Journal of Cancer Research and Clinical Oncology (2025) 151:125
https://doi.org/10.1007/s00432-025-06172-3

RESEARCH

^{18}F-FDG PET radiomics score construction by automatic machine learning for treatment response prediction in elderly patients with diffuse large B-cell lymphoma: a multicenter study

Jincheng Zhao[1] · Wenzhuo Zhao[2] · Man Chen[1] · Jian Rong[2] · Yue Teng[3] · Jianxin Chen[2] · Jingyan Xu[1]

Received: 25 November 2024 / Accepted: 14 March 2025 / Published online: 28 March 2025

Abstract

Purpose To explore the development and validation of automated machine learning (AutoML) models for ^{18}F-FDG PET imaging-based radiomics signatures to predict treatment response in elderly patients with diffuse large B-cell lymphoma (DLBCL).

Methods A retrospective analysis was conducted on 175 elderly (≥60 years) DLBCL patients diagnosed between March 2015 and March 2023 at two medical centers, with a total of 1010 lesions. The baseline PET imaging-based radiomics features of the training cohort were processed using AutoML model AutoGluon to generate a radiomics score (radscore) and predict treatment response at the lesion and patient levels. Furthermore, a multivariable logistic analysis was used to design and evaluate a multivariable model in the training and validation cohorts.

Results ROC curve analysis showed that the radscore generated by AutoML exhibited higher accuracy in predicting treatment response at the lesion level compared to metabolic parameters (SUVmax, MTV, and TLG) in both the training group (AUC: 0.791, 0.542, 0.667, 0.651, respectively) and the validation group (AUC: 0.712, 0.616, 0.639, 0.657, respectively). Multivariable logistic analysis indicated that NCCN-IPI (OR = 5.427, 95% CI: 1.163–25.317), BCL-2 (OR = 3.714, 95% CI: 1.406–9.816), TMTV (OR = 4.324, 95% CI: 1.095–17.067), and avg-radscore (OR = 3.176, 95% CI: 1.313-7. 686) were independent predictors of treatment response. The multivariable model comprising NCCN-IPI, BCL-2, TMTV, and avg-radscore outperformed conventional models and clinical-pathological models in predicting treatment response. (P<0.05).

Conclusion The radscore generated by AutoML can predict the treatment response of elderly DLBCL patients, potentially aiding in clinical decision-making.

Keywords ^{18}F-FDG PET/CT · Diffuse large B-cell lymphoma · Automated machine learning · Treatment response · Radiomics

Jincheng Zhao and Wenzhuo Zhao are co-first authors. They contributed equally to the work.

✉ Yue Teng
18304636833@163.com

✉ Jianxin Chen
chenjx@njupt.edu.cn

✉ Jingyan Xu
xjy1967@sina.com

Jincheng Zhao
3321092170@stu.cpu.edu.cn

Wenzhuo Zhao
1223014143@njupt.edu.cn

Man Chen
3293879768@qq.com

Jian Rong
1221014315@njupt.edu.cn

[1] Department of Hematology, School of Basic Medicine and Clinical Pharmacy, Nanjing Drum Tower Hospital, China Pharmaceutical University, Nanjing, China

[2] The Key Laboratory of Broadband Wireless Communication and Sensor Network Technology (Ministry of Education), Nanjing University of Posts and Telecommunications, Nanjing, China

[3] Department of Nuclear Medicine, Affiliated Hospital of Medical School, Nanjing Drum Tower Hospital, Nanjing University, Nanjing, China

Introduction

Lymphoma is a malignant tumor within the hematopoietic system, characterized by significant biological and clinical heterogeneity. Diffuse large B-cell lymphoma (DLBCL), the most common type of non-Hodgkin lymphoma (NHL), accounts for 45. 8% of NHL cases globally, with a particular prevalence in the elderly population, having a median age of onset of 57 years (Li et al. 2018). DLBCL in elderly patients exhibits significant molecular differences compared to younger patients, including a high proportion of MYC/BCL2 dual expression status and specific gene mutations, which are closely associated with poor disease prognosis(Klapper et al. 2012; Bohers et al. 2019). The standard treatment regimen for DLBCL is rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) chemotherapy, which, despite its widely validated efficacy, faces challenges in terms of treatment effectiveness and tolerability in the elderly population due to physiological changes (Song et al. 2021).

^{18}F-FDG PET/CT is essential for lymphoma diagnosis and treatment monitoring, but its susceptibility to false positives due to pathological heterogeneity and uptake in inflammation and other malignancies highlights the need for more precise imaging methods(Casali et al. 2021). The development of radiomics technology and artificial intelligence (AI) offers new perspectives for the diagnosis and treatment of lymphoma. Radiomics can extract a vast array of quantitative information from medical images, revealing the spatiotemporal heterogeneity of tumors and providing deeper biological insights into clinical oncology(Lambin et al. 2012). The application of AI in medical image analysis aids in the automatic identification and quantification of tumor heterogeneity, thereby providing significant value for tumor diagnosis, relapse detection, and personalized treatment. Automated machine learning (AutoML) platforms, such as AutoGluon, further simplify this process by automating data preprocessing and model training, enabling non-expert users to easily apply these advanced technologies(Erickson et al. 2020; Wu et al. 2020). Recent studies have demonstrated the immense potential of ^{18}F-FDG PET radiomics in tumor differential diagnosis, prognosis prediction, and the formulation of treatment plans(Kirienko et al. 2018; Lee et al. 2019; Kong et al. 2019). This study aims to develop a novel tool based on ^{18}F-FDG PET radiomics features to predict the therapeutic efficacy of elderly DLBCL patients treated with the R-CHOP regimen, providing important basis for the realization of personalized treatment.

Materials and methods

Patient data collection

A retrospective analysis was conducted on the ^{18}F-FDG PET/CT imaging data and clinical data of 175 elderly patients diagnosed with DLBCL from March 2015 to March 2023. The data were obtained from two medical institutions: Hospital A (95 cases) and Hospital B (80 cases). The study was approved by the Ethics Committee of Hospital A (Ethics Number: 2021-624-02) and the Ethics Committee of Hospital B (Ethics Number: 2023－954), and it adhered to the principles of the Declaration of Helsinki. The inclusion criteria were as follows: (1) confirmed elderly DLBCL by histopathological examination; (2) baseline examination and treatment response assessment by ^{18}F-FDG PET/CT (6–8 cycles after chemotherapy; training cohort: 74 patients with 6 cycles, 7 patients with 7 cycles, 14 patients with 8 cycles; validation cohort: 53 patients with 6 cycles, 9 patients with 7 cycles, 18 patients with 8 cycles); (3) chemotherapy with R-CHOP-like regimens; (4) no prior treatment; (5) no history of other tumors; (6) complete medical record. The study collected clinical data and pathological information, including gender, B symptoms, LDH levels, age, Ann Arbor staging, ECOG PS score, number of extranodal involvements, bone marrow involvement, IPI, NCCN-IPI, bulky disease, pathological type, MYC, BCL-2, double expression, BCL-6, Ki-67. Exclusion criteria: (1) liver SUVmean outside the range of 1. 3–3(Boellaard et al. 2015, p. 2); (2) no positive lesions on baseline PET/CT. The flowchart of patient selection in Fig. 1.

PET/CT scanning protocol

All patients were required to fast for at least 6 h before the scan to ensure blood glucose levels were below 11.1 mmol/L. Subsequently, ^{18}F-FDG (5. 18 MBq/kg) with a radiopharmaceutical dose of 185–370 MBq was administered intravenously. PET/CT scanning was performed 60 min post-injection, covering from the base of the skull to the upper thigh. At each bed position, emission data were acquired for 2 min. All patients underwent scanning using one of the following PET/CT systems: in the training cohort, Gemini GXL (reconstruction method: OSEM, PET slice thickness: 4 mm, PET pixel spacing: 4 mm); in the external validation cohort, Gemini GXL (reconstruction method: OSEM, PET slice thickness: 4 mm, PET pixel spacing: 4 mm), UM780PET/CT (reconstruction method: OSEM, PET slice thickness: 2.76 mm, PET pixel spacing: 2.9 mm), and GE Discovery PET/CT Clarity 710 (reconstruction method: VUE Point FX, PET slice thickness: 3.75 mm, PET pixel spacing: 5.47 mm). CT acquisition data

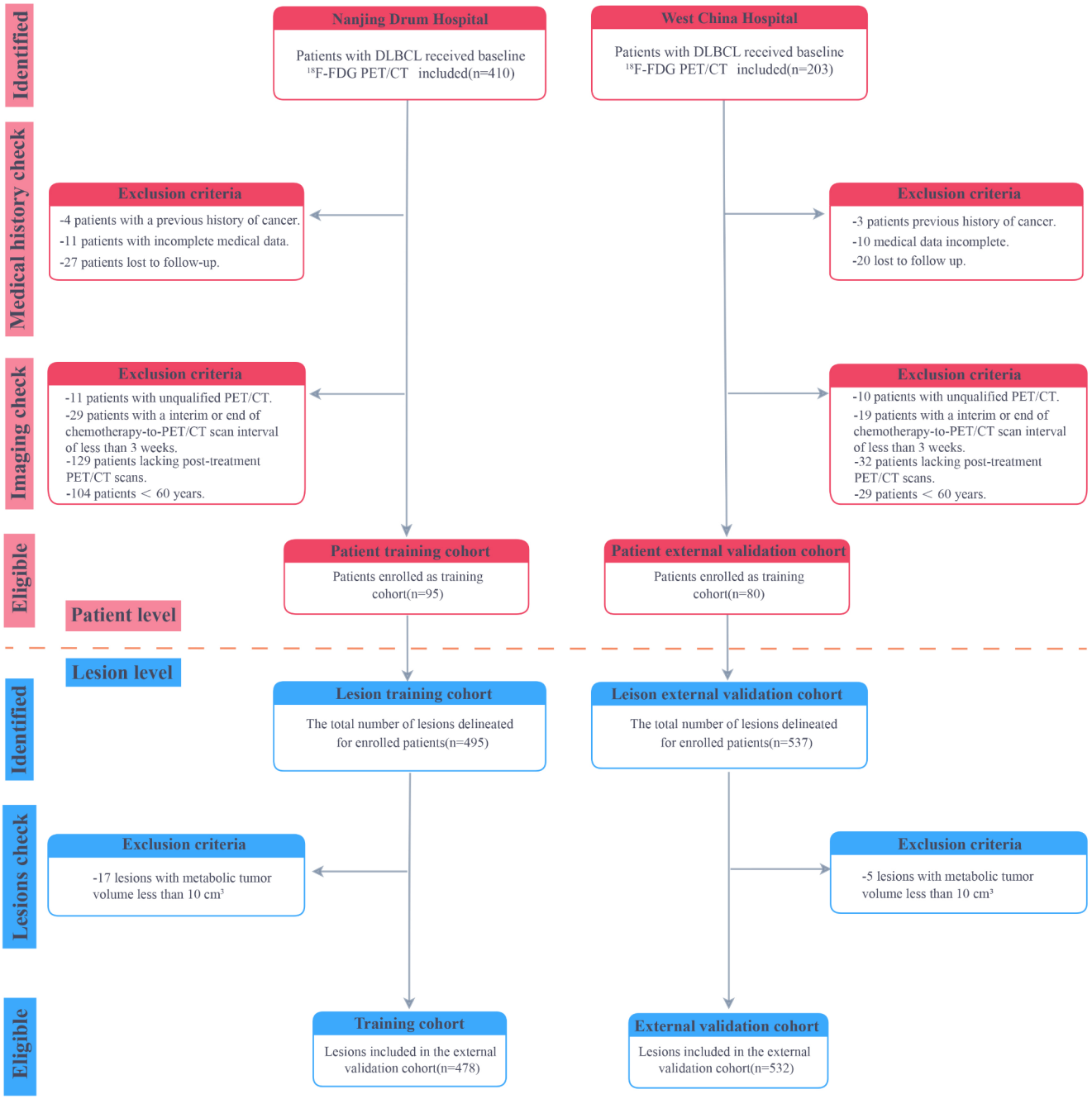

Fig. 1 Flowchart of patient selection

were used for attenuation correction. After image acquisition, the response line images were reconstructed to obtain transverse, coronal, and sagittal images of CT and PET.

ROI drawing and radiomics processing

The LIFEx-7.3.0 software was utilized for delineation(Nioche et al. 2018). A semi-automated method using a 41% maximum standardized uptake value (SUVmax) threshold was employed to outline the region of interest (ROI) on ^{18}F-FDG PET/CT images(Boellaard et al. 2010), from which the software automatically derived the tumor's SUVmax, metabolic tumor volume (MTV), and total lesion glycolysis (TLG). The total metabolic tumor volume (TMTV) was calculated by summing the MTVs of all lesions, with TLG defined as TMTV multiplied by the mean standardized uptake value (SUVmean). During the delineation process, care was taken to avoid misidentifying tissues with physiological ^{18}F-FDG uptake (such as the brain, bladder, kidneys, intestines, etc.) as lesions. If a lesion area was

adjacent to these physiologically active tissues and incorrectly categorized within the same ROI, the boundaries of the ROI needed to be redefined to ensure accuracy. When local or focal uptake of ^{18}F-FDG was observed in the liver and bone marrow, it should be recognized as a lesion area and delineated accordingly in the ROI. For the spleen, if its ^{18}F-FDG uptake exceeded 1.5 times the background uptake of the liver, whether focal or diffuse, the spleen should be considered an involved area and reflected in the ROI delineation. All PET/CT images were jointly reviewed by two nuclear medicine experts with 8 years of experience. In cases of disagreement, a senior nuclear medicine physician was involved to make the final decision.

Radscore construction

In this study, treatment outcomes were classified based on the Lugano classification, which is widely used in lymphoma research to evaluate treatment response. Patients were categorized into two groups: response and non-response. The response group comprised patients who achieved complete response (CR), while the non-response group included those with partial response (PR), stable disease (SD), and progressive disease (PD). This categorization reflects a standard practice in clinical research, enabling consistent and meaningful comparisons across studies. By employing these criteria, we ensured a robust and standardized evaluation of treatment efficacy. we encountered a highly imbalanced data ratio issue between the treatment response and non-response groups. To address this issue, we employed the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) to balance the distribution within the training cohort. The SMOTE technique narrows the disparity between minority and majority classes by creating synthetic data, applied solely to the training cohort to prevent data leakage, while the validation cohort remains unaltered to more accurately reflect actual clinical scenarios(Corino et al. 2018). The working principle of SMOTE includes the identification of minority samples, determination of K-nearest neighbors, and the generation of synthetic samples, thereby generating new samples within the feature space to achieve class balance. To construct a robust radscore, we utilized AutoGluon (version 0.7.0, available at https://auto.gluon. a.i.) for feature selection and model building. During the feature selection phase, features were ranked according to the weights calculated by AutoGluon, and the top 10 features were selected for radiomics analysis, a process performed in the training cohort and validated in an external validation cohort. The radiomics workflow is shown in Fig. 2.

Development and validation of the models

For patient-level analysis, the radscore of all lesions belonging to the same patient were summed and averaged. This average value represents the patient's average radiomics score (avg-radscore). Univariable and multivariable logistic analyses were employed to identify significant clinical risk factors, pathological risk factors, and PET metabolic indicators. These factors were then combined with the avg-radscore to construct a multi-parameter model. Calibration curves and receiver operating characteristic (ROC) curves were generated for these models to compare AUC. Additionally, clinical decision curve analysis (DCA) was performed

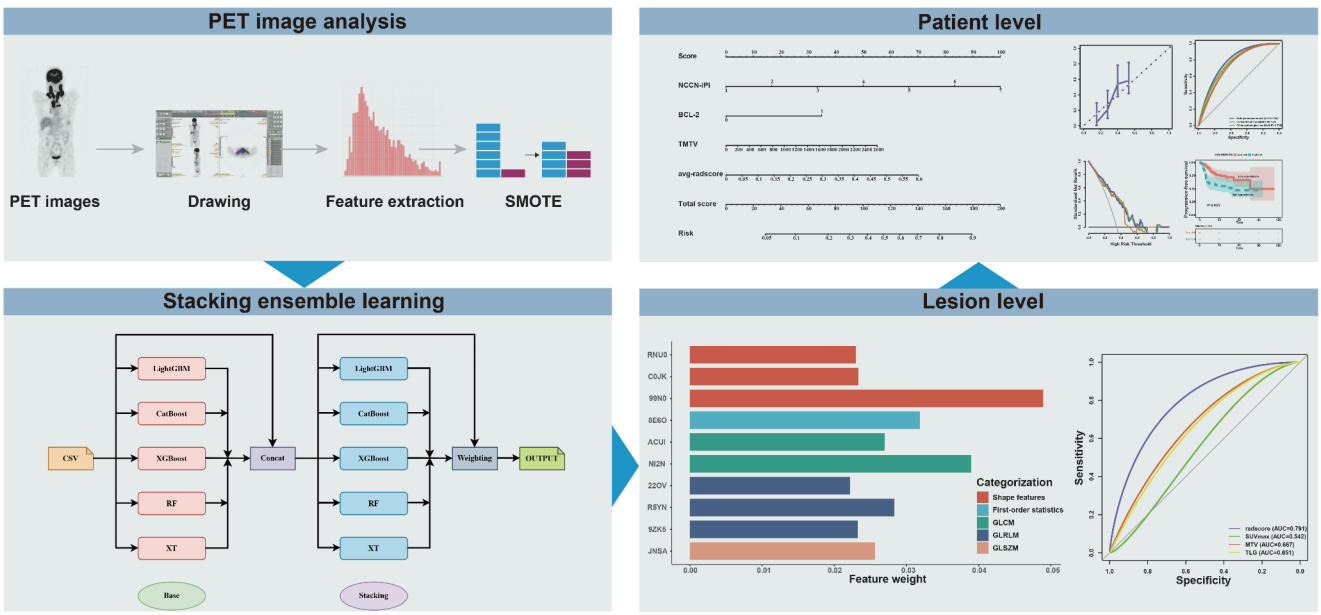

Fig. 2 Radiomics workflow

to estimate the false-positive rate of the models. Ultimately, the performance of these models was assessed in an external validation cohort. To evaluate the incremental predictive value of avg-radscore in response assessment, three models were constructed. The first model, referred to as the conventional model, combined metabolic indicators, pathological factors, and clinical variables. The second model, referred to as the clinic-pathological model, included only pathological and clinical variables. The third model, referred to as the multivariable model, integrated avg-radscore alongside the variables included in the conventional model.

Statistical analysis

Data analysis was performed using IBM SPSS 25 (https://www.ibm.com/products/spss-statistics), R software (version 4.2.2, www.R-project. org), and MedCalc software (version 20.216-64-bit, https://www.medcalc.org/). Quantitative data were tested for normality; those conforming to a normal distribution were expressed as x ± s, while those not conforming were expressed as M (Q1, Q3). Qualitative data were represented as frequency (percentage). Chi-square test and Fisher's exact test were used to assess differences in clinical characteristics, pathological characteristics, and metabolic parameters between the training and validation groups. The discriminative ability was measured by estimating the AUC. Univariable and multivariable logistic regression analyses were employed to explore the predictive value of potential independent response predictors and to construct models. Model calibration curves and decision analysis curves were plotted. In all analyses, a P-value less than 0. 05 was considered statistically significant.

Results

Patient characteristics

In this study, the training cohort included 95 patients, comprising 58 males and 37 females, with a median age of 69.00 (63.00, 73.00) years. The validation cohort consisted of 80 patients, with 46 males and 34 females, and a slightly lower median age of 68. 50 (65.00, 73.00) years. At the end of the follow-up period, a total of 53 patients experienced relapse, and 23 patients died in both the training and validation cohorts(Table 1). The interval between the end of therapy and the final PET/CT scan was 34. 00 (24. 5–54. 00) days.

Evaluation and external validation of Radscore for lesion response prediction

The median radscore in the training cohort was 0.077 (0.021, 0.161), and in the validation cohort, it was 0.063 (0.028, 0.135). The results of the ROC curve analysis indicated that the radscore generated by AutoML demonstrated higher accuracy in predicting treatment response at the lesion level compared to metabolic parameters (SUVmax, MTV, and TLG), with AUC values in the training group being 0.791, 0.542, 0.667, and 0.651, respectively, and in the validation group being 0.712, 0.616, 0.639, and 0.657, respectively. Using MedCalc software, the ROC curves for SUVmax, TMTV, TLG, and avg-radscore in the training group were plotted. The optimal threshold for each metric, corresponding to the maximum Youden's index, was automatically calculated by the software. The optimal thresholds for SUVmax, TMTV, TLG, and avg-radscore were 6.285, 198.144 cm^3, 1037.343, and 0.119, respectively, with AUCs of 0.529 (0.424–0.632), 0. 761 (0.662–0.842), 0.737 (0.636–0.822), and 0.647 (0.542–0.742), respectively.

Univariable and multivariable analysis results

Univariable logistic analysis indicated that in the prediction of treatment response, B symptoms (OR = 2.611, 95% CI:1.063–6.413), Ann Arbor stage (OR = 2.842, 95%CI: 1.023–7.894), IPI (OR = 2.593, 95%CI: 1.074–6.256), NCCN-IPI (OR = 5.791, 95%CI: 1.874–19.024), BCL-2 (OR = 3.714, 95%CI: 1.406–9.816), TMTV (OR = 10.051, 95%CI: 3.772–26.786), TLG (OR = 9.450, 95%CI: 3.365–26.536), and avg-radscore (OR = 3.176, 95%CI: 1.313–7.686) were significant predictors (all $P < 0.05$). Statistically significant clinical, pathological, metabolic, and radiomic factors were subjected to multivariable logistic regression analysis separately. Multivariable logistic analysis identified independent predictors of treatment response, including NCCN-IPI (OR = 5.427, 95% CI: 1.163–25.317), BCL-2 (OR = 3.714, 95% CI: 1.406–9.816), TMTV (OR = 4.324, 95% CI: 1.095–17.067), and avg-radscore (OR = 3.176, 95% CI: 1.313–7.686) (Table 2).

Evaluation and external validation of the multi-parameter model

The results of the logistic regression were utilized to construct a multi-parameter model for predicting treatment response, encompassing clinical factors (NCCN-IPI), pathological factors (BCL-2), PET metabolic parameters (TMTV), and avg-radscore(Fig. 3a). Calibration curves indicated a good agreement between the predicted and actual complete response rates in both the training and validation

Table 1 Demographic and clinical characteristics of the study population

Categorization	Variable	Training($n=95$)	Validation($n=80$)	p-value
Clinical factor	Sex			
	Female	37(38.9%)	34(42.5%)	0.634
	Male	58(61.1%)	46(57.5%)	
	B symptoms			
	No	65(68.4%)	63(78.8%)	0.125
	Yes	30(31.6%)	17(21.3%)	
	LDH level			
	Normal	50(52.6%)	49(61.3%)	0.252
	Elevated	45(47.4%)	31(38.8%)	
	Age			
	<80	81(85.3%)	77(96.3%)	0.014*
	≥80	14(14.7%)	3(3.8%)	
	Ann Arbor stage			
	I~II	30(31.6%)	35(43.8%)	0.097
	III~IV	65(68.4%)	45(56.3%)	
	ECOG PS			
	0~1	89(93.7%)	70(87.5%)	0.157
	≥2	6(6.3%)	10(12.5%)	
	Extranodal involvement			
	0–1	32(33.7%)	27(33.8%)	0.993
	≥2	63(66.3%)	53(66.3%)	
	Bone marrow involvement			
	No	91(95.8%)	71(88.8%)	0.077
	Yes	4(4.2%)	9(11.3%)	
	IPI			
	0–2	46(48.4%)	41(51.2%)	0.709
	≥3	49(51.6%)	39(48.8%)	
	NCCN-IPI			
	0–3	32(33.7%)	30(37.5%)	0.599
	≥4	63(66.3%)	50(62.5%)	
	Bulky disease			
	No	88(92.6%)	73(91.3%)	0.737
	Yes	7(7.4%)	7(8.8%)	
	CIRS-G score 1–2			
	<5	85(89.5%)	68(85.0%)	0.374
	≥5	10(10.5%)	12(15.0%)	
	CIRS-G score 3–4			
	0	65(68.4%)	65(81.3%)	0.053
	≥1	30(31.6%)	15(18.8%)	

Table 1 (continued)

Categorization	Variable	Training($n=95$)	Validation($n=80$)	p-value
Pathological factor	Pathological type			
	Non-GCB	70(73.7%)	66(82.5%)	0.163
	GCB	25(26.3%)	14(17.5%)	
	MYC			
	<40%	58(61.1%)	42(52.5%)	0.255
	≥40%	37(38.9%)	38(47.5%)	
	BCL-2			
	<50%	38(40.0%)	32(40.0%)	1.000
	≥50%	57(60.0%)	48(60.0%)	
	Double expression			
	Negative	62(65.3%)	49(61.3%)	0.583
	Positive	33(34.7%)	31(38.8%)	
	BCL-6			
	Negative	45(47.4%)	3(3.8%)	<0.001*
	Positive	50(52.6%)	77(96.3%)	
	Ki-67			
	<70%	29(30.5%)	21(26.3%)	0.533
	≥70%	66(69.5%)	59(73.8%)	
PET factor	SUVmax			
	≤6.285	7(7.4%)	1(1.3%)	0.117
	>6.285	88(92.6%)	79(98.8%)	
	TMTV			
	≤198.144	58(61.1%)	48(60.0%)	0.887
	>198.144	37(38.9%)	32(40.0%)	
	TLG			
	≤1037.343	48(50.5%)	45(56.3%)	0.450
	>1037.343	47(49.5%)	35(43.8%)	
Radiomics factor	avg-radscore			
	≤0.119	60(63.2%)	49(61.3%)	0.795
	>0.119	35(36.8%)_	31(38.8%)	

*p<0. 05

cohorts, demonstrating that the model's predictions are consistent with the actual observed outcomes(Fig. 3b, c). ROC curve analysis revealed that the multi-parameter model outperformed the conventional model (comprising NCCN-IPI, BCL-2, TMTV) and the clinic-pathological model (comprising NCCN-IPI, BCL-2) in predicting treatment response, in both the training set (AUC: 0.784 vs. 0.768 vs. 0.739) and the validation set (AUC: 0.789 vs. 0.771 vs. 0.765) (Fig. 3d, e). DCA showed that the multi-parameter model used to predict treatment response provided superior clinical benefit to patients across most threshold ranges(Fig. 3f, g).

Discussion

We selected 478 lesions as a training cohort to observe their response to chemotherapy via PET imaging and validated the key radiomic features generated radscore on an additional 532 independent lesions. The study results indicated that radscore effectively predicted the response of lesions to chemotherapy. Furthermore, we applied these radscores to patient-level analysis, using the method of calculating the avg-radscore of all lesions from a patient as a response characteristic. The results showed that these features effectively predicted the response of DLBCL patients to treatment.

AutoGluon, as an efficient AutoML framework, significantly enhanced the model's performance and generalization capability by automating key steps such as hyperparameter optimization, algorithm selection, and feature engineering. It can intelligently identify data types and select the most suitable model, while automatically building model ensembles to enhance predictive accuracy, greatly simplifying the application process of machine learning. This allows users, even those lacking in-depth professional knowledge, to quickly obtain high-quality models. In this study, the application of AutoGluon significantly improved the predictive accuracy of the response of elderly DLBCL patients after the end of treatment. In the field of radiomics, Zhao et al. (Zhao et al. 2023) successfully predicted the 2-year PFS and OS of DLBCL patients by adopting a stacked ensemble

Table 2 Univariable and multivariable logistic regression analyses for the prediction of end-of-treatment response in the training group

Categorization	Variable	Univariable analysis		Multivariable analysis	
		p-value	OR (95%CI)	*p*-value	OR (95%CI)
Clinical factor	Sex, Female/Male	0.092	2.196 (0.880–5.483)	-	-
	B symptoms, No/Yes	0.036*	2.611 (1.063–6.413)	0.309	1.651 (0.629–4.338)
	LDH level, Normal/Elevated	0.148	1.879 (0.799–4.419)	-	-
	Age, <80/≥80	0.491	1.500 (0.473–4.761)	-	-
	Ann Arbor stage, I-II/III-IV	0.045*	2.842(1.023–7.894)	0.509	1.501 (0.450–5.011)
	ECOG PS, 0–1/≥2	0.941	0.935 (0.162–5.397)	-	-
	Extranodal involvement, No/Yes	0.687	0.833 (0.343–2.023)	-	-
	Bone marrow involvement, No/Yes	0.124	6.100 (0.609–61.147)	-	-
	IPI, 0–2/≥3	0.034*	2.593 (1.074–6.256)	0.623	0.732 (0.210–2.544)
	NCCN-IPI, 0–3/≥4	0.003*	5.791(1.874–19.024)	0.031*	5.427 (1.163–25.317)
	Bulky disease, No/Yes	0.264	0.292 (0.034–2.532)	-	-
	CIRS-G score 1–2, <5/≥5	0.740	0.786 (0.189–3.263)	-	-
	CIRS-G score 3–4, 0/≥1	0.100	2.118 (0.866–5.182)	-	-
Pathological factor	Pathological type, Non-GCB/GCB	0.194	0.503 (0.178–1.418)	-	-
	MYC, <40%/≥40%	0.166	1.838 (0.776–4.350)	-	-
	BCL-2, <50%/≥50%	0.008*	3.714 (1.406–9.816)	0.008*	3.714 (1.406–9.816)
	Double expression, Negative/Positive	0.112	2.037 (0.847–4.899)	-	-
	BCL-6, Negative/Positive	0.482	1.357 (0.579–3.179)	-	-
	Ki-67, <70%/≥70%	0.665	0.818 (0.330–2.029)	-	-
PET factor	SUVmax, low/high	-	-	-	-
	TMTV, low/high	<0.001*	10.051 (3.772–26.786)	0.037*	4.324 (1.095–17.067)
	TLG, low/high	<0.001*	9.450 (3.365–26.536)	0.113	3.246 (0.758–13.906)
Radiomics factor	avg-radscore, low/high	0.010*	3.176 (1.313–7.686)	0.010*	3.176 (1.313–7.686)

*$p<0.05$

learning method, combining PET imaging radiomic features with clinical parameters, achieving AUC values of 0.791 and 0.843, respectively. This result not only verified the importance of radiomics in the evaluation of DLBCL prognosis but also highlighted the advantage of ensemble learning methods in dealing with such high-dimensional data.

Building upon the work of Zhao et al., this study further employed AutoGluon's multi-layer stacked ensemble model, which integrates the predictive results of multiple models at different levels, achieving in-depth mining and utilization of original data features. Unlike traditional single-layer models, multi-layer stacked ensemble models can re-evaluate and combine features at each level, thereby enhancing the model's ability to capture complex data structures. In addition, the AutoGluon framework ensures the stability and reliability of the model by conducting end-to-end training on the entire dataset and implementing k-fold cross-validation. In summary, the application of the AutoGluon framework in this study not only improved the predictive accuracy of the model for the treatment response of DLBCL patients but also provided a new methodological approach for radiomic data analysis. Through this multi-level, automated model fusion strategy, we can more effectively extract valuable information from complex radiomic data, providing more precise guidance for clinical decision-making.

Recent research by Ferrández et al(Ferrández et al. 2023) used CNNs with 18 F-FDG PET MIP images to predict treatment outcomes in DLBCL patients, achieving an AUC of 0.74 in external validation. In contrast, our study, which incorporates avg-radscore into a multivariable model using automated machine learning, achieved a higher AUC of 0.789 in the validation set. By integrating clinical, pathological, and imaging features, our model offers improved predictive performance and better clinical interpretability compared to the CNN approach, highlighting the potential complementarity of these methods for future research. Consistent with previous studies(Bishton et al. 2016; Go et al. 2018), our work also confirmed NCCN-IPI as an independent predictor of treatment response, with a higher non-CR rate in the high-risk group (NCCN-IPI≥4) compared to the low-risk group (NCCN-IPI = 0–3) (non-CR rate 41. 6% vs. 6. 5%, $P<0.001$). BCL-2 protein is one of the common pro-apoptotic inhibitors in the body, which can promote the development and differentiation of B lymphocytes. It has been found that BCL-2 is overexpressed in some DLBCL patients and is associated with lower patient survival rates(Yamagishi et al. 2015). In this study, the non-CR rate in the BCL-2 positive group was higher than that in the BCL-2 negative group (non-CR rate 36. 2% vs. 18. 6%, $P=0.012$).

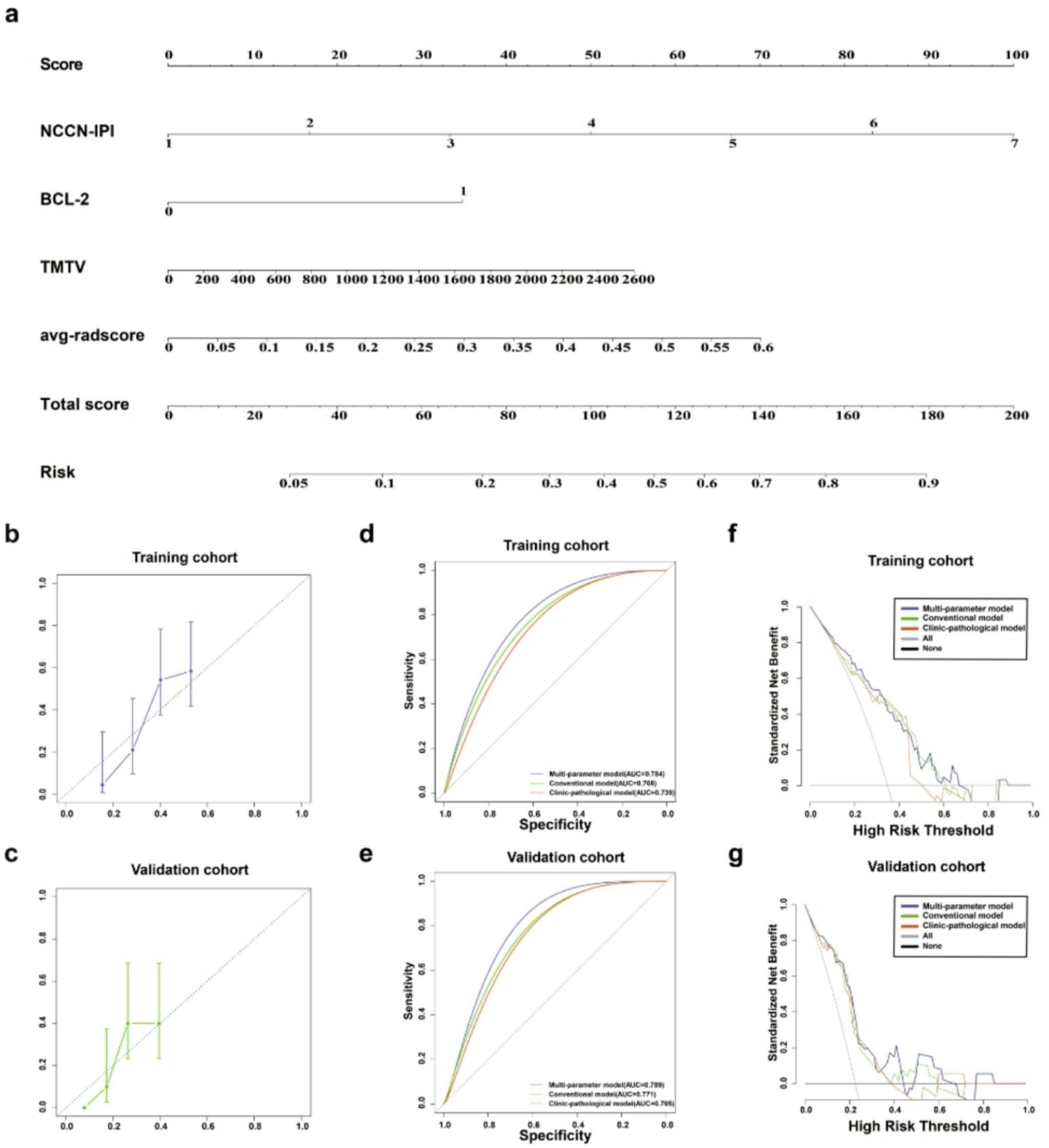

Fig. 3 Nomogram predicting treatment response in 95 patients (≥60 years) with DLBCL (**a**). Combined model calibration curves, ROC curves, DCA curves for training (**b**, **d**, **e**) and validation (**c**, **g**, **h**) group

TMTV reflects the metabolic burden of the tumor and is a more comprehensive indicator than the metabolic activity of a single lesion. Our results are similar to previous studies, which show that TMTV can significantly predict treatment response(Albano et al. 2021; Reed et al. 2021; Reinert et al. 2022). In the field of oncology research, single-dimensional information often provides only limited insights into the biological behavior of tumors. To deeply reveal the essential characteristics of tumors, a multidisciplinary and multi-angled approach is necessary, integrating key factors from

clinical, pathological, and metabolic levels. This comprehensive approach can build a more complete tumor biology model, providing more precise guidance for clinical decision-making. In this study, we established a multi-parameter model by selecting key factors such as clinical parameters, pathological features, and metabolic parameters, and combining them with radscore. The model aims to reveal the intrinsic characteristics and clinical behavior of tumors by comprehensively analyzing multi-dimensional data, thereby providing more accurate treatment strategies for patients. Through this interdisciplinary data fusion and analysis, we can more effectively predict treatment effects and provide a scientific basis for personalized medicine. The calibration curve showed a strong correlation between the predictions of the multi-parameter model and the actual results. Moreover, the ROC results showed that the multi-parameter model has superior predictive performance compared to conventional models and clinic-pathological models. By introducing the DCA method to consider the impact of false positives and false negatives on response judgment, we observed that the nomogram response prediction model is superior to other models in clinical application.

There are certain limitations to this study. First, this study adopted a retrospective design with a limited sample size. Therefore, future studies should validate the models established in this study on a broader dataset and consider adopting a prospective cohort study design to enhance the robustness and applicability of the model. This methodological improvement can further enhance the accuracy of the model's predictive ability and provide more reliable decision support for clinical practice. Second, the design framework of this study is based on extracting radiomic features from all lesions included in the study patients. Given that the process of manually outlining lesion boundaries is time-consuming and labor-intensive, it greatly limits the application potential of this method in clinical practice. Nevertheless, as an exploratory study, we have preliminarily verified the effectiveness of the research concept. In this multicenter study, the use of different scanners in the training and validation cohorts might have introduced variability in the results due to technical differences such as resolution and calibration methods. To mitigate this effect, we used a standardized software threshold (41% SUVmax) for lesion delineation and parameter calculation. Additionally, patients with liver SUVmean values outside the range of 1.3–3were excluded based on the predefined criteria to minimize systematic bias. However, we acknowledge that residual technical differences among scanners may still have influenced the results. Future studies should focus on developing more robust imaging data standardization protocols to further enhance the reliability of multicenter radiomics research outcomes.

In future studies, we plan to combine this radiomic analysis method with AI-based automated lesion delineation technology, aiming to narrow the gap between clinical research and daily clinical practice, thereby improving the efficiency and clinical operability of radiomic analysis. Finally, it should be clearly stated that ^{18}F-FDG, as a PET imaging agent, is not tumor-specific. Its accumulation in inflammatory lesions may lead to misjudgment of treatment effects. Existing studies have shown that the inflammatory response induced by treatment in lymphoma patients may persist for up to two weeks after chemotherapy, which may increase the risk of false-positive results in efficacy evaluation(Jerusalem et al. 2001). Therefore, when interpreting PET imaging results, it is essential to consider the accumulation of FDG in non-tumorous inflammatory processes to avoid misinterpretation of treatment effects.

Conclusion

This study developed and validated a PET radiomics signature using one of the AutoML models, AutoGluon, which can be used to predict the treatment response of elderly patients with DLBCL. The multi-parameter model, composed of NCCN-IPI, BCL-2, TMTV, and avg-radscore, demonstrated superior predictive efficacy in treatment outcome after completion compared to conventional models and clinic-pathological models.

Acknowledgements The authors would like to thank Chong Jiang for providing the data from West China Hospital, Sichuan University, which was instrumental to this study.

Author contributions Jincheng Zhao contributed to the study design, performed data analysis and interpretation, drafted the manuscript, approved the final version for publication, and assumed responsibility for all aspects of the work. Wenzhuo Zhao and Jian Rong contributed to the development of artificial intelligence models, assisted in manuscript drafting, approved the final version for publication, and took responsibility for all aspects of the work. Man Chen participated in image analysis, contributed to discussions of the results, and approved the final version for publication. Yue Teng, Jianxin Chen, and Jingyan Xu were involved in the study's conception and design, participated in image analysis and result discussions, and approved the final version of the manuscript. All authors reviewed and approved the final manuscript for submission.

Funding This work was partially supported by fundings for Clinical Trials from the Affiliated Drum Tower Hospital, Medical School of Nanjing University under Grant No. 2022-LCYJ-PY-44 and 2024-LCYJ-MS-25. This work was also partially supported by fundings for the Key Project of Medical Science and Technology of Nanjing under Grant No. ZKX21011.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Albano D, Dondi F, Mazzoletti A et al (2021) Prognostic impact of pretreatment 2-[18F]-FDG PET/CT parameters in primary gastric DLBCL. Medicina 57:498. https://doi.org/10.3390/medicina57050498

Bishton MJ, Hughes S, Richardson F et al (2016) Delineating outcomes of patients with diffuse large B cell lymphoma using the National comprehensive cancer network-international prognostic index and positron emission tomography-defined remission status; a population-based analysis. Br J Haematol 172:246–254. https://doi.org/10.1111/bjh.13831

Boellaard R, O'Doherty MJ, Weber WA et al (2010) FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. Eur J Nucl Med Mol Imaging 37:181–200. https://doi.org/10.1007/s00259-009-1297-4

Boellaard R, Delgado-Bolton R, Oyen WJG et al (2015) FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging 42:328–354. https://doi.org/10.1007/s00259-014-2961-x

Bohers E, Viailly P-J, Ruminy P et al (2019) Molecular characterisation of diffuse large B cell lymphoma in patients of 80 years old or more: clinical relevance in a multicentric randomized phase III study of the Lysa (SENIOR study). Blood 134:2765–2765. https://doi.org/10.1182/blood-2019-124444

Casali M, Lauri C, Altini C et al (2021) State of the Art of 18F-FDG PET/CT application in inflammation and infection: a guide for image acquisition and interpretation. Clin Transl Imaging 9:299–339. https://doi.org/10.1007/s40336-021-00445-w

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority Over-sampling technique. Jair 16:321–357. https://doi.org/10.1613/Jair.953

Corino VDA, Montin E, Messina A et al (2018) Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. Magn Reson Imaging 47:829–840. https://doi.org/10.1002/jmri.25791

Erickson N, Mueller J, Shirkov A et al (2020) AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. https://doi.org/10.48550/ARXIV.2003.06505

Ferrández MC, Golla SSV, Eertink JJ et al (2023) An artificial intelligence method using FDG PET to predict treatment outcome in diffuse large B cell lymphoma patients. Sci Rep 13:13111. https://doi.org/10.1038/s41598-023-40218-1

Go S-I, Park S, Kim JH et al (2018) A new prognostic model using the NCCN-IPI and neutrophil-to-lymphocyte ratio in diffuse large B-cell lymphoma. Tumori 104:292–299. https://doi.org/10.5301/tj.5000694

Jerusalem G, Beguin Y, Najjar F et al (2001) Positron emission tomography (PET) with 18F-fluorodeoxyglucose (18F-FDG) for the staging of low-grade non-Hodgkin's lymphoma (NHL). Ann Oncol 12:825–830. https://doi.org/10.1023/A:1011169332265

Kirienko M, Cozzi L, Rossi A et al (2018) Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions. Eur J Nucl Med Mol I 45:1649–1660. https://doi.org/10.1007/s00259-018-3987-2

Klapper W, Kreuz M, Kohler CW et al (2012) Patient age at diagnosis is associated with the molecular characteristics of diffuse large B-cell lymphoma. Blood 119:1882–1887. https://doi.org/10.1182/blood-2011-10-388470

Kong Z, Li J, Liu Z et al (2019) Radiomics signature based on FDG-PET predicts proliferative activity in primary glioma. Clin Radiol 74. https://doi.org/10.1016/j.crad.2019.06.019.:815.e15-815.e23

Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446. https://doi.org/10.1016/j.ejca.2011.11.036

Lee H, Lee D, Park S et al (2019) Predicting response to neoadjuvant chemotherapy in patients with breast cancer: combined statistical modeling using clinicopathological factors and FDG PET/CT texture parameters. Clin Nucl Med 44:21–29. https://doi.org/10.1097/RLU.0000000000002348

Li S, Young KH, Medeiros LJ (2018) Diffuse large B-cell lymphoma. Pathology 50:74–87. https://doi.org/10.1016/j.pathol.2017.09.006

Nioche C, Orlhac F, Boughdad S et al (2018) LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 78:4786–4789. https://doi.org/10.1158/0008-5472.CAN-18-0125

Reed JD, Masenge A, Buchner A et al (2021) The utility of metabolic parameters on baseline F-18 FDG PET/CT in predicting treatment response and survival in paediatric and adolescent hodgkin lymphoma. J Clin Med 10:5979. https://doi.org/10.3390/jcm10245979

Reinert CP, Perl RM, Faul C et al (2022) Value of CT-Textural features and Volume-Based PET parameters in comparison to serologic markers for response prediction in patients with diffuse large B-Cell lymphoma undergoing CD19-CAR-T cell therapy. JCM 11:1522. https://doi.org/10.3390/jcm11061522

Song Y, Zhou H, Zhang H et al (2021) Efficacy and safety of the biosimilar IBI301 plus standard CHOP (I-CHOP) in comparison with rituximab plus CHOP (R-CHOP) in patients with previously untreated diffuse large B-Cell lymphoma (DLBCL): A randomized, Double-Blind, Parallel-Group, phase 3 trial. Adv Ther 38:1889–1903. https://doi.org/10.1007/s12325-020-01603-8

Wu B, Yuan S, Li P et al (2020) Radar emitter signal recognition based on One-Dimensional convolutional neural network with attention mechanism. Sensors 20:6350. https://doi.org/10.3390/s20216350

Yamagishi M, Katano H, Hishima T et al (2015) Coordinated loss of MicroRNA group causes defenseless signaling in malignant lymphoma. Sci Rep 5:17868. https://doi.org/10.1038/srep17868

Zhao S, Wang J, Jin C et al (2023) Stacking ensemble Learning–Based [^{18} F]FDG PET radiomics for outcome prediction in diffuse large B-Cell lymphoma. J Nucl Med Jnumed. https://doi.org/10.2967/jnumed.122.265244..122.265244