PREDICTIVE MODELLING FOR OVERALL SURVIVAL IN PATIENTS WITH HEPATOCELLULAR CARCINOMA USING INTEGRATED BIOINFORMATICS AND CLINICOPATHOLOGICAL FEATURES

MUSA, A. O.[1], ISAH, A.[2], & YAKUBU, Y.[3]
[1]Mathematics and Statistics Department,
Confluence University of Science and Technology, Osara, Nigeria.
[2,3]Mathematics Department, Federal University of Technology Minna, Nigeria
E-mail: musaoyiza002@gmail.com

Abstract
The study aimed to develop a predictive model for overall survival (OS) in patients diagnosed with hepatocellular carcinoma (HCC) by integrating bioinformatics techniques with clinicopathological data. Survival analysis method was employed, including the Cox proportional hazards model and Kaplan-Meier survival curves, to analyse a retrospective cohort of patients. Feature selection was carried out using Least Absolute Shrinkage and Selection Operator (LASSO) regression to identify key prognostic variables. The final model included seven predictors: gender, alcohol consumption, molecular hallmark, encephalopathy, mean corpuscular volume, total protein, and survival time. Patients were stratified into high- and low-risk groups based on the median risk score derived from the model. The model demonstrated good predictive performance with an accuracy of 76.7% and a specificity of 90%. Kaplan-Meier analysis revealed significant differences in survival probabilities between the two groups. The findings suggested that integrating molecular and clinical features can improve prognostic accuracy and aid in risk stratification of HCC patients.

Keywords: Cancer, Predictive, Prognostic, Survival, Tumour

Introduction

Hepatocellular carcinoma (HCC), the most common form of primary liver cancer, is a major global health burden, accounting for over 700,000 deaths annually (Bray *et al.,* 2018). The disease is typically diagnosed at advanced stages due to its asymptomatic progression, which limits treatment options and contributes to poor overall survival (OS) rates. Although clinical prognostic models such as the tumour-node-metastasis (TNM) staging system and markers like vascular invasion provide useful information, they do not adequately capture the molecular heterogeneity among HCC patients.

Survival analysis has emerged as a robust approach for evaluating time-to-event outcomes, enabling researchers to quantify survival probabilities and hazard risks. The integration of bioinformatics techniques with traditional clinical data presents an opportunity to develop more accurate prognostic models. These integrated models can identify high-risk patients more effectively and support personalized treatment strategies.

Kakos *et al.* (2022) pooled the data from the worldwide literature and showed that survival after liver transplantation for pediatric hepatocellular carcinoma is favourable and many children do well even if their tumours exceed certain potentially restrictive criteria originally developed to select adults with hepatocellular carcinoma for liver transplantation. Abstract Liver transplantation (LT) is the only potentially curative option for children with unresectable hepatocellular carcinoma (HCC). Systematic review of the MEDLINE, Scopus, Cochrane Library, and Web of Science databases (end-of-search date: 31 July 2020) were performed. The outcomes were overall survival (OS) and disease-free survival (DFS) and the effect of clinically relevant variables on outcomes using the Kaplan–Meier method and log-rank test were evaluated. Sixty-seven studies reporting on 245 children undergoing LT for HCC were included.

DFS data were available for 150 patients and the 1-, 3-, and 5-year DFS rates were 92.3%, 89.1%, and 84.5%, respectively. Sixty of the two hundred and thirty-eight patients (25.2%) died over a mean follow-up of 46.8 ± 47.4 months. OS data were available for 222 patients and the 1-, 3-, and 5-year OS rates were 87.9%, 78.8%, and 74.3%, respectively. Although no difference was observed between children transplanted within vs. beyond Milan criteria ($p = 0.15$), superior OS was observed in children transplanted within vs. beyond UCSF criteria ($p = 0.02$). LT can yield favourable outcomes for pediatric HCC beyond Milan but not beyond UCSF criteria.

The cancerization of hepatocytes may lead to changes of the cell microenvironment, active substances, and enzymes. Viscosity is one of the important parameters of the cell microenvironment. Therefore, the study of the change in the viscosity of hepatocytes is very important for the detection and treatment of liver cancer. However, the hepatocyte-specific fluorescent probes which can detect viscosity have not been developed yet. Herein, the first hepatocyte-specific fluorescent probe (HT-V) for viscosity detection was designed and synthesized, which exhibited excellent optical properties for biological imaging studies. By using the unique probe HT-V, compared with the normal liver cells, a significant increase of viscosity in the liver cancer cells was observed in the cell imaging experiment. The organ imaging experiments showed that the probe HT-V could be successfully used to diagnose and image hepatocellular carcinoma in vivo. In addition, in situ imaging revealed that the new probe HT-V can specifically target and image hepatocellular carcinoma in mice.

Despite these advances, there are still several limitations and challenges in HCC research. To prevent overfitting, Guo *et al.* (2021) used a limited number of markers from a small dataset, which may affect the prediction model's outcomes. Additionally, (Steyerberg *et al.,* 2013) observed that parameters and weights in predictions were solely based on statistical modelling of patient data without input from clinical experts. These prognostic factors are more likely to be obtained when the dataset used is high in quality and relatively large, employs a sound statistical analysis strategy, and proposed model validation is done with an independent dataset (Steyerberg *et al.,* 2013).

Recent advancements in genomic profiling have revealed immune-related gene signatures and other molecular features that can serve as prognostic biomarkers. However, many existing models lack external validation, fail to integrate clinical and molecular features, and are not generalizable. Therefore, there is a growing need for comprehensive models that incorporate both data types.

Objective of the Study

This study developed and evaluated a predictive model for overall survival in patients with hepatocellular carcinoma by integrating clinicopathological data with bioinformatics-derived features. The goal is to identify independent prognostic factors and stratify patients into risk groups to inform clinical decision-making and improve prognostic accuracy.

Literature Review

Historically, survival prediction in HCC has relied on classical statistical methods such as logistic regression and the Cox proportional hazards model. Logistic regression is frequently employed for binary classification tasks, such as predicting one-year survival, due to its simplicity and interpretability. However, it assumes linearity between predictors and the log odds of the outcome, which can be overly simplistic in the presence of complex clinical data (Makary *et al.,* 2020).

The Cox model, widely used for survival analysis, models the hazard function and accommodates censored data, making it suitable for clinical settings where not all outcomes are observed within the study period. However, it relies on the proportional hazards assumption, which may not hold in heterogeneous patient populations (Ko *et al.*, 2020). Moreover, both models struggle with multicollinearity, non-linear relationships, and high-dimensional data, common challenges in modern medical datasets.

Navarrete *et al.,* (2023) reviewed the case of a 22-year-old male farmer who presents burning pain in the epigastrium with an intensity of 2/10 on the VAS scale, on palpation with the presence of a tumor in the right hypochondrium with ultrasound of the liver and bile ducts with evidence of liver mass in the segment VIII, in addition to exploratory laparoscopy with the discovery of multiple implants in 70% of the liver surface. In the histopathological analysis with morphological findings compatible with fibrolamellar carcinoma. In the laboratory with AST 356 UI/L, ALT 205 UI/L, GGT 930 U/L, LDH 343 UI/L AF 638 UI/L, for which treatment was started with atezolizumab 1200 mg and bevacizumab 1000 mg continuing to follow up. Fibrolamellar hepatocellular carcinoma is a variant of hepatocellular carcinoma, which usually presents with non-specific symptoms. The result showed that imaging approach is fundamental in the diagnosis, being the anatomopathological confirmation. Despite being low-frequency tumor, they should be suspected in all young patients with a palpable mass in the liver.

Toh *et al.*, (2023) evaluated the epidemiological trends and risk factors of Hepatocellular Carcinoma (HCC) and discussed the genetics of HCC including monogenic diseases, single-nucleotide polymorphisms, gut microbiome, and somatic mutations. Rebouissou and Nault (2020) Conducted a study on Hepatocellular Carcinoma and conclude that Hepatocellular carcinoma (HCC) arises from hepatocytes through the sequential accumulation of multiple genomic and epigenomic alterations resulting from Darwinian selection. Genes from various signalling pathways such as telomere maintenance, Wnt/β-catenin, P53/cell cycle regulation, oxidative stress, epigenetic modifiers, AKT/mTOR and MAP kinase are frequently mutated in HCC. Several subclasses of HCC have been identified based on transcriptomic dysregulation and genetic alterations that are closely related to risk factors, pathological features and prognosis. Undoubtedly, integration of data obtained from both preclinical models and human studies can help to accelerate the identification of robust predictive biomarkers of response to targeted biotherapy and immunotherapy.

Pinato *et al.,* (2022) discussed how the functional characteristics of the liver microenvironment can potentially be harnessed for the treatment of Hepatocellular Carcinoma (HCC). We will review the evidence supporting a therapeutic role for vaccines, cell-based therapies and immune-checkpoint inhibitors and discuss the potential for patient stratification in an attempt to overcome the series of failures that has characterised drug development in this disease area.

Santhakumar *et al.,* (2020) Hepatocellular carcinoma (HCC) is a heterogeneous inflammation-driven malignancy, which, despite significant advances in management, continues to portend a poor prognosis. Recent advances in basic and translational research have increasingly defined the role of the tumor microenvironment in the development and progression of HCC and facilitated the development of novel molecular targets. The hepatoma microenvironment is characterized by an immunosuppressive milieu of immune cells and tumor vasculature that is both structurally and functionally abnormal. Normalizing the tumor microenvironment by adopting a multipronged approach that targets both carcinogenic processes and the immunosuppressive milieu has been supported by pre-clinical and clinical data. In this review, we summarize the current understanding of the hepatoma microenvironment, its influences and dynamic interactions with tumor cells, the vasculature and the gut. They also conclude on

how manipulating the tumor microenvironment continues to shape the evolving landscape of HCC therapy.

Methods

Study Design and Population

This retrospective cohort study included patients diagnosed with hepatocellular carcinoma (HCC) between January 2015 and December 2023. Patients were identified from electronic health records (EHRs) and cancer registries across selected medical centres.

The inclusion criteria were: Histologically confirmed HCC diagnosis, Availability of clinical data (demographics, medical history, lab results, imaging), Minimum follow-up of 12 months or until death, and Availability of tumour tissue samples for genomic analysis

The exclusion criteria were: Diagnosis of non-HCC primary liver cancers, Incomplete clinical or follow-up data, and Absence of tissue samples for genomic profiling.

Data Preprocessing

Raw clinical and molecular data were pre-processed for quality and completeness. Missing values were handled using multiple imputations. Categorical variables were one-hot encoded, while continuous variables were normalised using z-score transformation:

$$Z = \frac{X - \mu}{\delta} \quad (1)$$

where x is the original value, μ is the mean, and σ is the standard deviation of the variable. Survival times were calculated from diagnosis to death or last follow-up. Patients who were alive at the end of the study were considered censored.

Feature Selection

To reduce model complexity and identify key prognostic variables, the Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied using the glmnet package in R. The LASSO optimization objective is:

$$Min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (2)$$

Where y_i is the outcome, x_i is the vector of predictors for the ith observation, β is the vector of coefficients, n is the number of observations, p is the number of predictors, and λ is the regularisation parameter that controls the strength of the penalty. The LASSO model was implemented using the glmnet package in R. The regularisation parameter λ was selected using cross-validation to minimize prediction error.

Model Development

The predictive model for overall survival was built using the Cox proportional hazards regression model, defined as:

$$h(t/\mathrm{X}) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\right) \quad (3)$$

where $h(t/\mathrm{X})$ is the hazard rate at time t for an individual with predictor values X, $h_0(t)$ is the baseline hazard function, β_j are the regression coefficients, and X_j are the predictor variables.

The Cox model was fitted using the survival package in R. The proportional hazards assumption was checked using Schoenfeld residuals.

Model Assumptions

The Cox model assumes

i. the hazard ratios are constant over time
ii. log hazard is linearly related to covariates
iii. Independence of Observations

Validation of Assumptions

The proportional hazards assumption was assessed using Schoenfeld residuals via the cox.zph function in R. No significant violations were detected ($p > 0.05$ for all covariates), confirming model validity.

Risk Stratification

A risk score for each patient was computed as a linear combination of the selected predictors and their coefficients

$$Risk\ score\ =\ \sum_{j=1}^{p} \beta_j X_j \quad (4)$$

Patients were classified into high-risk and low-risk groups based on the median risk score. Kaplan-Meier survival curves were plotted for both groups, and differences in survival were evaluated using the log-rank test.

Results and Discussion

Feature Selection Using LASSO Regression

LASSO regression was applied to select the most significant prognostic features from the clinical and genomic datasets. Seven variables with non-zero coefficients were retained:
Gender, Alcohol Consumption, Molecular Hallmark, Encephalopathy, Mean Corpuscular Volume (MCV), Total Protein (TP) and Survival Time.

These were subsequently used in the Cox regression model to build the survival prediction model.

Predictive Model and Performance Evaluation

The final Cox proportional hazards model was constructed using the seven selected variables. The risk score was calculated for each patient based on the model coefficients. Patients were then stratified into high-risk and low-risk groups using the median risk score.

Model Equation

$$\text{Risk Score} = -0.1166 \cdot \text{Gender} + 0.6679 \cdot \text{Alcohol} + 0.9957 \cdot \text{Hallmark} - 1.7490 \cdot \text{Encephalopathy} - 0.0012 \cdot \text{MCV} - 0.0438 \cdot \text{TP} \quad (5)$$

This equation implies that alcohol use and the presence of molecular hallmark genes are positively associated with risk, while encephalopathy, MCV, and TP are negatively associated with risk.

Model Performance Metrics

Table 1: Performance Metrics of the Predictive Model

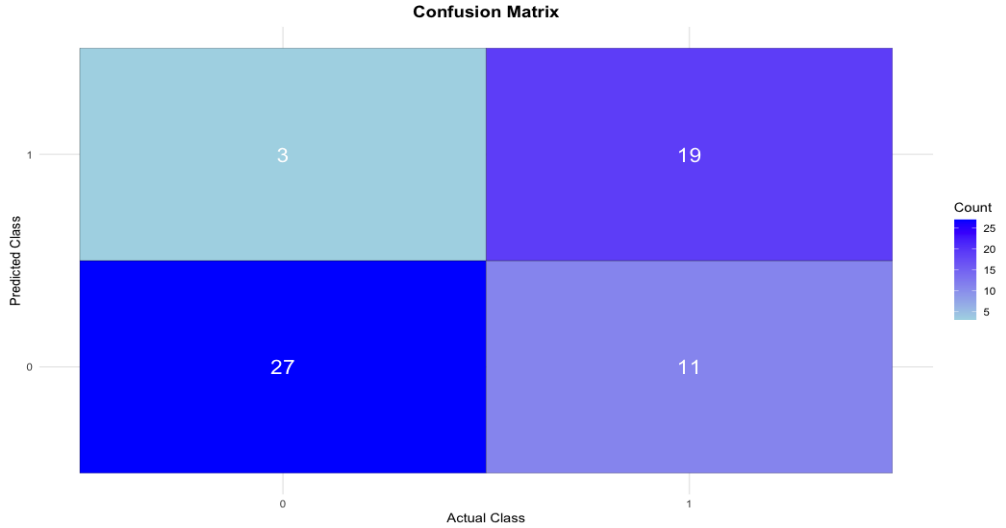
Metric	Value
Accuracy	0.7667
95% Confidence Interval	(0.6396, 0.8662)
No Information Rate	0.5
Sensitivity	63.33%
Specificity	90%

The model correctly classified 76.7% of patients. A sensitivity of 63.3% indicates that it detected high-risk patients moderately well, while a specificity of 90% demonstrates a strong ability to correctly classify low-risk patients.

Table 2: Confusion Matrix of the Predictive Model

	Reference 0 (Low Risk)	Reference 1 (High Risk)
Prediction 0 (Low Risk)	27	11
Prediction 1 (High Risk)	3	19

True Positives (TP): 19 (correctly predicted high-risk), True Negatives (TN): 27 (correctly predicted low-risk), False Positives (FP): 3, False Negatives (FN): 11



This confusion matrix confirms the model is more precise in identifying low-risk patients (few false positives), with slightly reduced performance on high-risk classification.

Kaplan-Meier Survival Analysis

Kaplan-Meier survival curves for the high- and low-risk groups demonstrated clear separation. The survival probability decreased more rapidly in the high-risk group, with the log-rank test confirming a significant difference ($p < 0.05$). This validates the discriminatory power of the model in identifying patient survival outcomes.

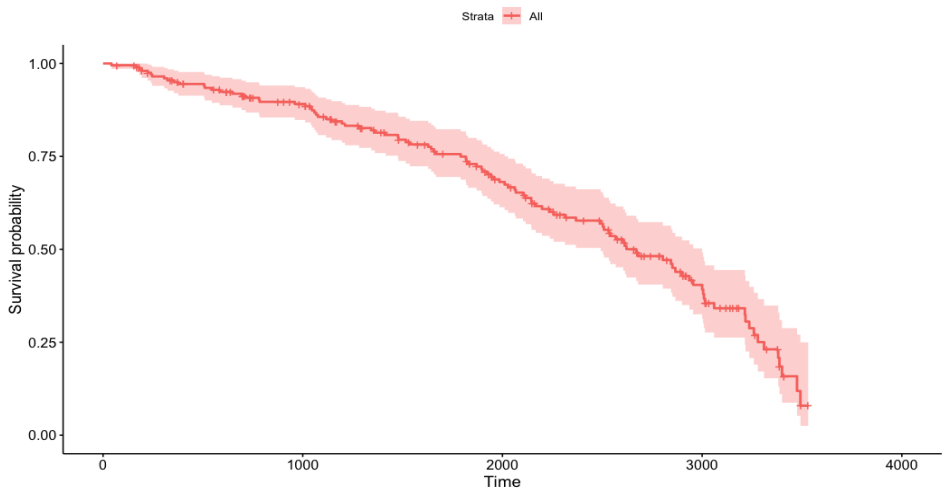


Figure 1: Kaplan-Meier survival curve

The Kaplan-Meier survival analysis results indicate a gradual decline in survival probability over time. At the start of the observation period, survival is nearly 100%, but as time progresses, the survival probability decreases, reaching approximately 50% at later time points. The standard error increases over time, suggesting greater uncertainty in survival estimates as the number of at-risk patients decreases. The 95% confidence interval narrows at the beginning and widens over time, reflecting reduced precision in estimates for longer follow-up periods. This trend helps in identifying patients at higher risk of poor survival, emphasizing the need for closer monitoring and potential intervention strategies for those with lower survival probabilities.

Conclusion

This study developed a predictive model for overall survival in patients with hepatocellular carcinoma by integrating clinicopathological and molecular features. Using LASSO regression for feature selection and Cox proportional hazards modelling, seven key prognostic variables were identified: gender, alcohol consumption, molecular hallmark, encephalopathy, mean corpuscular volume (MCV), total protein (TP), and survival time. The resulting model demonstrated good predictive performance, with an accuracy of 76.7% and a specificity of 90%. Risk stratification based on the model effectively differentiated between high-risk and low-risk patient groups, as confirmed by Kaplan-Meier survival analysis. These findings highlight the potential of combining clinical and molecular data to improve survival prediction in HCC management.

References

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424.

Guo, D. Z., Huang, A., Wang, Y. P., Cao, Y., Fan, J., Yang, X. R., & Zhou, J. (2021). Development of an eightgene prognostic model for overall survival prediction in patients with hepatocellular carcinoma. *JClinTranslHepatol.* 9(6), 898-908.

Kakos, C. D., Vlachaki, A., Tzilves, D., Triantafyllou, V., & Tzakis, A. G. (2022). Liver transplantation for pediatric hepatocellular carcinoma: A systematic review. *Transplantation*, *106*(1), 64–73.

Ko, K. L., Mak, L. Y., Cheung, K. S., & Yuen, M. F. (2020). Hepatocellular carcinoma: Recent advances and emerging medical therapies. *F1000Res.* 9, F1000. doi:10.12688/f1000research.24543.1

Makary, M., Khandpur, U., Cloyd, J., Mumtaz, K., & Dowell, J. (2020). Locoregional therapy approaches for hepatocellular carcinoma: Recent advances and management strategies. *Cancers*, 12. https://doi.org/10.3390/cancers12071914.

Navarrete, J. H. S., Torres, F. A. C., Zapata, D. P., & Granados, J. C. (2023). Fibrolamellar hepatocellular carcinoma: Case report. *Archivos Venezolanos de Farmacología y Terapéutica*, *42*(1), 61-64.

Pinato, D. J., Arizumi, T., Kudo, M. & Park, J. W. (2022). Immune checkpoint inhibitors in hepatocellular carcinoma: An update. *Liver International*, *42*(5), 935–948.

Rebouissou, S. & Nault, J. C. (2020). Advances in molecular classification and precision oncology in hepatocellular carcinoma. *Journal of Hepatology*, *72*(2), 297–315.

Santhakumar, C., Gane, E., Liu, K., & McCaughan, G. (2020). Current perspectives on the tumor microenvironment in hepatocellular carcinoma. *Hepatology International*, 14, 947 - 957. https://doi.org/10.1007/s12072-020-10104-3

Steyerberg, E. W., Vickers, A. J., Pencina, M. J., Kattan, M. W. & Briggs, A. H. (2013). Assessing the incremental value of diagnostic and prognostic markers: A review and illustration. *Biomarker Insights*, *8*, 129-139.

Toh, M., Ting, E., Hei, S., Tian, A., Lit-Hsin, L., Chow, P., & Yie, J. (2023). Global epidemiology and genetics of hepatocellular carcinoma. *Gastroenterology*. https://doi.org/10.1053/j.gastro.2023.01.033