Heliyon 10 (2024) e32788

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The development of a prediction model based on random survival forest for the prognosis of non- Hodgkin lymphoma: A prospective cohort study in China

Xiaosheng Li [a], Zailin Yang [a], Jieping Li [a], Guixue Wang [b], Anlong Sun [a], Ying Wang [a], Wei Zhang [a,**], Yao Liu [a,***], Haike Lei [c,*]

[a] *Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, Chongqing, 400030, China*

[b] *MOE Key Lab for Biorheological Science and Technology, State and Local Joint Engineering Laboratory for Vascular Implants, College of Bioengineering Chongqing University, Chongqing, 400030, China*

[c] *Chongqing Cancer Multi-omics Big Data Application Engineering Research Center, Chongqing University Cancer Hospital, Chongqing, 400030, China*

ARTICLE INFO

Keywords:
Machine learning
Non-Hodgkin lymphoma
Survival analysis
Prognosis

ABSTRACT

Background and objective: The pathological staging of non-Hodgkin lymphoma (NHL) is complex, the clinical manifestations are varied, and the prognosis differ considerably. To provide a useful reference for early detection and effective treatment of NHL, we developed a random survival forest (RSF) prognostic model based on machine learning (ML) algorithms using prospective cohort data collected from Chongqing Cancer Hospital from Jan 1, 2017 to Dec 31, 2019 (n = 1449) to compare with the traditional cornerstone method Cox proportional hazards (CPH) model and evaluate the predictability of the model.

Methods: Patients were randomly split into a training cohort (TC) and validation cohort (VC) based on 65/35 ratio. The least absolute shrinkage and selection operator (LASSO) regression analysis was used to extracted the important features. And the RSF was modeled to explore the prognostic factors impacting the overall survival (OS) of patients with NHLs in the TC and validated in the VC. The C-index, the Integrated Brier Score (IBS), Kaplan-Meir method, the receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC) were selected to measure performances and discriminations of the models. In addition, individual survival probability predicted for NHL patients.

Results: According to the features extracted by LASSO model and univariable Cox model, 16 variables were selected to develop the RSF model with log-rank splitting rule, which were age, ethnicity, medical insurance, Ann Arbor stage, pathology, targeted-therapy, chemo-therapy, peripheral blood neutrophil count to lymphocyte count ratio (NLR), peripheral blood platelet count to lymphocyte count ratio (PLR), serum lactate dehydrogenase (LDH), CD4/CD8, platelet (PLT), absolute neutrophil count (ANC), lymphocyte (LYM), B-symptoms, and (CPR) were important prognostic factors. Compared to the CPH model (C-index = 0.748, IBS = 0.166), the RSF model (C-index = 0.786, IBS = 0.165) is outperformed in predictability and accuracy. The AUC of the

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: cqchzhangwei@163.com (W. Zhang), liuyao77@cqu.edu.cn (Y. Liu), tohaike@163.com (H. Lei).

https://doi.org/10.1016/j.heliyon.2024.e32788
Received 10 July 2023; Received in revised form 7 June 2024; Accepted 10 June 2024
Available online 19 June 2024

RSF model to estimate the 1-, 3-, and 5-year OS in TC were 0.847, 0.847, and 0.809, respectively; while those in the CPH were 0.816, 0.803, and 0.750, respectively.
Conclusions: To provide practical implications for the implementation of individualized therapy, the study constructed a high-performed RSF model and reveal that it outperformed the traditional model CPH. And the RSF model ranked the risk variables. In addition, we stratified the risk of NHL patients and estimated individual survival probability based on the RSF model.

1. Introduction

Non-Hodgkin lymphoma (NHL) is lymphatic malignancy with a variety of biological and clinical behaviors [1]. NHL is characterized by rapid disease progression, multiple nodal invasions, and aggressiveness, causing severe threats to the survival of patients [2]. In China, the incidence of NHL is increasing annually [3], and the prognosis of patients with NHLs depends on the specific subtype of lymphoma, the stage, the immune dysfunction and the age [4–7]. Either indolent or aggressive NHL targeted therapy such as combination chemotherapy provides positive survival prospects for patients with NHLs [8,9].

The accurate prediction could help physicians deliver appropriate treatment plans to patients with NHLs, improving their opportunities for survival and alleviating suffering. The main prognostic prediction tool currently available for NHL is Cox proportional hazards (CPH) to explore the correlation between survival time and covariates. Wang et al. explored the links between T2DM, metformin, and the NHL in Women's Health Initiative based on multivariable-adjusted CPH models [10]. Furthermore, Lu et al. developed univariate and multivariate CPH models to predict overall survival (OS) in primary gastric diffuse large B-cell lymphoma, a kind of NHL, using 1716 patients from the SEER database [11]. However, the restrictive assumptions and unstably-diagnostic efficiency limit models' performance.

Several methods in machine learning (ML) have been employed to predict the prognosis of various cancers. The most commonly used ML for survival analysis is the random survival forest (RSF). RSF is trained with a large number of survival trees, and the ultimate prediction is weighted and elected from among the individual trees with a voting procedure. With the advantages of strong generalization, fast training, and great performance, the RSF model is an excellent choice to be considered for survival analysis. Li et al. applied RSF to structured electronic health record data to predict the survival of patients with follicular lymphoma (FL), one subtype of NHL [12]. However, to our knowledge, no studies have been reported on the application of RSF in prognostic models of NHL (both aggressive and inert NHL).

There is still controversy regarding which ML and traditional approaches can perform better in survival analysis. As we know, no studies have investigated the differences between CPH and RSF for modeling comparisons in the OS of NHL patients. Therefore, we used retrospectively collected demographic variables, therapeutic methods, clinical characteristics, and inflammatory indexes of patients with NHLs to develop CPH and RSF models to predict OS in the population and compared the model performance of the two. Moreover, this study provides clinical practitioners with risk stratification and individual prediction based on the RSF.

The aim of this study was to establish an ISF prognostic model for overall survival of patients with NHLS based on a variety of clinical characteristics, fully considering new clinical prognostic factors, to guide individualized treatment and management.

2. Methods and materials

2.1. Study population and data Source

All observers collected in this research were first diagnosed with NHLs at Chongqing University Cancer Hospital (CUCH) between Jan 1, 2017 and Dec 31, 2019 with histologically confirmed to have primary NHLs and were no younger than 18 years old. The exclusion criteria were uncompleted the entire treatment course of chemotherapy (CT), targeted therapy (TT), and Immuno-therapy (IT), unprovided complete baseline, relevant laboratory, and follow-up information.

Data including demographics, therapeutic methods, clinical characteristics, and inflammatory indexes of patients with NHLs were extracted from the CUCH tumor database platform. NHL was staged using the Ann Arbor staging system and pathology (diffuse large B-cell lymphoma (DLBCL), Marginal zone lymphomas (NZL), follicular lymphoma (FL), chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL), mantle cell lymphoma (MCL), and others) was collected. Demographic variable collected included sex, age, ethnicity, and medical insurance types. The therapeutic method included CT, TT, and IT. Clinical characteristics were considered in this study including peripheral blood neutrophil count to lymphocyte count ratio (NLR), peripheral blood platelet count to lymphocyte count ratio (PLR), serum lactate dehydrogenase (LDH), platelet (PLT), absolute neutrophil count (ANC), lymphocyte (LYM), and B-symptoms. Inflammatory indexes included CD4/CD8 and C-reactive protein (CRP). Furthermore, the continuous variables such as age and clinical characteristics cut to categorical variables based on the cutoff point calculated by X-tile, which were NLR ($\leq$7.09, >7.09), PLR ($\leq$72.19, >72.19), LDH ($\leq$245, >245), PLT ($\leq$300, >300), ANC ($\leq$2.02, >2.02), LYM ($\leq$0.67, >0.67), CD4/CD8 ($\leq$0.58, >0.58), and CPR ($\leq$10, >10). The OS of a patient with NHL was calculated from the day the NHL was diagnosed until the date the patient died or the date the final follow-up was conducted, whichever happened first.

The study's inclusion criteria were as follows [1]: age $\geq$18 years [2]; At least one hospitalization record [3]; Newly diagnosed NHLs (according to ICD-*O*-3 oncology codes) [4]; no previous history of other malignant tumors. Exclusion criteria included [1]: death within 48 h after admission [2]; significant missing and incomplete clinical data (such as treatment and blood test data). The study's

flowchart is illustrated in Fig. 1. The present study was performed according to the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of The Chongqing University Cancer Hospital. Written informed consent was obtained from all subjects (Approval Number -CZLS2021252-A)

2.2. The development of models

The models' construction and validation were fellow and reported the TRIPOD guidelines [13]. The patients were assigned randomly to the training cohort (TC) and the validation cohort (VC). Aiming to avoid over-fitting and find the essential variables, the least absolute shrinkage and selection operator (LASSO) regression analysis was modeled to select the features before the model development. And based on the optimal features extracted by LASSO, the univariate and multivariate CPH models and RSF model were used to identify the vital prognostic predictors.

Construction of the CPH model: According to the LASSO analysis results of features in the TC, the univariate Cox prognostic prediction model was developed. Considering the results of univariate analysis, LASSO analysis and the clinical significance, the multivariable Cox model was constructed.

Construction of the RSF model: The RSF model is not required to identify the distribution of the parameters beforehand, and it can evaluate the predictive power of each predictor variable for the outcome by ranking its importance; concurrently, it can evaluate the prediction error rate using internal cross-validation and ensure a relatively high level of accuracy. The splitting rule is a crucial part of the survival tree and is essential to the survival forest's performance [14]. In this study, four splitting criterion approaches, respectively log-rank, log-rank score, bs-gradient, and random, were implemented to model a random survival forest, and the optimal splitting rules were determined using prediction error and Integrated Brier Score (IBS).

2.3. The evaluation of models

The models' performance was evaluated in the TT and VT. The following evaluation indexes were adopted: the concordance index (C-index), the 1-, 3-, and 5-year receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC). IBS is a calibration indicator; the lower the value, the higher the model's accuracy. The Kaplan-Meier (K-M) method was used to assess survival risk, and the differences of risk groups in survival were carried out through a stratified log-rank test. The individual prognosis was

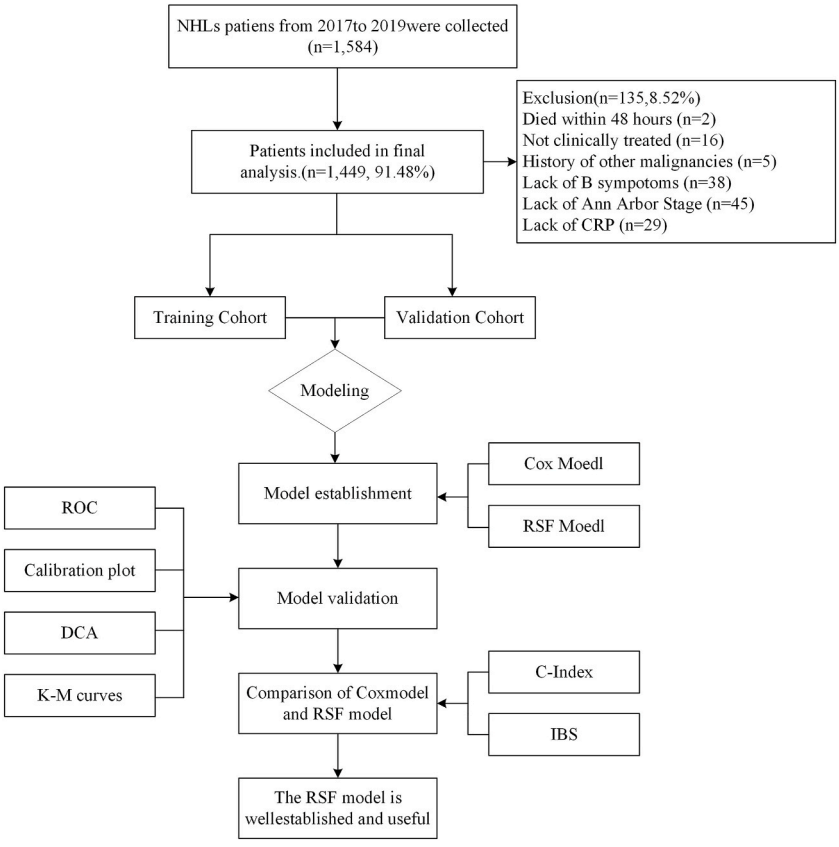

Fig. 1. Flow diagram of study design.

carried out using the model with the highest predictive accuracy and precision for individualized survival probability estimation.

2.4. Statistical analysis

The data analysis was conducted with R version 4.2.1 (Institute for Statistics and Mathematics, Austria). The R packages 'randomForest' (version 4.6.14) were utilized for developing and validating the RSF models. And the packages 'ezcox' (version 1.2.0) and 'rms' (version 5.0.1) were employed for development and evaluation of the CPH model. Besides, the discrimination, calibration of the models were measured via the packages 'pec' (version 2022.05.04), 'survival' (version 3.3–1), and 'timeROC' (version 0.4). The statistical significance of the two-sided p was set at ≤ 0.05.

3. Results

3.1. The characteristics of patients

The study included 1449 patients with NHLs who enrolled in follow-up visit between Jan 1, 2017, and Dec 31, 2022, from the CUCH tumor database platform. The entire cohort was randomly split into two exclusive set, which were training set of 1087 patients (75 %) and 362 (25 %) patients, respectively. And there were no significant differences between the two cohorts (Table 1.). The median survival time was 37.53 (0.10–133.70) months for the overall cohort. Table 1 summarizes the characteristics of patients. Overall, the considerable number of patients were below 55 years of age (651, 44.93 %), Han (1413, 97.52 %), medical insurance with residents (795, 54.87 %) male (886, 61.15), with premetastatic (1493,91.31 %), stage IV (556, 38.37 %) and pathological performance status of

Table 1
The clinical and demographic characteristics for NHL patients in the TC and VC.

Variables	Characteristics	Training set (n = 1087)	Validation Set (n = 362)	Overall (n = 1449)	P Value
Sex (%)	Female	424 (39.01)	139 (38.40)	563 (38.85)	0.886
	Male	663 (60.99)	223 (61.60)	886 (61.15)	
Age (%)	≤55	500 (46.00)	151 (41.71)	651 (44.93)	0.105
	56–65	289 (26.59)	117 (32.32)	406 (28.02)	
	≥66	298 (27.41)	94 (25.97)	392 (27.05)	
Ethnicity (%)	Han	1056 (97.15)	357 (98.62)	1413 (97.52)	0.173
	Others	31 (2.85)	5 (1.38)	36 (2.48)	
Medical Insurance (%)	Residents	608 (55.93)	187 (51.66)	795 (54.87)	0.341
	Employees	325 (29.90)	116 (32.04)	441 (30.43)	
	Others	154 (14.17)	59 (16.30)	213 (14.70)	
Ann Arbor Stage (%)	I-II	359 (33.03)	132 (36.46)	491 (33.89)	0.455
	III	308 (28.33)	94 (25.97)	402 (27.74)	
	IV	420 (38.64)	136 (37.57)	556 (38.37)	
Pathology (%)	DLBCL	580 (53.36)	203 (56.08)	783 (54.04)	0.588
	NZL	119 (10.95)	33 (9.12)	152 (10.49)	
	FL	82 (7.54)	25 (6.91)	107 (7.38)	
	CLL/SLL	49 (4.51)	10 (2.76)	59 (4.07)	
	MCL	54 (4.97)	18 (4.97)	72 (4.97)	
	OTHERS	203 (18.68)	73 (20.17)	276 (19.05)	
Targeted-therapy (%)	NO	719 (66.15)	254 (70.17)	973 (67.15)	0.178
	YES	368 (33.85)	108 (29.83)	476 (32.85)	
Immuno-therapy (%)	NO	1082 (99.54)	361 (99.72)	1443 (99.59)	1.000
	YES	5 (0.46)	1 (0.28)	6 (0.41)	
Chemotherapy (%)	NO	749 (68.91)	247 (68.23)	996 (68.74)	0.862
	YES	338 (31.09)	115 (31.77)	453 (31.26)	
NLR (%)	≤7.09	981 (90.25)	328 (90.61)	1309 (90.34)	0.922
	>7.09	106 (9.75)	34 (9.39)	140 (9.66)	
PLR (%)	≤72.19	150 (13.80)	44 (12.15)	194 (13.39)	0.480
	>72.19	937 (86.20)	318 (87.85)	1255 (86.61)	
LDH (%)	≤245	633 (58.23)	218 (60.22)	851 (58.73)	0.546
	>245	454 (41.77)	144 (39.78)	598 (41.27)	
CD4/CD8 (%)	≤0.58	129 (11.87)	39 (10.77)	168 (11.59)	0.640
	>0.58	958 (88.13)	323 (89.23)	1281 (88.41)	
PLT (%)	≤300	246 (22.63)	70 (19.34)	316 (21.81)	0.215
	>300	841 (77.37)	292 (80.66)	1133 (78.19)	
ANC (%)	≤2.02	114 (10.49)	42 (11.60)	156 (10.77)	0.621
	>2.02	973 (89.51)	320 (88.40)	1293 (89.23)	
LYM (%)	≤0.67	116 (10.67)	38 (10.50)	154 (10.63)	1.000
	>0.67	971 (89.33)	324 (89.50)	1295 (89.37)	
B-symptoms (%)	NO	1048 (96.41)	350 (96.69)	1398 (96.48)	0.937
	YES	39 (3.59)	12 (3.31)	51 (3.52)	
CRP (%)	≤10	839 (77.18)	276 (76.24)	1115 (76.95)	0.767
	>10	248 (22.82)	86 (23.76)	334 (23.05)	

DLBCL (783, 54.04 %). With regard to therapy, the majority of patients refused targeted therapy (973, 67.15 %), immune-therapy (1443,99.59 %), nor chemotherapy (996, 68.74 %) (Table 1.).

3.2. Survival analysis of the entire cohort of patients

A Kaplan-Meier survival curve was developed to explore the difference of OS between training and testing datasets. The results revealed that statistical difference in OS was absent between the two sets as well (Fig. 2.)

3.3. The development of models

3.3.1. Feature selection and Cox model construction

The LASSO regression analysis was developed to select features based on the 10-fold cross-validation to find the best λ. The 16 vital variables with non-zero coefficients were selected as the prognostic variables when $\lambda = 0.01$, which were: Sex, Age, Ethnic, Medical insurance, Ann Arbor Stage, pathologic-type, targeted-therapy, chemo-therapy, NLR, PLR, LDH, CD4/CD8, PLT, ANC, LYM, B-symptoms, and CPR (Fig. 3(A and B)).

The links between variables and OS were investigated to determine the independent risk factors. Age, Ann Arbor stage, pathologic.-type, targeted-therapy, chemo-therapy, NLR, PLR, LDH, CD4/CD8, PLT, ANC, LYM, CRP were significant predictors of OS on univariable analysis(Fig. 4A). Considering the results of univariate analysis, LASSO analysis and the clinical significance, the multivarible Cox model were developed. On the multivariable Cox analysis, the following variables demonstrated as independent predictors of OS: age (56 − 65 vs. ≤ 55, HR: 1.35; 95 %, CI: 1.08–1.68; ≥ 66 vs. ≤ 55, HR: 2.54, CI: 2.07–3.13), Ann Arbor stage (IV vs.I-II,HR: 1.83; 95 % CI: 1.46–2.29), pathologic-type (CLL/SLL vs. DLBCL HR: 0.51, 95 % CI: 0.31–0.83; MCL vs. DLBCL HR: 1.50, 95 % CI: 1.05–2.14), targeted-therapy (HR: 0.73; 95 % CI: 0.58–0.92), chemo-therapy (HR: 0.73; 95 % CI: 0.58–0.92), NLR (HR: 1.15; 95 % CI: 0.86–1.55), PLR (HR: 0.77; 95 % CI: 0.60–0.99), LDH (HR: 1.87; 95 % CI: 1.56–2.24), CD5/CD8 (HR: 0.8295 % CI: 0.64–1.05),PLT (HR: 0.91; 95 % CI: 0.74–1.10), ANC (HR: 0.64; 95 % CI: 0.49–0.83), LYM (HR: 0.77; 95 % CI: 0.56–1.02),and CRP (HR: 1.26; 95 % CI: 0.83–1.89) (Fig. 4B).

3.3.2. RSF model

To find the most suitable split rule for the RSF model, the model with four types of split rule, which were log-rank, log-rank score, bs-gradient, and random, were constructed. And the RSF with log-rank split rule was the best model based on the lowest prediction error (0.272) and IBS (0.165) among four splitting rules (Table 3). According to the features extracted by LASSO regression model and univariable Cox model, 16 variables were selected to develop the RSF model of log-rank splitting rule with combination of minimum depth method and variable importance (VIMP). Table 2 shows the value of the minimal depth and VIMP. Fig. 5(A) illustrates that as the number of survival trees increases, their out-of-bag (OOB) prediction error rate decreases noticeably, and the OOB error rate plateaus when the survival tree reaches 450, and displays the importance ranking of the variables. The result demonstrated that Age, Ethnic, Medical insurance, Ann Arbor Stage, pathologic-type, targeted-therapy, chemo-therapy, NLR, PLR, LDH, CD4/CD8, PLT, ANC, LYM, B-symptoms, and CPR were important prognostic factors (Fig. 5(B)).

3.4. The evaluation and interpretation of the models

The prediction performance of different models was compared in the training cohort and validation cohort with C-index, error rate, IBS, and AUC. The results of the C-index, error rate, and IBS of models are shown in Table 3. The 5 models' IBS were all under 0.25, indicating that all models had great calibration. And the RSF with log-rank splitting rule model outperformed the Cox model, according

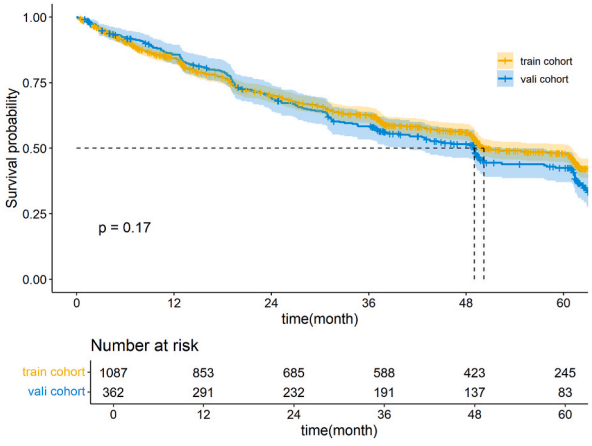

Fig. 2. Kaplan-Meier survival curved of overall survival for the TC and VC.

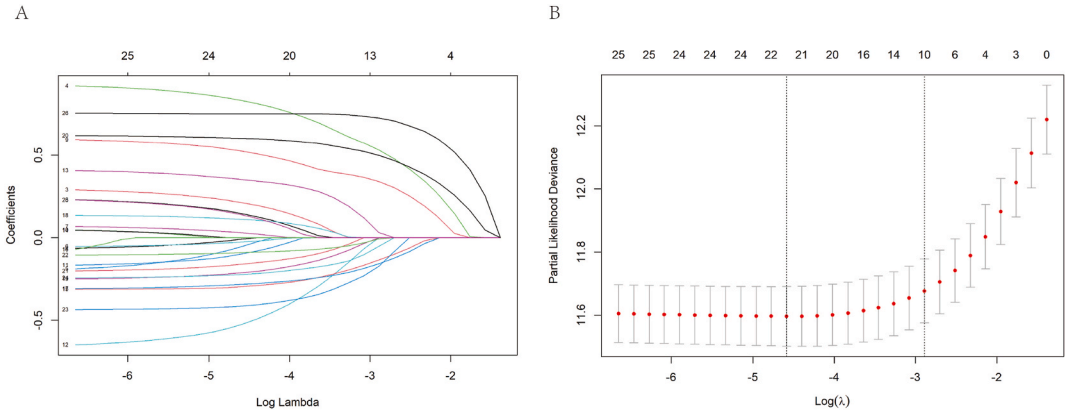

Fig. 3. Identification of important features for OS in NHLs patients with LASSO. (A) LASSO coefficient profiles of the expression of 17 variables. (B) Selection of the λ in the LASSO analysis via 10-fold cross-validation.

A

Univariate Model

Variable	n	HR (95% CI)	p.value
Sex			
Female	424		
Male	663	1.160 (0.97 to 1.37)	0.101
Age			
≤55	500		
56-65	289	1.340 (1.08 to 1.66)	0.007
≥66	298	2.490 (2.05 to 3.02)	<0.001
Ethnic			
Han	1056		
Others	31	0.644 (0.35 to 1.17)	0.149
Medical insurance			
Residents	608		
Employees	325	0.844 (0.69 to 1.03)	0.09
Others	154	1.120 (0.89 to 1.42)	0.323
Ann Arbor Stage			
I-II	359		
III	308	1.640 (1.30 to 2.07)	<0.001
IV	420	2.430 (1.97 to 3.01)	<0.001
Pathologic-type			
DLBCL	580		
NZL	119	0.983 (0.75 to 1.30)	0.905
FL	82	0.691 (0.47 to 1.01)	0.054
CLL/SLL	49	0.584 (0.37 to 0.93)	0.023
MCL	54	1.610 (1.15 to 2.24)	0.006
OTHERS	203	1.090 (0.88 to 1.35)	0.419
Targeted-therapy			
NO	719		
YES	368	0.571 (0.47 to 0.69)	<0.001
Immuno-therapy			
NO	1082		
YES	5	0.278 (0.04 to 1.98)	0.201
Chemo-therapy			
NO	749		
YES	338	0.547 (0.44 to 0.67)	<0.001
NLR			
≤7.09	981		
>7.09	106	1.670 (1.29 to 2.14)	<0.001
PLR			
≤72.19	150		
>72.19	937	0.695 (0.56 to 0.87)	0.001
LDH			
≤245	633		
>245	454	2.470 (2.09 to 2.92)	<0.001
CD4/CD8			
≤0.58	129		
>0.58	958	0.644 (0.51 to 0.82)	<0.001
PLT			
≤300	246		
>300	841	0.656 (0.54 to 0.79)	<0.001
ANC			
≤2.02	114		
>2.02	973	0.657 (0.51 to 0.84)	0.001
LYM			
≤0.67	116		
>0.67	971	0.517 (0.41 to 0.65)	<0.001
B-symptoms			
NO	1048		
YES	39	1.150 (0.78 to 1.71)	0.476
CRP			
≤10	839		
>10	248	3.270 (2.74 to 3.90)	<0.001

B

Multivariable Model

Variable	n	HR (95% CI)	p.value
Age			
≤55	500		
56-65	289	1.351 (1.08 to 1.68)	0.007
≥66	298	2.543 (2.07 to 3.13)	<0.001
Ethnic			
Han	1056		
Others	31	0.805 (0.44 to 1.48)	0.149
Medical insurance			
Residents	608		
Employees	325	0.940 (0.77 to 1.15)	0.09
Others	154	1.072 (0.84 to 1.36)	0.323
Ann Arbor Stage			
I-II	359		
III	308	1.270 (1.00 to 1.61)	0.05
IV	420	1.826 (1.46 to 2.29)	<0.001
Pathologic-type			
DLBCL	580		
NZL	119	1.060 (0.79 to 1.41)	0.6916
FL	82	0.837 (0.57 to 1.23)	0.3616
CLL/SLL	49	0.511 (0.31 to 0.83)	0.007
MCL	54	1.497 (1.05 to 2.14)	0.0265
OTHERS	203	1.051 (0.83 to 1.33)	0.6742
Targeted-therapy			
NO	719		
YES	368	0.733 (0.58 to 0.92)	<0.001
Chemo-therapy			
NO	749		
YES	338	0.729 (0.58 to 0.92)	<0.001
NLR			
≤7.09	981		
>7.09	106	1.153 (0.86 to 1.55)	<0.001
PLR			
≤72.19	150		
>72.19	937	0.772 (0.60 to 0.99)	<0.001
LDH			
≤245	633		
>245	454	1.871 (1.56 to 2.24)	<0.001
CD4/CD8			
≤0.58	129		
>0.58	958	0.823 (0.64 to 1.05)	<0.001
PLT			
≤300	246		
>300	841	0.906 (0.74 to 1.10)	<0.001
ANC			
≤2.02	114		
>2.02	973	0.636 (0.49 to 0.83)	<0.001
LYM			
≤0.67	116		
>0.67	971	0.774 (0.59 to 1.02)	<0.001
B-symptoms			
NO	1048		
YES	39	2.123 (1.75 to 2.57)	0.476
CRP			
≤10	839		
>10	248	1.255 (0.83 to 1.89)	<0.001

Fig. 4. Forest plots of univariate (A) and multivariate (B) LASSO-Cox model for OS in the training cohort.

Table 2
Minimal depth, VIMP and relative importance for patients with NHLs.

Variable	Minimal depth	Importance
CRP	1.444	0.125
LDH	1.756	0.073
Age	1.776	0.079
Ann Arbor Stage	2.154	0.041
Pathology	3.038	0.021
Chemo-therapy	3.414	0.013
Targeted-therapy	3.524	0.013
Medical insurance	3.530	0.006
ANC	3.798	0.013
LYM	3.830	0.006
PLR	4.458	0.004
CD4/CD8	4.592	0.007
NLR	4.676	0.008
PLT	4.764	0.001
B-symptoms	5.566	0.006
Ethnicity	5.802	0.006

Table 3
Evaluation indications of the models used.

Model	Training cohort			Validation cohort		
	C-index	Error rate	IBS [0; 124.6]	C-index	Error rate	IBS [0; 124.6]
Cox model	0.748	0.252	0.166	0.720	0.280	0.172
RSF (log-rank)	0.786	0.272	0.165	0.721	0.297	0.172
RSF (bs-gradient)	0.774	0.272	0.166	0.724	0.304	0.174
RSF (log-rank-score)	0.772	0.273	0.169	0.729	0.300	0.175
RSF (random)	0.753	0.273	0.171	0.720	0.301	0.176

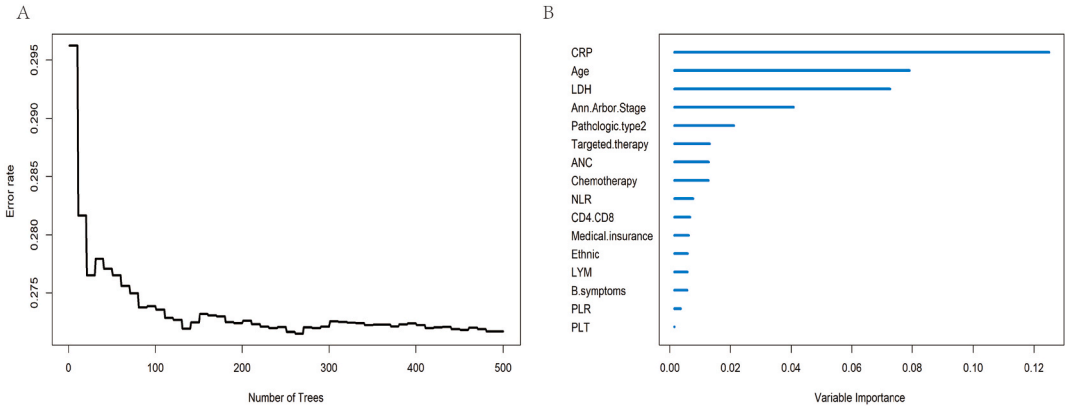

Fig. 5. Out-of-bag error rate of RSF(A) and variable importance of RSF(B).

to its higher C-index (0.786) and lower IBS (0.165) in the training cohort. Fig. 6(A-D) illustrated the AUC of the models for OS in 1-, 3-, and 5- year. And the plots show that the AUC of RSF with the log-rank splitting rule was the highest at all year in training and validation cohort.

3.5. The risk stratification of patients

According to the predictive risk scores of the RSF model with log-rank splitting rule, the research used K-M survival curves to stratify the patients into two risk groups by the threshold calculated based on a maximally selected rank statistic. The results of K-M survival analyses for the two models found that NHLs patients in low-risk groups had significantly longer OS than the patients with high-risk scores in the TC and VC (Fig. 7(A and B)). And the results of log-rank test between the high- and low-risk group demonstrated that significant difference between the two groups were existed in the RSF model.

We validated the applicability of the RSF model to different pathological subtypes of NHLs using data from patients with diffuse large B-cell lymphoma, marginal zone lymphoma, and follicular lymphoma. The results showed that the C-index of the RSF model in

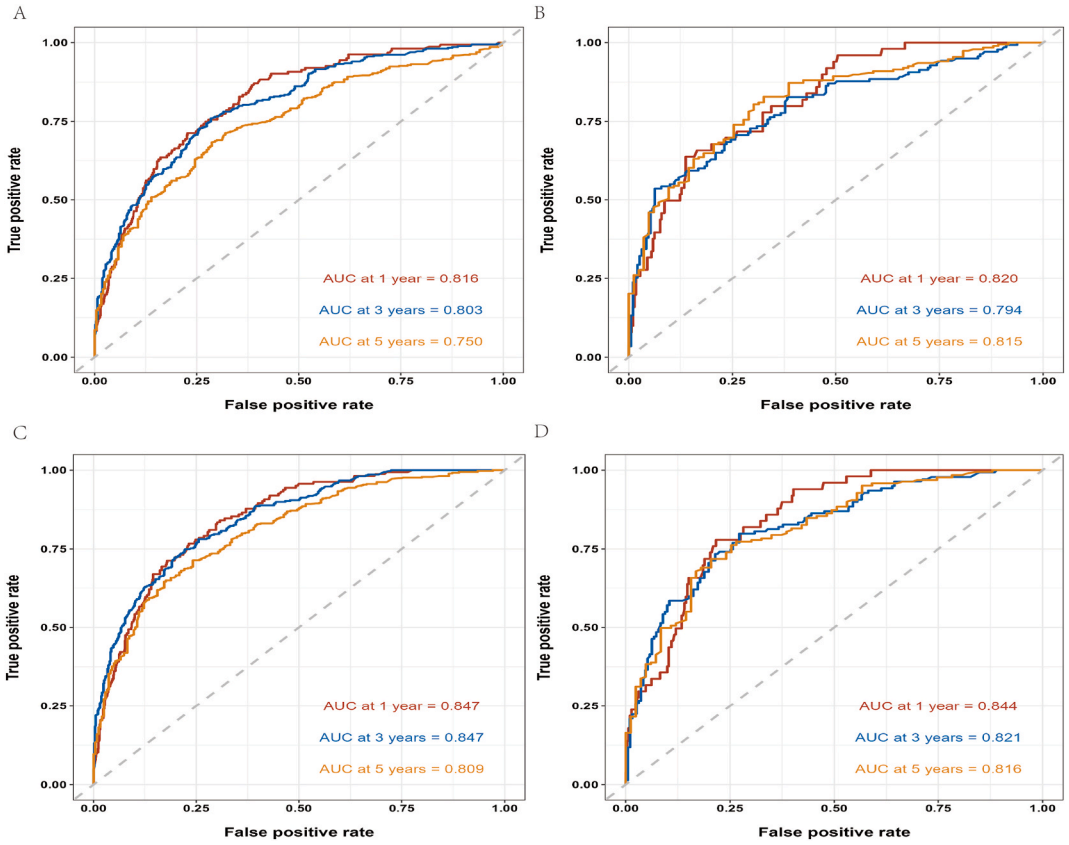

Fig. 6. ROC curve of each model in the TC and VC. (A) ROC curve of Cox model in TC; (B) ROC curve of Cox model in VC; (C) ROC curve of RSF model in TC; (D) ROC curve of RSF model in VC.

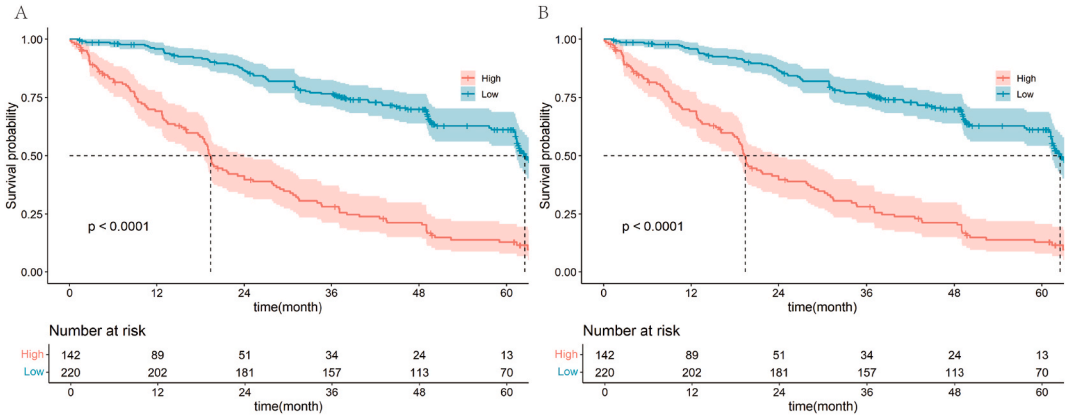

Fig. 7. K-M curves depicting OS of low-risk versus high-risk group in TC (A) and VC (B).

the three pathological subtypes was 0.728 (0.704–0.753), 0.797 (0.750–0.843), and 0.781 (0.716–0.845), respectively, indicating that the model constructed in this study has good predictive accuracy in different pathological subtypes and can effectively identify patient risks.

3.6. The probability of OS for each individual based on RSF model

Considering the model's effectiveness and accuracy and the scope of application, the random survival forest model is superior to the Cox model. Therefore, the results of the random survival forest to predict the probability of OS for each patient with NHL. Fig. 8(A and

B) presented the prognostic survival curves for 10 randomized patients with NHLs.

4. Discussion

The significance of prognostic prediction for clinical and personal decision-making in patients with NHLs cannot be overstated. The CPH and RSF models are the most widely used prognostic modeling approaches for cancers, and both can handle survival data. This study's primary purpose is to develop the RSF model with the optimal splitting rule and compare its performance in predicting OS with that of the traditional statistical model CPH. In terms of accuracy, precision, and generalizability, the results revealed that the RSF model is more appropriate for individual prognosis prediction of patients with NHLs. CPR, age, LDH, Ann Arbor stage, and pathology were the top five most influential risk factors, as determined by the RSF model with the log-rank split rule based on the results of VIMP. In addition, a model for individual prognostic prediction was developed based on the RSF model with log-rank rule to benefit clinicians and patients in actual practice.

The RSF is based on randomness utilizing bootstrap and splitting rules to develop the survival tree, which will be the set of all trees into RSF. Not only does the RSF model not rely on p-values, but it also permits prognostic risk prediction of survival data and selection of significant factors based on the linear or nonlinear correlations between the variables evaluated in the data. RSF with a non-parametric structure is gradually becoming the preferred predictive prognostic approach, given its superior accuracy, computing speed, and generalization capacity, as well as its ability to overcome the CPH model's inability to meet certain assumptions. In addition, as demonstrated by the present study, the RSF model has provided excellent results in both model fitting performance and interpretation as the lower prediction error (0.272) and IBS (0.165) than CPH.

Based on the RSF model with the log-rank splitting rule, CRP has a considerable influence on patient prognosis and is an independent risk factor for NHL patients, in accordance with previous clinical results [15–17]. CRP, as a valuable and easy prognostic biomarker of NHL [18], is an acute time-phase response protein whose synthesis is regulated by pro-inflammatory cytokines, and its serum level is extremely low under normal conditions [19]. As NHL progresses, the patient's immune function becomes defective, resulting in dysregulation of the body's immune response, which causes malignant tumor cells to multiply rapidly and secrete enormous amounts of CRP to promote tumor infiltration and metastasis.

Meanwhile, the RSF model identifies clinical characteristics as potential prognostic factors for NHL patients, including LDH, NLR, PLR, PLT, ANC, and LYM [20]. [21] [22]. [23].In addition, LDH, which ranked third according to the RSF model, was investigated further to confirm that high levels of LDH had a negative impact on survival in NHL patients and a substantial prognostic ability in the OS [24]. The Ann Arbor Staging Classification is used routinely to classify the extent of disease and remains the best method available for the anatomic staging of NHL, although shortcomings exist [25]. Consistent with the findings in the literature, our study discovered that the Ann Arbor stage provided predictive value [26], indicating that early detection is crucial for the treatment of NHL patients. Indeed, different survival outcomes differ by histological subtype [8] and are well-recognized as highly prognostic. Prognosis depends greatly on the type of NHL, with differences in prognosis and clinical presentation among subtypes, and usually aggressive NHL has a worse prognosis and shorter survival than inert NHL [27], with differences in prognosis and clinical presentation between subtypes; aggressive NHL often has a poorer prognosis and shorter survival than inert NHL. Moreover, the data demonstrated a decline in patient survival with older ages. This condition may be associated with immunological insufficiency and somatic decline in elderly people. Regarding the prognosis of patients with certain risk factors, clinicians should pay closer attention.

Notably, the B symptom (fever $> 38^{\circ}C$ for three consecutive days, weight loss exceeding 10 % of body weight in 6 months, night sweats) is a newly updated predictor to the RSF model, indicating that it is a poor prognostic factor for NHL. However, this marker is excluded from the standard CPH model. Studies have revealed that individuals with B symptoms have a poor prognosis [28,29]. Increasing evidence that RSF is superior to CPH.

This study predicts the survival probability of NHL patients based on the final RSF model and visualizes their survival probability, effectively providing a more accurate perspective and visualization of patients' prognoses. In addition, it emphasizes the significance of prompt diagnosis, practical treatment approaches for different subtypes, and personalized medicine improves prognosis.

Our study is not devoid of limitations. The study's potential inclusion/exclusion criteria bias and absence of an external validation cohort are limitations. However, internal validation was undertaken using bootstrap approaches to evaluate the model's discriminatory ability and accuracy and to sustain their performance. However, subsequent external validation is still needed to verify its generalizability.

5. Conclusion

NHL is a malignant immune system tumor whose prognosis varies significantly among pathological subtypes. Accurate prognostication is clinically helpful for developing appropriate treatment options, improving outcomes, and minimizing treatment-related toxicities. To predict the prognosis of NHL patients, we developed an RSF model with a log-rank splitting rule and demonstrated that it outperformed the traditional statistical model COX. Also ranked were the risk variables affecting the individuals. Using the RSF model, we also stratified the risk of NHL patients and estimated individual survival probability. This may have practical implications for the implementation of individualized therapy.

Ethics approval and consent to participate

The study conformed to the Declaration of Helsinki's principles for medical research involving human beings and informed consent

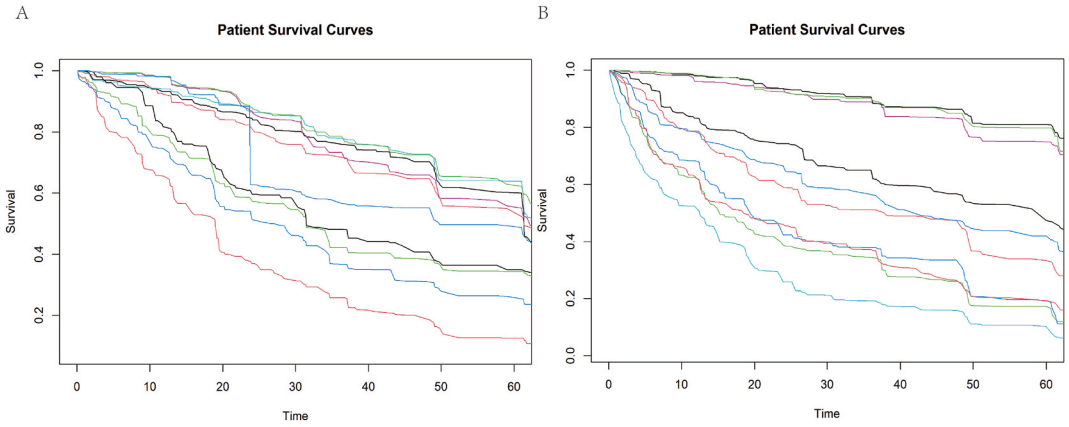

Fig. 8. The individual prognostic survival prediction in TC (A) and VC (B).

was obtained from all participants. The Ethics Committee at Chongqing University Cancer Hospital reviewed and authorized the human participant studies(Approval Number: CZLS2021252-A).

Consent for publication

Not applicable.

Funding

This study was supported by grants from the Chongqing Science and Technology Bureau (CSTB2022TIAD-GPX0066).

Data availability statement

The authors have made the raw data supporting the conclusion of this article available to all qualified researchers without undue reservation.

CRediT authorship contribution statement

Xiaosheng Li: Writing – original draft, Software. **Zailin Yang:** Resources, Conceptualization. **Jieping Li:** Visualization, Validation. **Guixue Wang:** Methodology, Data curation. **Anlong Sun:** Writing – review & editing. **Ying Wang:** Writing – review & editing. **Wei Zhang:** Writing – review & editing, Supervision. **Yao Liu:** Writing – review & editing, Supervision. **Haike Lei:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Ying Wang reports financial support was provided by Chongqing Science and Technology Bureau. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all participants for their willingness to participate.

References

[1] S.M. Ansell, Non-Hodgkin lymphoma: diagnosis and treatment, Mayo Clin. Proc. 90 (8) (2015) 1152–1163.
[2] K.M. McCarten, H.R. Nadel, B.L. Shulkin, S.Y. Cho, Imaging for diagnosis, staging and response assessment of Hodgkin lymphoma and non-Hodgkin lymphoma, Pediatr. Radiol. 49 (11) (2019) 1545–1564.
[3] W. Liu, J. Liu, Y. Song, X. Zeng, X. Wang, L. Mi, et al., Burden of lymphoma in China, 2006-2016: an analysis of the Global burden of disease study 2016, J. Hematol. Oncol. 12 (1) (2019) 115.
[4] B.D. Cheson, R.I. Fisher, S.F. Barrington, F. Cavalli, L.H. Schwartz, E. Zucca, et al., Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification, J. Clin. Oncol. : official journal of the American Society of Clinical Oncology 32 (27) (2014) 3059–3068.
[5] J.A.B. Bispo, P.S. Pinheiro, E.K. Kobetz, Epidemiology and etiology of leukemia and lymphoma, Cold Spring Harbor perspectives in medicine 10 (6) (2020).

[6] L.R. Teras, K.A. Bertrand, E.L. Deubler, C.R. Chao, J.V. Lacey Jr., A.V. Patel, et al., Body size and risk of non-Hodgkin lymphoma by subtype: a pooled analysis from six prospective cohorts in the United States, Br. J. Haematol. 197 (6) (2022) 714–727.
[7] A. Miranda-Filho, M. Piñeros, A. Znaor, R. Marcos-Gragera, E. Steliarova-Foucher, F. Bray, Global patterns and trends in the incidence of non-Hodgkin lymphoma, Cancer causes & control : CCC (Cancer Causes Control) 30 (5) (2019) 489–499.
[8] D. Chihara, Y. Oki, M.A. Fanale, J.R. Westin, L.J. Nastoupil, S. Neelapu, et al., Stage I non-Hodgkin lymphoma: no plateau in disease-specific survival, Ann. Hematol. 98 (5) (2019) 1169–1176.
[9] L. Zanoni, D. Bezzi, C. Nanni, A. Paccagnella, A. Farina, A. Broccoli, et al., PET/CT in non-hodgkin lymphoma: an update, Semin. Nucl. Med. 53 (3) (2023) 320–351.
[10] Z. Wang, L.S. Phillips, T.E. Rohan, G.Y.F. Ho, A.H. Shadyab, A. Bidulescu, et al., Diabetes, metformin use and risk of non-Hodgkin's lymphoma in postmenopausal women: a prospective cohort analysis in the Women's Health Initiative, Int. J. Cancer. 152 (8) (2022) 1556–1569.
[11] G. Lu, Z. Lin, Y. Ruan, H. Huang, J. Lin, J. Pan, A novel prognostic model for patients with primary gastric diffuse large B-cell lymphoma, Journal of oncology 2022 (2022) 9636790.
[12] C. Li, V. Patil, K.M. Rasmussen, C. Yong, H.C. Chien, D. Morreall, et al., Predicting survival in veterans with follicular lymphoma using structured electronic health record information and machine learning, Int. J. Environ. Res. Publ. Health 18 (5) (2021).
[13] K.G. Moons, D.G. Altman, J.B. Reitsma, J.P. Ioannidis, P. Macaskill, E.W. Steyerberg, et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration, Annals of internal medicine 162 (1) (2015) W1–W73.
[14] H. Ishwaran, The effect of splitting on random forests, Mach. Learn. 99 (1) (2015) 75–118.
[15] L. Zhang, J. Zhang, H. He, X. Ling, F. Li, Z. Yang, et al., Increased cytokine levels assist in the diagnosis of respiratory bacterial infections or concurrent bacteremia in patients with non-hodgkin's lymphoma, Front. Cell. Infect. Microbiol. 12 (2022) 860526.
[16] L. Vercellino, R. Di Blasi, S. Kanoun, B. Tessoulin, C. Rossi, M. D'Aveni-Piney, et al., Predictive factors of early progression after CAR T-cell therapy in relapsed/refractory diffuse large B-cell lymphoma, Blood advances 4 (22) (2020) 5607–5615.
[17] P. Sesques, E. Ferrant, V. Safar, F. Wallet, J. Tordo, A. Dhomps, et al., Commercial anti-CD19 CAR T cell therapy for patients with relapsed/refractory aggressive B cell lymphoma in a European center, Am. J. Hematol. 95 (11) (2020) 1324–1333.
[18] R.W. Merryman, R. Houot, P. Armand, C. Jacobson, Immune and cell therapy in non-hodgkin lymphoma, Cancer journal (Sudbury, Mass) 26 (3) (2020) 269–277.
[19] J. Lu, Y. Wu, B. Li, X. Luo, W. Zhang, Y. Zeng, et al., Predictive value of serological factors, maximal standardized uptake value and ratio of Ki67 in patients diagnosed with non-Hodgkin's lymphoma, Oncol. Lett. 20 (4) (2020) 47.
[20] M. Le, Y. Garcilazo, M.J. Ibáñez-Juliá, N. Younan, L. Royer-Perron, M. Benazra, et al., Pretreatment hemoglobin as an independent prognostic factor in primary central nervous system lymphomas, Oncol. 24 (9) (2019) e898–e904.
[21] J. Jung, H. Lee, T. Yun, E. Lee, H. Moon, J. Joo, et al., Prognostic role of the neutrophil-to-lymphocyte ratio in patients with primary central nervous system lymphoma, Oncotarget 8 (43) (2017) 74975–74986.
[22] Y. Feng, Y. Liu, M. Zhong, L. Wang, Complete blood count score model predicts inferior prognosis in primary central nervous system lymphoma, Frontiers in oncology 11 (2021) 618694.
[23] M. Okay, O. Meletli, E. Kelkitli, U.Y. Malkan, M. Turgut, Y. Buyukasik, et al., Mantle cell lymphoma: a Turkish multi-center study, Journal of BUON : official journal of the Balkan Union of Oncology 24 (5) (2019) 2084–2089.
[24] M. Geva, A. Pryce, R. Shouval, J.A. Fein, I. Danylesko, N. Shem-Tov, et al., High lactate dehydrogenase at time of admission for allogeneic hematopoietic transplantation associates to poor survival in acute myeloid leukemia and non-Hodgkin lymphoma, Bone Marrow Transplant. 56 (11) (2021) 2690–2696.
[25] P. Klener, M. Klanova, Drug resistance in non-hodgkin lymphomas, Int. J. Mol. Sci. 21 (6) (2020).
[26] D.L. Guevara, S. Bernard, S. Manhood, S. Melani, F. Yerovi, M. Rodríguez, [Prognostic value of interim PET/CT in non-hodgkin lymphoma], Rev. Med. Chile 148 (11) (2020) 1558–1567.
[27] S. Pratap, T.S. Scordino, Molecular and cellular genetics of non-Hodgkin lymphoma: diagnostic and prognostic implications, Exp. Mol. Pathol. 106 (2019) 44–51.
[28] M. Iqbal, L. Jiang, K.D. Li, M.A. Moustafa, E.O. Kimbrough, S.M. Ansell, et al., Poly-lymphomatous syndrome with concurrent or sequential hodgkin and non-hodgkin lymphoma, Clin. Lymphoma, Myeloma & Leukemia 23 (2) (2023) 138–144.
[29] A. Dogan, The real risk of secondary non-Hodgkin lymphoma following classical Hodgkin lymphoma, Haematologica 108 (5) (2023) 1220–1221.