Journal of Clinical and Nursing Research, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JCNR
Online ISSN: 2208-3693
Print ISSN: 2208-3685

Prognostic Classification of Hepatocellular Carcinoma Under Incomplete Data Conditions Using a Public Reference Cohort: A Weakly Supervised Multi-Omics Study

Hanlin Lv[1], Xiao Wang[1], Kesong Wu[2], Lei Li[3]*, Lei Wang[1]*

[1]RYTECH, Wuhan 430074, Hubei, China
[2]Ian Frazer Centre for Children's Immunotherapy Research, The University of Queensland, Brisbane, QLD 4101, Australia
[3]School of Life Science and Technology, Southeast University, Nanjing 210096, Jiangsu, China

***Corresponding authors:** Lei Li, lei.li@seu.edu.cn; Lei Wang, raywong2121@gmail.com

Abstract: *Background*: Prognostic stratification of hepatocellular carcinoma (HCC) remains difficult because the disease is highly heterogeneous and complete matched multi-omics data are not always available in clinical cohorts. *Objective*: To develop a weakly supervised multi-omics framework that derives prognostic subtype labels from a public reference cohort and transfers them to cohorts with incomplete data. *Methods*: This study has analyzed 363 patients in TCGA-LIHC with matched mRNA, miRNA, DNA methylation, and clinical data. Overall-survival-related features were selected by univariate Cox regression, integrated by similarity network fusion (SNF), and clustered by spectral clustering to generate pseudo-labels. TCGA-LIHC was then split 6:4 into training and test sets for supervised modeling. External validation used LIRI-JP, GSE14520, GSE54236, and GSE31384, with models rebuilt on features shared with each cohort. Prognostic performance was evaluated by Kaplan-Meier analysis, log-rank testing, and the concordance index (C-index). *Results*: A total of 3,890 mRNA features, 150 miRNA features, and 1,889 methylation features were retained. SNF plus spectral clustering identified two subtypes: S1 (n = 257) and S2 (n = 106). S2 had significantly worse overall survival than S1 (log-rank $P = 3.891 \times 10^{-9}$; C-index = 0.866). In internal validation, XGBoost showed the highest AUC (0.983). In external validation, the predicted subtypes remained prognostically informative, with C-index values of 0.857 in LIRI-JP, 0.875 in GSE14520, 0.930 in GSE54236, and 0.883 in GSE31384. *Conclusions*: In the public datasets included in this study, this weakly supervised framework identified two prognostically distinct HCC subtypes and retained prognostic discrimination after transfer to external cohorts with incomplete data.

Keywords: Hepatocellular carcinoma; Prognostic classification; Machine learning; Incomplete data; Multi-omics

Online publication: Apr 30, 2026

1. Introduction

Hepatocellular carcinoma is the most common form of primary liver cancer, and survival remains unsatisfactory despite advances in surgery and systemic therapy [1–3]. Marked biological heterogeneity complicates prognosis assessment and limits the precision of conventional clinical classifications [3].

Multi-omics studies have improved characterization of HCC heterogeneity, but many prognostic models still depend on complete matched data and are therefore difficult to transfer to local cohorts with incomplete multi-omics profiles [3,4]. Previous gene-expression- and miRNA-based studies have proposed several HCC subtyping frameworks, but transferability across cohorts remains challenging [5–7]. In this study, TCGA-LIHC was used as a public reference cohort to derive prognostic pseudo-labels through Cox filtering, SNF integration, and spectral clustering, and then tested whether the learned subtype structure could be transferred to four external cohorts with different available feature spaces.

2. Methods

2.1. Data sources and study design

The discovery dataset was TCGA-LIHC from UCSC Xena and included 363 HCC cases with matched RNA-seq, miRNA-seq, DNA methylation, and clinical data. External validation datasets were LIRI-JP from ICGC (232 tumors), GSE14520 (221 tumors), GSE54236 (81 tumors), and GSE31384 (166 matched HCC and noncancerous liver specimens). All analyzed data were open access and de-identified.

The study had two stages. First, Cox filtering, SNF, and spectral clustering were applied to the full TCGA-LIHC cohort to identify prognostic subtypes and assign pseudo-labels. Second, TCGA-LIHC was randomly divided 6:4 into training and test sets, supervised models were trained with the pseudo-labels, and external validation was performed after rebuilding models on the features shared between TCGA-LIHC and each validation cohort.

2.2. Preprocessing and subtype labeling

Features and samples with more than 20% missing values were removed, remaining missing values were imputed with the mean, and the lowest 20% of biological features by mean value were discarded. For methylation data, CpG sites within 1,500 base pairs upstream of the transcription start site were aggregated by gene symbol. Clinical categorical variables were one-hot encoded.

To reduce noise in high-dimensional unsupervised integration, univariate Cox proportional-hazards models were first used to retain overall-survival-related features ($p < 0.05$). Each omics layer and the clinical data were then converted to patient-similarity matrices and integrated by SNF [8]. Spectral clustering was evaluated for k = 2–10, and the optimal cluster number was selected by the Silhouette score and Calinski-Harabasz index.

2.3. Supervised transfer and evaluation

Four classifiers (logistic regression, random forest, XGBoost, and neural network) were trained on the TCGA-LIHC training set with the pseudo-labels as weak supervision. Ten-fold cross-validation was used to summarize accuracy, precision, recall, and AUC, and the classifier with the highest AUC was selected.

For external validation, models were rebuilt on the TCGA-LIHC training set after restricting features

to the intersection with each validation dataset. The shared feature counts were 15,989 mRNAs for LIRI-JP, 10,132 mRNAs for GSE14520, 15,175 mRNAs for GSE54236, and 338 miRNAs for GSE31384. After cross-cohort normalization, the best model for each dataset was applied to predict subtypes. Prognostic discrimination was assessed by Kaplan-Meier analysis, log-rank testing, and C-index estimation.

3. Results

3.1. Identification of two prognostic subtypes in TCGA-LIHC

After preprocessing and survival filtering, 3,890 mRNA features, 150 miRNA features, and 1,889 methylation features remained. SNF followed by spectral clustering supported a two-cluster solution and identified S1 (n = 257) and S2 (n = 106) (**Figure 1A**). Overall survival differed significantly between the subtypes, with worse outcome in S2 (**Figure 1B**; log-rank $p = 3.891 \times 10^{-9}$; C-index = 0.866).

Compared with S1, S2 showed a less favorable distribution of pathologic T stage, overall stage, histologic grade, and Child-Pugh classification, supporting the clinical relevance of the pseudo-labels.

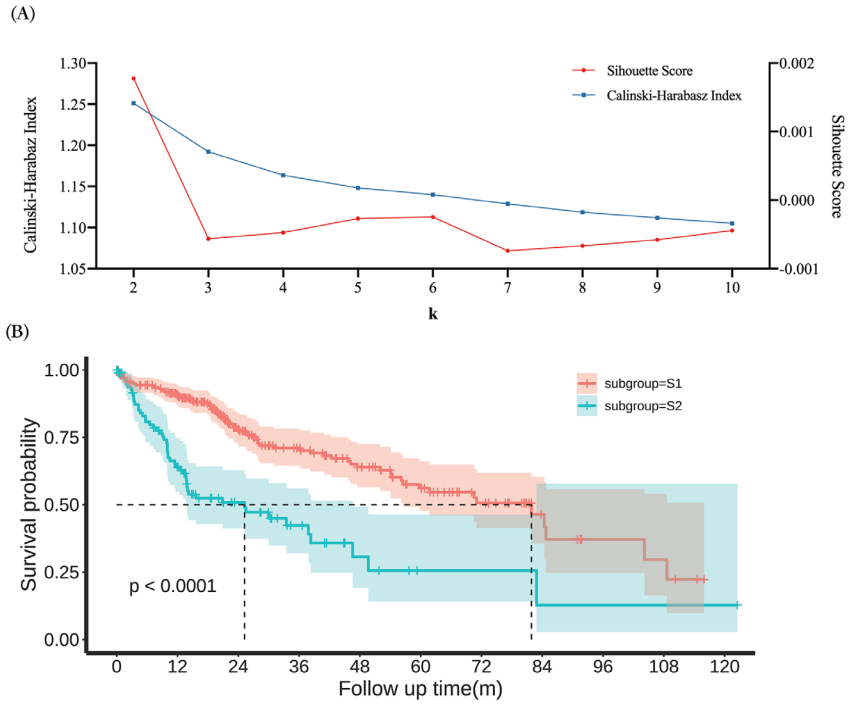

Figure 1. Subtype identification and survival difference in TCGA-LIHC. (A) Selection of cluster number using the Silhouette score and Calinski-Harabasz index. (B) Kaplan-Meier survival curves for the two subtypes.

3.2. Transferability in incomplete external cohorts

In internal validation, all four classifiers performed well, and XGBoost had the highest AUC (0.983), with accuracy of 0.97, precision of 0.91, and recall of 0.90 (**Table 1**). In external validation, logistic regression had the highest AUC in LIRI-JP, GSE14520, and GSE54236, with AUC values of 0.988, 0.990, and 0.992, respectively, whereas the neural network had the highest AUC in GSE31384 (0.988) (**Table 1**).

The predicted subtypes remained prognostically informative in all external cohorts (**Figure 2**). The C-index was 0.857 in LIRI-JP, 0.875 in GSE14520, 0.930 in GSE54236, and 0.883 in GSE31384, with

corresponding log-rank p values of 2.30×10^{-7}, 2.30×10^{-4}, 1.768×10^{-6}, and 0.029.

Table 1. Summary of internal and external validation

Cohort	Features	Model	AUC	C-index	p
TCGA-LIHC (internal)	mRNA + miRNA + methylation + clinical	XGBoost	0.983	0.866	3.891×10^{-9}
LIRI-JP	mRNA	Logistic regression	0.988	0.857	2.30×10^{-7}
GSE14520	mRNA	Logistic regression	0.990	0.875	2.30×10^{-4}
GSE54236	mRNA	Logistic regression	0.992	0.930	1.768×10^{-6}
GSE31384	miRNA	Neural network	0.988	0.883	0.029

Note: The classifier with the highest AUC in each cohort is shown. In internal TCGA-LIHC validation, XGBoost achieved an accuracy of 0.97, precision of 0.91, and recall of 0.90.

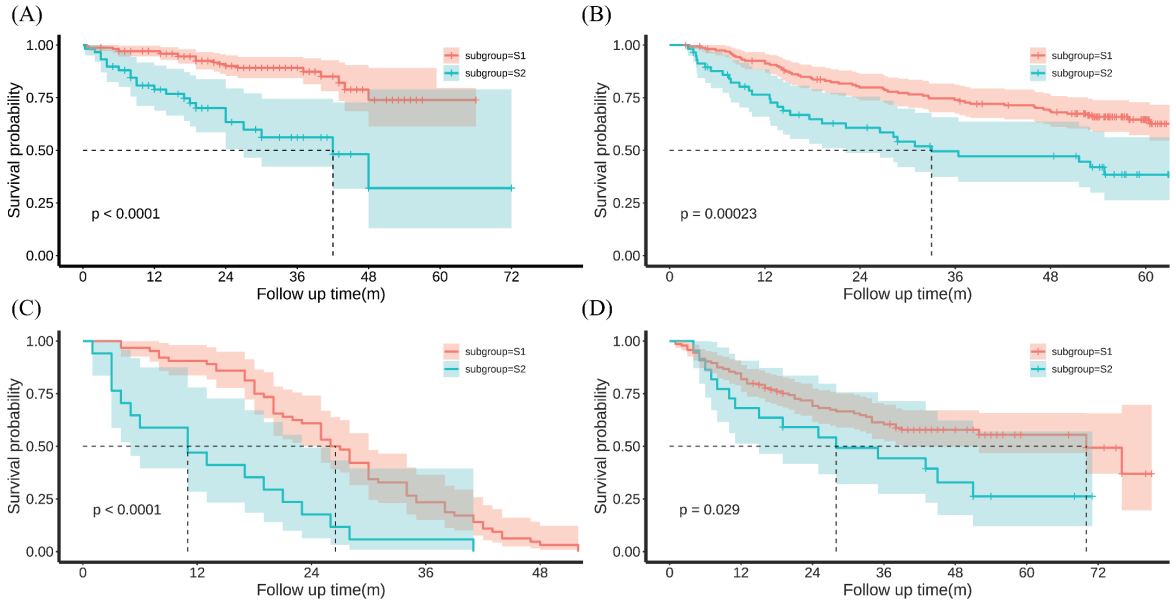

Figure 2. Kaplan-Meier survival curves in external validation cohorts. (A) LIRI-JP; (B) GSE14520; (C) GSE54236; (D) GSE31384.

4. Discussion

This study used a weakly supervised strategy to derive prognostic subtype labels from a public reference cohort and then transfer that structure to external cohorts with incomplete data. The central evidence is the consistent survival separation observed in TCGA-LIHC and across four external cohorts, which is broadly consistent with previous transcriptomic- and miRNA-based HCC subtyping studies [5–7].

The practical value of the framework lies in its use when complete matched multi-omics data are unavailable locally, because subtype labels are learned from a public reference cohort and the classifier can be rebuilt within the shared feature space of each target dataset.

This study has several limitations. The discovery and validation cohorts were modest in size. All analyses were retrospective and based on public datasets, which may introduce selection and platform bias. In addition, the biological interpretation remains computational and requires further empirical validation.

5. Conclusion

In the public datasets included in this study, a weakly supervised, SNF-based multi-omics framework identified two clinically relevant prognostic subtypes of HCC and retained prognostic discrimination across external cohorts with incomplete data. These findings support the use of public multi-omics resources for HCC prognostic classification when local data are not fully matched.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Bertuccio P, Turati F, Carioli G, et al., 2017, Global Trends and Predictions in Hepatocellular Carcinoma Mortality. Journal of Hepatology, 67(2): 302–309.

[2] Villanueva A, 2019, Hepatocellular Carcinoma. New England Journal of Medicine, 380(15): 1450–1462.

[3] Wu Y, Liu Z, Xu X, 2020, Molecular Subtyping of Hepatocellular Carcinoma: A Step Toward Precision Medicine. Cancer Communications, 40(12): 681–693.

[4] Zhang Q, Lou Y, Yang J, et al., 2019, Integrated Multiomic Analysis Reveals Comprehensive Tumour Heterogeneity and Novel Immunophenotypic Classification in Hepatocellular Carcinomas. Gut, 68(11): 2019–2031.

[5] Lee J, Chu I, Heo J, et al., 2004, Classification and Prediction of Survival in Hepatocellular Carcinoma by Gene Expression Profiling. Hepatology, 40(3): 667–676.

[6] Hoshida Y, Nijman S, Kobayashi M, et al., 2009, Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma. Cancer Research, 69(18): 7385–7392.

[7] Toffanin S, Hoshida Y, Lachenmayer A, et al., 2011, MicroRNA-Based Classification of Hepatocellular Carcinoma and Oncogenic Role of miR-517a. Gastroenterology, 140(5): 1618–1628.

[8] Wang B, Mezlini A, Demir F, et al., 2014, Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. Nature Methods, 11(3): 333–337.