HEMATOL TRANSFUS CELL THER. 2025;xxx(xx):103847

HEMATOLOGY, TRANSFUSION AND CELL THERAPY

www.htct.com.br

Letter to the Editor

# Multi-cohort gene expression model enhances prognostic stratification in diffuse large B-cell lymphoma

*Dear Editor,*

Diffuse large B-cell lymphoma (DLBCL), the most common type of lymphoma, in most cases is marked by significant heterogeneity and aggressive clinical behavior. While standard chemotherapy often achieves initial responses, these are short-lived, and resistance and relapse are frequent challenges.[1] Traditionally, risk stratification has relied on clinical tools, including the International Prognostic Index (IPI) and its variation.[2] However, molecular stratification is promising to predict outcomes with greater accuracy, though gene-based approaches are still preliminary.[3] Progress in this field is hindered by limited sample sizes and the substantial intra- and inter-regional variability of DLBCL.[4,5] Consequently, large-scale studies are essential to refine risk stratification and optimize patient outcomes.

This study aimed to establish a prognostic gene expression signature for patients with DLBCL based on tumor transcriptome patterns. To achieve this, we analyzed transcriptome and survival data from 11 diverse cohorts worldwide. Given the variability in RNA sequencing or microarray platforms across the 11 datasets, we focused on the genes common to all datasets, resulting in a panel of 11,425 genes. Detailed information regarding the datasets can be found in Supplementary Table 1. Due to platform-specific differences in scale, the gene expression values were transformed into z-scores. Datasets with fewer than 100 patients were combined into a cohort referred to as the Merged Cohort. In total, six cohorts were used in this study: the National Cancer Institute Cohort (GSE10846), University of York Cohort (GSE181063), University of York II Cohort (GSE32918), Universitätsmedizin Berlin Cohort (GSE4475), University of Leeds Cohort (GSE69053), and the Merged Cohort (GSE69053, E_TABM_346, GSE11318, GSE21846, GSE23501, GSE57611, and TCGA-DLBC).

For each cohort, a univariate Cox regression was performed employing all genes in the panel, identifying those with a p-value <0.05 as prognostic. Genes were defined as core prognostic genes (CPGs) if they consistently predicted either favorable prognosis in at least 5 out of 6 cohorts or unfavorable prognosis in at least 5 out of 6 cohorts, with no conflicting outcomes.

This process led to the identification of 50 CPGs. To mitigate the risk of overfitting, a penalized Cox regression was applied using the Least Absolute Shrinkage and Selection Operator (Lasso-Cox), thereby allowing for the selection of only the most significant CPGs. The University of York cohort had the largest number of patients and was therefore used to train the Lasso-Cox model, while the other cohorts were used for validation. The final risk score was developed based on the expression levels of 22 CPGs selected through the Lasso-Cox regression (Figure 1A). The formula for calculating the risk score is as follows:

$$\text{Risk Score} = (\beta 1 \times \text{Gene1}) + (\beta 2 \times \text{Gene2}) + \ldots + (\beta 22 \times \textit{Gene22})$$

where '$\beta$X' represents the coefficients derived from the Lasso-Cox regression, and 'GeneX' refers to the z-score of the expression of each gene for a given sample. The list of selected genes and their corresponding coefficients can be found in Supplementary Table 2.

Patients were then divided into High Risk (> median) and Low Risk (≤ median) Groups based on the risk score. Survival analysis using Kaplan-Meier curves was conducted, revealing that the developed risk groups were significant predictors of overall survival in all cohorts (Figure 1B). Additionally, the risk score demonstrated high predictive accuracy, achieving great (≥0.69) areas under the receiver operating characteristic curve (AUC) across all cohorts (Figure 1C). By pooling the hazard ratios (HR) from the cohorts using a random effects model, the HR for death of being in the High Risk Group was 2.73 (range: 2.43–3.05; Figure 1D), further validating the risk score as a strong predictor of survival.

To ensure the prognostic value of the risk groups, even when assessed alongside clinical data, we conducted multivariable Cox regressions for each cohort. The results demonstrated that the risk groups remained strong predictors of survival. Figure 2A presents the clinical characteristics of the cohorts analyzed in this study, along with the results of the multivariate Cox regression analysis.

To integrate the established risk groups with other clinical variables, we developed a nomogram (Figure 2B) using a meta-cohort of patients who provided complete information on sex,

Please cite this article as: et al., Multi-cohort gene expression model enhances prognostic stratification in diffuse large B-cell lymphoma, Hematology, Transfusion and Cell Therapy (2025), https://doi.org/10.1016/j.htct.2025.103847

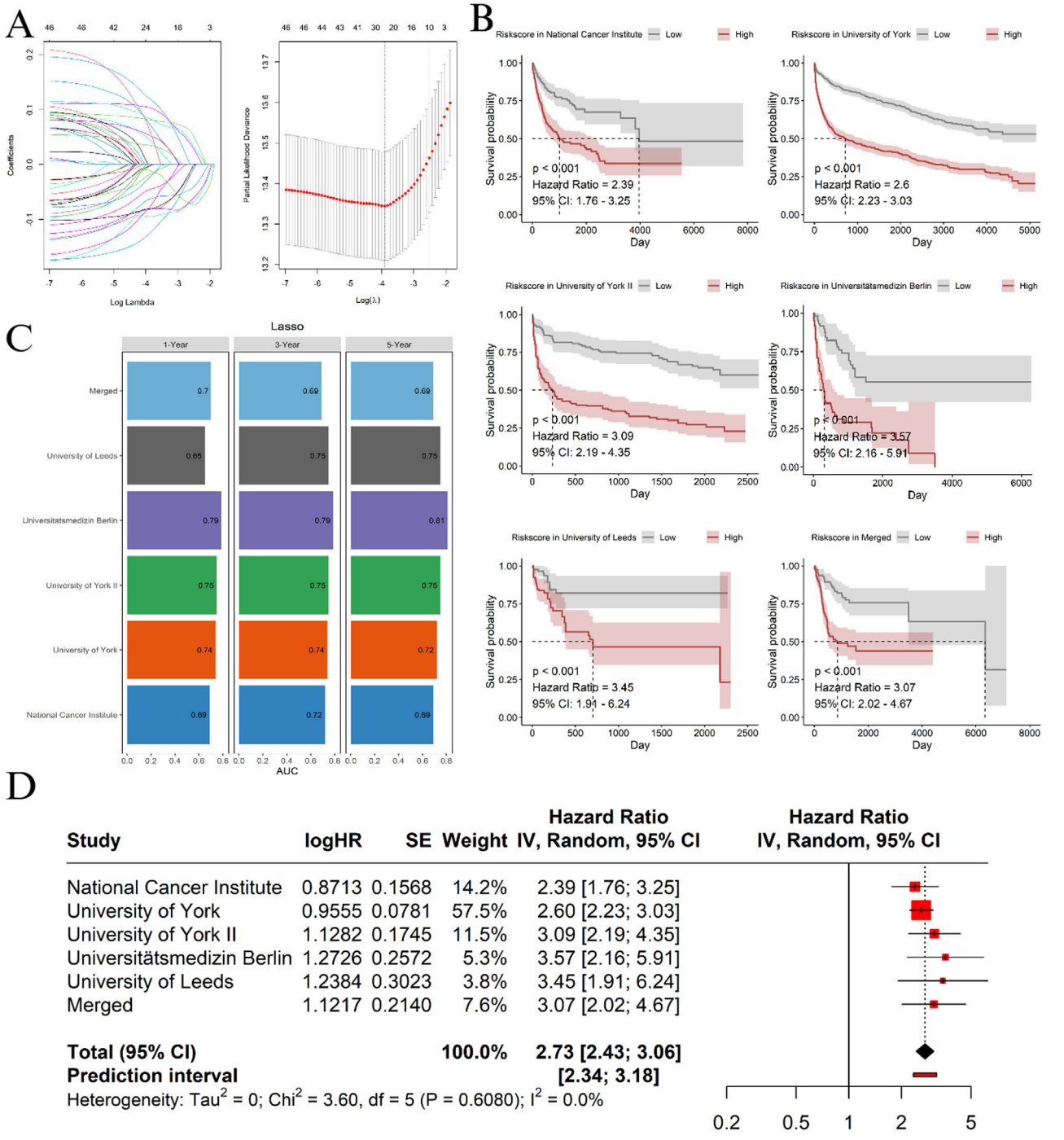
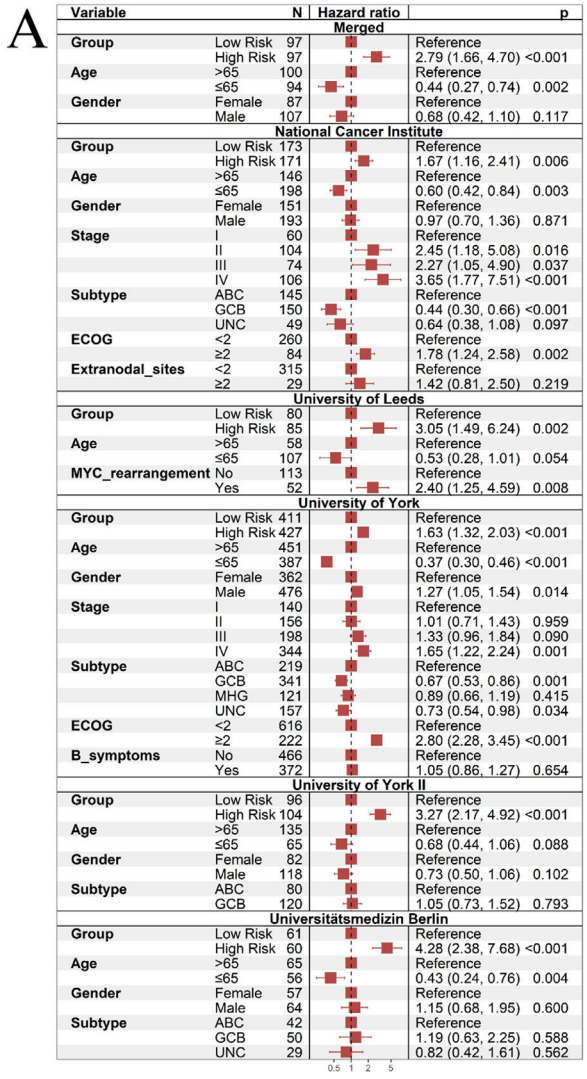



**Figure 1 – A: least absolute shrinkage and selection operator penalized cox regression feature selection. B: Kaplan-Meier analysis of the different cohorts used in this study comparing Low Risk to High Risk Groups. C: Area under the receiver operating characteristic curve (AUC) for the model in different cohort and time-point evaluations. D: Pooled analysis of the hazard ratio of being in the High Risk Group.**

Lasso: least absolute shrinkage and selection operator.

age (over 65 years or 65 years and younger), and DLBCL subtype (germinal center B-cell-like, activated B-cell-like, molecular high-grade B-cell lymphoma, and unclassified), comprising a total of 2102 patients. The nomogram showed an excellent AUC for survival prediction at 1, 3, and 5 years (Figure 2C) and generated survival predictions that closely matched observed outcomes as determined by the calibration plot (Figure 2D). Moreover, the nomogram attained the highest c-index for survival prediction when compared to risk groups and clinical variables alone (Figure 2E). A free online platform has been developed and made accessible at https://costafilhoetal.shinyapps.io/CoreProgDLBCL/ to enhance the applicability of the nomogram.

Please cite this article as: et al., Multi-cohort gene expression model enhances prognostic stratification in diffuse large B-cell lymphoma, Hematology, Transfusion and Cell Therapy (2025), https://doi.org/10.1016/j.htct.2025.103847



| Variable | | N | Hazard ratio | p |
|---|---|---|---|---|
| **Merged** | | | | |
| **Group** | Low Risk | 97 | Reference | |
| | High Risk | 97 | 2.79 (1.66, 4.70) | <0.001 |
| **Age** | >65 | 100 | Reference | |
| | ≤65 | 94 | 0.44 (0.27, 0.74) | 0.002 |
| **Gender** | Female | 87 | Reference | |
| | Male | 107 | 0.68 (0.42, 1.10) | 0.117 |
| **National Cancer Institute** | | | | |
| **Group** | Low Risk | 173 | Reference | |
| | High Risk | 171 | 1.67 (1.16, 2.41) | 0.006 |
| **Age** | >65 | 146 | Reference | |
| | ≤65 | 198 | 0.60 (0.42, 0.84) | 0.003 |
| **Gender** | Female | 151 | Reference | |
| | Male | 193 | 0.97 (0.70, 1.36) | 0.871 |
| **Stage** | I | 60 | Reference | |
| | II | 104 | 2.45 (1.18, 5.08) | 0.016 |
| | III | 74 | 2.27 (1.05, 4.90) | 0.037 |
| | IV | 106 | 3.65 (1.77, 7.51) | <0.001 |
| **Subtype** | ABC | 145 | Reference | |
| | GCB | 150 | 0.44 (0.30, 0.66) | <0.001 |
| | UNC | 49 | 0.64 (0.38, 1.08) | 0.097 |
| **ECOG** | <2 | 260 | Reference | |
| | ≥2 | 84 | 1.78 (1.24, 2.58) | 0.002 |
| **Extranodal_sites** | <2 | 315 | Reference | |
| | ≥2 | 29 | 1.42 (0.81, 2.50) | 0.219 |
| **University of Leeds** | | | | |
| **Group** | Low Risk | 80 | Reference | |
| | High Risk | 85 | 3.05 (1.49, 6.24) | 0.002 |
| **Age** | >65 | 58 | Reference | |
| | ≤65 | 107 | 0.53 (0.28, 1.01) | 0.054 |
| **MYC_rearrangement** | No | 113 | Reference | |
| | Yes | 52 | 2.40 (1.25, 4.59) | 0.008 |
| **University of York** | | | | |
| **Group** | Low Risk | 411 | Reference | |
| | High Risk | 427 | 1.63 (1.32, 2.03) | <0.001 |
| **Age** | >65 | 451 | Reference | |
| | ≤65 | 387 | 0.37 (0.30, 0.46) | <0.001 |
| **Gender** | Female | 362 | Reference | |
| | Male | 476 | 1.27 (1.05, 1.54) | 0.014 |
| **Stage** | I | 140 | Reference | |
| | II | 156 | 1.01 (0.71, 1.43) | 0.959 |
| | III | 198 | 1.33 (0.96, 1.84) | 0.090 |
| | IV | 344 | 1.65 (1.22, 2.24) | 0.001 |
| **Subtype** | ABC | 219 | Reference | |
| | GCB | 341 | 0.67 (0.53, 0.86) | 0.001 |
| | MHG | 121 | 0.89 (0.66, 1.19) | 0.415 |
| | UNC | 157 | 0.73 (0.54, 0.98) | 0.034 |
| **ECOG** | <2 | 616 | Reference | |
| | ≥2 | 222 | 2.80 (2.28, 3.45) | <0.001 |
| **B_symptoms** | No | 466 | Reference | |
| | Yes | 372 | 1.05 (0.86, 1.27) | 0.654 |
| **University of York II** | | | | |
| **Group** | Low Risk | 96 | Reference | |
| | High Risk | 104 | 3.27 (2.17, 4.92) | <0.001 |
| **Age** | >65 | 135 | Reference | |
| | ≤65 | 65 | 0.68 (0.44, 1.06) | 0.088 |
| **Gender** | Female | 82 | Reference | |
| | Male | 118 | 0.73 (0.50, 1.06) | 0.102 |
| **Subtype** | ABC | 80 | Reference | |
| | GCB | 120 | 1.05 (0.73, 1.52) | 0.793 |
| **Universitätsmedizin Berlin** | | | | |
| **Group** | Low Risk | 61 | Reference | |
| | High Risk | 60 | 4.28 (2.38, 7.68) | <0.001 |
| **Age** | >65 | 65 | Reference | |
| | ≤65 | 56 | 0.43 (0.24, 0.76) | 0.004 |
| **Gender** | Female | 57 | Reference | |
| | Male | 64 | 1.15 (0.68, 1.95) | 0.600 |
| **Subtype** | ABC | 42 | Reference | |
| | GCB | 50 | 1.19 (0.63, 2.25) | 0.588 |
| | UNC | 29 | 0.82 (0.42, 1.61) | 0.562 |



**Figure 2 – A: Multivariate cox-regression analysis of the risk groups and other available clinical information. B: Nomogram integrating our risk groups with clinical information. E: Comparison of the concordance index of our model and other variables. F: Gene set enrichment analysis plot comparing high-risk and low-risk groups in the University of York cohort.**

GCB: Germinal center B-cell-like; ABC: Activated B-cell-like; MHG: Molecular high-grade B-Cell lymphoma; UNC: Unclassified.

Please cite this article as: et al., Multi-cohort gene expression model enhances prognostic stratification in diffuse large B-cell lymphoma, Hematology, Transfusion and Cell Therapy (2025), https://doi.org/10.1016/j.htct.2025.103847

We performed a Gene Set Enrichment Analysis (GSEA) using raw data from the University of York cohort and the Hallmark of Cancer gene sets from the Molecular Signatures Database (MSigDB) to better understand the biological processes distinguishing the risk groups. Notably, the GSEA (Supplementary Table 3) results revealed that the High Risk Group was predominantly enriched for E2F targets, MYC targets, and G2M checkpoint pathways, while showing downregulation of inflammatory response, interferon-gamma response, and epithelial-mesenchymal transition pathways (Figure 2F).

This study introduces a promising approach to prognostic stratification in DLBCL, utilizing gene expression data to identify CPGs and develop a validated risk score. While the IPI and its variations remain widely used for stratification in DLBCL, their discriminative power is often limited, with various studies reporting suboptimal overall survival prediction when used alone.[6,7] Nonetheless, the European Society for Medical Oncology (ESMO) currently endorses age-adjusted IPI, which has a reported c-index of 0.613, for stratifying under 60-year-old patients who may benefit from involved-field radiotherapy or autologous stem-cell transplantation.[7]

Furthermore, neither the ESMO nor the National Comprehensive Cancer Network guidelines have incorporated transcriptomic and exome stratification in patient management.[8,9] By outperforming traditional approaches focused on histopathology, our model was able to refine risk stratification by integrating precision oncology and shows promise in aiding treatment decisions, addressing the urgent need for improved stratification in a context where 30–50 % of DLBCL patients are not cured by standard chemotherapy.[10] In conclusion, our global multi-cohort study represents a significant advancement in the prognostic stratification of DLBCL. The integration of this model with clinical variables enabled the development of an accurate nomogram for survival prediction. Future studies should aim to validate this model in large prospective cohorts and explore its integration into clinical practice to enhance patient outcomes.

## Conflicts of interest

The author declares no conflicts of interest

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.htct.2025.103847.

## REFERENCES

1. Bakhshi TJ, Georgel PT. Genetic and epigenetic determinants of diffuse large B-cell lymphoma. Blood Cancer J. 2020;10(12):123. https://doi.org/10.1038/s41408-020-00389-w.
2. Jelicic J, Larsen TS, Maksimovic M, Trajkovic G. Available prognostic models for risk stratification of diffuse large B cell lymphoma patients: a systematic review. Crit Rev Oncol Hematol. 2019;133:1–16. https://doi.org/10.1016/j.critrevonc.2018.10.006.
3. Mendeville MS, Janssen J, Los-de Vries GT, van Dijk E, Richter J, Nijland M, et al. Integrating genetic subtypes with PET scan monitoring to predict outcome in diffuse large B-cell lymphoma. Nat Commun. 2025;16(1):109. https://doi.org/10.1038/s41467-024-55614-y.
4. Atallah-Yunes SA, Khurana A, Maurer M. Challenges identifying DLBCL patients with poor outcomes to upfront chemoimmunotherapy and its impact on frontline clinical trials. Leuk Lymphoma. 2024;65(4):430–9. https://doi.org/10.1080/10428194.2023.2298705.
5. Magnes T, Wagner S, Thorner AR, Neureiter D, Klieser E, Rinnerthaler G, et al. Spatial heterogeneity in large resected diffuse large B-cell lymphoma bulks analysed by massively parallel sequencing of multiple synchronous biopsies. Cancers (Basel). 2021;13(4):650. https://doi.org/10.3390/cancers13040650.
6. Eertink JJ, Zwezerijnen GJC, Heymans MW, Pieplenbosch S, Wiegers SE, Dührsen U, et al. Baseline PET radiomics outperforms the IPI risk score for prediction of outcome in diffuse large B-cell lymphoma. Blood [Internet]. 2023;141(25):3055–64. https://doi.org/10.1182/blood.2022018558.
7. Jelicic J, Juul-Jensen K, Bukumiric Z, Roost Clausen M, Ludvigsen Al-Mashhadi A, Pedersen RS, et al. Prognostic indices in diffuse large B-cell lymphoma: a population-based comparison and validation study of multiple models. Blood Cancer J [Internet]. 2023;13(1):157. https://doi.org/10.1038/s41408-023-00930-7.
8. NCCN [Internet]. [cited https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1480; 2025.
9. Tilly H, Gomes da Silva M, Vitolo U, Jack A, Meignan M, Lopez-Guillermo A, et al. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol [Internet]. 2015;26 Suppl 5(suppl 5):v116–25. https://doi.org/10.1093/annonc/mdv304.
10. Coiffier B, Sarkozy C. Diffuse large B-cell lymphoma: R-CHOP failure—What to do? Hematology Am Soc Hematol Educ Program [Internet]. 2016(1):366–78. https://doi.org/10.1182/asheducation-2016.1.366.

Valbert Oliveira Costa Filho [a], Felipe Pantoja Mesquita [a], Erick Figueiredo Saldanha [b], Pedro Robson Costa Passos [a], Mariana Macambira Noronha [c], Silvia Helena Barem Rabenhorst [c,*]

[a] *Center of Research and Drug Development (NPDM), Federal University of Ceara, Fortaleza, Brazil*

[b] *Division of Medical Oncology and Hematology, Princess Margaret Cancer Center, University Health, Biomedical Research Center, Federal University of Ceara, Fortaleza, Brazil*

[c] *Biomedical Research Center, Federal University of Ceara, Fortaleza, Brazil*

*Corresponding author.
*E-mail address:* srabenhorst@hotmail.com (S.H.B. Rabenhorst).

Received 9 February 2025
Accepted 9 March 2025
Available online xxx

https://doi.org/10.1016/j.htct.2025.103847
2531-1379/