Elevate Series

RESEARCH ARTICLE

🔓 OPEN ACCESS | Check for updates

Establishing and validating models integrated with hematological biomarkers and clinical characteristics for the prognosis of non-esophageal squamous cell carcinoma patients

Ning Xue[a#] (ID), Xiaoyan Wen[b#], Qian Wang[c#], Yong Shen[a], Yuanye Qu[a], Qingxia Xu[a], Shulin Chen[d,e] (ID) and Jing Chen[a] (ID)

aDepartment of Clinical Laboratory, The Affiliated Cancer Hospital of Zhengzhou University and Henan Cancer Hospital, Zhengzhou Key Laboratory of Digestive System Tumor Marker Diagnosis, Zhengzhou, P. R. China; bCentral Sterilization Supply Department, The Guanghua Stomatological College of Sun Yat-sen University, Hospital of Stomatology, SunYat-sen University, Guangzhou, P. R. China; cDepartment of radiation oncology, China–Japan Union Hospital of Jilin University, Changchun, P.R. China; dState Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Collaborative Innovation Center for Cancer Medicine, Sun Yat-Sen University Cancer Center, Guangzhou, China; eResearch Center for Translational Medicine, First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, P. R. China

ABSTRACT

Background: This study aimed to construct a novel model and validate its predictive power in non-esophageal squamous cell carcinoma (NESCC) patients.

Methods: This retrospective study included 151 patients between October 2006 and September 2016. The LASSO Cox and Random Survival Forest (RSF) models were developed with the help of hematological biomarkers and clinical characteristics. The concordance index (C-index) was used to assess the prognostic power of the LASSO Cox model, RSF model, and TNM staging. Based on the risk scores of the LASSO Cox and RSF models, we divided patients into low-risk and high-risk subgroups.

Results: We constructed two models in NESCC patients according to LASSO Cox regression and RSF models. The RSF model reached a C-index of 0.841 (95% CI: 0.792–0.889) in the primary cohort and 0.880 (95% CI: 0.830–0.930) in the validation cohort, which was higher than the C-index of the LASSO Cox model 0.656 (95% CI: 0.580–0.732) and 0.632 (95% CI: 0.542–0.720) in the two cohorts. The integrated C/D area under the ROC curve (AUC) values for the LASSO Cox and RSF models were 0.701 and 0.861, respectively. In both two models, Kaplan-Meier survival analysis and the estimated restricted mean survival time (RMST) values indicated that the low-risk subgroup had a better prognostic outcome than the high-risk subgroup ($p < 0.05$).

Conclusions: The RSF model has better prediction power than the LASSO Cox and the TNM staging models. It has a guiding value for the choice of individualized treatment in patients with NESCC.

Introduction

Esophageal cancer (EC) is one of the leading causes of tumor-induced morbidity and mortality worldwide [1,2]. The most common histological subtype of EC is esophageal squamous cell carcinoma (ESCC), which also includes adenocarcinoma, adenosquamous carcinoma, neuroendocrine carcinoma, and esophageal sarcomatoid carcinoma, among others., and non-esophageal squamous cell carcinoma (NESCC) accounts for approximately 10% of all ECs [3]. In developing countries including China, NESCC is rare,

and more than 90% patients with EC are diagnosed with ESCC [4]. However, esophageal adenocarcinoma represents 80% of all cases of ECs in United States [5]. EC is usually diagnosed at a later stage, with high mortality and a 5-year overall survival (OS) rate of 15–20% [6]. One previous study showed that patients with ESCC had better OS outcomes than patients with adenocarcinoma of the esophagus after the same treatment [7]. Many studies have focused on the prognosis of ESCC. However, only few studies have focused on the prognosis of NESCC. Thus, it is indispensable to implement a novel model for the prognosis of patients with NESCC, which will then allow the patients to receive more precise and individualized treatment.

With the popularity of electronic medical records and improvements in laboratory examinations, numerous prognostic models have been developed and widely applied in clinical practice. A clinical prognostic model based on albumin-to-fibrinogen ratio (AFR) and neutrophil–lymphocyte ratio (NLR) can be used to predict the prognosis of ESCC patients after radical resection [8]. Luo et al. established a novel model based on serum IgA, CRP, and cTNM stage to evaluate the prognosis of ESCC patients receiving neoadjuvant chemoradiotherapy plus surgery [9]. Further, Wang et al. constructed a nomogram that included five variables: age, N stage, location, tumor response, and monocyte/lymphocyte ratio [10]. This nomogram can serve as an effective tool for patients with locally advanced ESCC who are undergoing concurrent chemoradiotherapy. Furthermore, a nomogram based on hemoglobin (HB), C-reactive protein-to-albumin ratio (CAR), and platelet-to-lymphocyte ratio (PLR) may accurately and effectively predict OS in ESCC [11]. However, all existing models have only been used to evaluate the prognosis of patients with ESCC.

Random survival forest (RSF) is an ensemble tree method, which is used to analyze right-censored survival data [12,13]. It adopts iterating the missing data algorithm, and its accuracy can be substantially improved even with missing data [13]. Moreover, RSF can easily be applied to real data settings to reveal highly complex interrelationships between variables [14]. In this study, RSF and LASSO Cox regression were performed to screen for the potential factors affecting the prognosis of patients with NESCC and to construct novel prediction models. We further compared the predictive power of the two prognostic models with regard to OS in compared with the 8th edition of the Union for International Cancer Control (UICC)/American Joint Committee on Cancer (AJCC) TNM staging system.

Materials and methods

Participants

This study is a retrospective analysis of patients with NESCC who were undergoing treatment at the Sun-Yat Sen University Cancer Center. We constructed the LASSO Cox and RSF models based on patients clinical factors. Subsequently, we compared the prediction ability of LASSO Cox, RSF, and the TNM staging models. In total, 151 patients with NESCC were enrolled in this study. All patients were treated from October 2006 to September 2016 at the Sun-Yat Sen University Cancer Center. The inclusion criteria were as follows: (a) NESCC diagnosed by two independent pathologists, (b) absence of any second carcinoma, (c) no anti-tumor therapy or anti-inflammatory treatment, (d) complete clinical and laboratory findings. The patients were followed up every 3 months. The last follow-up was conducted in March 2024. The follow-up end-point was cancer-specific death or March 2024. Follow-up data were obtained from medical records, clinics, and telephone inquiries. The TNM stage was based on the 8th edition of the Union for International Cancer Control (UICC)/American Joint Committee on Cancer (AJCC) TNM staging system.

This study was approved by the Clinical Research Ethics Committee of Sun Yat-Sen University Cancer Center and was conducted in accordance with the Declaration of Helsinki. Given that this is a retrospective study, the Clinical Research Ethics Committee of Sun Yat-sen University Cancer Center waived the need for informed consent.

Clinical data collection

We collected the following clinical data for each patient: Age, Gender, location, Histological type, white blood cell (WBC) count, neutrophil (NEU) count, lymphocyte (LYM) count, monocyte (MO) count, hemoglobin (HGB), platelet count (PLT), C-reactive protein (CRP), body mass index (BMI), Glucose(GLU), the advanced lung cancer inflammation index (ALI), nutritional risk index (NRI), NLR, LYM/MO ratio (LMR), PLT/LYM ratio (PLR), NEU/LYM ratio (NLR), LYM/CRP ratio (LCR), GLU/LYM ratio (GLR), prognostic nutritional index (PNI), alanine aminotransferase (ALT), aspartate aminotransferase (AST), ALT/AST ratio (LSR), alkaline phosphatase (ALP), glutamyl transpeptidase (GGT), lactic dehydrogenase (LDH), LDH/CRP ratio (LDCR), albumin (ALB), CRP/ALB ratio (CAR), total protein (TP), globulin (GLOB), ALB/GLOB ratio (AGR), total bilirubin (TBIL), direct bilirubin (DBIL), blood urea nitrogen (BUN), creatinine (CRE), uric acid (UA), cystatin C (CYSC),

triglyceride (TG), cholesterol (CHO), high density lipo-protein (HDL), TG/HDL ratio (THR), low density lipoprotein (LDL), LDL/HDL ratio (LHR), apolipoprotein A (APOA), apolipoprotein B (APOB), APOA/APOB ratio (ABR), systemic immune-inflammation index (SII), systemic inflammation response index (SIRI), aggregate index of systemic inflammation (AISI), and prognostic impact index. The relative calculation formulae were as follows: $SII = NEU \times PLT/LYM$ [15]. $NRI = 1.487 \times ALB$ (g/L) $+ 41.7 \times$ preoperative weight/ideal body weight (kg) [16]. Ideal body weight $= 22 \times$ height $(m)^2$. $PNI = ALB$ (g/L) $+ 5 \times LYM \times 10^9/L$ [17]. $SIRI = NEU \times MO/LYM$. $AISI = NEU \times MO \times PLT/LYM$ [18].

Statistical analysis

All statistical analyses were performed using the R software (version 3.6.2) and SPSS 25.0 (SPSS, Chicago, USA). LASSO Cox regression ('glmnet' R package) and RSF ('randomForestSRC' R package) were used to construct models in the primary cohort, and potential factors were screened through cross-validation to determine the final factors. LASSO Cox model used L1 regularization to reduce the regression coefficients toward zero, and reduce the risk of overfitting. The RSF algorithm can measure the importance of factors, and rank factors according to their importance. Performance evaluation for the LASSO Cox model, RSF model, and TNM staging model was performed using the integrated Brier score (IBS), concordance index (C-index), and receiver operating characteristic curve (ROC). The C-index and area under the ROC curve (AUC) were calculated using the R package 'survcomp' and R package 'survivalROC'. Based on the cutoff values ('survminer' R package), patients were divided into low-risk and high-risk subgroups in the LASSO Cox, RSF, and TNM staging models. The OS in the risk subgroups was evaluated via the Kaplan–Meier method. Data are shown as the mean ± SD. Statistical tests were two-sided, and significant differences were considered at $p < 0.05$.

Results

Patient characteristics

A total of 151 patients with NESCC were randomly divided into the primary cohort ($n = 90$) and the validation cohort ($n = 61$). The clinical characteristics of the primary and validation cohorts are shown in Table 1. There were 90 patients in the primary cohort, comprising 75 (83.33%) males and 15 (16.67%) females. The mean age (SD) of the patients was 58.44 (10.22) years.

There were 36 (40.00%) and 54 (60.00%) I/II and III/IV stage patients, respectively. The 1-, 3-, and 5-year OS rates were 65.56%, 30.00%, and 22.22%, respectively. In the validation cohort, there were 47 (77.05%) males and 14 (22.95%) females. The mean age (SD) of these patients was 60.26 (9.28) years. There were 35 (57.38%) and 26 (42.62%) I/II and III/IV stage patients, respectively. The 1-, 3-, and 5-year OS rates in this cohort were 72.13%, 34.43%, and 27.87%, respectively.

Prognostic models for OS

In the primary cohort, TBIL, LHR, NEU, LDCR, LMR BMI and SIRI were identified as candidate markers from 52 clinical factors by LASSO Cox regression analysis. The trajectory changes for each factor (Figure 1A) and confidence intervals for each λ (Figure 1B) are shown in Figure 1. RSF was performed to select potential factors from these clinical features. We constructed an RSF model that included DBIL, TP, BUN, PLT, and SII levels. Figure 1 shows the variable importance (VIMP) values permutation (Figure 1C), minimal depth variable importance (Figure 1D) and variable importance value rank (Figure 1E) for patients with NESCC. Variables with lower values of minimal depth variable importance are implied to be more important variables. In Figure 1E, the blue dot represents a VIMP score greater than zero and the red dot represents a VIMP score less than zero. The area above the red diagonal dotted line represents a higher rank of the VIMP score, and that below the diagonal dotted line represents the minimum depth of higher rank.

### Variable interpretation and comparison of prognostic models

Because of the complexity of the algorithms, the Survex R package was used to compare the LASSO Cox model and the RSF model and explain their global features. In both the primary and validation cohorts, The C-index, the IBS loss after permutation (Supplemental Figure 1A,C) and integrated C/D AUC loss after permutation (Supplemental Figure 1B,D) were evaluated to determine the time-dependent variable importance for the LASSO Cox model and RSF model. The IBS and integrated C/D AUCs of the LASSO Cox and RSF model were 0.141, 0.701 and 0.166, 0.861, respectively. In the primary cohort, the C-index of the RSF model was 0.841(95% CI: 0.792–0.889), which was significantly higher than that of the LASSO Cox model (0.656; 95% CI: 0.580–0.732; $p < 0.001$) and TNM staging model (0.518; 95% CI: 0.444–0.592; $p < 0.001$). For the validation cohort, the C-index values of these

Table 1. Demographics and clinical characteristics of NESCC patients in the primary and validation cohorts.

Characteristic	Primary cohort n = (90) No. (%) or mean ± SD	Validation cohort n = (61) No. (%) or mean ± SD	p value
Gender			0.336[a]
Female	15 (16.67%)	14 (22.95%)	
Male	75 (83.33%)	47 (77.05%)	
Age (years)	58.44 ± 10.22	60.26 ± 9.28	0.256[b]
Histologic type			0.861[a]
Adenocarcinoma	11 (12.22%)	6 (9.84%)	
Adenosquamous carcinoma	12 (13.33%)	11 (18.03%)	
Neuroendocrine carcinoma	17 (18.89%)	9 (14.75%)	
Esophageal sarcomatoid carcinoma	41 (45.56%)	30 (8.20%)	
Other	9 (10.00%)	5 (16.67%)	
TNM stage			0.036[a]
I + II	36 (40.00%)	35 (57.38%)	
III + IV	54 (60.00%)	26 (42.62%)	
Tumor location			0.411[a]
Upper	10 (11.11%)	3 (4.92%)	
Middle	52 (57.78%)	38 (62.30%)	
Lower	28 (31.11%)	20 (32.79%)	
BMI (kg/m^2)			0.532[a]
<18.00	12 (13.33%)	5 (8.20%)	
18.00–24.00	59 (65.56%)	40 (65.57%)	
>24.00	19 (21.11%)	16 (26.23%)	
WBC (10^9/L)	8.26 ± 3.40	7.75 ± 2.99	0.376[b]
Neutrophils (10^9/L)	5.63 ± 3.32	5.11 ± 2.95	0.341[b]
Lymphocyte (10^9/L)	1.74 ± 0.73	1.93 ± 0.68	0.059[b]
Monocytes (10^9/L)	0.61 ± 0.28	0.54 ± 0.22	0.116[b]
HGB (g/L)	129.41 ± 17.76	134.37 ± 15.99	0.098[b]
Platelet (10^9/L)	257.74 ± 88.88	227.13 ± 71.10	0.054[b]
NRI	101.01 ± 11.22	104.03 ± 11.14	0.045[b]
NLR	4.48 ± 5.18	3.53 ± 4.70	0.129[b]
LMR	3.48 ± 2.17	4.22 ± 2.63	0.040[b]
PLR	176.13 ± 112.86	130.62 ± 58.50	0.008[b]
LCR	1.04 ± 1.63	1.81 ± 2.69	0.003[b]
GLR	4.09 ± 2.94	3.38 ± 2.12	0.061[b]
TBIL (μmol/L)	12.70 ± 9.71	11.83 ± 5.56	0.862[b]
DBIL (μmol/L)	4.46 ± 5.42	4.00 ± 2.60	0.685[b]
TP (g/L)	70.73 ± 5.86	70.72 ± 5.52	0.743[b]
ALB (g/L)	40.64 ± 6.07	41.91 ± 6.29	0.059[b]
GLB (g/L)	30.24 ± 5.96	28.17 ± 4.92	0.019[b]
AGR	1.40 ± 0.31	1.57 ± 0.38	0.020[b]
ALT (U/L)	18.84 ± 16.40	22.98 ± 29.74	0.678[b]
AST (U/L)	20.37 ± 10.37	22.67 ± 18.85	0.897[b]
LSR	0.89 ± 0.30	0.92 ± 0.30	0.422[b]
ALP (U/L)	76.67 ± 35.53	78.84 ± 37.40	0.634[b]
GGT (U/L)	40.95 ± 63.08	47.34 ± 114.34	0.675[b]
LDH (U/L)	179.80 ± 72.36	172.45 ± 73.76	0.256[b]
LDCR	111.45 ± 242.18	148.27 ± 202.48	0.008[b]
CRP (mg/L)	21.42 ± 35.72	6.79 ± 0.90	0.005[b]
CAR	2.56 ± 19.11	0.18 ± 0.33	0.006[b]
CRE (μmol/L)	76.77 ± 21.21	76.69 ± 20.67	0.958[b]
BUN (mmol/L)	9.55 ± 43.28	5.27 ± 2.12	0.844[b]
UA (μmol/L)	332.26 ± 100.79	338.52 ± 80.80	0.553[b]
CYSC (mg/L)	1.70 ± 7.03	0.93 ± 0.29	0.262[b]
GLU (mmol/L)	5.65 ± 1.29	5.49 ± 1.44	0.448[b]
TG (mmol/L)	1.22 ± 0.55	1.22 ± 0.58	0.971[b]
CHO (mmol/L)	5.06 ± 1.00	4.97 ± 1.04	0.522[b]
HDL (mmol/L)	1.12 ± 0.30	1.21 ± 0.33	0.110[b]
LDL (mmol/L)	3.71 ± 3.75	3.18 ± 0.90	0.171[b]
THR	1.23 ± 0.82	1.14 ± 0.92	0.418[b]
LHR	3.41 ± 2.58	2.78 ± 0.93	0.014[b]
APOA (g/L)	1.12 ± 0.22	1.25 ± 0.25	0.002[b]
APOB (g/L)	1.01 ± 0.21	1.03 ± 0.58	0.191[b]
ABR	1.16 ± 0.35	1.36 ± 0.56	0.007[b]
ALI	373.27 ± 272.34	460.88 ± 294.48	0.046[b]
SII	1217.86 ± 1624.63	746.22 ± 881.18	0.037[b]
AISI	980.30 ± 1824.88	449.89 ± 682.80	0.030[b]
SIRI	3.45 ± 6.01	2.25 ± 4.43	0.070[b]
PNI	49.36 ± 7.56	51.53 ± 8.12	0.018[b]

Note. a. Chi-squared test, b Wilcoxon test.
BMI, body mass index; WBC, white blood cell; HGB, hemoglobin; NRI, the nutritional risk index; NLR, neutrophil/lymphocyte ratio; LMR, lymphocyte/monocyte ratio; PLR, platelet/lymphocyte ratio; CRP, C-reactive protein; LCR, lymphocyte/CRP; GLR, glucose/lymphocyte; TBIL, bilirubin total; DBIL, direct bilirubin; TP, total protein; ALB, albumin; GLOB, globulin; AGR, ALB/GLOB ratio; ALT, alanine aminotransferase; AST, aspartate aminotransferase; SLR, AST/ALT ratio; ALP, alkaline phosphatase; GGT, glutamyl transpeptidase; LDH, lactic dehydrogenase; LDCR, LDH/CRP ratio; CAR, CRP/ALB; CRE, creatinine; BUN, blood urea nitrogen; UA, uric acid; CYSC, cystatin C; GLU, Glucose; TG, triglyceride; CHO, cholesterol; HDL, high density lipoprotein; LDL, low density lipoprotein; THR, TG/HDL ratio; LHR, LDL/HDL ratio; APOA, apolipoprotein AI; APOB, apolipoprotein B; ABR, APOA/APOB ratio; ALI, the advanced lung cancer inflammation index; SII, systemic immune-inflammation index; AISI, aggregate index of systemic inflammation; SIRI, systemic inflammation response index; PNI, prognostic impact index.
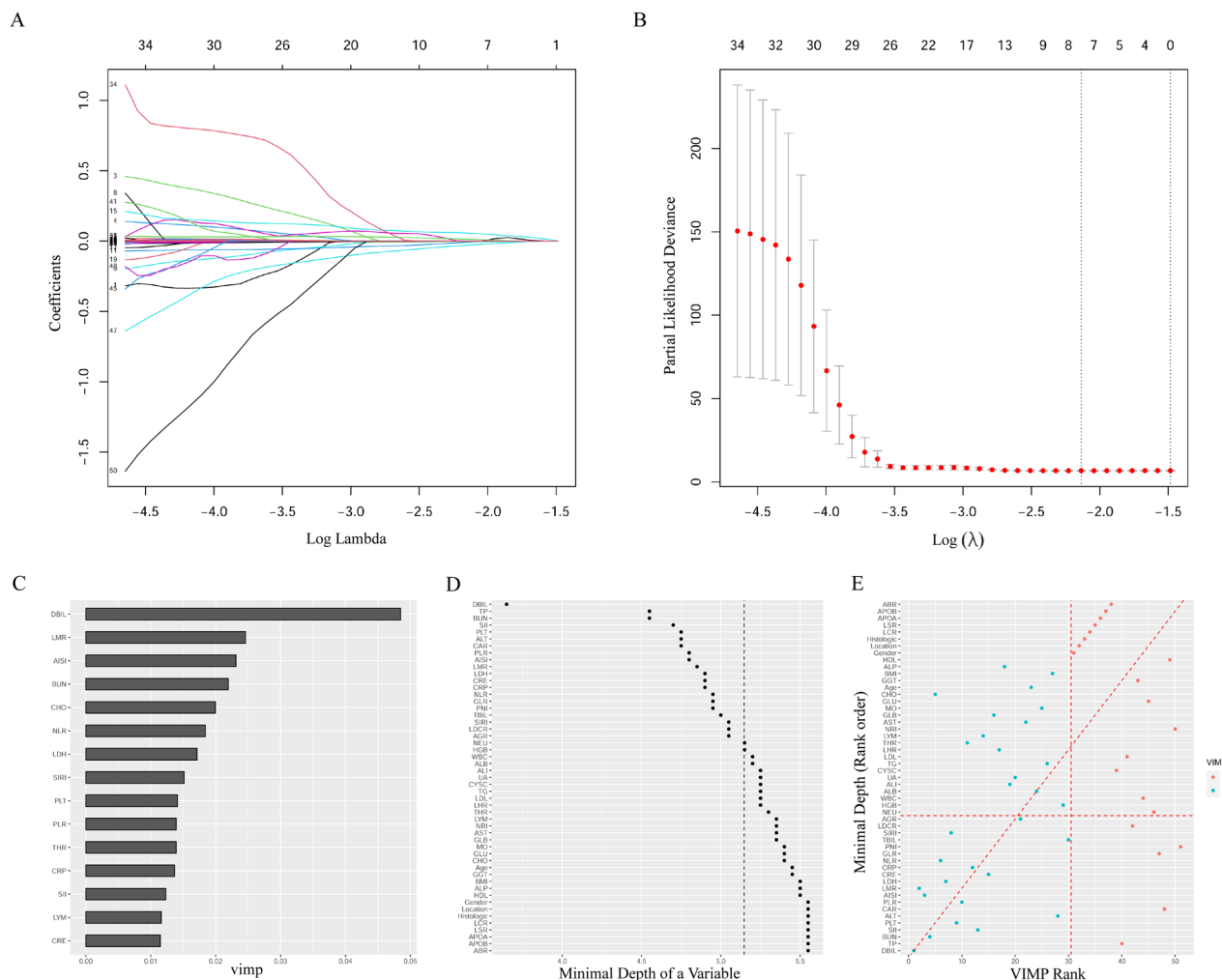
Figure 1. Potential features selection (A) and tenfold cross validation (B) of LASSO Cox model. Variable importance values permutation (C), minimal depth variable importance (D) and variable importance values rank (E) for patients of NESCC.

models (RSF model, LASSO Cox model, and TNM staging model) were 0.880 (95% CI: 0.830–0.930), 0.632 (95% CI: 0.542–0.720), and 0.655 (95% CI: 0.569–0.741), respectively (Table 2). When it comes to global interpretation, the results show that the influence of the LHR feature in the LASSO Cox model gradually increased. In the RSF model, SII and DBIL are important features. Supplemental Figure 2 shows the partial-dependence survival profiles of the LASSO Cox and RSF models. LHR was the main influencing factor in the LASSO Cox model (Supplemental Figure 2A). Further, in the RSF model, DBIL and SII were the most influential factors (Supplemental Figure 2B). The AUCs of the RSF model in the primary cohort for 1-, 3-, and 5-year OS were 0.871, 0.975, and 0.970, respectively. Meanwhile, the AUCs of the LASSO Cox model in the primary cohort for 1-, 3-, and 5-year OS were 0.664, 0.771, and 0.757, respectively. The AUCs of the TNM staging model in the primary cohort for 1-, 3-, and 5-year OS were 0.533, 0.494, and 0,478, respectively

(Supplemental Figure 3). Similarly, the AUCs of the RSF model for 1-, 3-, and 5-year OS were higher than those of the LASSO Cox model and TNM staging model in the validation cohort. These results demonstrate that the prediction performance of the RSF model is superior to that of the LASSO Cox model.

Model performance at stratifying risk

Based on the risk scores, we adopted the Survminer and Survival R package to determine cutoff values. In the LASSO Cox model, patients with a risk score less than 1.27 were in the low-risk subgroup, and the others were in the high-risk subgroup. Moreover, in the RSF model, patients with a risk score less than 15.31 were in the low-risk subgroup, and others were in the high-risk group. In the LASSO Cox model, the low-risk subgroup had longer OS than the high-risk subgroup in both the primary cohort ($p < 0.001$) (Figure 2A) and the validation cohort ($p = 0.011$) (Figure 2D). In the RSF

model, the low-risk subgroup showed better OS than the high-risk subgroup in both cohorts ($p < 0.001$) (Figure 2B,E). The OS of each patient in the TNM staging group was significantly different in the validation cohort ($p = 0.002$) (Figure 2F). However, there was no significant difference in the OS between different TNM stages in the validation cohort ($p = 0.980$) (Figure 2C).

**Table 2.** The C-index of OS for LASSO Cox model, RSF model and TNM staging model.

Survival prediction	C-index	95 CI%	p
For primary cohort			
LASSO Cox model	0.656	0.580–0.732	
RSF model	0.841	0.792–0.889	
TNM staging model	0.518	0.444–0.592	
LASSO Cox model vs TNM staging model			0.003
RSF model vs TNM staging model			<0.001
LASSO Cox model vs RSF			<0.001
For validation cohort			
LASSO Cox model	0.632	0.542–0.720	
RSF model	0.880	0.830–0.930	
TNM staging model	0.655	0.569–0.741	
LASSO Cox model vs TNM staging model			0.699
RSF model vs TNM staging model			<0.001
LASSO Cox model vs RSF			<0.001

C-index = concordance index; p values were calculated based on normal approximation using the rcorrp.cens function in the Hmisc package.

In the primary cohort, the estimated restricted mean survival time (RMST) values in the high-risk subgroup and low-risk subgroup in the LASSO Cox were 18.9 and 58.61 months, respectively (Figure 3A). The RMST values in the high-risk and low-risk subgroups in the RSF model were 13.87 and 94.37 months, respectively (Figure 3B). For the validation cohort, the RMST values for the high-risk and low-risk subgroups in of the LASSO Cox model and RSF model were 22.65, 53.82 months and 12.01, 79.21 months, respectively (Figure 3C,D). Meanwhile, we divided the patients into I/II and III/IV groups based on their TNM stage. In the primary cohort of the LASSO Cox model, the low-risk patients had significantly longer OS than the high-risk patients in both group I/II ($p = 0.063$) (Figure 4A) and group III/IV ($p < 0.001$) (Figure 4B). In the validation cohort of the LASSO Cox model, although the OS of low-risk and high-risk patients was significantly different in the I/II group ($p = 0.024$) (Figure 4E), there was no significant difference between the OS of low-risk and high-risk patients in stage III/IV ($p = 0.650$) (Figure 4F). In both the primary and validation cohorts, the low-risk patients showed better OS benefit than high-risk patients in both the I/II and III/IV groups in the RSF model ($p < 0.001$) (Figure 4C,D,G,H).
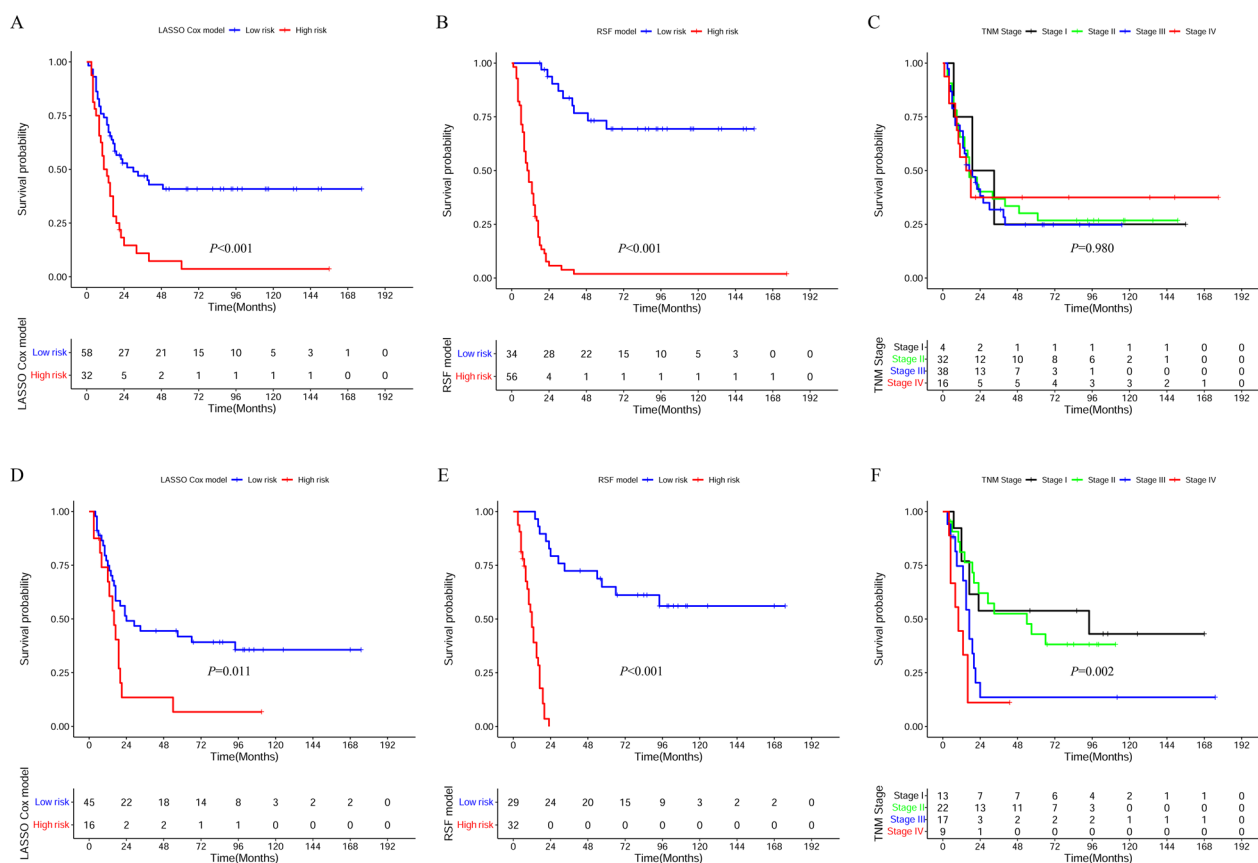
Figure 2. Kaplan–Meier curves of high risk and low risk subgroups for the LASSO Cox model, the RSF model, and the TNM stage in the primary cohort (A–C) and the validation cohort (D–F).
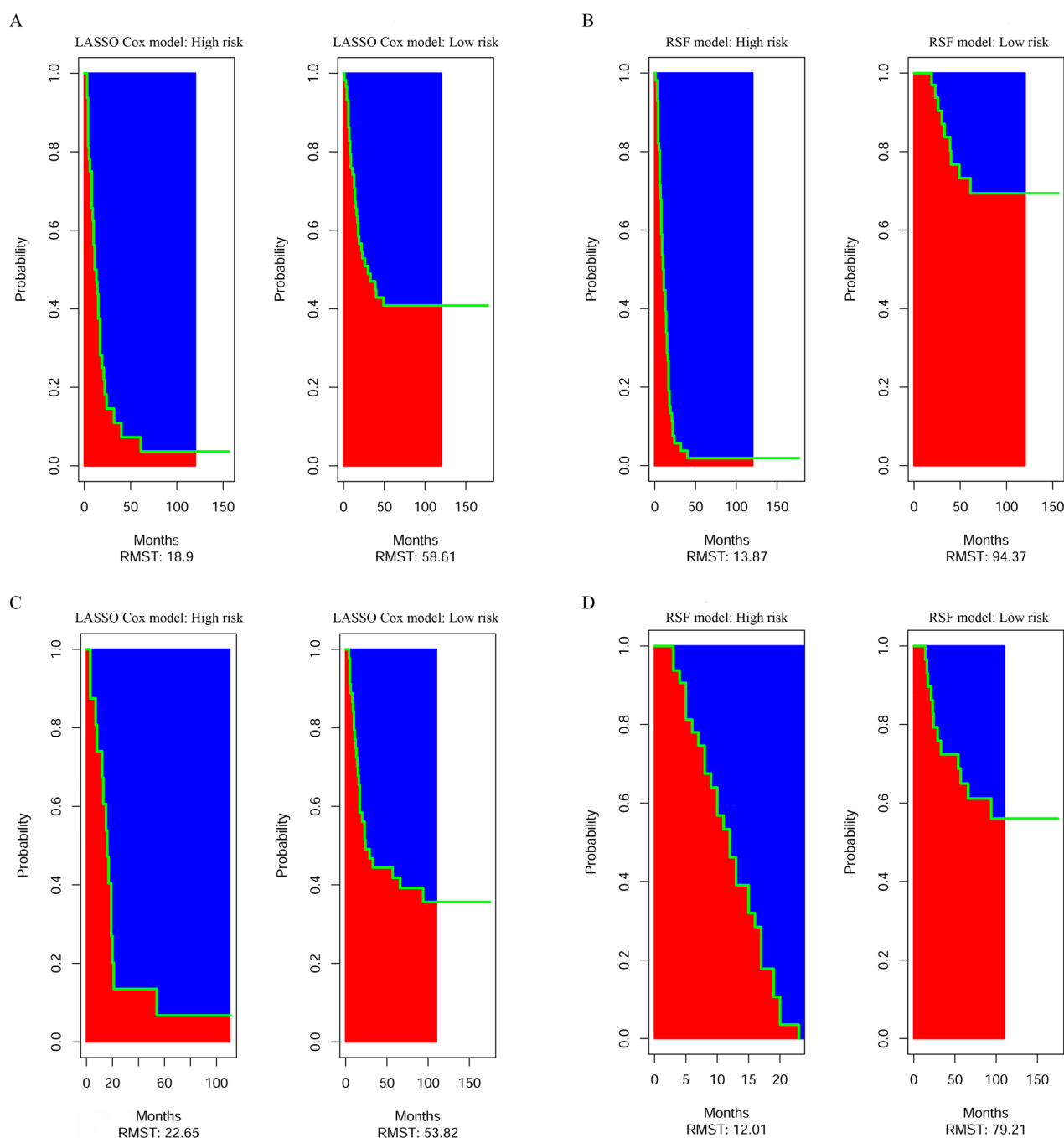
Figure 3. The estimate restricted mean survival time (RMST) values of the LASSO Cox model (A, C) and the RSF model in the primary cohort and the validation cohort (B, D).

Correlations among prognostic models

Figure 5 shows the correlations among the LASSO model, RSF, and TNM staging models. The Pearson correlation coefficient (PCC) ranged from −1 to +1. Red and Blue colors represent negative and positive correlations, respectively. In the primary cohort, the LASSO Cox model weakly correlated with the RSF model [Pearson correlation coefficient (PCC): 0.38]. Meanwhile, the LASSO Cox model moderately correlated with the RSF model in the validation cohort (PCC: 0.43). Moreover, there was a weak correlation between the TNM staging model and RSF model in the validation cohort (PCC: 0.35).

Discussion

EC with different histological types and differences in epidemiology, pathogenesis, and risk factors. Therefore, the survival times of patients with EC with different histological types are different. Most prognostic

Figure 4. Kaplan–Meier curves of I/II and III/IV patients with NESCC in the primary and validation cohorts. LASSO Cox model: I/II patients (A) and III/IV patients (B) in the primary cohort, I/II patients (E) and III/IV patients (F) in the validation cohort; RSF model: I/II patients (C) and III/IV patients (D) in the primary cohort, I/II patients (G) and III/IV patients (H) in the validation cohort.
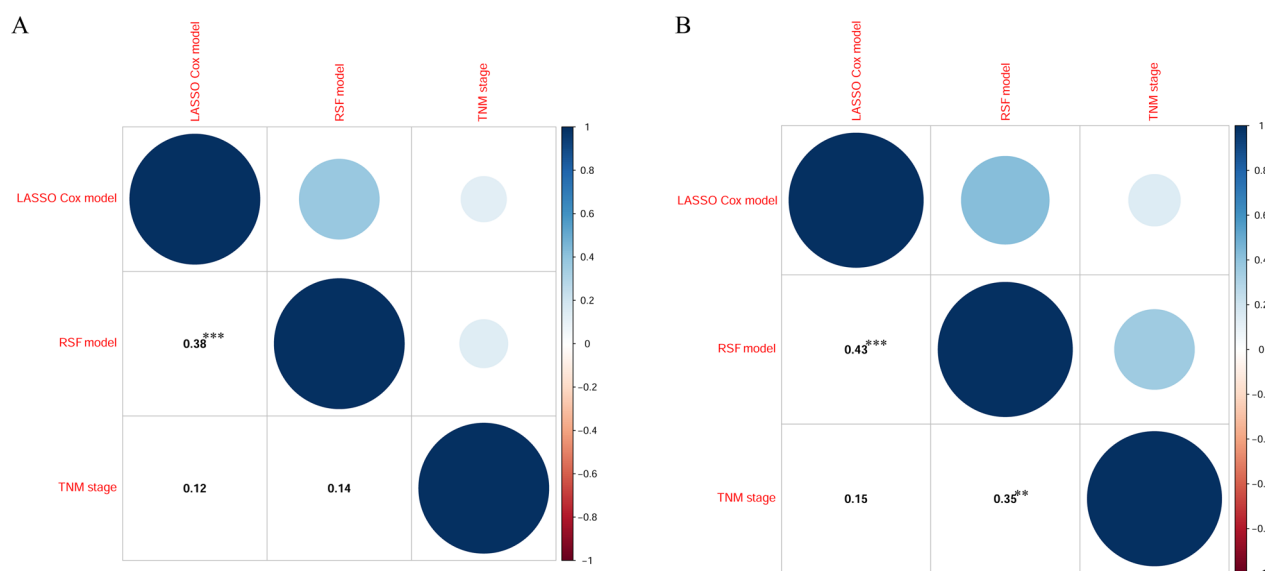
Figure 5. The correlations among the LASSO Cox model, RSF model, and TNM stage in the primary cohort (A) and the validation cohort (B).

models for EC focus on ESCC, and those for NESCC are rare. In the present study, the LASSO Cox and RSF models were developed for patients with NESCC. Moreover, the LASSO Cox and RSF models were found to have better prediction performances than the traditional TNM staging system.

In this study, we constructed LASSO Cox and RSF models for patients with NESCC. The LASSO Cox model included seven factors (TBIL, LHR, NEU, LDCR, LMR BMI and SIRI) and the RSF model included five factors (DBIL, TP, BUN, PLT, and SII). The LASSO Cox model based on the LASSO Cox regression algorithm achieved a C-index of 0.656 (95% CI: 0.580–0.732), IBS of 0.166, and integrated C/D AUC of 0.701. Similarly, the RSF model using the random forest algorithm reached a C-index of 0.841 (95% CI: 0.792–0.889), IBS of 0.141, and integrated C/D AUC of 0.861. In both the primary and validation cohorts, the RSF model achieved a higher AUC than the LASSO Cox model and TNM staging model in terms of 1-, 3-, and 5-year OS of patients with NESCC. According to the respective cutoff values, NESCC patients were divided into low risk and high-risk subgroups. In both the primary and validation cohorts of the LASSO Cox model, patients in the low-risk subgroup had a longer OS than those in the high-risk subgroup. Similar results were observed in the RSF model.

To the best of our knowledge, this is the first study to construct prognostic models for NESCC based on the RSF and LASSO Cox regression. This study developed an RSF prognostic model superior to the LASSO Cox model and the traditional TNM staging system for patients with NESCC. The

advantage of RSF is that it is not constrained by proportional hazards assumption, log-linear assumption and other conditions [19,20]. The prediction by RSF was at least equal to or better than that by traditional survival analysis methods such as the Cox proportional hazards regression model [21–23]. RSF can also perform survival analysis and variable screening of high-dimensional data and can be applied to the analysis of competing risks [24–26]. Moreover, RSF has been widely used for the analysis of time-to-event outcomes and the construction of clinical prediction models [27–31]. Based on interpretable genomic inputs, an RSF model was proposed and used to predict relapse/death in high-risk pediatric patients with acute lymphoblastic leukemia [32]. Zhang et al. constructed an RSF model to predict the 30-day mortality among elderly patients with sepsis. The RSF model showed better predictive ability than the traditional scoring system [33].

In this study, the RSF model included five factors (DBIL, TP, BUN, PLT, and SII). Studies have demonstrated that the levels of DBIL are associated with systemic oxidative stress (SOS), and a low SOS index is an important factor for better prognosis in patient with ESCC [34]. TP and ALB reflect the nutritional status of patients, and nutritional status significantly affects disease progression and survival in patients with EC [35]. The combination of HGB, ALB, LYM, and PLT score (HALP) reflects the status of inflammation and nutrition. A low HALP score has shown to be an independent risk factor for poor OS in patients with ESCC [36]. BUN is a nitrogenous end product of protein metabolism, and it can estimate the function of renal [37].

A previous study found that BUN decreased in patients with advanced oral cancer [38]. SII was an indicator of host inflammatory status. Evidence has shown that inflammatory responses were considered as important factors associated with clinical prognosis in cancers [39,40]. Thus, SII may be a promising predictor of cancer survival and tumor progression in EC. Moreover, elevated SII was significantly associated with deeper infiltration depth, positive lymph node metastasis, and advanced clinical stage [41].

This study has several limitations. First, we constructed multiple-models for patients with NESCC. However, we did not perform a detailed analysis on each histological type (adenocarcinoma, adenosquamous carcinoma, neuroendocrine carcinoma, esophageal sarcomatoid carcinoma, etc.). Second, the data of this study came from a single center and this was a retrospective study. Therefore, potential biases cannot be completely ruled out. Third, this study included a small number of patients with NESCC. We intend to collect data on a larger number of patients from other institutions and validate the RSF model.

In conclusion, we found that the LASSO Cox and RSF models were better than the TNM staging model for predicting OS in patients with NESCC. Moreover, the RSF model showed a superior discrimination ability compare with the LASSO Cox model. Our RSF model has guiding values for clinicians, which will benefit their choice of individualized treatment and follow-up prognosis for patients with NESCC.

Acknowledgments

Jing Chen, Shu-Lin Chen, and Qing-Xia Xu conceived of the study. Shu-Lin Chen, Xiao-Yan Wen, Yong Shen, and Yuan-Ye Qu accessed and verified the data. Ning Xue, Qian Wang, Qing-Xia Xu, and Shu-Lin Chen analyzed the data. Ning Xue, Xiao-Yan Wen, and Jing Chen drafted the manuscript. The review and editing were conducted by Jing Chen, Shu-Lin Chen, Qing-Xia Xu, Ning Xue, Xiao-Yan Wen, and Qian Wang. All authors read and approved the final manuscript.

Authors contributions

CRediT: **Ning Xue**: Formal analysis, Funding acquisition, Writing – original draft, Writing – review & editing; **Xiaoyan Wen**: Data curation, Resources, Writing – original draft, Writing – review & editing; **Qian Wang**: Formal analysis, Funding acquisition, Writing – review & editing; **Yong Shen**: Data curation; **Yuanye Qu**: Data curation; **Qingxia Xu**: Conceptualization, Formal analysis, Writing – review & editing; **Shulin Chen**: Conceptualization, Data curation, Formal analysis, Writing – review & editing; **Jing Chen**: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Ethics statement

This study was approved by the Clinical Research Ethics Committee of Sun Yat-sen University Cancer Center. (protocol code SL-B2024-562-01) and was conducted in accordance with the Declaration of Helsinki.

Consent to participate

This is a retrospective study and the Clinical Research Ethics Committee of Sun Yat-sen University Cancer Center waived the need for informed consent.

Disclosure statement

The authors declare no conflict of interests.

ORCID

Ning Xue http://orcid.org/0000-0003-4748-851X
Shulin Chen http://orcid.org/0000-0003-0816-2664
Jing Chen http://orcid.org/0009-0004-2508-3380

Data availability statement

The datasets analyzed during the current study are not publicly available because of patient privacy concerns, but are available from the corresponding author upon reasonable request.

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–249. doi: 10.3322/caac.21660.

[2] Arnold M, Abnet CC, Neale RE, et al. Global burden of 5 major types of gastrointestinal cancer. Gastroenterology. 2020;159(1):335–349, e315. doi: 10.1053/j.gastro.2020.02.068.

[3] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. doi: 10.3322/caac.21492.

[4] Deng HY, Wang YC, Ni PZ, et al. Long noncoding RNAs are novel potential prognostic biomarkers for esophageal squamous cell carcinoma: an overview. J Thorac Dis. 2016;8(8):E653–659. doi: 10.21037/jtd.2016.07.01.

[5] Napier KJ, Scheerer M, Misra S. Esophageal cancer: a review of epidemiology, pathogenesis, staging workup and treatment modalities. World J Gastrointest Oncol. 2014;6(5):112–120. doi: 10.4251/wjgo.v6.i5.112.

[6] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7–30. doi: 10.3322/caac.21590.

[7] Shapiro J, van Lanschot JJB, Hulshof M, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol. 2015;16(9):1090–1098. doi: 10.1016/S1470-2045(15)00040-6.

[8] Zheng Z, Lin D, Chen Q, et al. Prognostic value of combined detection of preoperative albumin-to-fibrinogen ratio and neutrophil-to-lymphocyte ratio in operable esophageal squamous cell carcinoma patients without neoadjuvant therapy. Cancer Manag Res. 2021;13:2359–2370. doi: 10.2147/CMAR.S296266.

[9] Luo Y, Weng XF, Huang JT, et al. Nomogram constructed by immunological and inflammatory indicators for predicting prognosis of patients with esophageal squamous cell carcinoma treated with neoadjuvant chemoradiotherapy plus surgery. Front Oncol. 2022;12:882900. doi: 10.3389/fonc.2022.882900.

[10] Wang C, Cheng X, Jin L, et al. Development and validation of a nomogram for predicting overall survival to concurrent chemoradiotherapy in patients with locally advanced esophageal squamous cell carcinoma. Biomed Res Int. 2022;2022(1):6455555. doi: 10.1155/2022/6455555.

[11] Feng JF, Wang L, Chen QX, et al. Development and validation of a new integrative score based on various systemic inflammatory and nutritional indicators in predicting prognosis in patients with resectable esophageal squamous cell carcinoma: a retrospective cohort study. Cancer Control. 2022;29:10732748221091394. doi: 10.1177/10732748221091394.

[12] Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Statist Surv. 2011;5(none):44–71. doi: 10.1214/09-SS047.

[13] Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Ann Appl Stat. 2008;2(3):841–860. doi: 10.1214/08-AOAS169.

[14] Hsich E, Gorodeski EZ, Blackstone EH, et al. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circ Cardiovasc Qual Outcomes. 2011;4(1):39–45. doi: 10.1161/CIRCOUTCOMES.110.939371.

[15] Yang R, Chang Q, Meng X, et al. Prognostic value of systemic immune-inflammation index in cancer: a meta-analysis. J Cancer. 2018;9(18):3295–3302. doi: 10.7150/jca.25691.

[16] Lin F, Xia W, Chen M, et al. A prognostic model based on nutritional risk index in operative breast cancer. Nutrients. 2022;14(18):3783. doi: 10.3390/nu14183783.

[17] Okada S, Shimada J, Kato D, et al. Clinical significance of prognostic nutritional index after surgical treatment in lung cancer. Ann Thorac Surg. 2017;104(1):296–302. doi: 10.1016/j.athoracsur.2017.01.085.

[18] Tuzimek A, Dziedzic EA, Beck J, et al. Correlations between acute coronary syndrome and novel inflammatory markers (systemic immune-inflammation index, systemic inflammation response index, and aggregate index of systemic inflammation) in patients with and without diabetes or prediabetes. J Inflamm Res. 2024;17:2623–2632. doi: 10.2147/JIR.S454117.

[19] Zhou Y, McArdle JJ. Rationale and applications of survival tree and survival ensemble methods. Psychometrika. 2015;80(3):811–833. doi: 10.1007/s11336-014-9413-1.

[20] Gerds TA, Kattan MW, Schumacher M, et al. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med. 2013;32(13):2173–2184. doi: 10.1002/sim.5681.

[21] Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. BMC Res Notes. 2017;10(1):459. doi: 10.1186/s13104-017-2775-6.

[22] Li LW, Liu X, Shen ML, et al. Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the SEER database and an external validation cohort. Am J Cancer Res. 2024;14(4):1609–1621. doi: 10.62347/OJTY4008.

[23] Lin J, Yin M, Liu L, et al. The development of a prediction model based on random survival forest for the postoperative prognosis of pancreatic cancer: a SEER-based study. Cancers (Basel). 2022;14(19):4667. doi: 10.3390/cancers14194667.

[24] Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 2010;26(14):1752–1758. doi: 10.1093/bioinformatics/btq257.

[25] Chen Z, Xu HM, Li ZX, et al. [Random survival forest: applying machine learning algorithm in survival analysis of biomedical data]. Zhonghua Yu Fang Yi Xue Za Zhi. 2021;55(1):104–109. doi: 10.3760/cma.j.cn112150-20200911-01197.

[26] Adham D, Abbasgholizadeh N, Abazari M. Prognostic factors for survival in patients with gastric cancer using a random survival forest. Asian Pac J Cancer Prev. 2017;18(1):129–134. doi: 10.22034/APJCP.2017.18.1.129.

[27] Segal MR, Bloch DA. A comparison of estimated proportional hazards models and regression trees. Stat Med. 1989;8(5):539–550. doi: 10.1002/sim.4780080503.

[28] Nasejje JB, Mwambi H, Dheda K, et al. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. BMC Med Res Methodol. 2017;17(1):115. doi: 10.1186/s12874-017-0383-8.

[29] Nachit M, Horsmans Y, Summers RM, et al. AI-based CT body composition identifies myosteatosis as key mortality predictor in asymptomatic adults. Radiology. 2023;307(5):e222008. doi: 10.1148/radiol.222008.

[30] Qi X, Ge Y, Yang A, et al. Potential value of mitochondrial regulatory pathways in the clinical application of clear cell renal cell carcinoma: a machine learning-based study. J Cancer Res Clin Oncol. 2023;149(19):17015–17026. doi: 10.1007/s00432-023-05393-8.

[31] Passera R, Zompi S, Gill J, et al. Explainable machine learning (XAI) for survival in bone marrow transplantation trials: a technical report. BioMedInformatics. 2023;3(3):752–768. doi: 10.3390/biomedinformatics3030048.

[32] Bohannan ZS, Coffman F, Mitrofanova A. Random survival forest model identifies novel biomarkers of event-free survival in high-risk pediatric acute lymphoblastic leukemia. Comput Struct Biotechnol J. 2022;20:583–597. doi: 10.1016/j.csbj.2022.01.003.

[33] Zhang L, Huang T, Xu F, et al. Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). BMC Emerg Med. 2022;22(1):26. doi: 10.1186/s12873-022-00582-z.

[34] Feng J, Wang L, Yang X, et al. The systemic oxidative stress index predicts clinical outcomes of esophageal squamous cell carcinoma receiving neoadjuvant immunochemotherapy. Front Immunol. 2025;16:1535507. doi: 10.3389/fimmu.2025.1535507.

[35] Antasouras G, Papadopoulou SK, Tolia M, et al. May nutritional status positively affect disease progression and prognosis in patients with esophageal and pharyngeal cancers? A scoping review of the current clinical studies. Med Sci (Basel). 2023;11(4):64. doi: 10.3390/medsci11040064.

[36] Hu SJ, Zhao XK, Song X, et al. Preoperative maximal voluntary ventilation, hemoglobin, albumin, lymphocytes and platelets predict postoperative survival in esophageal squamous cell carcinoma. World J Gastroenterol. 2021;27(4):321–335. doi: 10.3748/wjg.v27.i4.321.

[37] Keshet R, Szlosarek P, Carracedo A, et al. Rewiring urea cycle metabolism in cancer to support anabolism. Nat Rev Cancer. 2018;18(10):634–645. doi: 10.1038/s41568-018-0054-z.

[38] Caruntu A, Moraru L, Ciubotaru DA, et al. Assessment of serum urea, creatinine and uric acid in oral cancer. J Clin Med. 2022;11(12):3459. doi: 10.3390/jcm11123459.

[39] Hu B, Yang XR, Xu Y, et al. Systemic immune-inflammation index predicts prognosis of patients after curative resection for hepatocellular carcinoma. Clin Cancer Res. 2014;20(23):6212–6222. doi: 10.1158/1078-0432.CCR-14-0442.

[40] Diakos CI, Charles KA, McMillan DC, et al. Cancer-related inflammation and treatment effectiveness. Lancet Oncol. 2014;15(11):e493-503–e503. doi: 10.1016/S1470-2045(14)70263-3.

[41] Zhang Y, Xiao G, Wang R. Clinical significance of systemic immune-inflammation index (SII) and C-reactive protein-to-albumin ratio (CAR) in patients with esophageal cancer: a meta-analysis. Cancer Manag Res. 2019;11:4185–4200. doi: 10.2147/CMAR.S190006.