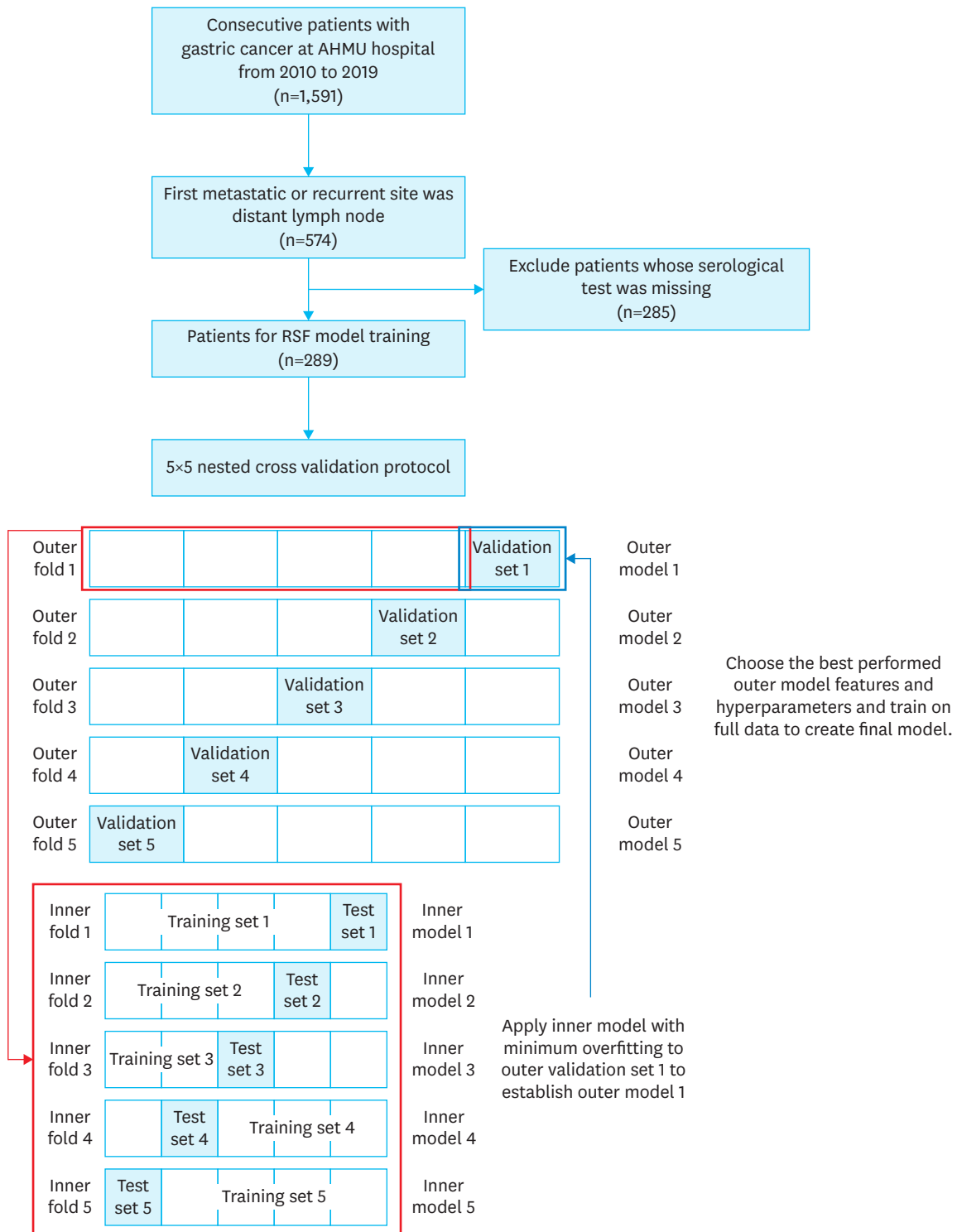


Fig. 1. Flow diagram of patient selection and model building. AHMU = Anhui Medical University; RSF = random survival forest.

an inner training set (4-folds) and a validation set (1-fold). In each inner model, data imputation, feature selection, and hyperparameter tuning were conducted on the training set and tested on the validation set. The inner model with the best performance was evaluated on the outer test set. Data imputation was based on the random forest algorithm [20]. A random search technique with 100 iterations was employed to select features and tune 3 important hyperparameters: *ntree* (the number of individual trees in the forest), *mtry* (the number of features to randomly sample at each node), and *nodesize* (the minimum number of cases allowed in a leaf). The hyperparameters and features of the best outer model were selected to train the entire dataset and create the final model. The protocol of the 5×5 nCV is illustrated in **Fig. 1**. Model performance was measured using Harrell's concordance index (C-index), which was calculated as a 1-prediction error. A value of 0.5 implies random guessing, and a value of 1 is a perfect prediction. The sensitivity, specificity, and area under receiver operating characteristic (ROC) curve at 12 months were also determined. This process was implemented using R packages "randomForestSRC" [20], "mlr3" [21], and "ranger" [22].

Model interpretation

The final RSF model was interpreted using variable importance and partial dependence plots [23]. A variable was considered important if breaking the relationship between it and survival resulted in an increased prediction error and a degraded model performance. In brief, the ten most important variables were sequentially displayed in a bar chart. A partial dependence plot was used to illustrate the partial effect of the important variables on the predicted 12-month survival probability. A contour plot was drawn to interpret a two-variable partial effect, which contained three-dimensional information (variable A, variable B, and predicted survival probability). This process was implemented using R package "randomForestSRC" [20].

Other statistical analysis

The median, 25th percentile, and 75th percentile were used to describe the various variables, and density plots were used to show their distribution. Dot plots with Spearman's correlation tests were employed to identify the possible association between two variables. Numbers and percentages were used to describe categorical variables. Survival curves were drawn using the Kaplan–Meier (KM) method and compared using the log-rank method. The PH assumption was tested using a graph of the scaled Schoenfeld residuals drawn using R package "survminer." Cox regression with restricted cubic spline (RCS) analysis was performed to explore the nonlinear associations between the hazard ratio (HR) and numerous variables employing R package "rms." RStudio 1.4.1717 was used, and a two-sided P-value less than 0.05 was considered statistically significant.

RESULTS

Characteristics of patients

A total of 574 patients with advanced GC whose first metastatic site was a distant lymph node were enrolled. By the last follow-up (December 2020), 547 patients (95.3%) had reached the endpoint. The median follow-up time was 10.1 months. The median survival time was 10.1 months. Detailed baseline characteristics are listed in **Table 1**. After excluding the patients with massive missing values of clinical and laboratory features, 289 patients were finally included in the model building. In this population, 275 patients (95.2%) reached the endpoint. The median follow-up time was 9.7 months. The median survival time was 9.2 months (**Fig. 2A**). The median survival of metachronous cases was identical to that of synchronous cases (9.1 vs.

Table 1. Characteristics of included and excluded patients

Variables	Cohort inside model (n=289)	Cohort outside model (n=285)	P-value	Total cohort (n=574)
Median age at metastasis (yr)	62 (53–68)	63 (55–69)	0.189	62 (53–69)
Sex			0.261	
Male	205 (0.71)	215 (0.75)		420 (0.73)
Female	84 (0.29)	70 (0.25)		154 (0.27)
Tumor location			0.830	
Cardia	148 (0.51)	143 (0.50)		291 (0.51)
Body	54 (0.19)	59 (0.21)		113 (0.20)
Pylorus	87 (0.30)	83 (0.29)		170 (0.30)
Histological subtype			0.585	
Adenocarcinoma, NOS	247 (0.85)	235 (0.82)		482 (0.84)
Mucinous	27 (0.09)	34 (0.12)		61 (0.11)
Signet-ring cell	15 (0.05)	14 (0.05)		29 (0.05)
Tumor grade			0.018	
G1–G2	37 (0.13)	56 (0.20)		93 (0.16)
G3–G4	176 (0.61)	175 (0.61)		351 (0.61)
Gx	75 (0.26)	52 (0.18)		127 (0.22)
NA	1 (<0.01)	2 (0.01)		3 (0.01)
Tumor stage			0.008	
I	6 (0.02)	9 (0.03)		15 (0.03)
II	22 (0.08)	37 (0.13)		59 (0.10)
III	103 (0.36)	120 (0.42)		223 (0.39)
IV	150 (0.52)	110 (0.39)		260 (0.45)
NA	8 (0.03)	9 (0.03)		17 (0.03)
Gastrectomy			0.002	
Yes	150 (0.52)	185 (0.65)		335 (0.58)
No	139 (0.48)	100 (0.35)		239 (0.42)
Adjuvant chemotherapy			0.002	
Yes	92 (0.32)	127 (0.45)		219 (0.38)
No	197 (0.68)	158 (0.55)		355 (0.62)
Palliative chemotherapy			0.415	
Yes	256 (0.89)	245 (0.86)		501 (0.87)
No	33 (0.11)	40 (0.14)		73 (0.13)
Radiotherapy for nodes			0.666	
Yes	12 (4.15)	15 (5.26)		27 (4.70)
No	277 (95.85)	270 (94.74)		547 (95.30)
Other distant metastases			0.114	
Yes	161 (0.56)	139 (0.49)		300 (0.52)
No	128 (0.44)	146 (0.51)		274 (0.48)
Median follow-up (mon)	9.7 (4.7–17.9)	10.7 (5.2–19.4)	0.084	10.1 (5.0–18.3)
Median survival (mon)	9.2 (8.0–10.7)	10.7 (9.8–12.6)	0.143	10.1 (9.1–11.2)
One-year survival rate (%)	38.7 (33.5–44.8)	45.0 (39.6–51.2)	0.185	42.0 (38.2–46.3)

Values are presented as number (%) or median (range).
 NOS = not otherwise specified; NA = not available.

9.2 months, $P_{\log\text{-rank}}=0.449$). The baseline characteristics and median survival of the populations inside and outside the model were comparable (**Table 1, Fig. 2B**). The only exception was that more individuals with high-stage cancer were in the model cohort, which might be attributed to the higher probability of availability of serological tests in metastasis in this population. Therefore, the participants in the model cohort were less likely to undergo surgery and adjuvant chemotherapy.

Building model using RSF algorithm

Based on the nCV results, the optimal hyperparameters were $n\text{tree}=500$, $n\text{odesize}=4$, and $m\text{try}=29$. **Fig. 2C** shows that the OOB error of the model stabilizes as the number of trees increases to 500. The ten most important features are shown in **Fig. 1D**. Prealbumin and the prothrombin time (PT) were the most crucial and second most crucial survival predictors,

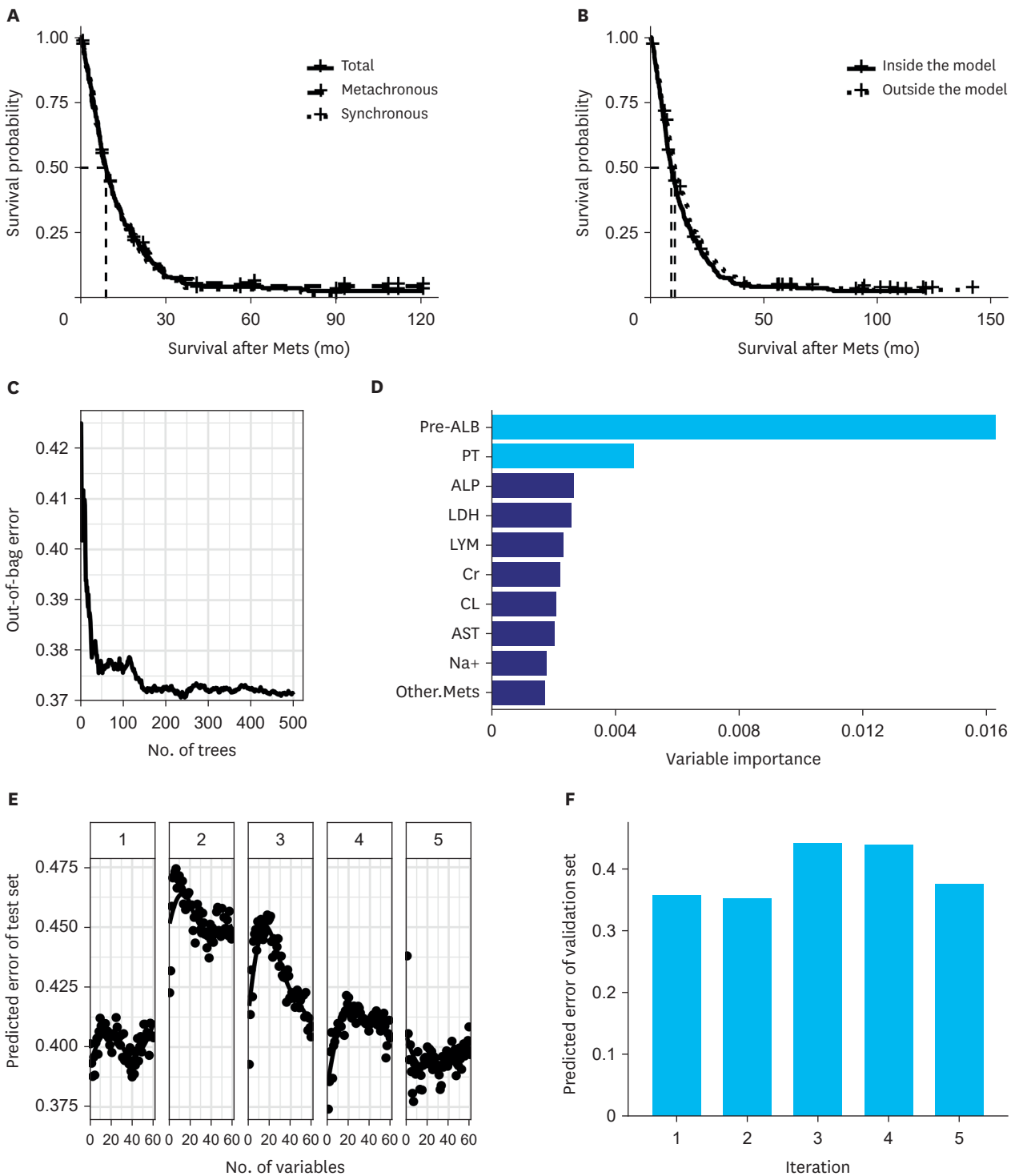


Fig. 2. Process of RSF model building by 5 × 5 nested cross validation. (A) Survival curves of metachronous and synchronous gastric cancer. (B) Survival curves of population inside and outside RSF model. (C) Process of tuning number of trees. Out-of-bag error stabilizes and declines as number of trees increases to 500. (D) Bar chart of 10 most important features in model. Pre-ALB and PT contribute most to reduce prediction error and are only 2 features included in best performance model by feature selection. (E) Relationship between number of included variables and prediction error on test set. In each iteration, stepwise inclusion of most important variables in model based on training set varies model performance. (F) Prediction error on validation set. In each iteration, best performing model with lowest prediction error on test set is further confirmed on validation set. Model with lowest prediction error on validation set is optimal (second iteration model).

Mets = metastasis; Pre-ALB = prealbumin; PT = prothrombin time; ALP = alkaline phosphatase; LDH = lactate dehydrogenase; LYM = lymphocyte count; Cr = creatinine; CL = chloridion; AST = aspartate transaminase.

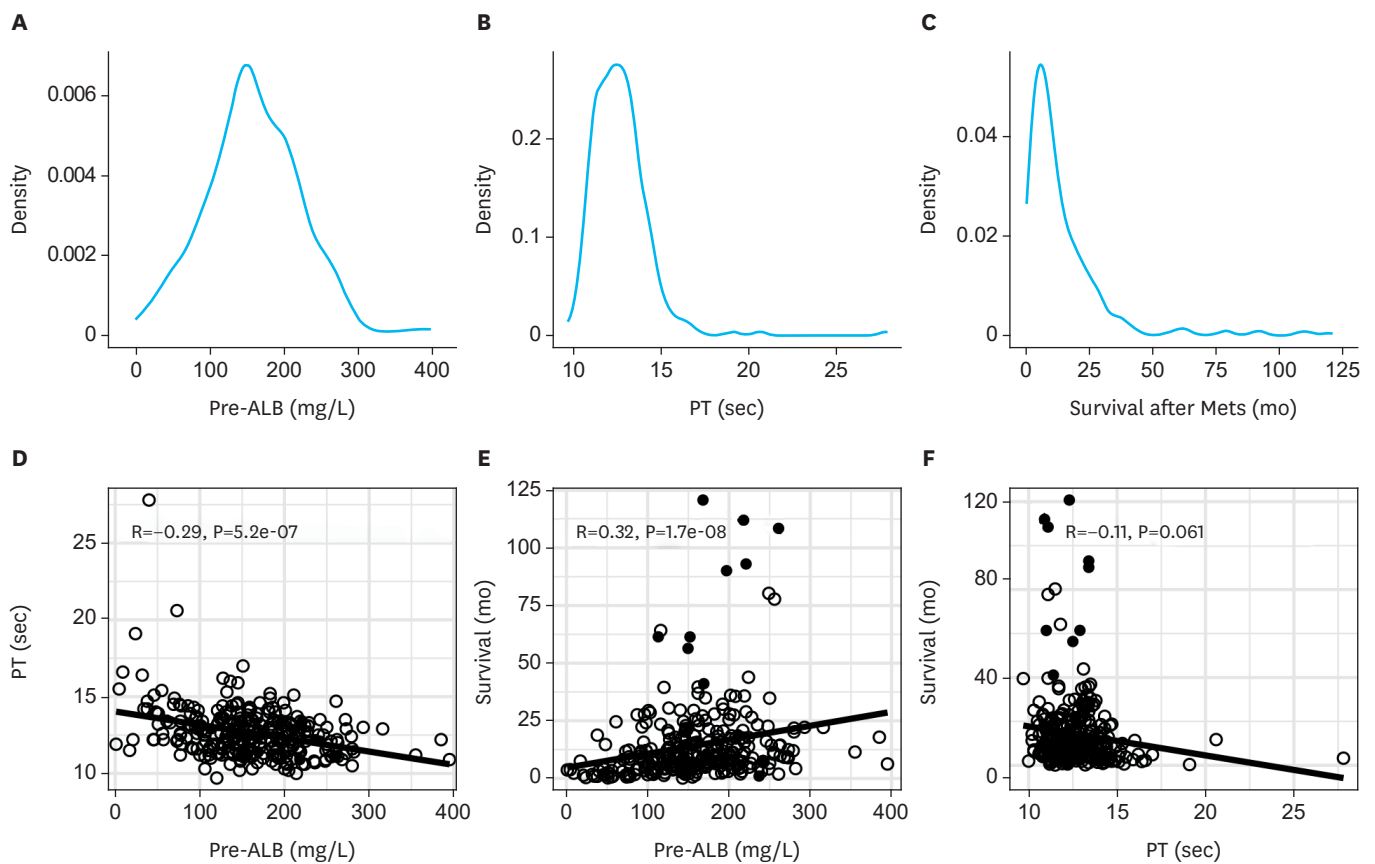


Fig. 3. Distributions and correlation analysis of Pre-ALB, PT, and survival time after metastasis. Density plots show distributions of (A) Pre-ALB, (B) PT, and (C) survival time. (D-F) Dot plots with linear correlation of Pre-ALB, PT, and survival time. Solid point denotes censored or alive individual; hollow point represents deceased individual. Pre-ALB = prealbumin; PT = prothrombin time; Mets = metastasis.

respectively. The other features contributed little, thereby not remarkably improving the model. **Fig. 2E** shows that stepwise inclusion of the important variables changes the prediction error. The best performance was the result of a second iteration (**Fig. 2F**), yielding a prediction error of 0.353 (C-index of 0.647) when the two most important features (prealbumin and PT) were included (**Fig. 2E**). The area under ROC curve, sensitivity, and specificity at 12 months were 0.613, 0.229, and 0.653, respectively. Finally, with the optimal hyperparameters, the prealbumin level and the PT on metastasis were chosen to train the entire dataset and build the final RSF model.

Correlation between important features and survival time

Because the prealbumin level and the PT were identified as the most crucial predictors, we further explored their inter-correlation. As shown in **Fig. 3A**, the distributions of the PT and the survival time are left-skewed. The prealbumin level is normally distributed, and it is negatively correlated with the PT ($R = -0.29$) and positively correlated with the survival time ($R = 0.32$). The linear association between the PT and the survival is not statistically significant (**Fig. 3B**).

Nonlinear association of prealbumin and PT with survival

The dot plots (**Fig. 3B**) clearly show that the relations of the prealbumin level and the PT with the survival time are nonlinear; therefore, we subsequently investigated these nonlinear associations using partial dependence plots. The predicted one-year survival increased

sharply as the prealbumin level increased, and it peaked when the prealbumin level was approximately 150–170 mg/L. Subsequently, the predicted survival started to gradually decline (Fig. 4A). The predicted survival maximized when the PT was 12–13 seconds, and decreased substantially when the PT rose from 14 to 15 seconds. After the PT became longer than 15 seconds, the predicted survival gradually declined (Fig. 4A).

To confirm the nonlinear association, the PH assumptions of the prealbumin level and the PT were tested (Supplementary Fig. 1) and the Cox regression with RCS analysis was performed

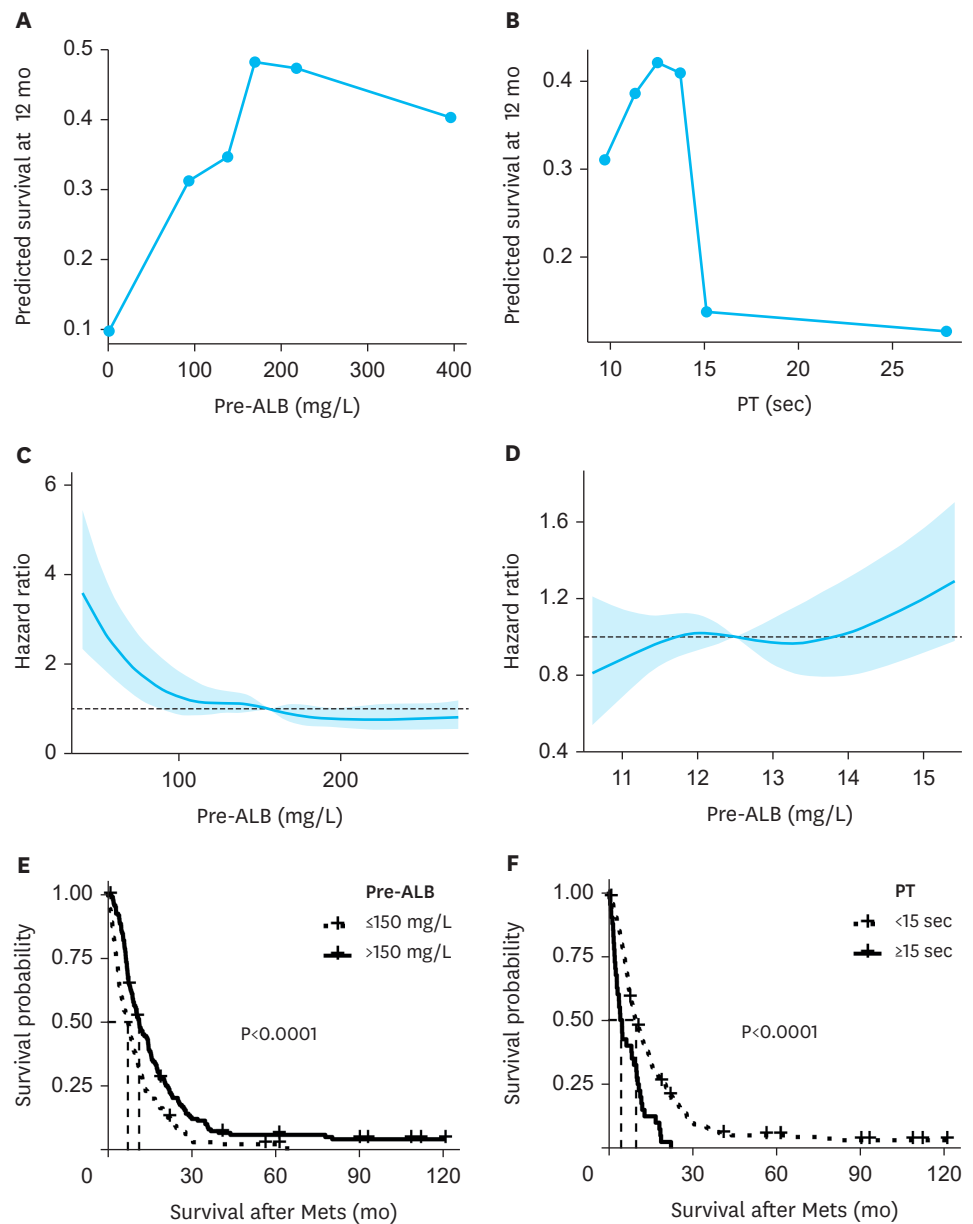


Fig. 4. Nonlinear association of prealbumin and PT with survival after metastasis. Partial dependent plot shows predicted one-year survival against prealbumin (A) and PT (B) level on metastasis. Nonlinear association of hazard ratio with prealbumin (C) and PT (D) level on metastasis. Reference (dot) line represents hazard ratio of 1. Patients with (E) high prealbumin or (F) shorter PT on metastasis have significantly better post-metastasis survival. Pre-ALB = prealbumin; PT = prothrombin time; Mets = metastasis.

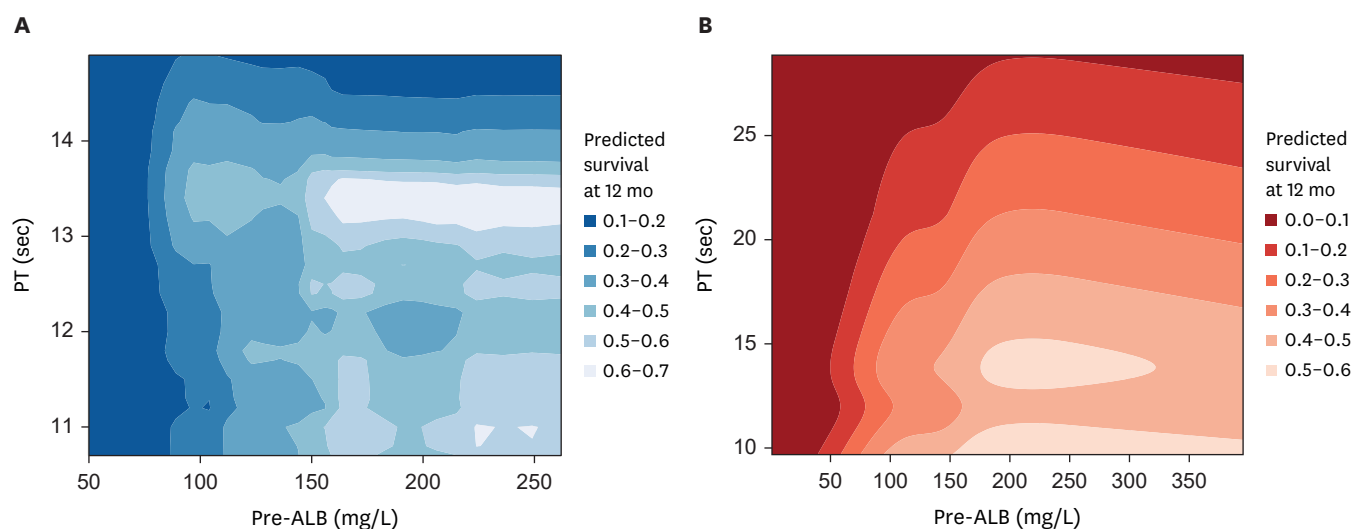


Fig. 5. Contour plots of joint effect of prealbumin and PT on predicted one-year survival after metastasis by (A) random survival forest algorithm and (B) restricted cubic spline Cox proportional hazard function. Pre-ALB = prealbumin; PT = prothrombin time.

(**Fig. 4B**). The obtained L-shaped line plot showed that the HR first steeply declined and subsequently remained stable as the prealbumin level increased to 150 mg/L. Following this, the upper limit of the HR decreased to approximately below 1. The plot of the PT was also a curved line, in which a PT above 15 seconds significantly increased the death risk.

Based on the visual inspection of **Fig. 4A and B**, we select 150 mg/L and 15 seconds as the cutoff values of the prealbumin level and the PT, respectively. The KM plots show that the prealbumin level and the PT are significant predictors of survival after metastasis. Higher prealbumin and shorter PT subgroups had an apparently better outcome (**Fig. 4C**).

Contour plotting of effect of prealbumin and PT on survival

To further illustrate the joint effect of the prealbumin level and the PT on survival, a contour plot is presented in **Fig. 5A**. For a given combination of prealbumin and PT, the contour plot reflects the interval of the estimated one-year survival rate. For example, for a patient with a prealbumin level above 150 mg/L and a PT between 13.0 and 13.5 seconds, the predicted 1-year survival rate is 60%–70% based on the RSF algorithm. Another plot based on the Cox regression with RCS is also shown in **Fig. 5B** to provide additional evidence of the relationship of the prealbumin level and the PT with the survival.

DISCUSSION

GC is well-known for its heterogeneity. GC with distant lymph node recurrence or metastasis is a distinct subgroup of advanced GC, and only a few studies have investigated its survival determinants. This study was based on an ambispective GC cohort at a tertiary hospital in a real-world clinical practice setting. A machine learning algorithm was used to identify important features affecting the survival of GC with nodal metastasis. The prealbumin level and the PT were found to be the most crucial factors, and a contour plot was depicted accordingly as a clinically applicable survival prediction tool. These findings are expected to

assist oncologists in predicting outcomes and devising decision-making strategies for the distinct subgroup of patients with advanced GC.

In this study, 574 GC patients with distant node metastasis were consecutively enrolled, of whom 289 were included for the RSF exploration. Compared to previous reports [4-9,24,25], our study collected a greater number of samples in this subgroup of advanced GC. For each individual, 81 variables were documented, based on which a data frame containing 289×81 pieces of information was created. Traditional analysis methods perform poorly on such high-dimensional datasets because of noisy variables, missing values, collinearity, and restriction of data distribution. In contrast, tree-based algorithms can easily address these problems. The RSF algorithm makes no assumption about the data distribution, satisfactorily handles missing values and numerous variables with different scales, and improves model performance using ensemble techniques [20].

Model validation by cross-validation is indispensable for evaluating the performance of a model on unseen data. Compared with holdout, leave-one-out, and k-fold cross-validation, nCV provides an almost unbiased estimate of the true error [26] and is the key technique in building a machine learning algorithm-based model. Theoretically, a small number of samples left out at each step of the outer loop implies high reliability of the true error estimation [26]. However, nCV is computationally expensive; therefore, we split the inner and outer loops by five folds, instead of ten folds, at the cost of a more biased evaluation. Notably, we included data-dependent preprocesses (i.e., missing value imputation, hyperparameter tuning, and feature selection) in the nCV procedure, to ensure that the data used for the final evaluation of the model had not been seen by the model at all.

The performance of a machine learning model is substantially affected by the quality of the selected features, and this phenomenon is called “bias-variance dilemma.” Specifically, incorporating excessive irrelevant features in the training model may lead to predictions that do not generalize well on validation data (overfitting). By contrast, excluding important features from the training model may lead to predictions with low accuracy (underfitting) [27]. In this study, the prealbumin level and the PT were identified as the predictors of top priority. Their importance was much greater than those of the other factors. However, the inclusion of other factors with the two variables would result in an increased prediction error. In the settings of the RSF algorithm and survival prediction of GC with distant node involvement, the prealbumin level and the PT on metastasis were the most crucial factors that outperformed the other variables, including pathological, therapeutic, and other biochemical features.

We previously reported that preoperative prealbumin and coagulation parameters indicate long-term outcomes of resectable GC [28]. In this study, the prealbumin level and the PT were recognized as the most important determinants of post-metastasis survival. Some studies have transformed these two variables into integrated scores to predict the prognosis of nonmetastatic GC [29-31]; however, clinical investigations seldom observe their impact on metastasis or recurrence. The RCS and the KM plots consistently indicated the existence of a threshold value that could classify patients into high- and low-risk subgroups. Nevertheless, the partial dependence plot did not support “the higher the better” pattern regarding the prealbumin level: a moderate decline of survival was clearly observed after the peak. Regarding the PT, the predicted survival rate decreased dramatically to approximately 15% once it exceeded 15 seconds, compared to a normal level reflecting a rate of over 40%.

Prealbumin is a sensitive and early indicator of malnutrition, a condition that predicts poor clinical outcomes in people with cancer [32]. The underlying mechanism may involve immune function. Enteral nutrition improves prealbumin levels in postoperative cancer patients, and increased prealbumin is strongly correlated with higher CD4 and CD8 T cell counts and immunoglobulin levels [33]. In addition, prealbumin is predominantly synthesized in the liver; therefore, a decline in prealbumin may reflect a hepatic dysfunction and abnormal anticancer drug metabolism, which is a key factor that predicts worse outcomes in GC [34]. A prolonged PT is a sign of hemostatic system activation, and some tumors can express coagulation factors [35]. Alternatively, hemostasis may affect tumor progression by influencing the proliferation rate, angiogenesis, invasion, and metastasis [35]. GC patients with tumor thrombi [36] or disseminated intravascular coagulation [37] frequently have a severely poor prognosis.

Clinicians are expected to use prediction models in practice; however, machine learning models are difficult to interpret meaningfully, and they are also commonly described as black-box models [38]. The most well-known tool for risk prediction in clinical practice is a nomogram, which is based on the beta coefficient of each variable [39]. Thus, this tool is indirectly compatible with the RSF. Some studies have integrated nomograms with RSF as a dimension reduction technique for omics data; nevertheless, the estimated risk is still based on traditional methods [40,41]. In addition, these studies apply the RSF technique to a specific proportion of the entire database (typically numerous variables of a single domain); therefore, they do not utilize the strength that an RSF can also deal with categorized factors. Different from previous studies, we created a contour plot-based clinically applicable prediction system directly developed using an RSF. Contour plots are useful for displaying a three-dimensional data frame (i.e., prealbumin, PT, and predicted survival rate).

This study had some limitations. First, selection bias was probable because of the single-center retrospective nature of the study. Second, heterogeneity of primary staging and treatment existed in the population owing to the inclusion of both synchronous and metachronous GC with nodal involvement. In fact, the prognosis after metastasis is the same for synchronous and metachronous cases as long as the cancer is in an advanced stage [4]. Third, the key predictors were closely correlated with physical status; therefore, the Eastern Cooperative Oncology Group score or the Karnofsky performance score is an important covariate. Here, we adjusted the results by the body mass index on metastasis, because the performance score was not documented for recurrence or metastasis diagnosis. Fourth, inclusion of more features or analysis of a cohort with different characteristics would result in different model constructions because machine learning learns from data, and changes in data would result in different outcomes. Our model performed best when only the prealbumin level and the PT were considered, whereas the other features were excluded because they were mathematically noisy. However, the exclusion of a variable from this model does not indicate its clinical insignificance.

In conclusion, machine learning with nCV is useful for identifying important determinants of cancer survival from high-dimensional datasets. In this study, the constructed RSF algorithm model revealed the prealbumin level and the PT as the two most crucial factors for predicting survival after GC with distant node metastasis is diagnosed. A contour plot was also depicted to intuitively display the joint relationship between these two parameters and the survival outcome.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Mr. Zhenhui He for his help in data electronization.

SUPPLEMENTARY MATERIALS

Supplementary Table 1

Number and mean post-metastasis survival times of patients with different metastatic sites

[Click here to view](#)

Supplementary Table 2

All variables included in analysis

[Click here to view](#)

Supplementary Fig. 1

Scaled Schoenfeld residual graph and smooth curve to test proportional hazard assumption of prealbumin and PT. Assumption holds for both variables when restricted cubic spline technique is implemented.

[Click here to view](#)

REFERENCES

1. Qiu MZ, Shi SM, Chen ZH, Yu HE, Sheng H, Jin Y, et al. Frequency and clinicopathological features of metastasis to liver, lung, bone, and brain from gastric cancer: a SEER-based study. *Cancer Med* 2018;7:3662-3672.
[PUBMED](#) | [CROSSREF](#)
2. In H, Solsky I, Palis B, Langdon-Embry M, Ajani J, Sano T. Validation of the 8th edition of the AJCC TNM staging system for gastric cancer using the national cancer database. *Ann Surg Oncol* 2017;24:3683-3691.
[PUBMED](#) | [CROSSREF](#)
3. Tan HL, Chia CS, Tan GH, Choo SP, Tai DW, Chua CW, et al. Metastatic gastric cancer: Does the site of metastasis make a difference? *Asia Pac J Clin Oncol* 2019;15:10-17.
[PUBMED](#) | [CROSSREF](#)
4. Patel PR, Yao JC, Hess K, Schnirer I, Rashid A, Ajani JA. Effect of timing of metastasis/disease recurrence and histologic differentiation on survival of patients with advanced gastric cancer. *Cancer* 2007;110:2186-2190.
[PUBMED](#) | [CROSSREF](#)
5. Sawaki K, Kanda M, Ito S, Mochizuki Y, Teramoto H, Ishigure K, et al. Survival times are similar among patients with peritoneal, hematogenous, and nodal recurrences after curative resections for gastric cancer. *Cancer Med* 2020;9:5392-5399.
[PUBMED](#) | [CROSSREF](#)
6. Sato S, Kunisaki C, Tanaka Y, Sato K, Miyamoto H, Yukawa N, et al. Curative-intent surgery for stage IV advanced gastric cancer: Who can undergo surgery and what are the prognostic factors for long-term survival? *Ann Surg Oncol* 2019;26:4452-4463.
[PUBMED](#) | [CROSSREF](#)
7. Kim JH, Lee HH, Seo HS, Jung YJ, Park CH. Stage-specific difference in timing and pattern of initial recurrence after curative surgery for gastric cancer. *Surg Oncol* 2019;30:81-86.
[PUBMED](#) | [CROSSREF](#)

8. Agnes A, Biondi A, Laurino A, Strippoli A, Ricci R, Pozzo C, et al. A detailed analysis of the recurrence timing and pattern after curative surgery in patients undergoing neoadjuvant therapy or upfront surgery for gastric cancer. *J Surg Oncol* 2020;122:293-305.
[PUBMED](#) | [CROSSREF](#)
9. Takahashi R, Ohashi M, Kano Y, Ida S, Kumagai K, Nunobe S, et al. Timing and site-specific trends of recurrence in patients with pathological stage II or III gastric cancer after curative gastrectomy followed by adjuvant S-1 monotherapy. *Gastric Cancer* 2019;22:1256-1262.
[PUBMED](#) | [CROSSREF](#)
10. Fang WL, Lan YT, Huang KH, Liu CA, Hung YP, Lin CH, et al. Clinical significance of circulating plasma DNA in gastric cancer. *Int J Cancer* 2016;138:2974-2983.
[PUBMED](#) | [CROSSREF](#)
11. Zhang C, Jing LW, Li ZT, Chang ZW, Liu H, Zhang QM, et al. Identification of a prognostic 28-gene expression signature for gastric cancer with lymphatic metastasis. *Biosci Rep* 2019;39:BSR20182179.
[PUBMED](#) | [CROSSREF](#)
12. Zeng D, Zhou R, Yu Y, Luo Y, Zhang J, Sun H, et al. Gene expression profiles for a prognostic immunoscore in gastric cancer. *Br J Surg* 2018;105:1338-1348.
[PUBMED](#) | [CROSSREF](#)
13. Izuishi K, Mori H. Recent strategies for treating stage IV gastric cancer: roles of palliative gastrectomy, chemotherapy, and radiotherapy. *J Gastrointestin Liver Dis* 2016;25:87-94.
[PUBMED](#) | [CROSSREF](#)
14. Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *Br J Cancer* 2018;119:1456-1463.
[PUBMED](#) | [CROSSREF](#)
15. Greenwood CJ, Youssef GJ, Letcher P, Macdonald JA, Hagg LJ, Sanson A, et al. A comparison of penalised regression methods for informing the selection of predictive markers. *PLoS One* 2020;15:e0242730.
[PUBMED](#) | [CROSSREF](#)
16. Adeli E, Li X, Kwon D, Zhang Y, Pohl KM. logistic regression confined by cardinality-constrained sample and feature selection. *IEEE Trans Pattern Anal Mach Intell* 2020;42:1713-1728.
[PUBMED](#) | [CROSSREF](#)
17. Dong D, Fang MJ, Tang L, Shan XH, Gao JB, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann Oncol* 2020;31:912-920.
[PUBMED](#) | [CROSSREF](#)
18. Li J, Dong D, Fang M, Wang R, Tian J, Li H, et al. Dual-energy CT-based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer. *Eur Radiol* 2020;30:2324-2333.
[PUBMED](#) | [CROSSREF](#)
19. Cai WY, Dong ZN, Fu XT, Lin LY, Wang L, Ye GD, et al. Identification of a tumor microenvironment-relevant gene set-based prognostic signature and related therapy targets in gastric cancer. *Theranostics* 2020;10:8633-8647.
[PUBMED](#) | [CROSSREF](#)
20. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008;2:841-860.
[CROSSREF](#)
21. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, et al. mlr3: a modern object-oriented machine learning framework in R. *J Open Source Softw* 2019;44:1903.
[CROSSREF](#)
22. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 2017;77:1-17.
[CROSSREF](#)
23. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29:1189-1232.
[CROSSREF](#)
24. Mokadem I, Dijksterhuis WP, van Putten M, Heuthorst L, de Vos-Geelen JM, Haj Mohammad N, et al. Recurrence after preoperative chemotherapy and surgery for gastric adenocarcinoma: a multicenter study. *Gastric Cancer* 2019;22:1263-1273.
[PUBMED](#) | [CROSSREF](#)
25. Saka M, Katai H, Fukagawa T, Nijjar R, Sano T. Recurrence in early gastric cancer with lymph node metastasis. *Gastric Cancer* 2008;11:214-218.
[PUBMED](#) | [CROSSREF](#)

26. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91.
[PUBMED](#) | [CROSSREF](#)
27. Parvande S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. *Bioinformatics* 2020;36:3093-3098.
[PUBMED](#) | [CROSSREF](#)
28. Wu ZJ, Xu H, Wang R, Bu LJ, Ning J, Hao JQ, et al. Cumulative score based on preoperative fibrinogen and pre-albumin could predict long-term survival for patients with resectable gastric cancer. *J Cancer* 2019;10:6244-6251.
[PUBMED](#) | [CROSSREF](#)
29. Wang Z, Zhang L, Wang J, Wang Y, Dong Q, Piao H, et al. Prealbumin-to-globulin ratio can predict the chemotherapy outcomes and prognosis of patients with gastric cancer receiving first-line chemotherapy. *J Immunol Res* 2020;2020:6813176.
[PUBMED](#) | [CROSSREF](#)
30. Lu J, Xu BB, Zheng ZF, Xie JW, Wang JB, Lin JX, et al. CRP/prealbumin, a novel inflammatory index for predicting recurrence after radical resection in gastric cancer patients: post hoc analysis of a randomized phase III trial. *Gastric Cancer* 2019;22:536-545.
[PUBMED](#) | [CROSSREF](#)
31. Wang N, Xi W, Lu S, Jiang J, Wang C, Zhu Z, et al. A novel inflammatory-nutritional prognostic scoring system for stage III gastric cancer patients with radical gastrectomy followed by adjuvant chemotherapy. *Front Oncol* 2021;11:650562.
[PUBMED](#) | [CROSSREF](#)
32. Bullock AF, Greenley SL, McKenzie GA, Paton LW, Johnson MJ. Relationship between markers of malnutrition and clinical outcomes in older adults with cancer: systematic review, narrative synthesis and meta-analysis. *Eur J Clin Nutr* 2020;74:1519-1535.
[PUBMED](#) | [CROSSREF](#)
33. Chen X, Zhao G, Zhu L. Home enteral nutrition for postoperative elderly patients with esophageal cancer. *Ann Palliat Med* 2021;10:278-284.
[PUBMED](#) | [CROSSREF](#)
34. Yin HM, He Q, Chen J, Li Z, Yang W, Hu X. Drug metabolism-related eight-gene signature can predict the prognosis of gastric adenocarcinoma. *J Clin Lab Anal* 2021;35:e24085.
[PUBMED](#) | [CROSSREF](#)
35. Ünlü B, Versteeg HH. Effects of tumor-expressed coagulation factors on cancer progression and venous thrombosis: is there a key factor? *Thromb Res* 2014;133 Suppl 2:S76-S84.
[PUBMED](#) | [CROSSREF](#)
36. Eom BW, Lee JH, Lee JS, Kim MJ, Ryu KW, Choi IJ, et al. Survival analysis of gastric cancer patients with tumor thrombus in the portal vein. *J Surg Oncol* 2012;105:310-315.
[PUBMED](#) | [CROSSREF](#)
37. Rhee J, Han SW, Oh DY, Im SA, Kim TY, Bang YJ. Clinicopathologic features and clinical outcomes of gastric cancer that initially presents with disseminated intravascular coagulation: a retrospective study. *J Gastroenterol Hepatol* 2010;25:1537-1542.
[PUBMED](#) | [CROSSREF](#)
38. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019;19:146.
[PUBMED](#) | [CROSSREF](#)
39. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol* 2008;26:1364-1370.
[PUBMED](#) | [CROSSREF](#)
40. Yu Y, Tan Y, Xie C, Hu Q, Ouyang J, Chen Y, et al. Development and validation of a preoperative magnetic resonance imaging radiomics-based signature to predict axillary lymph node metastasis and disease-free survival in patients with early-stage breast cancer. *JAMA Netw Open* 2020;3:e2028086.
[PUBMED](#) | [CROSSREF](#)
41. Zhou R, Zhang J, Zeng D, Sun H, Rong X, Shi M, et al. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer Immunol Immunother* 2019;68:433-442.
[PUBMED](#) | [CROSSREF](#)