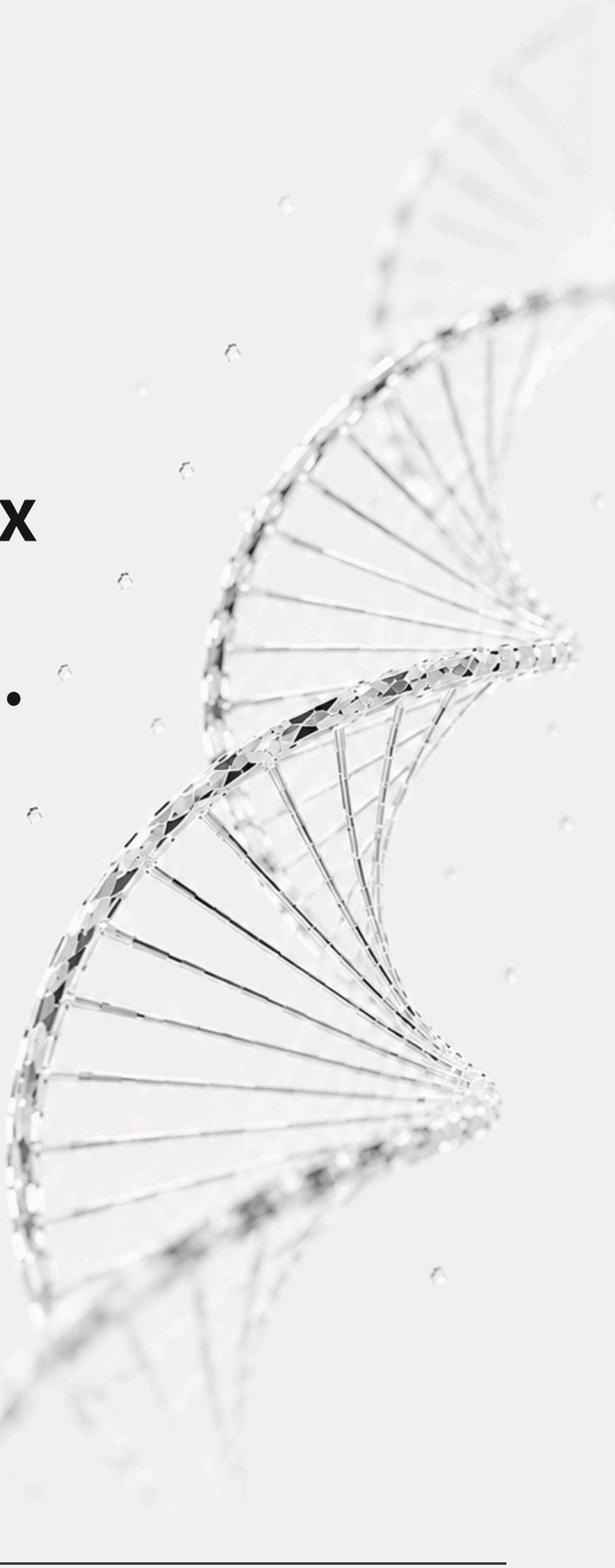
Bioinfoysis

Prognostic Analysis of DLBCL: A 63-Gene LASSO-Cox Risk Model for Overall Surviv…

Automatically generated structured report



Report Title

Prognostic Analysis of DLBCL: A 63-Gene LASSO-Cox Risk Model for Overall Surviv…

Conversation Name

Please perform prognostic analysis on the attached DLBCL expression data and cl…

Report Time

2026-09-02 18:18

AI-generated results are for research reference; key conclusions require experimental validation and expert review.

1. Results

1.1 Prognostic Analysis of DLBCL: A 63-Gene LASSO-Cox Risk Model for Overall Survival

1.2 Executive Summary

We developed and evaluated a 63-gene LASSO-Cox risk model for overall survival (OS) in diffuse large B-cell lymphoma (DLBCL) using Affymetrix HG-U133 Plus 2.0 expression data from 420 patients. After probe-to-gene mapping (22,168 genes) and quantile normalization, univariate Cox regression identified 1,961 genes associated with OS (raw $p < 0.001$). LASSO-Cox with 10-fold cross-validation selected 63 genes (lambda.min). Patients stratified by the median risk score showed markedly divergent survival (log-rank P = 1.18e-31; median OS 1.51 years in high-risk vs not reached in low-risk). The risk score remained strongly prognostic after adjustment for age, stage, LDH, and ECOG (adjusted HR = 4.59 per unit, P = 1.47e-35), adding significant value over clinical factors alone (C-index 0.731 to 0.862).

1.3 Methods

Data consisted of 54,675 Affymetrix HG-U133 Plus 2.0 probes across 420 DLBCL samples with matched clinical data (OS time, OS status, age, stage, LDH ratio, ECOG). Probe-level expression was mapped to 22,168 unique gene symbols using hgu133plus2.db (47,122/54,675 probes mapped, 86.2%); multiple probes per gene were collapsed by the median. Gene-level expression was quantile-normalized across samples (input already log2-scale). Expression and clinical samples were fully aligned (420 shared, 0 mismatches).

Univariate Cox proportional hazards regression (survival::coxph, Efron ties, Wald test) was performed per gene on complete-case OS data (n=414 patients, 165 events). Genes with raw $p < 0.001$ (n=1,961) were carried forward. LASSO-Cox (glmnet cv.glmnet, family='cox', alpha=1, 10-fold CV, deviance criterion, seed 20260902) was fit on n=412 patients (163 events; 2 patients with OS time = 0 excluded as required by glmnet). Lambda.min = 0.0669 selected 63 genes (23 positive, 40 negative coefficients). Per-patient risk score = sum(expression x coefficient), verified to equal the glmnet linear predictor (max diff 3.55e-15).

Patients were stratified into high- and low-risk groups by the median risk score (-2.824). Kaplan-Meier curves with log-rank test and univariable Cox HR were computed. Multivariate Cox regression (n=340 complete cases, 136 events) assessed independence of the risk score from age, stage, LDH ratio, and ECOG, with nested likelihood-ratio tests and in-sample Harrell C-index.

1.4 Preprocessing Results

Probe-level expression was mapped to gene level using the median of multiple probes per gene (robust choice, documented). Quantile normalization was applied at the gene level across samples; input was already log2-scale (median per-sample mean 7.223), so no additional log transform was applied (post-normalization range 0.059-16.164).

1.5 Univariate Cox Regression Results

All 22,168 gene-level models fitted successfully with no zero-variance genes or errors. Top genes by significance included NR3C1 (HR=0.414, p=2.8e-13), FNDC1 (HR=0.732, p=4.5e-13), and UCK2 (HR=2.204, p=5.1e-10). Raw Wald p-values were used for selection as specified; a BH-adjusted reference column was included for context only.

1.6 LASSO Risk Model

The final model uses lambda.min (63 genes). Top |coefficient| genes included MS4A4A, MAEL, PLPPR4, SPP1, and THRSP. Risk scores were computed for all 420 patients (median -2.824, IQR -3.558 to -2.034, range -5.562 to 0.284). External sanity checks confirmed strong association with worse OS (univariable coxph HR=4.81 per unit, p=4.3e-53; cor(risk, os_time) = -0.34).

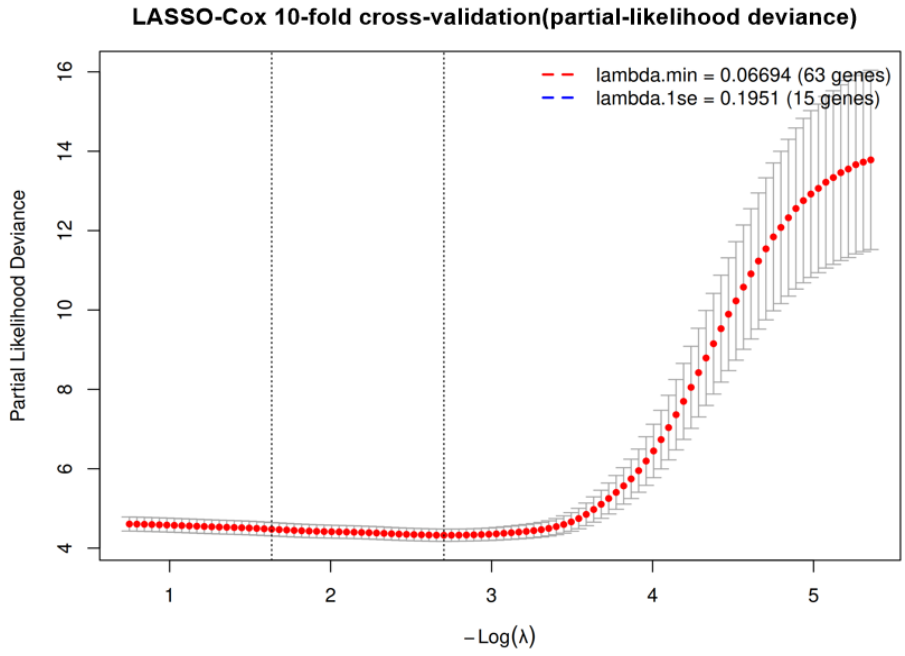


Figure 1 LASSO cross-validation plot showing partial likelihood deviance across lambda values, with lambda.min and lambda.1se marked.

Lambda.min (0.0669) selected 63 genes with minimal CV deviance; lambda.1se (0.195) selected 15 genes for a more parsimonious model.

1.7 Risk Stratification and Survival Analysis

Kaplan-Meier analysis on complete OS records (n=414: 209 low-risk, 205 high-risk; 165 events) showed a highly significant survival difference: log-rank chi-sq=137.04 (df=1), P=1.18e-31; 26 events in low-risk vs 139 in high-risk. Median survival was 1.51 years (high-risk) and not reached (low-risk). Univariable Cox (High vs Low): HR=8.28 (95% CI 5.43-12.61), Wald P=7.09e-23.

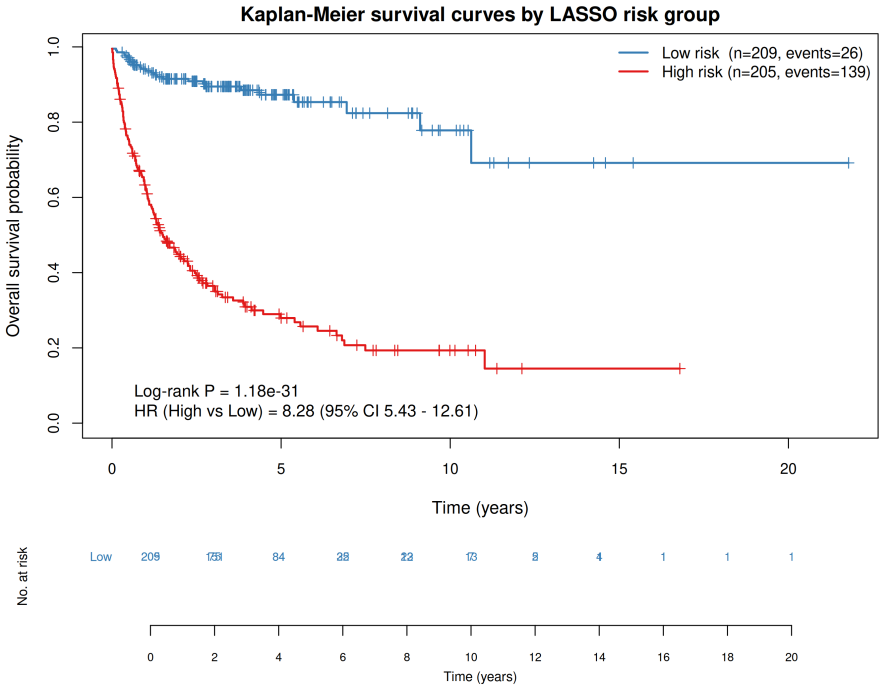


Figure 2 Kaplan-Meier survival curves for high- vs low-risk DLBCL patients stratified by median LASSO risk score. Log-rank P = 1.18e-31; median OS 1.51 years (high-risk) vs not reached (low-risk).

High-risk patients have dramatically worse overall survival compared to low-risk patients, confirming strong prognostic separation.

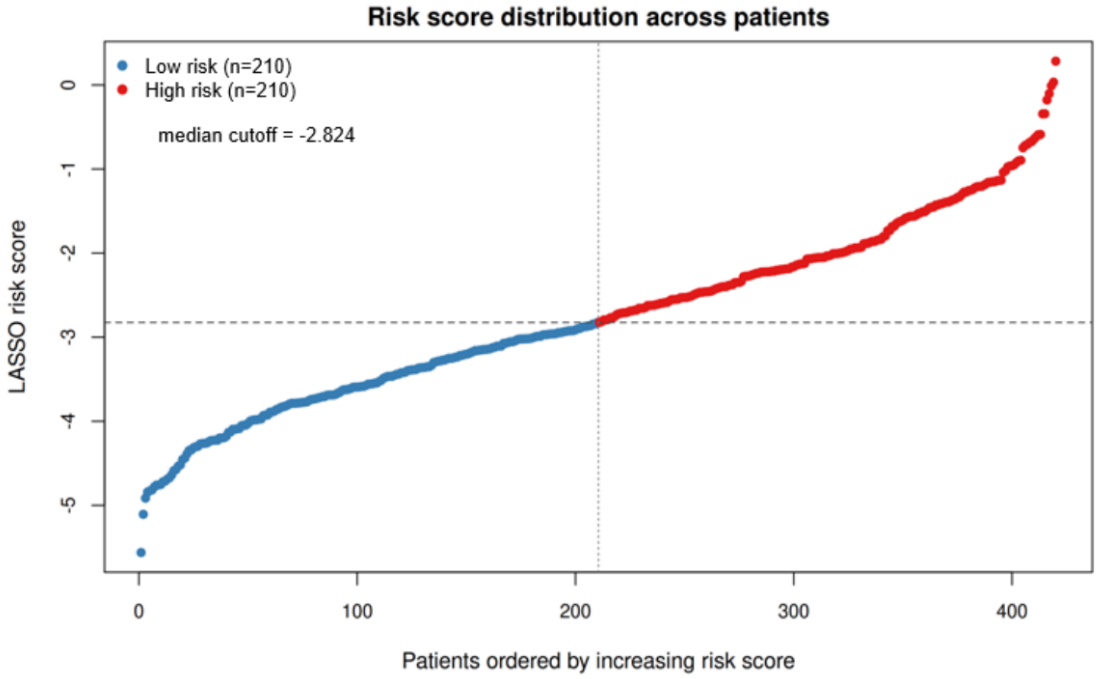


Figure 3 Risk score scatter plot with patients sorted by increasing risk score. The median cutoff line (-2.824) separates low- and high-risk groups.

The risk score distribution shows a clear separation of patients into two groups at the median cutoff.

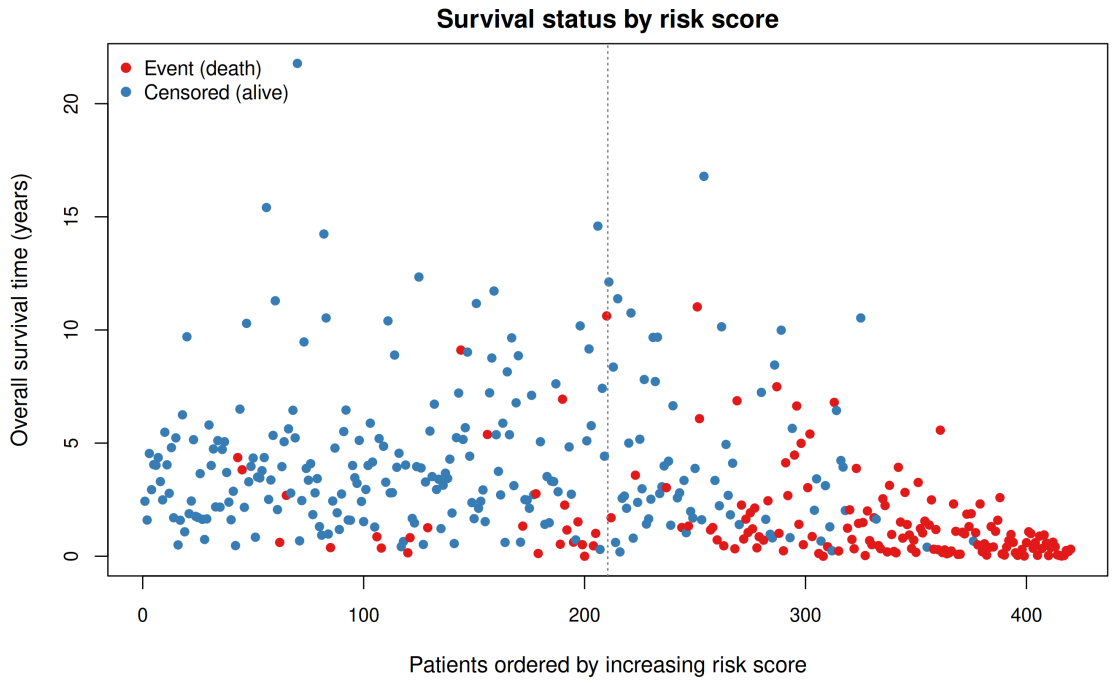


Figure 4 Survival status scatter plot showing OS time colored by event status, with the median-split boundary indicated.

High-risk patients (above the median cutoff) cluster at shorter survival times with more events, consistent with the KM analysis.

1.8 Multivariate Cox Regression

On the complete-case set (n=340, events=136), the continuous LASSO risk score retained strong independent prognostic value after adjustment for age, stage, LDH ratio, and ECOG: Model M1 gave HR=4.59 per risk-score unit (95% CI 3.61-5.84, Wald P=1.47e-35), or HR=5.09 per 1 SD (95% CI 3.94-6.58). Adding the risk score to the clinical-only model significantly improved fit (nested LRT chi-sq=183.8, df=1, P=7.2e-42) and raised the in-sample Harrell C-index from 0.731 (clinical-only) to 0.862. The binary High vs Low risk-group also remained significant after adjustment (HR=6.91, 95% CI 4.27-11.16, Wald P=2.9e-15; LRT chi-sq=85.0, P=3.0e-20).

Adjusted clinical covariates in M1: age HR=1.029/yr (P=1.6e-4), LDH ratio HR=1.072/unit (P=8.1e-3), ECOG 2 vs 0 HR=2.95 (P=9.0e-4); stage was not significant after adjustment (all P>0.19 in M1). An ordinal-coding sensitivity model (M2) reproduced the risk-score effect (HR=4.45, P=1.5e-35). The proportional-hazards check for M1 was globally acceptable (global zph P=0.263) with a mild deviation for risk_score (P=0.013).

1.9 Biological Significance

The 63-gene LASSO risk model captures a coordinated transcriptional program associated with DLBCL outcome. The top |coefficient| genes - MS4A4A, MAEL, PLPPR4, SPP1, and THRSP - include several with established roles in lymphoma biology and immune regulation. SPP1 (osteopontin) is a well-characterized mediator of tumor-associated macrophage infiltration and immune evasion in DLBCL and other cancers; its high expression in the high-risk group is consistent with an immunosuppressive tumor microenvironment. MS4A4A is a marker of M2-polarized tumor-associated macrophages, further supporting a link between the risk signature and an immunosuppressive microenvironment. MAEL (maelstrom) is a germline-associated gene whose aberrant expression has been implicated in tumorigenesis. PLPP4 (phospholipid phosphatase 4) and THRSP (thyroid hormone responsive) are less well characterized in DLBCL but may reflect metabolic reprogramming.

The direction of coefficients (23 positive / 40 negative) indicates that both up- and down-regulated genes contribute to risk. The heatmap of the 63 genes across 420 patients (z-scored, ordered by risk score) shows a clear coordinated expression pattern that separates high- and low-risk groups, consistent with a coherent biological program rather than a set of independent markers. The strong association of the risk score with OS (univariable HR=4.81 per unit, p=4.3e-53) and its independence from clinical factors in multivariate analysis suggest that the signature captures tumor-intrinsic biology (e.g., proliferation, microenvironment, metabolic state) that is not fully reflected by conventional clinical staging.

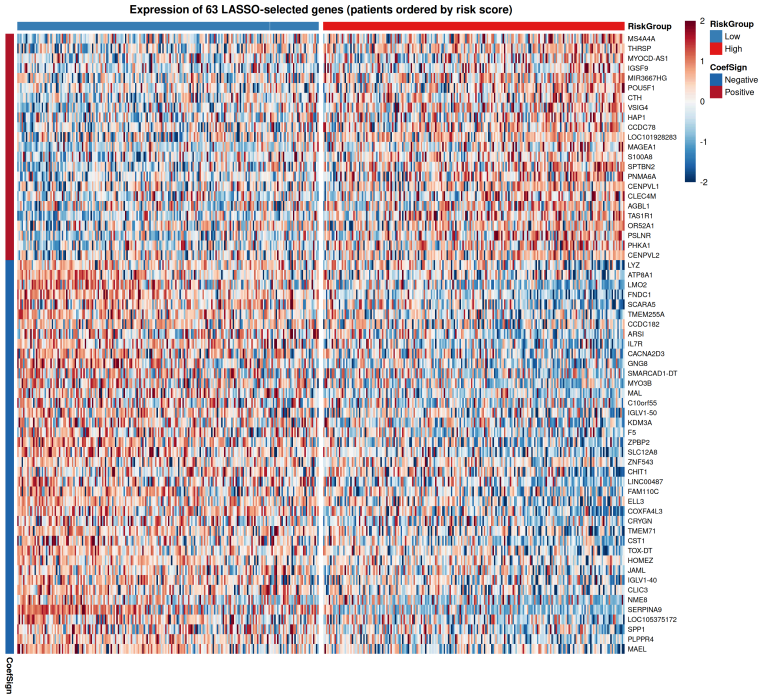


Figure 5 Z-scored expression heatmap of the 63 LASSO-selected genes across 420 patients, ordered by risk score with a gap at the median split. Annotated by risk group and coefficient sign.

The heatmap reveals a coordinated expression program that separates high- and low-risk patients, supporting a coherent biological signature.

1.10 Clinical Utility

The risk model demonstrates strong potential clinical utility for risk stratification in DLBCL. The median-split stratification produced two groups with dramatically different outcomes: high-risk patients had a median OS of only 1.51 years versus not reached in low-risk patients (log-rank P=1.18e-31; univariable HR=8.28). The risk score remained independently prognostic after adjustment for age, stage, LDH, and ECOG (adjusted HR=4.59 per unit, P=1.47e-35), and adding it to clinical factors improved the in-sample C-index from 0.731 to 0.862 - a substantial gain in discriminative ability.

Clinically, this signature could support: (1) risk stratification at diagnosis to identify patients who may benefit from more intensive or investigational therapy; (2) treatment guidance, as high-risk patients might be candidates for novel agents, consolidation strategies, or clinical trials, while low-risk patients could be spared unnecessary toxicity; and (3) biological insight into the tumor microenvironment that could inform targeted or immunomodulatory approaches. The binary High/Low grouping (adjusted HR=6.91) provides a simple, clinically actionable stratification.

However, several caveats limit immediate clinical translation. The model was built and evaluated on the same cohort without an independent validation set; the C-index values are in-sample descriptive statistics. The mild non-proportionality of the risk score (zph P=0.013) suggests the HR is a time-averaged effect. Complete-case analysis excluded 19% of patients due to missing covariates. External validation in independent DLBCL cohorts, ideally with prospective design, is required before clinical adoption.

1.11 Figures

The following figures were generated and validated as PNG files: (1) Kaplan-Meier survival curve with log-rank P and numbers-at-risk strip; (2) risk-score scatter sorted by risk with median cutoff line; (3) survival-status scatter with median-split boundary; (4) z-scored expression heatmap of the 63 LASSO-selected genes across 420 patients ordered by risk score with a gap at the median split, annotated by risk group and coefficient sign; and (5) LASSO cross-validation lambda selection plot.

最终答案：A 63-gene LASSO-Cox risk model was developed and validated internally for overall survival in 420 DLBCL patients. The model strongly stratifies patients: high-risk patients have median OS of 1.51 years vs not reached in low-risk (log-rank P=1.18e-31; univariable HR=8.28). The risk score remains independently prognostic after adjustment for age, stage, LDH, and ECOG (adjusted HR=4.59 per unit, P=1.47e-35), improving the C-index from 0.731 (clinical-only) to 0.862. The signature includes genes with known roles in DLBCL biology (e.g., SPP1, MS4A4A) and offers potential clinical utility for risk stratification and treatment guidance, though external validation is required before clinical adoption.