Original Article

Comparison of Cox regression and generalized Cox regression models to machine learning in predicting survival of children with diffuse large B-cell lymphoma

Jia-Jia Qin[1]^, Xiao-Xiao Zhu[1]^, Xi Chen[1]^, Wei Sang[2]^, Ying-Liang Jin[1]^

[1]Department of Medical Public Health, Center for Medical Statistics and Data Analysis of Xuzhou Medical University, Xuzhou, China; [2]Department of Hematology, Affiliated Hospital of Xuzhou Medical University, Xuzhou, China

Contributions: (I) Conception and design: JJ Qin,YL Jin; (II) Administrative support: YL Jin, W Sang; (III) Provision of study materials or patients: JJ Qin, XX Zhu; (IV) Collection and assembly of data: JJ Qin, X Chen; (V) Data analysis and interpretation: JJ Qin, YL Jin, XX Zhu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Ying-Liang Jin, MD. Department of Medical Public Health, Center for Medical Statistics and Data Analysis of Xuzhou Medical University, Dongdian, 209 Huangshan Street, Yunlong District, Xuzhou 221004, China. Email: spark9809@126.com.

Background: The incidence of diffuse large B-cell lymphoma (DLBCL) in children is increasing globally. Due to the immature immune system in children, the prognosis of DLBCL is quite different from that of adults. We aim to use the multicenter large retrospective analysis for prognosis study of the disease.

Methods: For our retrospective analysis, we retrieved data from the Surveillance, Epidemiology and End Results (SEER) database that included 836 DLBCL patients under 18 years old who were treated at 22 central institutions between 2000 and 2019. The patients were randomly divided into a modeling group and a validation group based on the ratio of 7:3. Cox stepwise regression, generalized Cox regression and eXtreme Gradient Boosting (XGBoost) were used to screen all variables. The selected prognostic variables were used to construct a nomogram through Cox stepwise regression. The importance of variables was ranked using XGBoost. The predictive performance of the model was assessed by using C-index, area under the curve (AUC) of receiver operating characteristic (ROC) curve, sensitivity and specificity. The consistency of the model was evaluated by using a calibration curve. The clinical practicality of the model was verified through decision curve analysis (DCA).

Results: ROC curve demonstrated that all models except the non-proportional hazards and non-log linearity (NPHNLL) model, achieved AUC values above 0.7, indicating high accuracy. The calibration curve and DCA further confirmed strong predictive performance and clinical practicability.

Conclusions: In this study, we successfully constructed a machine learning model by combining XGBoost with Cox and generalized Cox regression models. This integrated approach accurately predicts the prognosis of children with DLBCL from multiple dimensions. These findings provide a scientific basis for accurate clinical prognosis prediction.

Keywords: Diffuse large B-cell lymphoma (DLBCL); DLBCL in children; generalized Cox; eXtreme Gradient Boosting (XGBoost); prognosis

Submitted Dec 24, 2023. Accepted for publication Jun 04, 2024. Published online Jul 26, 2024.
doi: 10.21037/tcr-23-2358
View this article at: https://dx.doi.org/10.21037/tcr-23-2358

^ ORCID: Jia-Jia Qin, 0009-0002-9281-2251; Xiao-Xiao Zhu, 0009-0005-5612-8980; Xi Chen, 0009-0007-5309-7271; Wei Sang, 0009-0006-9889-3290; Ying-Liang Jin, 0009-0001-8819-7287.

Introduction

Lymphoma is among the eight most common malignant tumors worldwide. Its incidence has been increasing annually. According to statistics from the World Health Organization, the annual growth rate of lymphoma incidence was 5% to 7% between 2000 and 2019 with over 200,000 deaths occurring each year (1). Globally, diffuse large B-cell lymphoma (DLBCL) is the most prevalent form of non-Hodgkin lymphoma (NHL), accounting for 20% of NHL in children (2). According to the data from the National Cancer Institute, the incidence of DLBCL in children showed an upward trend from 2011 to 2022. White children have the highest incidence of DLBCL, followed by black children and Asian children (3). A current study has revealed that DLBCL in children are more invasive and exhibits distinct prognosis differences (4). However, due to the relative rarity of children with DLBCL in clinical practice, a previous study only analyzed mixed cases involving both children and other kinds of DLBCL cases, introducing various confounding factors (5). Currently, the international prognostic index (IPI) has limited utility in guiding prognostic stratification of pediatric DLBCL patients (6). As a result, there is a lack of multicenter and extensive sample data to construct more refined prognostic models to aid in predicting the prognosis of such patients.

Although the Cox regression model is a standard method for tumor prognosis analysis, more methods are needed to analyze the significance of prognostic factors and to conduct precision analysis of the model (7). The Cox regression calculates crude mortality and requires the following for constructing a prognostic model with continuous variables (8): (I) a linear relationship between the logarithm of the hazard ratio and the covariates; (II) independence of the logarithm of the risk ratio from time, with only a relation to the linear combination of covariates. In other words, Cox regression needs to satisfy the requirements of equal proportional hazards (PH) and log linearity (LL). However, in a real clinical study, DLBCL in children exhibits high heterogeneity and invasiveness (9). Data that meet both requirements of Cox regression are scarce. Therefore, there is an urgent need for more flexible models that accurately reflect the prognostic characteristics of childhood DLBCL in clinical practice. The generalized Cox regression offers a flexible modeling approach for relative survival. It allows for analysis of the nonlinear influence of variables on survival risk, especially for prognostic variables including continuous variables (8). This means that the data do not need to adhere strictly to the PH assumption or to the assumption of linearity on the logarithmic scale. Therefore, the constructed prognostic model is more consistent with the actual data. The eXtreme Gradient Boosting (XGBoost) model, as an efficient machine learning approach, is an integrated algorithm that has been widely used in artificial intelligence, data mining, and statistical analysis. It improves up on gradient boosting decision trees by offering loose data requirements, fast training speed and accurate training results (10). XGBoost is utilized to construct the prediction model. Once the boosting tree is created, the importance score of each attribute can be obtained quite directly (11). This allows for ranking the importance of prognostic variables. Based on the Surveillance, Epidemiology and End Results (SEER) database, this study aimed to enhance the prediction accuracy and precision of the model by analyzing the clinicopathological data of 836 children with DLBCL through a multicenter big data approach. The multidimensional analysis was carried out by using traditional Cox and generalized Cox regression, combined with XGBoost algorithm to investigate the clinical data of children patients. Decision curve analysis (DCA) was also employed to evaluate the clinical effectiveness of the model, aiming to enhance the clinical workers' awareness of the disease and to provide novel insights for clinical diagnosis and treatment of patients as well as the whole management

Highlight box

Key findings

- Using regression analysis, we identified several independent prognostic factors for overall survival (OS) in children with diffuse large B-cell lymphoma (DLBCL) namely age, sex, surgery, primary surgical procedure, radiotherapy, chemotherapy, systemic therapy, Ann Arbor stage and time from diagnosis to treatment.

What is known and what is new?

- The incidence of DLBCL in children is increasing globally. Due to the immature immune system in children, the prognosis of DLBCL is quite different from that of adults.
- By comparing our novel model to the traditional prognostic system, we used the decision curve analysis. The results indicated that within the threshold of 0.87, the net benefit rates of the novel models exceeded those of traditional international prognostic index prognostic score.

What is the implication, and what should change now?

- It can be demonstrated that the novel model has better clinical value and utility in evaluating the OS of DLBCL in children when compared to the traditional prognostic system. We could use novel models to assist clinical practitioners in predicting more accurate prognoses for children DLBCL patients.

of patients. We present this article in accordance with the TRIPOD reporting checklist (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-2358/rc).

Methods

Research subjects

We obtained authorization to collect clinicopathological data of patients diagnosed with DLBCL between 2000 and 2019 using SEER*Stat software (version 8.4.0.1). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Finally, we gathered detailed clinicopathological and follow-up data from 832 patients who met the inclusion and exclusion criteria. Inclusion criteria: (I) patients with histopathological diagnosis of DLBCL; (II) under 18 years old; (III) the year of diagnosis was 2000–2019; (IV) follow-up information such as patient survival data was complete. Exclusion criteria: (I) patients older than 18 years; (II) those with unknown tumor size; (III) those with unknown tumor stage.

Observation indicators

Lots of meaningful variables were extracted for subsequent analysis. These variables included the patient's age, sex, Ann Arbor stage, year of diagnosis, IPI, ethnicity, B-symptoms, histological grade, surgical method, primary tumor location, chemotherapy, radiotherapy, systemic therapy, time from diagnosis to treatment, survival, cause of death and survival time. The follow-up period was extended until December 31, 2019. The primary outcome measure considered was overall survival (OS), which was defined as the interval from the time of the patient's histopathological diagnosis until death.

Cox regression model

The Cox PH regression model is currently the most commonly utilized semi-parametric model for conducting multivariate survival analysis. It has been widely used because it combines the advantages of both parametric and non-parametric models, which can analyze the influence of survival time for incomplete data. Survival analysis is the collective term for a series of methods dealing with the statistics of time to event and variables and is often used to study disease occurrence, outcome, recovery and death (12). Among the essential multivariate analysis techniques in survival analysis, the Cox PH regression model is extensively employed in identifying risk factor and predicting clinical outcomes by using follow-up data (8,12).

XGBoost model

XGBoost is an ensemble learning algorithm based on gradient boosting and its principle is to achieve precise classification effect through iterative computation of weak classifiers (13). It is an optimization model that combines both linear model and boosted tree model, providing the benefits of fast processing speed and superior performance. As a result, the XGBoost model is extensively utilized by experts in various domains such as machine learning, data mining, statistics, and other fields such as artificial intelligence, data analysis, and statistical learning (4,11,13).

Generalized Cox regression model

Generalized Cox regression is an advanced predictor model that exhibits great flexibility compared to Cox regression. Cox regression employs the assumption of PH and LL (8). First, Cox regression assumes a linear relationship between the logarithm of the hazard ratio and the covariates, known as the PH assumption. Second, Cox regression also assumes that the logarithm of the hazard ratio is independent of time and only relates to the linear combination of covariates, referred to as the LL assumption. On the other hand, generalized Cox regression employs three flexible models to fit the data, allowing for independent testing and relaxation of both the PH and LL assumptions (8). The first non-proportional hazards (NPH) model, relaxes the PH assumption and does not require prognostic data of DLBCL patients to satisfy the PH requirements (8). The second model, non-log linearity (NLL), relaxes the LL assumption and does not assume the hazard ratio to be linear on the logarithmic scale (8). The third model, NPHNLL, relaxes both PH and LL assumptions (8). As a result, generalized Cox regression is more widely used and provides a more precise fit to DLBCL prognostic data.

Statistical analysis

Statistical analysis was performed using the R software (version 4.1.3). Various packages were utilized, including flexrsurv, ggpubr, rms, survival, MASS, survminer, ggplot2, survivalROC, ggforest, ggDCA, Hmisc, lattice, Formula, XGBoost, etc. Multivariate Cox regression, Generalized Cox regression and XGBoost were used to screen out the

prognostic characteristic variables of children with DLBCL respectively (Cox and Generalized Cox regression used the minimum Akaike Information Criterion as the basis for variable screening. The top ten characteristic variables were used as the basis for screening of XGBoost). Cox regression was used to construct the selected prognostic factors. XGBoost ranked the importance of prognostic feature variables. Generalized Cox regression determined the non-proportional risk and non-log-linear relationship between prognostic variables and survival risk. Receiver operating characteristic (ROC) curve, C-index, sensitivity and specificity were used to determine the predictive performance of model. Calibration curve and DCA were used to determine the consistency and clinical validity of the model. The analysis indicators were 1-, 3-, and 5-year OS of the tumor. The significance level was set at α=0.05.

Results

Basic patient information

There were 2,230 children diagnosed with DLBCL from 2000 to 2019 obtained from the SEER database by using SEER*Stat software. The clinicopathological information mainly included patient age, race, year of diagnosis, primary tumor location, histological type, grade, Ann Arbor stage, surgical method, chemotherapy, radiotherapy, systematic treatment, time from diagnosis to treatment, IPI score, B symptoms, survival status, cause of death and survival time. The obtained patient information was screened several times according to the inclusion and exclusion criteria. Finally, 836 cases that met the study criteria were included, which were randomly divided into a modeling group (n=585) and a validation group (n=251) by using the "rms" package of R software at a ratio of 7:3 according to a random number table method. The related detailed clinicopathological information is shown in *Table 1*.

Cox regression model results

To identify the key factors influencing survival in children with DLBCL, we conducted multivariate Cox stepwise regression analysis. This analysis helped us identify independent prognostic factors associated with OS. The findings revealed that Ann Arbor stages III and IV were significantly associated with poorer OS in patients aged under 18 years. Children aged under 2 years exhibited worse 3-, 5-year OS compared to those older than 2 years. In terms of treatment, patients who underwent chemotherapy or radiotherapy had a better prognosis than those who did not receive these treatments. In addition, Patients who received systemic therapy also had

Table 1 Demographic and clinical data of 836 patients with children DLBCL (n=836)

Variable	Modeling group (n=585)	Verification group (n=251)	P value
Age (years)			0.33
≤10	121 (20.7)	76 (30.3)	
>10	464 (79.3)	175 (69.7)	
Race			0.17
White	210 (35.9)	102 (40.6)	
Black	186 (31.8)	41 (16.3)	
Other	189 (32.3)	108 (43.1)	
Site of disease			0.15
Left	221 (37.8)	134 (53.4)	
Right	56 (9.6)	35 (13.9)	
Bilateral	308 (52.6)	82 (32.7)	
SEER stage			0.08
Localized	85 (14.5)	59 (23.5)	
Regional	197 (33.7)	108 (43.0)	
Distant	303 (47.8)	84 (33.5)	

Table 1 *(continued)*

Table 1 *(continued)*

Variable	Modeling group (n=585)	Verification group (n=251)	P value
Ann Arbor stage			0.64
Stage I	64 (10.9)	36 (14.3)	
Stage II	123 (21.0)	67 (26.7)	
Stage III	183 (31.3)	55 (21.9)	
Stage IV	215 (36.8)	93 (37.1)	
Surgery			0.79
Yes	323 (55.2)	134 (53.4)	
No	262 (44.8)	117 (46.6)	
Radiotherapy			0.06
Yes	321 (54.9)	167 (66.5)	
No	264 (45.1)	84 (33.5)	
Radiotherapy type			0.13
No radiotherapy	241 (41.2)	84 (33.5)	
Postoperative radiotherapy	161 (27.5)	67 (26.7)	
Preoperative radiotherapy	89 (15.2)	32 (12.7)	
Radiotherapy before and after surgery	94 (16.1)	68 (27.1)	
Chemotherapy			0.11
Yes	486 (83.1)	177 (70.5)	
No	99 (16.9)	74 (29.5)	
Systematic therapy			0.08
Yes	209 (35.7)	121 (48.2)	
No	376 (64.3)	130 (51.8)	
Time from diagnosis to treatment (month)			0.21
1	241 (41.2)	86 (34.3)	
2	190 (32.5)	79 (31.5)	
3	93 (15.9)	54 (21.5)	
4	61 (10.4)	32 (12.7)	
IPI score			0.42
1	103 (17.6)	71 (28.3)	
2	84 (14.4)	30 (11.9)	
3	179 (30.6)	69 (27.5)	
4	219 (37.4)	81 (32.3)	
B symptom			0.09
Yes	390 (66.7)	183 (72.9)	
No	195 (33.3)	68 (27.1)	

DLBCL, diffuse large B-cell lymphoma; SEER, Surveillance, Epidemiology and End Results; IPI, international prognostic index.

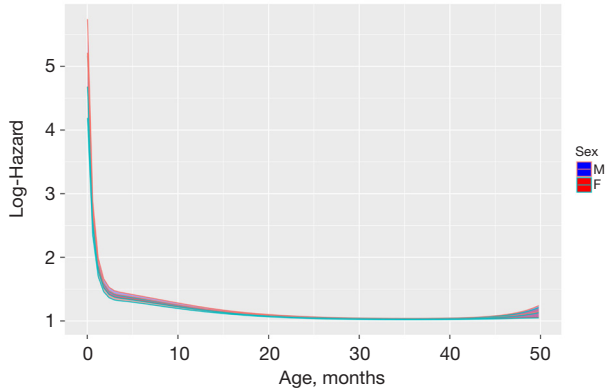

Figure 1 Non-proportional hazard relationships between age and overall survival. M, male; F, female.

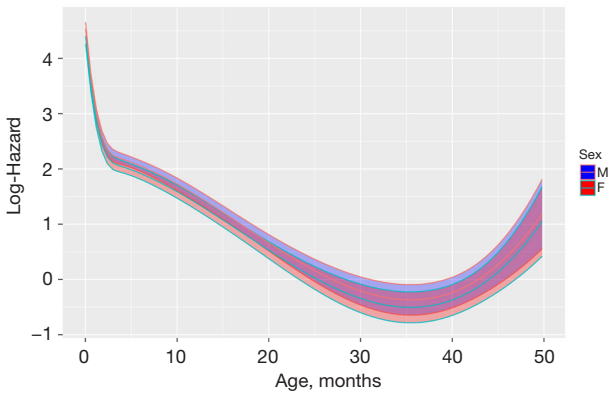

Figure 2 Non-log linearity relationships between age and overall survival. M, male; F, female.

a more favorable prognosis. According to the multivariate Cox stepwise regression analysis, the OS of patients who only had primary tumor surgery did not gain prolonged extension. In addition, we analyzed social variables such as family income status and marital status, but they did not emerge as independent prognostic factors for OS.

Generalized Cox regression model results

To uncover intricate NLL relationships between prognostic variables and OS, we analyzed all variables using generalized Cox regression. NLL model indicated that age, surgery, chemotherapy, Ann Arbor stage, sex and radiotherapy for OS exhibited NLL relationships (P<0.05). ROC curves were employed to assess the generalized Cox regression. In the training cohort, the area under the curve (AUC) for NLL model was 0.870 [95% confidence interval (CI): 0.818, 0.892]. In the validation cohort, the AUC for NLL was 0.869 (95% CI: 0.830, 0.872). The NPH model suggested that age, Ann Arbor stage, radiotherapy and chemotherapy for OS demonstrated NPH relationships (P<0.05). We showed in detail the NLL and NPH relationships between age and OS (*Figures 1*,*2*). In the training cohort, the AUC for NPH model was 0.813 (95% CI: 0.808, 0.862). In the validation cohort, the AUC for NPH model was 0.871 (95% CI: 0.860, 0.882). These findings demonstrated the exceptional effectiveness of NPH and NLL models in predicting the survival of children with DLBCL at OS. The NPHNLL model suggested that there were non-proportional risk and non-log-linear relationship among age, surgery and chemotherapy for OS (P<0.05). In the training cohort, the AUC for the NPHNLL model was 0.679 (95% CI: 0.308, 0.962). In the validation cohort, the AUC for the NPHNLL was 0.691 (95% CI: 0.460, 0.782).

Machine learning approach using XGBoost results

XGBoost analysis revealed the top 10 characteristic variables associated with OS in children with DLBCL. These prognostic factors were ranked based on their importance: age, systemic therapy, Ann Arbor stage, laterality, radiation type, stage, surgery type, chemotherapy, radiation, and time from diagnosis to treatment (*Figure 3*). Our XGBoost algorithm model demonstrated extraordinary efficiency in predicting OS in both the training and validation cohorts of children with DLBCL. Specifically, in the training cohort, the AUC was 0.892 (95% CI: 0.707, 0.939), and in the validation cohort, the AUC was 0.889 (95% CI: 0.801, 0.991). In comparison with other models, Cox regression model (train set AUC =0.799; test set AUC =0.770), NPH model (train set AUC =0.813; test set AUC =0.871), NLL model (train set AUC =0.870; test set AUC =0.869), NPHNLL model (train set AUC =0.679; test set AUC =0.691). The XGBoost model outperformed them all, demonstrating the best prediction performance.

Establishment and comparison of the prognostic models

The variables with P<0.05 in the multivariate Cox regression were incorporated into nomogram (*Figure 4*). The AUC under the ROC curve of the training group was 0.799, while the AUC under the ROC curve of the validation group was 0.770, both of which showed that the model had high prediction accuracy. The 1-, 3-, and 5-year calibration curves of the training cohort and the verification cohort (shown in

Figures 5-10) showed that the prediction model constructed in this study had a high consistency with the actual observed values. DCA curve analysis showed that the net benefit of the predictive model constructed in this study was generally superior to the traditional IPI prognostic model. It had high clinical practicability (*Figures 11,12*). Finally, we compared the predictive performance of all prognostic models (*Table 2*).

Discussion

Although children with DLBCL are rare, their incidence is gradually increasing (14). A growing number of studies have shown that the prognosis of DLBCL in children is quite different from that in adults (14,15). Therefore, more precise prognostic evaluation is needed for the individualized treatment of children with DLBCL. In this study, we constructed new prediction models for DLBCL in

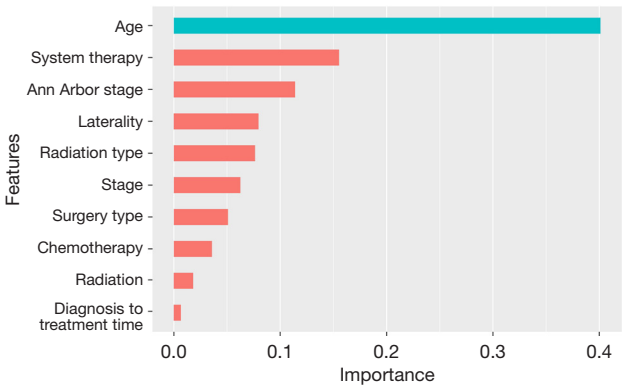

Figure 3 Feature importance evaluation of extreme gradient boosting model.

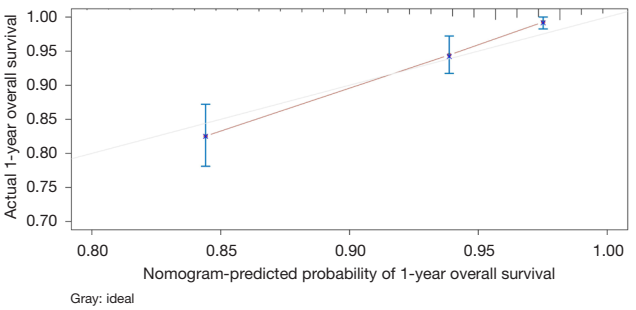

Figure 5 Calibration curve for 1-year overall survival in the training cohort.

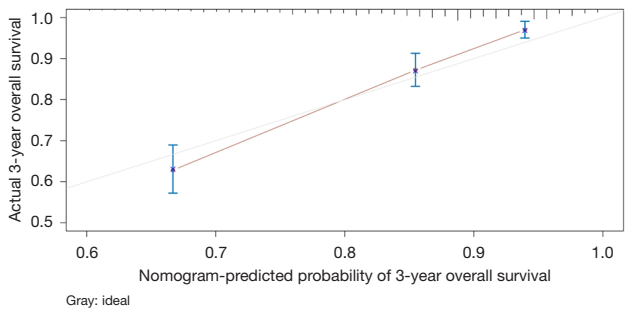

Figure 6 Calibration curve for 3-year overall survival in the training cohort.

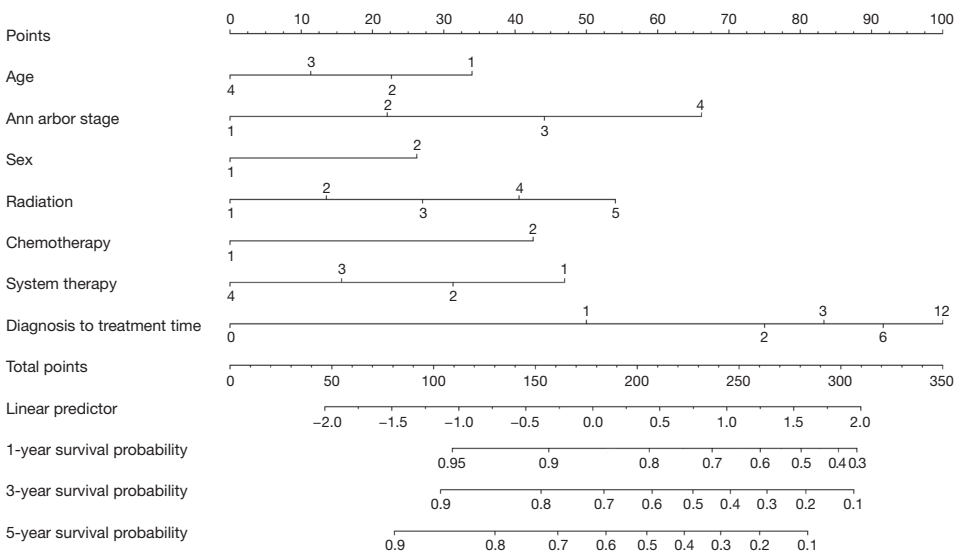

Figure 4 Nomogram for predicting overall survival.

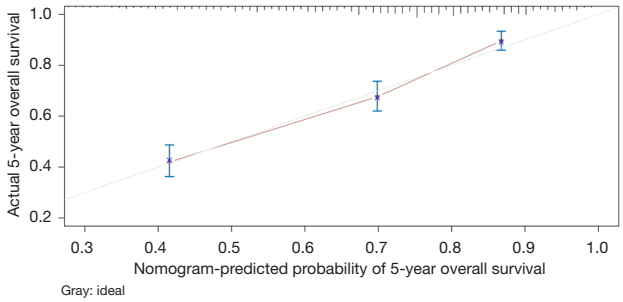

Figure 7 Calibration curve for 5-year overall survival in the training cohort

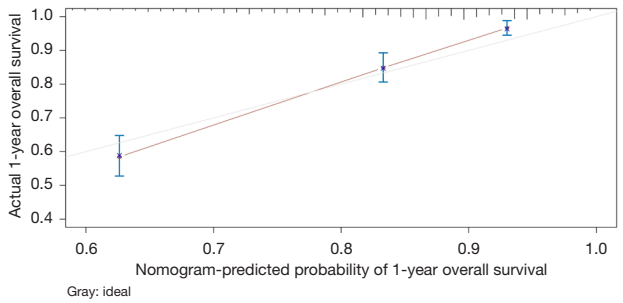

Figure 8 Calibration curve for 1-year overall survival in the validation cohort.

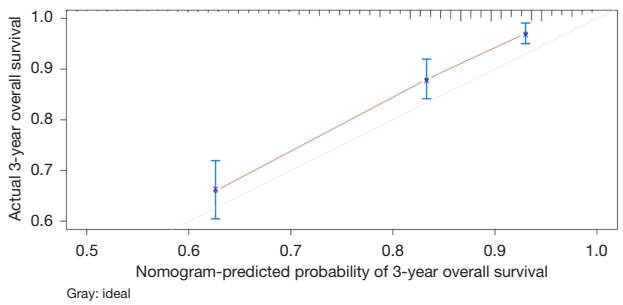

Figure 9 Calibration curve for 3-year overall survival in the validation cohort.

children using XGBoost, Generalized Cox and Cox stepwise regression. These models can comprehensively, rapidly and accurately assess the prognosis of children patients. The IPI scoring system has been the main prognostic assessment tool used in clinics for children with DLBCL (16,17), but it has limitations as it does not consider other important factors that impact survival (18). However, our model can accurately predict the significance of prognostic factors and their non-proportional and non-linear relationship with

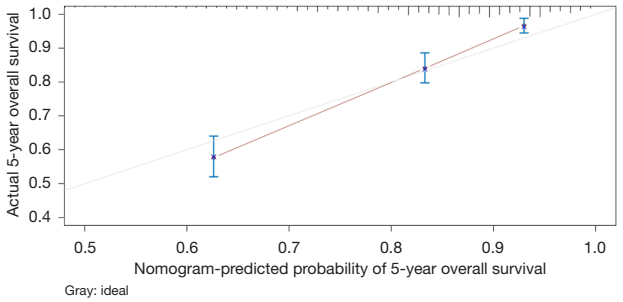

Figure 10 Calibration curve for 5-year overall survival in the validation cohort.

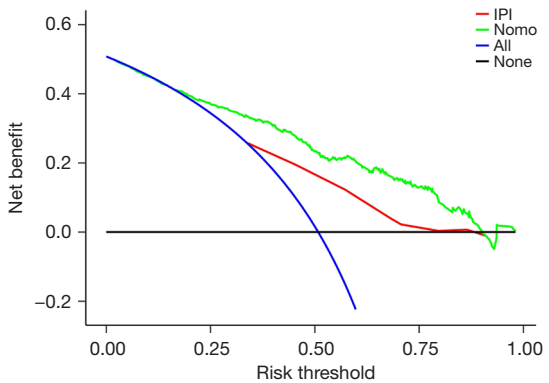

Figure 11 Decision curve analysis of the training cohort. IPI, international prognostic index; Nomo, nomogram.

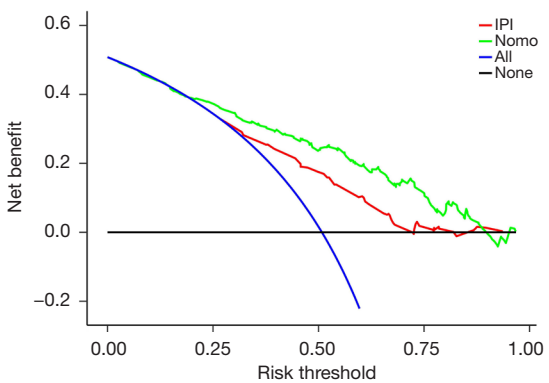

Figure 12 Decision curve analysis of the validation cohort. IPI, international prognostic index; Nomo, nomogram.

survival outcomes. Our analysis and validation demonstrate the contribution and effect of each prognostic variable on survival risk.

Previous studies showed that sex, age, stage, chemotherapy,

Table 2 Performance of the prognostic models

Models	Training cohort			Validation cohort		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Cox	0.799	0.799	0.801	0.770	0.799	0.748
NPH	0.813	0.871	0.819	0.871	0.802	0.847
NLL	0.870	0.872	0.871	0.869	0.868	0.876
NPHNLL	0.679	0.701	0.679	0.691	0.690	0.709
XGBoost	0.892	0.877	0.899	0.889	0.864	0.892

NPH, non-proportional hazards; NLL, non-log linearity; NPHNLL, non-proportional hazards and non-log linearity; XGBoost, extreme gradient boosting; AUC, area under curve.

B symptoms and large mass size were influencing factors for the prognosis of children with DLBCL (19,20). However, these studies were primarily retrospective analyses with small samples and being single-center. Our study collected data from 836 DLBCL patients of children, making it the largest multicenter retrospective study of DLBCL in children. Therefore, the findings were more accurate and comprehensive. Using regression analysis, we identified several independent prognostic factors for OS in children with DLBCL namely age, sex, surgery, primary surgical procedure, radiotherapy, chemotherapy, systemic therapy, Ann Arbor stage and time from diagnosis to treatment. Our predictive model demonstrated that male children with DLBCL, those who did not receive systemic therapy and those with a longer time from diagnosis to treatment experienced worse outcomes. Early detection and systematic diagnosis have been shown to be beneficial for treatments. Therefore, systemic therapy and a shorter time from diagnosis to treatment were associated with better prognosis. Earlier studies have also indicated that early detection and systemic treatment were independently associated with the 5-year relative survival of patients with DLBCL (21,22). Consistent with previous findings, our study found that male children with DLBCL generally had worse outcomes compared to their female counterparts (9,15,23). This difference in survival rates could be attributed to the genetic profiles and lifestyle habits that differ between males and females. We thought there might be a complex interaction among genetic factors, environmental factors, and cancer outcomes. In general, the results need to be verified with a large amount of research evidence from real-world data.

In addition, we constructed a novel predictive nomogram and validated it with a 251 patients internal validation cohort. Our findings revealed that patients' characteristics, such as age, sex, Ann Arbor stage, radiotherapy, chemotherapy, systemic therapy, and time from diagnosis to treatment were associated with prognosis. Additionally, this nomogram performed excellently as assessed by the calibration curve and AUC. Compared to the IPI scoring system, the calibration curve of the nomogram for predicting OS was more accurate both in the training and validation sets. Moreover, the AUC values of the nomogram for predicting 3- and 5-year OS were higher, providing clinicians with a more accurate prediction of individual patients. Furthermore, the DCA curve was used to identify high-risk patients to intervene while low-risk patients to avoid unnecessary intervention (24). This evaluation method allows us to assess the degree of patient benefit (25). Consequently, we utilized DCA curve to evaluate the clinical practicality of the newly constructed model. Our analysis of the modeling group revealed the predictive ability of the new nomogram for OS in children with DLBCL. In the validation group, the nomogram effectively incorporated clinical and demographic information. By comparing our novel model to the traditional IPI scoring system, we used the DCA curve. The results indicated that within the threshold of 0.87, the net benefit rates of the novel models exceeded those of traditional IPI prognostic score. It can be demonstrated that the novel model has better clinical value and utility in evaluating the OS of DLBCL in children when compared to the IPI scoring system.

In recent years, extensive retrospective studies have indicated that elderly patients are at high risk for DLBCL (26,27). Our research observed a complex nonlinear relationship between age and OS in children with DLBCL by using the generalized Cox regression. The NPH model suggested that there was a NPH relationship between age and OS. Hence, for age the PH assumption was relaxed.

The effect of the age variable on OS was not constant with time. Additionally, our NLL model suggested that there was a non-log-linear relationship between age and OS. That is to say, for age the LL assumption was relaxed. The effects of this variable were not constant with time and exhibited non-linear relationships between this variable and the log-hazard of OS. Consequently, the generalized Cox regression allows for greater flexibility in capturing changes in the impact of a predictor variable on survival risk of children with DLBCL over time. Furthermore, our study showed that the younger the age, the higher the prognostic risk of DLBCL in children. Compared to adult patients, DLBCL patients aged 1 to 2 years had a greater survival risk. The NPHNLL model also suggested that age, Ann Arbor stage, surgery, sex and chemotherapy exhibited both NPH and non-log-linear relationships with OS. However, the AUC value of the NPHNLL model was less than 0.7, indicating that further validation of the relationship between the prognostic factors revealed by the model and OS was necessary. This finding provides a direction for future research.

So far, XGBoost model has been widely utilized in the cancer field for survival analysis (28). However, its application in predicting the significance of prognostic factors in children with DLBCL has not been explored. In our study, we compared the predictive ability of Cox stepwise regression, generalized Cox regression and XGBoost, and found that XGBoost demonstrated the highest predictive accuracy in identifying prognostic factors in children with DLBCL. The XGBoost model was employed as an efficient machine learning tool to rank the significance of prognostic factors of DLBCL in children in our research. Interestingly, the model revealed that age was the most influential prognostic factor, validating its impact on DLBCL prognosis. Previous studies have indicated that the addition of radiotherapy to immunochemotherapy improves the prognosis of selected DLBCL patients (29-31). However, our XGBoost analysis did not identify chemotherapy and radiotherapy as the most important prognostic factors. Therefore, further research is required to confirm whether chemotherapy and radiotherapy are key prognostic factors for DLBCL in children.

Despite several promising findings, there are several limitations in our research. Firstly, the data from the SEER database were derived from retrospective studies, which might introduce selection bias and information bias. Secondly, there were missing data regarding chemotherapy regimens of patients and doses. As a result, we were unable to investigate the impact of specific chemotherapy regimens on OS of patients.

Conclusions

We constructed three novel models to predict DLBCL in children by using a large set of DLBCL samples from SEER database and an analysis of commonly used clinical indicators. These high-precision models effectively predict the intricate relationships between variables and survival risk in children with DLBCL. Our research aims to assist clinical practitioners in predicting more accurate prognoses for children DLBCL patients, enabling individual precision treatment and offering guidance for the management of patients.

Acknowledgments

We would like to thank the National Cancer Institute (NCI) for open access to their SEER database.
Funding: This study was supported by grant from Xuzhou Science and Technology Project (KC22128).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-2358/rc

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-2358/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-2358/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Kanas G, Ge W, Quek RGW, et al. Epidemiology of diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) in the United States and Western Europe: population-level projections for 2020-2025. Leuk Lymphoma 2022;63:54-63.
2. Dagar G, Gupta A, Masoodi T, et al. Harnessing the potential of CAR-T cell therapy: progress, challenges, and future directions in hematological and solid tumor treatments. J Transl Med 2023;21:449.
3. Kuczmarski TM, Tramontano AC, Mozessohn L, et al. Mental health disorders and survival among older patients with diffuse large B-cell lymphoma in the USA: a population-based study. Lancet Haematol 2023;10:e530-8.
4. Restivo GA, Farruggia P, Pillon M, et al. Pediatric gray zone lymphoma according to the 2022 WHO classification: An Italian cohort study. Pediatr Blood Cancer 2023. [Epub ahead of print]. doi: 10.1002/pbc.30481.
5. Shi Q, He Y, Yi HM, et al. Positron emission tomography-adapted therapy in low-risk diffuse large B-cell lymphoma: results of a randomized, phase III, non-inferiority trial. Cancer Commun (Lond) 2023;43:896-908.
6. Shen Y, Wang L, Ou J, et al. Loss of 5-hydroxymethylcytosine as a Poor Prognostic Factor for Primary Testicular Diffuse Large B-cell Lymphoma. Int J Med Sci 2022;19:225-32.
7. Liu Y, Sheng L, Hua H, et al. A Novel and Validated Inflammation-Based Prognosis Score (IBPS) Predicts Outcomes in Patients with Diffuse Large B-Cell Lymphoma. Cancer Manag Res 2023;15:651-66.
8. Goerdten J, Carrière I, Muniz-Terrera G. Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. Alzheimers Dement (N Y) 2020;6:e12041.
9. Forlenza CJ, Chadburn A, Giulino-Roth L. Primary Mediastinal B-Cell Lymphoma in Children and Young Adults. J Natl Compr Canc Netw 2023;21:323-30.
10. Yan FJ, Chen XH, Quan XQ, et al. Development and validation of an interpretable machine learning model-Predicting mild cognitive impairment in a high-risk stroke population. Front Aging Neurosci 2023;15:1180351.
11. Li B, Verma R, Beaton D, et al. Predicting Outcomes Following Endovascular Abdominal Aortic Aneurysm Repair Using Machine Learning. Ann Surg 2024;279:521-7.
12. Kinslow CJ, Rae AI, Taparra K, et al. MGMT Promoter Methylation Predicts Overall Survival after Chemotherapy for 1p/19q-Codeleted Gliomas. Clin Cancer Res 2023;29:4399-407.
13. Liu Y, Wang X, Wu Z, et al. Automated anatomical labeling of a topologically variant abdominal arterial system via probabilistic hypergraph matching. Med Image Anal 2022;75:102249.
14. Beishuizen A, Mellgren K, Andrés M, et al. Improving outcomes of childhood and young adult non-Hodgkin lymphoma: 25 years of research and collaboration within the framework of the European Intergroup for Childhood Non-Hodgkin Lymphoma. Lancet Haematol 2023;10:e213-24.
15. Abrahão R, Ribeiro RC, Lichtensztajn DY, et al. Survival after diffuse large B-cell lymphoma among children, adolescents, and young adults in California, 2001-2014: A population-based study. Pediatr Blood Cancer 2019;66:e27559.
16. Martelli M, Ferreri AJ, Agostinelli C, et al. Diffuse large B-cell lymphoma. Crit Rev Oncol Hematol 2013;87:146-71.
17. Liu Y, Barta SK. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. Am J Hematol 2019;94:604-16.
18. Ghione P, Ahsanuddin S, Luttwak E, et al. Diffuse large B-cell lymphoma involving osseous sites: utility of response assessment by PET/CT and good longterm outcomes. Haematologica 2024;109:200-8.
19. Rahiman EA, Bakhshi S, Deepam Pushpam, et al. Outcome and prognostic factors in childhood B non-Hodgkin lymphoma from India: Report by the Indian Pediatric Oncology Group (InPOG-NHL-16-01 study). Pediatr Hematol Oncol 2022;39:391-405.
20. Reiter A, Klapper W. Recent advances in the understanding and management of diffuse large B-cell lymphoma in children. Br J Haematol 2008;142:329-47.
21. Liu Y, Wang J, Shen X, et al. A novel angiogenesis-related scoring model predicts prognosis risk and treatment responsiveness in diffuse large B-cell lymphoma. Clin Exp Med 2023;23:3781-97.
22. Liu YZ, Xue K, Wang BS, et al. The size and depth of lesions measured by endoscopic ultrasonography are novel prognostic factors of primary gastric diffuse large B-cell lymphoma. Leuk Lymphoma 2019;60:934-9.
23. Gardenswartz A, Mehta B, El-Mallawany NK, et al.

Safety and efficacy of combinatorial therapy utilizing myeloablative conditioning and autologous stem cell transplantation, targeted immunotherapy, and reduced intensity conditioning and allogeneic stem cell transplantation in children, adolescents, and young adults with relapsed/refractory mature B-cell non-Hodgkin lymphoma. Leuk Lymphoma 2023;64:234-7.
24. Jiang C, Qian C, Jiang Z, et al. Robust deep learning-based PET prognostic imaging biomarker for DLBCL patients: a multicenter study. Eur J Nucl Med Mol Imaging 2023;50:3949-60.
25. Liu Y, Sheng L, Hua H, et al. An Externally Validated Nomogram for Predicting the Overall Survival of Patients With Diffuse Large B-Cell Lymphoma Based on Clinical Characteristics and Systemic Inflammatory Markers. Technol Cancer Res Treat 2023;22:15330338231180785.
26. Jiang L, Wang C, Tong Y, et al. Web-based nomogram and risk stratification system constructed for predicting the overall survival of older adults with primary kidney cancer after surgical resection. J Cancer Res Clin Oncol 2023;149:11873-89.
27. Gini G, Tani M, Tucci A, et al. Lenalidomide plus rituximab for the initial treatment of frail older patients with DLBCL: the FIL_ReRi phase 2 study. Blood 2023;142:1438-47.
28. Li W, Li Y, Liu X, et al. Machine learning-based radiomics for predicting BRAF-V600E mutations in ameloblastoma. Front Immunol 2023;14:1180908.
29. Lu T, Zhang J, Xu-Monette ZY, et al. The progress of novel strategies on immune-based therapy in relapsed or refractory diffuse large B-cell lymphoma. Exp Hematol Oncol 2023;12:72.
30. McCurry D, Flowers CR, Bermack C. Immune-based therapies in diffuse large B-cell lymphoma. Expert Opin Investig Drugs 2023;32:479-93.
31. Shen Z, Hu L, Zhang S, et al. Visceral fat area and albumin based nutrition-related prognostic index model could better stratify the prognosis of diffuse large B-cell lymphoma in rituximab era. Front Nutr 2022;9:981433.

Cite this article as: Qin JJ, Zhu XX, Chen X, Sang W, Jin YL. Comparison of Cox regression and generalized Cox regression models to machine learning in predicting survival of children with diffuse large B-cell lymphoma. Transl Cancer Res 2024;13(7):3370-3381. doi: 10.21037/tcr-23-2358

