Predicting prognosis outcomes of primary central nervous system lymphoma with high-dose methotrexate-based chemotherapeutic treatment using lipidomics

Yi Zhong[‡], Liying Zhou[‡], Jingshen Xu, and He Huang[◉]

All author affiliations are listed at the end of the article

Corresponding Author: He Huang, Mailing Add: 2005 Songhu Road, Shanghai City, 200438, China (he_huang@fudan.edu.cn).

[‡]These authors contributed equally.

Abstract

Background. Primary central nervous system lymphoma (PCNSL) is a rare extranodal lymphomatous malignancy which is commonly treated with high-dose methotrexate (HD-MTX)-based chemotherapy. However, the prognosis outcome of HD-MTX-based treatment cannot be accurately predicted using the current prognostic scoring systems, such as the Memorial Sloan-Kettering Cancer Center (MSKCC) score.

Methods. We studied 2 cohorts of patients with PCNSL and applied lipidomic analysis to their cerebrospinal fluid (CSF) samples. After removing the batch effects and features engineering, we applied and compared several classic machine-learning models based on lipidomic data of CSF to predict the relapse of PCNSL in patients who were treated with HD-MTX-based chemotherapy.

Results. We managed to remove the batch effects and get the optimum features of each model. Finally, we found that Cox regression had the best prediction performance (AUC = 0.711) on prognosis outcomes.

Conclusions. We developed a Cox regression model based on lipidomic data, which could effectively predict PCNSL patient prognosis before the HD-MTX-based chemotherapy treatments.

Key Points

1. Our results provide an applicable model that predicts the prognosis of HD-MTX-based treatment for PCNSL better than any other machine-learning models tested.

2. Our results highlight the potential of lipidomics as biomarkers in clinical and medical research.

Primary central nervous system lymphoma (PCNSL), an extranodal non-Hodgkin lymphoma (NHL),[1] is an aggressive neoplasm confined to the brain, eyes, cranial nerves, leptomeninges, or spinal cord in the absence of disease outside of the central nervous system.[2] It is a rare disease whose incidence is 0.4/100 000 individuals,[1] making up about 1% of NHL cases and 3%–4% of all intracranial tumors.[3] Notably, PCNSL incidence is increased in immunocompromised patients, in which the tumor cells are typically Epstein-Barr virus (EBV)-positive.[4]

PCNSL has been shown to be highly sensitive to chemotherapy and radiation,[5] with high doses of methotrexate-based treatments proving to be the most effective.[3] HD-MTX can also be combined with other chemotherapies.[6] Alternatively, rituximab, a monoclonal anti-CD20 antibody has been widely used as a treatment option,[4] but there is no consensus for a single superior regimen. Unfortunately, despite available treatments, 25%–50% relapse after initial response.[7] Chemoradiation treatments can prolong lifespan, but patients are often faced with debilitating side effects, such as neurotoxicity, psychomotor slowing, neurocognitive impairments, memory dysfunction, gait ataxia, behavioral changes, and incontinence associated with significant functional decline.[8]

Importance of the Study

MTX is the therapeutic backbone in the treatment of PCNSL. However, it could produce toxicity and side effects such as bone marrow suppression, pulmonary toxicity, nephrotoxicity, hematological toxicity, and an increased risk of infections. Following HD-MTX-based treatment, 60% of patients experience reversible hepatitis, with 25% developing hyperbilirubinemia. Additionally, lymphoma patients undergoing HD-MTX-based treatment have a 9.1% probability of experiencing nephrotoxicity. It is essential for patients to find the balance between the benefits and risks of HD-MTX-based treatment to better their quality of life. Faced with a scarcity of clinical samples, and the challenge of the high dimensionality of lipidomic data, we embarked on comparing several classic machine-learning models. Ultimately, our analysis revealed that the Cox regression model exhibited the most promising prediction performance and lipidomic data could offer guidance in selecting clinical therapeutic regimens.

Taking into account these risk factors, it would be beneficial to be able to predict the effectiveness of HD-MTX-based treatment. This would allow healthcare professionals and patients to make calculated and informed decisions when choosing treatment for PCNSL.

Several clinical prognostic models have been developed for PCNSL, including the International Extranodal Lymphoma Study Group (IELSG) score,[9] the Nottingham-Barcelona score,[10] and the Memorial Sloan-Kettering Cancer Center (MSKCC) classification.[11] Recently, several studies have utilized machine learning to assist the diagnosis[12,13] and survival prediction of PCNSL.[14] However, these models are limited to predict the prognosis only based on clinical and biochemical characteristics,[15,16] and are not designed to specifically consider the outcome of HD-MTX-based treatments. As MTX interferes with one-carbon metabolism[17] and metabolomics has been proven to be helpful in the diagnosis of PCNSL,[18] we have sufficient grounds to believe that the application of "omics" to machine-learning models can provide better and more accurate prognostic predictions.

As a disease with high heterogeneity and complex pathogenesis, PCNSL is a challenging yet essential disease to characterize. Alternatively, lipids have been established as robust biomarkers for cancerous diseases. Specific lipid profiles are emerging as unique disease biomarkers, with diagnostic, prognostic, and predictive potentials.[17]

Lipids serve crucial roles in cellular function, acting as structural components of membranes, energy sources for metabolic processes, and signaling molecules. As such, bioactive lipids play a significant role in controlling immunological processes, inflammation, and maintaining homeostasis.[18] The complete set of lipids present within a cell is referred to as the cell's "lipidome," potentially comprising over 1000 distinct lipid species.[19] In the context of cancer, both cancer cells and other cells within the tumor microenvironment undergo significant metabolic reprogramming. Consequently, alterations in lipid metabolism are closely linked to the pathogenesis of cancer, leading to the identification of distinct lipid profiles as potential biomarkers for diagnosis, prognosis, and prediction of therapeutic outcomes. Lipidomics have been applied to the studies on ovarian cancer,[20] lung cancer,[21,22] and pancreatic cancer.[23]

In cases of PCNSL, tumors are highly invasive and can infiltrate surrounding tissues along blood vessels.[1] The degradation and remodeling of these tumors' extracellular matrix release PCNSL-related lipid biomarkers to cerebrospinal fluid.[24] Using metabolomics data of patient CSF samples, several metabolic biomarkers have been identified, and a Logistic Regression (LR) model to predict the effectiveness of HD-MTX-based chemotherapeutic treatments has been established.[16]

MTX-resistant cells have been proven to have altered lipid metabolism compared to nonresistant cancer cells,[25] and identification of such cells could be possible using lipidomic analysis of CSF. Currently, there are no reports on predicting the prognosis of PCNSL patients undergoing MTX-based chemotherapy through CSF lipidome.[26,27]

In this work, we managed to avoid common pitfalls in machine learning, like batch effects and leaky preprocessing, and offered a standard operating procedure for processing mass spectrometric datasets gathered from multiple batches. Based on lipidomic characteristics that a relatively small number of clinical samples but relatively deep dimensional data structure, we constructed and compared several classic machine-learning models which included Cox regression, LR, K-nearest neighbor, Naive Bayes (NB), decision tree, Random Forest (RF), XGBoost and Support Vector Machine (SVM). Finally, we found that the Cox regression model had the best prediction performance, which may be helpful for the prognostic prediction of PCNSL patients with HD-MTX-based treatment.

Materials and Methods

Patients and Grouping Criteria

Progression-free survival (PFS) was used as the criterion for assessing the treatment effectiveness of HD-MTX chemotherapy. Patients with PFS longer than 12 months were classified as the good prognosis group, patients with PFS shorter than 12 months and experiencing relapse were classified as the bad prognosis group, and patients who did not relapse at the last follow-up and whose PFS were shorter than 12 months could not be categorized according to this criterion and were therefore excluded from the analysis.

All patients gave their informed consent for the utilization of samples, following the guidelines approved by

the respective Ethical Committees on Human Research at Huashan Hospital institution.

All patients received an HD-MTX-based chemotherapy as a combination regimen. But none of these patients received consolidation, eg, with high-dose chemotherapy plus–minus autologous stem cell transplant. Each HD-MTX treatment was administered as a 3-hour infusion. Prehydration and alkalinization protocols were followed at least 72 hours prior to MTX administration. Standard leucovorin rescue was initiated 24 hours after the start of each HD-MTX infusion. Eight cycles of treatment were repeated every 3 weeks, until tumor progression or toxicity occurred. The clinical features of all patients were collected from their medical records, including, but not limited to, age, gender, height, weight, performance status, time of diagnosis, surgical resection, biopsy type, lesion site, number of lesions, HIV status, and serum lactate dehydrogenase (LDH) levels. Magnetic resonance imaging (MRI) was used to assess the location and quantity of lesions in all patients. The continuity of treatment was evaluated by contrast-enhanced MRI scans following every treatment cycle, and by Positron Emission Tomography-Computed Tomography (PET-CT) after every 3 cycles of treatment and after all therapeutic procedures. PFS was calculated from the date of diagnosis to the date of disease progression, the first relapse, death from any cause, or the last follow-up.

Sample Collection

All patients underwent biopsy before CSF sampling and were diagnosed pathologically. Each PCNSL patient underwent a lumbar puncture to examine CSF at baseline before chemotherapy. An additional 2 mL of CSF was collected, centrifuged for 10 minutes to remove cells, and frozen at –80°C for future use.

Lipid Extraction

Lipid extraction was performed following a published protocol.[28] We transferred 50 μL CSF sample, 200 μL of water, and 1 mL of methanol at 4°C into a clean glass centrifuge tube. Five milliliters of methyl tert-butyl ether (MTBE) was then added and the mixture was vortexed for 1 minute. The glass centrifuge tube containing the homogenate was rocked on a shaker at 140 rpm for 1 hour at 23°C. In total, 1.25 mL of water was then added to the glass centrifuge tube followed by another minute of vortexing. The homogenate was centrifuged at 4°C at $1000 \times g$ for 10 minutes, after that 2-phase layers could be observed in the glass centrifuge tube. A total of 4 mL of the top lipid phase supernatant were collected and dried under a stream of nitrogen. The extracted lipid samples were stored at –80°C in preparation for liquid chromatography–mass spectrometry with tandem mass spectrometry (LC-MS/MS) analysis.

Untargeted Lipidomic Analysis

The lipidomic approach was adopted from a published method.[29] Lipid samples, extracted as per the previous section, were resuspended in 100 μL of isopropanol: acetonitrile: water (v/v/v, 30:65:5), and 10 μL was injected into an Orbitrap Exploris 480 LC-MS/MS (Thermo, USA) coupled to high-performance liquid chromatography (HPLC) system (Shimadzu, Kyoto, Japan). Lipids were eluted using a 3 μm, 2.1 × 150 mm C30 column (Waters, USA) with a flow rate of 260 μL/min using buffer A (10 mM ammonium formate at a 60:40 ratio with acetonitrile: water) and buffer B (10 mM ammonium formate at a 90:10 ratio with isopropanol: acetonitrile). Gradients were held in 32% buffer B for 0.5 minutes and run from 32% buffer B to 45% buffer B at 0.5–4 minutes; from 45% buffer B to 52% buffer B at 4–5 minutes; from 52% buffer B to 58% buffer B at 5–8 minutes; from 58% buffer B to 66% buffer B at 8–11 minutes; from 66% buffer B to 70% buffer B at 11–14 minutes; from 70% buffer B to 75% buffer B at 14–18 minutes; from 75% buffer B to 97% buffer B at 18–21 minutes; 97% buffer B was held from 21 to 25 minutes; from 97% buffer B to 32% buffer B at 25–25.01 minutes; and 32% buffer B was held from 25.01 to 33.01 minutes. All ions were acquired through non-targeted multiple reaction monitoring (MRM) transitions associated with their predicted retention time, switching between positive and negative modes. ESI voltage was +5500 and –4500 V in positive and negative modes, respectively.

Procedures for Data Processing and Modeling

One batch of data was generated from each of the 2 cohorts of patients subjected to untargeted lipidomic analysis. Firstly, we merged data from different batches, retaining only the lipids which were detected in both batches. Secondly, all the MS data were processed with Log10 transformation using Metaboanalyst (https://new.metaboanalyst.ca/) to get normalized data. Missing values were replaced by one-fifth of minimal positive values of their corresponding variables, which is a common strategy against missing values that are due to low abundance or deficiency of metabolites.[30] Thirdly, we used the Combat function of R package *sva*,[31] based on empirical Bayes methods to correct for the batch effect. Finally, in order to remove the dependence of the rank of the metabolites on the average concentration and the magnitude of the fold changes,[32] the data were auto-scaled using metaboanalyst. Specifically, all data were mean-centered and divided by the standard deviation of each variable.[33]

Using stratified sampling, we randomly divided the samples into a training set and a testing set in a 70:30 ratio. Feature engineering, model training, and hyperparameter optimization were based on data in the training set. The machine-learning models applied in our study include Cox regression, LR, K-nearest neighbor, NB, decision tree, RF, XGBoost, and SVM.

In the Cox regression model, features associated with prognosis were determined by univariate Cox analysis. The 15 features with the smallest *P*-value were included in the multivariate Cox analysis. Features of other models were determined by Sequential Selection Algorithms. Specifically, we used classification accuracy as the standard of model iteration, gradually increasing the number of features from 0. We selected the feature combination with the highest accuracy and the lowest number of features.

We tuned hyper-parameters of KNN, decision tree, and SVM models based on their classification AUC performance

Table 1. Demographic and Clinical Baseline Characteristic for 2 Batches of Patients

	Overall (*n* = 81)	Test cohort (*n* = 24)	Train cohort (*n* = 57)	*P*
Age, years	57.85 (11.52)	59.75 (11.45)	57.05 (11.55)	.339
Gender				.391
Male	43 (53.1)	15 (62.5)	28 (49.1)	
Female	38 (46.9)	9 (37.5)	29 (50.9)	
Weight, kg	65.80 (15.51)	65.69 (9.41)	65.84 (17.52)	.968
BMI, kg/m^2	1.70 (0.18)	1.72 (0.15)	1.69 (0.20)	.445
Group				.137
Good prognosis	42 (51.9)	16 (66.7)	26 (45.6)	
Bad prognosis	39 (48.1)	8 (33.3)	31 (54.4)	
Relapse				.369
Yes	45 (55.6)	11 (45.8)	34 (59.6)	
No	36 (44.4)	13 (54.2)	23 (40.4)	
CSF protein, g/L	0.78 (0.53)	0.76 (0.51)	0.78 (0.54)	.870
Deep brain involvement				.099
Yes	59 (72.8)	21 (87.5)	38 (66.7)	
No	22 (27.2)	3 (12.5)	19 (33.3)	
ECOG score				.327
0	18 (22.2)	5 (20.8)	13 (22.8)	
1	28 (34.6)	11 (45.8)	17 (29.8)	
2	17 (21.0)	6 (25.0)	11 (19.3)	
3	12 (14.8)	1 (4.2)	11 (19.3)	
4	6 (7.4)	1 (4.2)	5 (8.8)	
LDH, U/L	190.85 (41.39)	181.33 (36.80)	195.00 (42.90)	.179
No. of lesions				.689
0	20 (24.7)	5 (20.8)	15 (26.3)	
1	43 (53.1)	15 (62.5)	28 (49.1)	
2	7 (8.6)	2 (8.3)	5 (8.8)	
3	11 (13.6)	2 (8.3)	9 (15.8)	
Regimen				.084
MTX	46 (56.8)	14 (58.3)	32 (56.1)	
MTX + IDA	17 (21.0)	2 (8.3)	15 (26.3)	
MTX + IDA + R	2 (2.5)	0 (0.0)	2 (3.5)	
MTX + R	16 (19.8)	8 (33.3)	8 (14.0)	

Abbreviation: BMI, body mass index; CSF, cerebrospinal fluid; ECOG, Eastern Cooperative Oncology Group; LDH, serum lactate dehydrogenase; MTX, methotrexate; IDA, Idarubicin; R, Rituximab.
For continuous variables, data were presented as mean (SD) and the student *t*-test was applied; for categorical variables, data were presented as numbers (percentage) and the Chi-square test was applied.

using 3-fold cross-validation in the training dataset. The hyper-parameters of RF and XGBoost models were optimized using R packages RandomForest[34] and Caret,[35] respectively.

Statistical Analysis

The patient's baseline characteristics were summarized using descriptive statistics, and descriptive analyses were conducted for all variables. Survival curves were plotted using the Kaplan–Meier method.[36] All statistical analyses were performed using R 4.3.1, Metaboanalyst, Graphpad Prism version 9.0.0 (Graphpad Software).

Results

Baseline Characteristics of Patients

In this study, 2 batches of patient CSF samples were collected; their clinical baseline characteristics are shown in Table 1. The first cohort consisted of 52 patients diagnosed
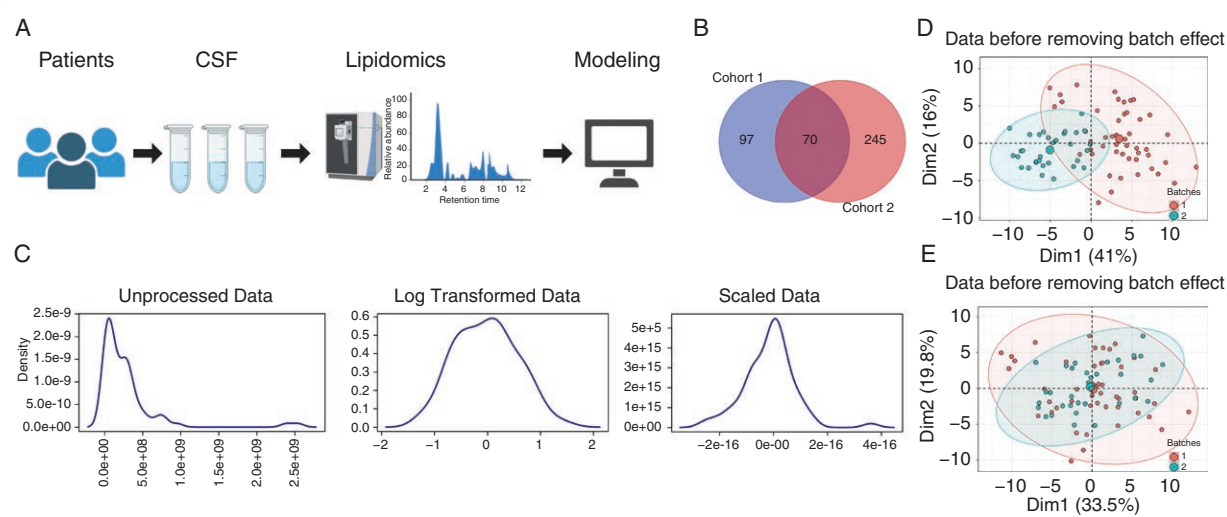
Figure 1. Results of data pre-processing: (A) The workflow of lipidomic analysis; (B) Venn diagram of lipid detected in two cohorts; (C) Data structure before/after Log10 transformation and autoscaling; (D) PCA plot of data before removing batch effect; (E) PCA plot of data after removing batch effect.

with PCNSL between 2018 and 2020. Diagnosis for all participants was confirmed based on the WHO classification of hematologic and lymphoid tissue tumors.[37] Patients with systemic lymphoma or immunodeficiency diseases were excluded from our study. Four additional patients from this cohort were excluded, because they had not relapsed by the date of the final follow-up appointment, or had a PFS period shorter than 12 months. The second cohort included 35 patients diagnosed with PCNSL between 2020 and 2021. Their samples were treated following the same inclusion–exclusion criteria and sample handling procedures as the first cohort of patients. Follow-ups were conducted until July 2022. Due to insufficient clinical information, 2 samples were removed, leaving a final sample size of 33 patients for the second cohort.

We then classified 81 patients into training sets and test sets as we mentioned in *Materials and Methods*. The patients exhibited relatively similar characteristics (in terms of Age, BMI, CSF protein, deep brain involvement, ECOG, gender distribution, prognosis grouping, lactate dehydrogenase (LDH), number of lesions, regimen, and weight) across the 2 groups.

Data Preprocessing

CSF was collected and processed by untargeted lipidomics analysis using an LC-MS-based technique, as illustrated in Figure 1A. We selected 70 lipids that were detected in both batches (Figure 1B). All mass spectrum data were Log10 transformed and normalized before the batch effect correction (Figure 1C). This accomplished several effects: Elimination of heteroscedasticity, removal of the dependence of metabolite rank on average concentration and fold change magnitude, and demonstration of biological significance. To fit a model that will generalize to new batches, we removed the batch effects. Principal component analysis indicated successful correction of batch effects (Figure 1D and E).

Feature Engineering and Hyper-Parameters Optimization

In the Cox regression model, features associated with prognosis were determined by univariate Cox analysis. Fifteen features with the lowest *P*-values were included in the multivariate Cox analysis, as shown in the forest plot (Figure 2A). Features of other applied models (Table 2) were determined by Sequential Selection Algorithms (Figure 2B–H). Details of Sequential Selection Algorithms are described in the *Materials and Methods* section.

To obtain a better model for HD-MTX-based PCNSL lipidome prognosis prediction, we also tuned hyperparameters (Table 3) for multiple machine-learning models, including KNN, decision tree, RF, XGBoost, and SVM.

Comparison of Machine-Learning Models

According to the model prediction values, we applied receiver operating characteristic curves analysis (Figure 3A–B). Area under the curve (AUC) > 0.70 indicates that the model discrimination is acceptable in a clinical-based model.[38] All models achieved good performance with AUC > 0.7 in the training dataset and the AUC value of the RF model even reached 1. However, when faced with the testing dataset, their performances were significantly decreased, with RF and K-nearest neighbor models performing worst of all (Figure 3A–B). From all tested models, Cox regression performed the best, reaching an AUC value of 0.711. It was closest to the AUC value of the training dataset, which indicates we did not overfit it.

In order to evaluate if the Cox regression model was overfitting, the tested calibration curves of the models were tested (Figure 3C–D). A good calibration degree was evaluated by the overlapping of the calibration curve to Y = X curve,[39] and the Cox regression calibration curve indicated a better fitting degree without overfitting. However,
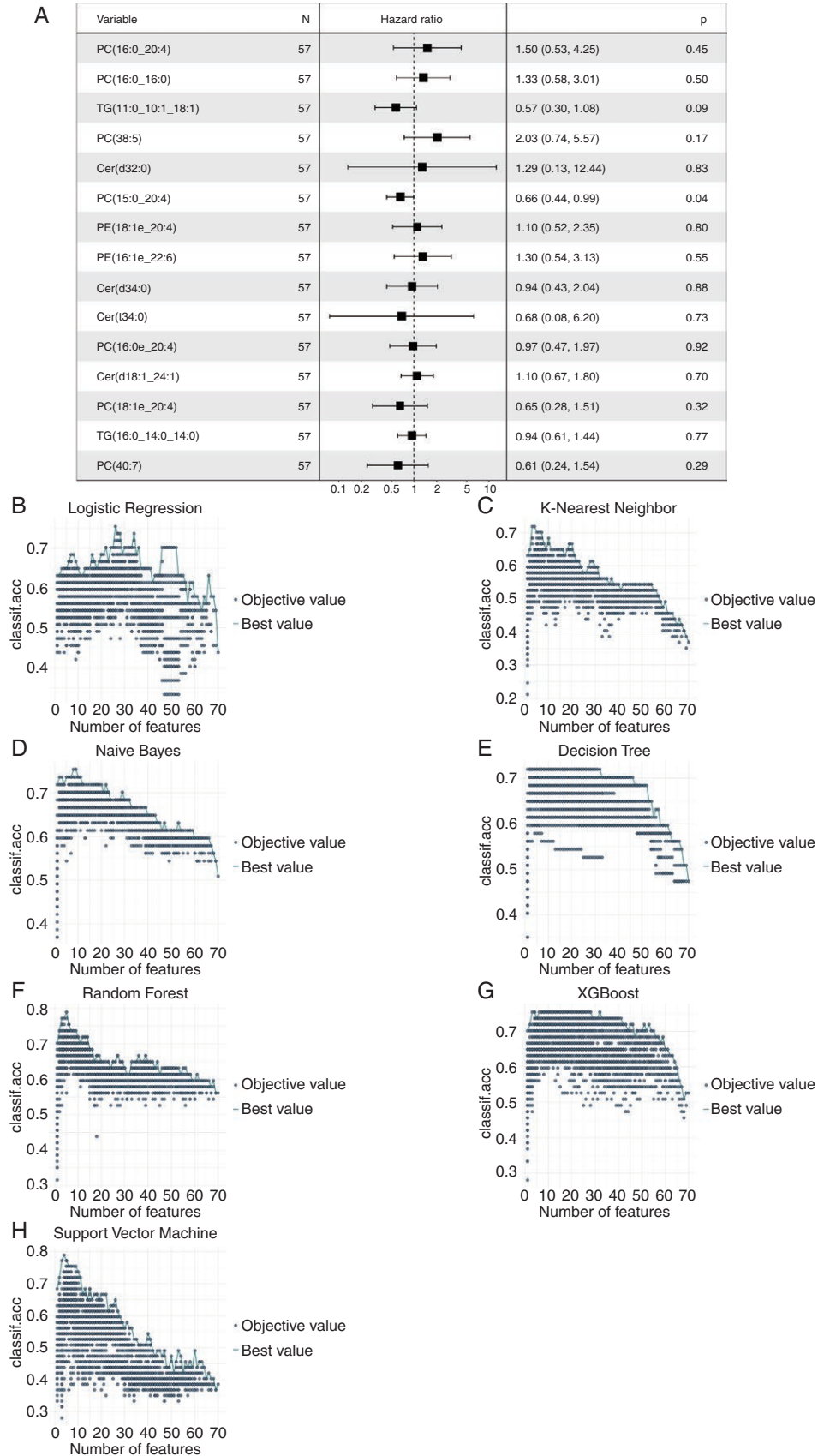
Figure 2. Results of feature engineering: (A) Features included in Cox Regression model; (B–H) Relationship between the number of features and classification accuracy in Logistic Regression, K-Nearest Neighbor, Naive Bayes, Decision tree, Random Forest, XGBoost, and Support Vector Machine.

Table 2. Features for Each Model

Models	Features
Cox regression	PC(16:0_20:4), PC(16:0_16:0),TG(11:0_10:1_18:1), PC(38:5), Cer(d32:0), PC(15:0_20:4), PE(18:1e_20:4), PE(16:1e_22:6), Cer(d34:0), Cer(t34:0), PC(16:0e_20:4), Cer(d18:1_24:1), PC(18:1e_20:4),TG(16:0_14:0_14:0), PC(40:7)
Logistic regression	Cer(d18:1_24:1), Cer(d32:0), Cer(d34:0), Cer(d36:0), Cer(d36:2), ChE, ChE(18:2), ChE(22:6), DG(18:0_18:0), DG(26:2e), DG(32:0e), DG(32:1e), LPC(16:0), LPC(18:0), LPC(18:2), MG(18:1), PC(15:0_20:4), PC(16:0_16:0), PC(18:1e_20:4), PC(36:4), PC(38:5), PC(38:6), PE(16:1e_22:6), PE(18:0_20:6), SM(d32:1),TG(16:0_16:1_16:1)
K-nearest neighbor	DG(26:2e),TG(16:0_18:1_18:2),TG(18:1_18:2_18:2)
Naive Bayes	Cer(t34.0), ChE(18:2), DG(32:0e), DG(6:0_14:1), PC(36:4), PE(16:0_22:6),TG(15:0_16:0_16:1), TG(18:1_18:2_18:2)
Decision tree	TG(18:1_18:2_18:2)
Random forest	Cer(d22:0_18:0), DG(34:4e), DG(36:2e), DG(6:0_14:1),TG(16:0_14:0_14:0)
XGBoost	DG(34:4e), LPC(18:1),TG(18:1_18:2_18:2)
Support vector machine	PC(15:0_20:4), PC(38:6),TG(18:1_18:1_18:2),TG(18:1_18:2_18:2)

Table 3. Hyper-Parameters for Models Which Need to Tune Hyper-Parameters

Algorithms	Hyper-parameters
KNN	k = 8 and other default parameters
Decision tree	minbucket = 2, maxdepth = 2 and other default parameters
Random forest	mtry = 1, ntree = 250 and other default parameters
XGBoost	max_depth = 1, eta = 0.3, colsample_bytree = 0.6, min_child_weight = 1, subsample = 1, gamma = 0, nround = 50
SVM	gamma = 10^-1, cost = 10^1, kernel="radial"

the calibration curves of the other machine-learning models were found to be overfitted.

To avoid varied model performance caused by different feature sets, we applied features which were generated by univariate Cox analysis to all models. Surprisingly, the performances of almost all models deteriorated (Figure S1), which indicated each model should undergo independent feature engineering to achieve the best model performance.

Performance of Cox Regression model

The results above showed that the Cox regression model was the most accurate model at predicting the effectiveness of high-dose methotrexate-based chemotherapeutic treatment on PCNSL. The formula used by the model can be written as follows:

$$H(t) = H_0(t) \times exp\,[\,0.4513 \times PC(16:0\setminus20:4) + 0.28182 \\ \times PC(16:0\setminus16:0) + 0.7067 \times PC(38:5) + 0.2509 \\ \times Cer(d32:0) + 0.09848 \times PE(18:1e\setminus20:4) \\ + 0.26526PE(16:1e\setminus22:6) + 0.9525 \\ \times Cer(d18:1e\setminus24:1) - 0.55617 \\ \times TG(11:0\setminus10:1\setminus18:1) - 0.44201 \\ \times PC(15:0\setminus20:4) - 0.06013 \times Cer(t34:0) - 0.38158 \\ \times Cer(t34:0) - 0.03433 \times PC(16:0e\setminus20:4) - 0.42982 \\ \times PC(18:1e\setminus20:4) - 0.49833 \times PC(40:7)\,]$$

The optimal cutoff was 0.0744. When $H_{score} < 0.0744$, it meant the patients would have good prognosis; while when $H_{score} \geq 0.0744$, it meant the patients would have bad prognosis. In this formula, lipids with negative coefficients include TG(11:0\10:1\18:1), PC(15:0\20:4), Cer(t34:0), PC(16:0e\20:4), PC(18:1e\20:4), Cer(d34:0), TG(16:0\14:0\14:0), PC (40:7), suggesting an association with good prognosis. On the other hand, lipids with positive coefficients include PC(16:0\20:4), PC(16:0\16:0), PC(38:5), PE(18:1e\20:4), Cer(d32:0), PE(16:1e\22:6), Cer(d18:1e\24:1), suggesting an association with poor prognosis.

We then used Kaplan–Meier analysis to compare the relapse of the patients (Figure 4). The model developed by Cox regression could accurately classify patients with different survival probabilities, and the median PFS of predicting the bad prognosis group was about 9 months, which also fitted our hypothesis.

Discussion

Using "omics" data to develop a prediction model for clinical utilization was always been a challenging task,[40] especially for clinical lipidomic results. A limited amount of existing lipidomics data, combined with the existence of batch effects, and the unique structure of MS data have made it challenging to construct machine-learning models. To address this, we have developed a standard operating procedure which avoids the common pitfalls in machine learning such as batch effect and leaky preprocessing. This has allowed us to utilize metabolomics and lipidomics data from different batches without the help of internal standards. Existing research on the application of lipidomic data to machine learning has not reached a consensus on model selection. Perakakis and colleagues[41] used SVM, K-nearest neighbors (KNN), and RF to diagnose nonalcoholic steatohepatitis (NASH), finding that SVM had the best performance. Villagrana and colleagues[42] have tested the performance of RF, LR, SVM, and NB in the prediction of Sudden Infant Death Syndrome, finding that RF had the
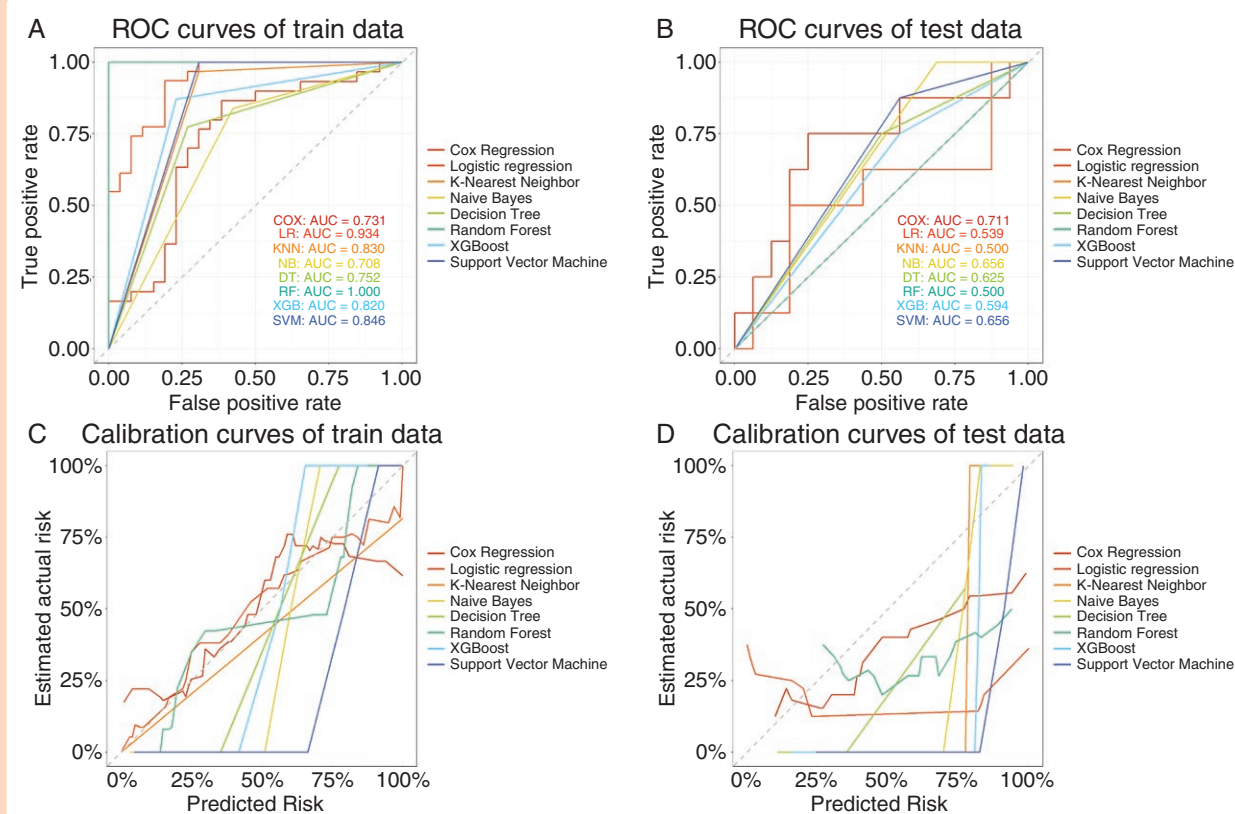
Figure 3. The ROC and calibration curves. (A) ROC curves of train data; (B) ROC curves of test data; (C) Calibration curves of train data; (D) Calibration curves of test data.

best performance. We think it was essential to try more models to find the model with the best performance.

From this study, we have found that Cox regression model could provide a better prediction performance, and 15 features were selected as the prognostic-related factors. In total, 86.7% of identified features of patients with PCNSL are membrane lipids, which not only indicates an altered lipid metabolism, but also implies that the membrane lipids play an important role in MTX-resistance. The plasma membrane exhibits non-uniformities characterized by variations in lipid composition, structure, and signaling activity.[21] Membrane lipid saturation, which is mediated by de novo lipogenesis, has been linked to the sensitivity of cancer cells to free radicals and chemotherapeutics.[43] Sphingolipids and cholesterol, which are involved in the formation of lipid rafts, are also critical for the pathological process of the tumor.[27,44–47]

Phosphoinositide 3-kinase (PI3K) exhibited at least 3 times as many distal translocations in PCNSL, which indicates the hyperactivation of PI3K signaling.[4] PI3K/ AKT/mTOR axis is critical to lipid metabolism[48–50] and its hyperactivation promotes membrane building via a supply of ample de novo lipogenesis molecules,[51] which promotes cell proliferation, growth, metabolism, and motility.[52] Notably, the NF-κB pathway, which is not only associated with resistance to chemotherapy,[53] but also essential to altered lipid metabolism via inflammation,[54] was found to be activated in patients with PCNSL.[4] As PCNSL is a rare

subtype of NHL and is confined to the central nervous system,[55] we suggest that the lipid features group found in CSF could become the biomarkers for prognosis prediction of HD-MTX-based treatment.

Compared to other machine-learning models whose advantages are estimation of the strength of the association between different predictor variables and categorical outcomes, Cox proportional hazards regression approach allows the use of multiple predictor variables when comparing time-to-event outcomes that include censored data.[56] As the prognosis grouping criteria is based on time-to-event outcome, Cox regression has its advantages in feature engineering. It is important to consider the characteristics of the model to select and develop a model for prediction. For example, the KNN model is very sensitive to the distance metric used to find the K-nearest neighbors,[26] which makes it incapable of dealing with high-dimensional data, such as MS data. SVMs failed to accurately scale with large datasets because of their computational complexity.[57] Tree-based models, including decision tree, RF, and XGBoost, could discover important combinations of predictors among a large pool of covariates, which may introduce noise due to limited sample size.[58] Although the NB classifier did not perform well in this study, this model showed unexpected robustness when applied to the testing dataset. Continued efforts are still needed to determine its value in making predictions based on MS data.[59]
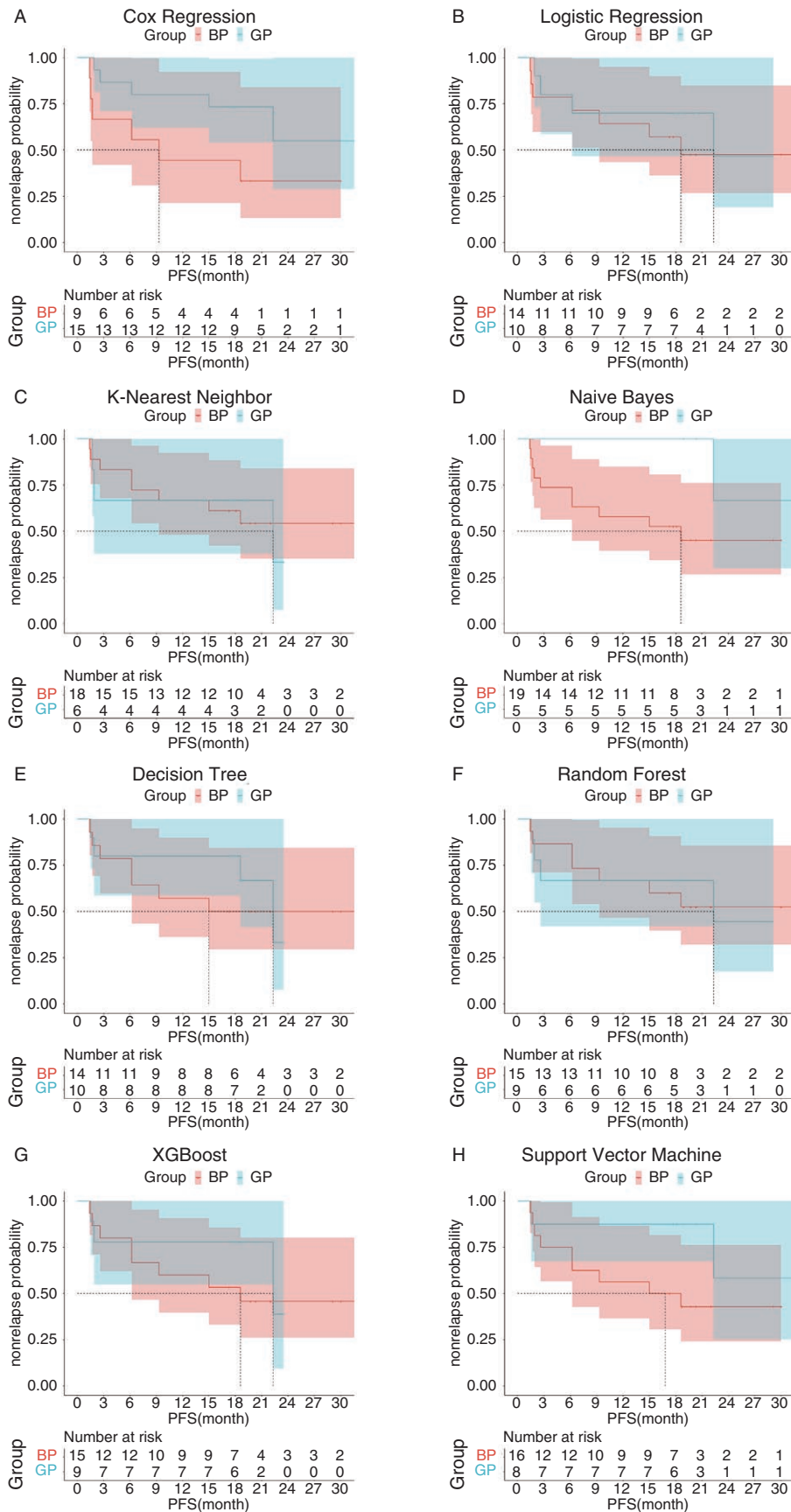
Figure 4. Kaplan–Meier survival analysis of patients in test set based on the prediction of each model. (A) Cox Regression, (B) Logistic Regression, (C) K-Nearest Neighbor, (D) Naive Bayes, (E) Decision tree, (F) Random Forest, (G) XGBoost, and (H) Support Vector Machine.

This study created the first prediction model based on lipidomic data that can be used to predict the prognosis behavior of PCNSL patients treated with HD-MTX-based chemotherapeutic treatment. The results shed light on the value of lipidomic data in predicting treatment effectiveness, allowing patients to weigh the advantages and disadvantages of HD-MTX-based treatment. As it could provide reliable relapse prediction, our model could reduce the suffering of patients faced with the considerable side effects of HD-MTX-based chemotherapy. Additionally, our procedure avoids batch effects and leaky processing, which has considerable value for future models based on mass spectrum data.

Supplementary material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords

HD-MTX-based chemotherapy | lipidomics | machine-learning | primary central nervous system lymphoma

Lay Summary

Primary central nervous system lymphoma is a rare and aggressive brain tumor. The usual treatment involves high doses of a chemotherapy drug called methotrexate. It is challenging to predict how long patients will survive after this treatment. To better understand this, the authors studied the different types of fats in the fluid surrounding the brain and spine of patients with this disease. They created computer models using information about these fats to predict patient survival. Their results showed that they could predict survival with about 70% accuracy.

Conflict of interest statement

None.

Authorship statement

H.H. initiated and supervised the project and Y.Z. wrote the manuscript. L.Z. and J.X. helped with data collection and analysis. L.Z. helped with manuscript revision.

Affiliations

Shanghai Key Laboratory of Metabolic Remodeling and Health, Institute of Metabolism and Integrative Biology, Fudan University, Shanghai, 200438, China (Y.Z., L.Z., J.X., H.H.)

References

1. Schaff LR, Grommes C. Primary central nervous system lymphoma. *Blood*. 2022;140(9):971–979.
2. Hoang-Xuan K, Deckert M, Ferreri AJM, et al. European Association of Neuro-Oncology (EANO) guidelines for treatment of primary central nervous system lymphoma (PCNSL). *Neuro Oncol*. 2023;25(1):37–53.
3. Chen T, Liu Y, Wang Y, et al. Evidence-based expert consensus on the management of primary central nervous system lymphoma in China. *J Hematol Oncol*. 2022;15(1):136.
4. Radke J, Ishaque N, Koll R, et al; ICGC MMML-Seq Consortium. The genomic and transcriptional landscape of primary central nervous system lymphoma. *Nat Commun*. 2022;13(1):2558.
5. Schaff LR, Mellinghoff IK. Glioblastoma and other primary brain malignancies in adults: A review. *JAMA*. 2023;329(7):574–587.
6. Holdhoff M, Mrugala MM, Grommes C, et al. Challenges in the treatment of newly diagnosed and recurrent primary central nervous system lymphoma. *J Natl Compr Canc Netw*. 2020;18(11):1571–1578.
7. Ferreri AJM, Calimeri T, Cwynarski K, et al. Primary central nervous system lymphoma. *Nat Rev Dis Primers*. 2023;9(1):29.
8. Grommes C, Rubenstein JL, DeAngelis LM, Ferreri AJM, Batchelor TT. Comprehensive approach to diagnosis and treatment of newly diagnosed primary CNS lymphoma. *Neuro Oncol*. 2019;21(3):296–305.
9. Ferreri AJ, Blay JY, Reni M, et al. Prognostic scoring system for primary CNS lymphomas: The International Extranodal Lymphoma Study Group experience. *J Clin Oncol*. 2003;21(2):266–272.
10. Bessell EM, Graus F, Lopez-Guillermo A, et al. Primary non-Hodgkin's lymphoma of the CNS treated with CHOD/BVAM or BVAM chemotherapy before radiotherapy: Long-term survival and prognostic factors. *Int J Radiat Oncol Biol Phys*. 2004;59(2):501–508.
11. Abrey LE, Ben-Porat L, Panageas KS, et al. Primary central nervous system lymphoma: The Memorial Sloan-Kettering Cancer Center prognostic model. *J Clin Oncol*. 2006;24(36):5711–5715.
12. Lu G, Zhang Y, Wang W, Miao L, Mou W. Machine learning and deep learning CT-Based Models for predicting the primary central nervous system lymphoma and glioma types: A multicenter retrospective study. *Front Neurol*. 2022;13:905227.
13. Vallée R, Vallée JN, Guillevin C, et al. Machine learning decision tree models for multiclass classification of common malignant brain tumors using perfusion and spectroscopy MRI data. *Front Oncol*. 2023;13:1089998.
14. Jiang Q, Zhan G, Jiang W, et al. Prognostic model and treatment choices for patients with primary intracranial central nervous system lymphoma: A population-based study. *Clin Neurol Neurosurg*. 2023;233:107912.
15. Liu CJ, Lin SY, Yang CF, et al. A new prognostic score for disease progression and mortality in patients with newly diagnosed primary CNS lymphoma. *Cancer Med*. 2020;9(6):2134–2145.
16. Zhou L, Li Q, Xu J, et al. Cerebrospinal fluid metabolic markers predict prognosis behavior of primary central nervous system lymphoma with high-dose methotrexate-based chemotherapeutic treatment. *Neurooncol Adv*. 2023;5(1):vdac181.

17. Butler LM, Perone Y, Dehairs J, et al. Lipids and cancer: Emerging roles in pathogenesis, diagnosis and therapeutic intervention. *Adv Drug Deliv Rev.* 2020;159:245–293.

18. Shimizu T. Lipid mediators in health and disease: Enzymes and receptors as therapeutic targets for the regulation of immunity and inflammation. *Annu Rev Pharmacol Toxicol.* 2009;49:123–150.

19. Wu Z, Bagarolo GI, Thoröe-Boveleth S, Jankowski J. "Lipidomics": Mass spectrometric and chemometric analyses of lipids. *Adv Drug Deliv Rev.* 2020;159:294–307.

20. Mukherjee A, Chiang CY, Daifotis HA, et al. Adipocyte-induced FABP4 expression in ovarian cancer cells promotes metastasis and mediates carboplatin resistance. *Cancer Res.* 2020;80(8):1748–1761.

21. Bi J, Ichu TA, Zanca C, et al. Oncogene amplification in growth factor signaling pathways renders cancers dependent on membrane lipid remodeling. *Cell Metab.* 2019;30(3):525–538.e8.

22. Wang G, Qiu M, Xing X, et al. Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci Transl Med.* 2022;14(630):eabk2756.

23. Naudin S, Sampson JN, Moore SC, et al. Lipidomics and pancreatic cancer risk in two prospective studies. *Eur J Epidemiol.* 2023;38(7):783–793.

24. Deckert M, Montesinos-Rongen M, Brunn A, Siebert R. Systems biology of primary CNS lymphoma: From genetic aberrations to modeling in mice. *Acta Neuropathol.* 2014;127(2):175–188.

25. Takashima Y, Hayano A, Yamanaka R. Metabolome analysis reveals excessive glycolysis via PI3K/AKT/mTOR and RAS/MAPK signaling in methotrexate-resistant primary CNS lymphoma-derived cells. *Clin Cancer Res.* 2020;26(11):2754–2766.

26. Wang X, Su W, Jiang Y, et al. Regulation of nucleotide metabolism with nutrient-sensing nanodrugs for cancer therapy. *Adv Sci (Weinh).* 2022;9(20):e2200482.

27. Kim JH, An YJ, Kim TM, et al. Ex vivo NMR metabolomics approach using cerebrospinal fluid for the diagnosis of primary CNS lymphoma: Correlation with MR imaging characteristics. *Cancer Med.* 2023;12(4):4679–4689.

28. Breitkopf SB, Ricoult SJH, Yuan M, et al. A relative quantitative positive/negative ion switching method for untargeted lipidomics via high resolution LC-MS/MS from any biological source. *Metabolomics.* 2017;13(3):30.

29. Song Z, Wang S, Lu L, et al. Lipidomics and transcriptomics differ liposarcoma differentiation characteristics that can be altered by Pentose Phosphate Pathway intervention. *Metabolites.* 2022;12(12):1227.

30. Reinhold D, Pielke-Lombardo H, Jacobson S, Ghosh D, Kechris K. Pre-analytic considerations for mass spectrometry-based untargeted metabolomics data. *Methods Mol Biol.* 2019;1978:323–340.

31. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28(6):882–883.

32. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics.* 2006;7:142.

33. Tokareva AO, Chagovets VV, Kononikhin AS, et al. Normalization methods for reducing interbatch effect without quality control samples in liquid chromatography-mass spectrometry-based studies. *Anal Bioanal Chem.* 2021;413(13):3479–3486.

34. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2002;2(3):18–22.

35. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008;28(5):1–26.

36. Dudley WN, Wickham R, Coombs N. An introduction to survival statistics: Kaplan-Meier analysis. *J Adv Pract Oncol* 2016;7(1):91–100.

37. Alaggio R, Amador C, Anagnostopoulos I, et al. The 5th edition of the World Health Organization classification of haematolymphoid tumours: Lymphoid neoplasms. *Leukemia.* 2022;36(7):1720–1748.

38. Lv J, Ren H, Guo X, et al. Nomogram predicting bullying victimization in adolescents. *J Affect Disord.* 2022;303:264–272.

39. Hoesseini A, van Leeuwen N, Sewnaik A, et al. Key aspects of prognostic model development and interpretation from a clinical perspective. *JAMA Otolaryngol Head Neck Surg.* 2022;148(2):180–186.

40. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet.* 2022;23(3):169–181.

41. Perakakis N, Polyzos SA, Yazdani A, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: A proof of concept study. *Metabolism.* 2019;101:154005.

42. Villagrana-Bañuelos KE, Galván-Tejada CE, Galván-Tejada JI, et al. Machine learning model based on lipidomic profile information to predict sudden infant death syndrome. *Healthcare (Basel).* 2022;10(7):1303.

43. Rysman E, Brusselmans K, Scheys K, et al. De novo lipogenesis protects cancer cells from free radicals and chemotherapeutics by promoting membrane lipid saturation. *Cancer Res.* 2010;70(20):8117–8126.

44. Huang B, Song BL, Xu C. Cholesterol metabolism in cancer: Mechanisms and therapeutic opportunities. *Nat Metab.* 2020;2(2):132–141.

45. Luo J, Yang H, Song BL. Mechanisms and regulation of cholesterol homeostasis. *Nat Rev Mol Cell Biol.* 2020;21(4):225–245.

46. Meng Q, Hu X, Zhao X, et al. A circular network of coregulated sphingolipids dictates lung cancer growth and progression. *EBioMedicine.* 2021;66:103301.

47. Muthusamy T, Cordes T, Handzlik MK, et al. Serine restriction alters sphingolipid diversity to constrain tumour growth. *Nature.* 2020;586(7831):790–795.

48. Koundouros N, Karali E, Tripp A, et al. Metabolic fingerprinting links oncogenic PIK3CA with enhanced arachidonic acid-derived eicosanoids. *Cell.* 2020;181(7):1596–1611.e27.

49. Petersen MC, Shulman GI. Mechanisms of insulin action and insulin resistance. *Physiol Rev.* 2018;98(4):2133–2223.

50. Saxton RA, Sabatini DM. mTOR signaling in growth, metabolism, and disease. *Cell.* 2017;169(2):361–371.

51. Dai W, Xiang W, Han L, et al. PTPRO represses colorectal cancer tumorigenesis and progression by reprogramming fatty acid metabolism. *Cancer Commun (Lond).* 2022;42(9):848–867.

52. Alzahrani AS. PI3K/Akt/mTOR inhibitors in cancer: At the bench and bedside. *Semin Cancer Biol.* 2019;59:125–132.

53. Woyach JA, Johnson AJ. Targeted therapies in CLL: Mechanisms of resistance and strategies for management. *Blood.* 2015;126(4):471–477.

54. Huang X, Yao Y, Hou X, et al. Macrophage SCAP contributes to metaflammation and lean NAFLD by activating STING-NF-κB Signaling Pathway. *Cell Mol Gastroenterol Hepatol.* 2022;14(1):1–26.

55. Sun J, Li Q, Ding Y, et al. Profiling phosphoproteome landscape in circulating extracellular vesicles from microliters of biofluids through functionally tunable paramagnetic separation. *Angew Chem Int Ed Engl.* 2023;62(29):e202305668.

56. Msaouel P, Jimenez-Fonseca P, Lim B, Carmona-Bayonas A, Agnelli G. Medicine before and after David Cox. *Eur J Intern Med.* 2022;98:1–3.

57. Galal A, Talal M, Moustafa A. Applications of machine learning in metabolomics: Disease modeling and classification. *Front Genet.* 2022;13:1017340.

58. Gueorguieva R, Wu R, Fucito LM, O'Malley SS. Predictors of abstinence from heavy drinking during follow-up in COMBINE. *J Stud Alcohol Drugs.* 2015;76(6):935–941.

59. Md Ghazi MGB, Chuen Lee L, Samsudin AS, Sino H. Comparison of decision tree and naïve Bayes algorithms in detecting trace residue of gasoline based on gas chromatography-mass spectrometry data. *Forensic Sci Res.* 2023;8(3):249–255.