Multivariable Survival Prediction In Hepatitis: Model Development And Internal Validation From A Retrospective Cohort

Chinelo Ijeoma Nnabude
Department of Statistics, Faculty of Physical Sciences
Nnamdi Azikiwe University, Nigeria

***Abstract*—Prognosis in hepatitis is heterogeneous, and risk prediction tools may improve early risk stratification especially in resource-limited settings. This study aimed to develop and internally validate a multivariable prediction model to identify clinical and biochemical predictors of survival in hepatitis patients. A retrospective cohort of 155 patients (32 deaths, 123 survivors) from the Kaggle Hepatitis Survival Dataset was analyzed using multivariable logistic regression. Predictors included age, bilirubin, albumin, and prothrombin time. Albumin emerged as the strongest independent predictor of survival (Adjusted OR = 0.210, 95% CI: 0.061–0.724, p = 0.016). The multivariable logistic regression model demonstrated excellent discrimination (AUC = 0.895), good calibration (Hosmer-Lemeshow p = 0.354), and high accuracy (Brier Score = 0.087). Albumin is an accurate and sensitive predictor of mortality in hepatitis patients. External validation in larger prospective African cohorts should be carried out before clinical implementation.**

***Index Terms*—Risk Stratification, Resource-Limited Settings, Internal Validation, Multivariable Logistic Regression, Hepatitis Prognosis**

I. INTRODUCTION

Hepatitis is inflammation of the liver, and is still a significant global health problem. It is a major contributor to morbidity and mortality globally, particularly in low- and middle-income countries due to incomplete vaccination, diagnosis and treatment. Hundreds of millions of people have chronic hepatitis B or C and viral hepatitis is still causing more than one million deaths each year, according to the Global Hepatitis report 2024 (World Health Organization, 2024). Apart from fatal consequences, hepatitis can be a cause of progressive damage or injury to the liver, leading to cirrhosis, hepatocellular carcinoma and disability (Asrani et al., 2019). The prognosis of hepatitis is not the same; some patients recover and some develop end-stage liver disease and for this reason the early assessment of the prognosis is essential.

Moreover, liver disease is still proven to be somewhat difficult to predict even with the current advancement in treatments. Ioannou et al. (2020) discovered that, despite treatment of hepatitis C with direct acting antivirals, there was still a higher risk of HCC and liver related death in the presence of advanced fibrosis. Tapper and Kanwal (2021) also addressed the importance of risk stratification and data-driven decision-making in the management of chronic liver disease. D'Amico et al. (2023) indicated that chronic liver disease does not have a linear clinical course, but rather consists of multiple clinical states and competing risks that ought to be modelled by individual-level prognostic models. In addition, Rinella et al. (2023) stated that with advances in the understanding of fatty liver disease and the spectrum of overlapping syndromes, prognosis has become more complicated. Overall, the studies illustrated clearly that the requirement of good predictive tools for managing high-risk patients is very important.

Clinical prediction models can be used to predict an individual patient's probability of developing a particular outcome in the future. In hepatology, prognostic tools practically play an established role. For instance, Kamath et al. (2001) developed the Model for End-Stage Liver Disease (MELD) which is one of the most frequently used multivariable prediction tools for liver disease. Strandberg et al. (2024) showed a good review of clinical prediction

modelling in hepatology recently, and concluded many prediction models remain subject to development, validation and reporting constraints. Arupzhanov et al. (2024) showed that prediction modelling methods can be beneficial to hepatitis prognosis and performed well in prediction of 1 year mortality of hepatitis patients using administrative health data. In spite of the progress described above for prognosis tools in Hepatology, there is still a need for simple and interpretable models based on routinely available variables. High prediction performance with complex machine learning methods is helpful, but traditional regression-based models are important for its clear and easy implementation. Steyerberg and Vergouwe (2014) highlighted that even a statistically developed and statistically valid prediction model may be useless if not understandable to clinicians. Additionally, Moons et al. (2015) highlighted the need for the development and transparent reporting of multivariable prediction models. Use of simple prediction models using readily available clinical and laboratory variables are highly beneficial and easier to implement in resource limited settings rather than complex algorithms which are technology intensive.

The predictors considered in this study were chosen because of their known clinical significance in liver disease and their routine availability in practice. Bilirubin reflects the hepatic excretory function and increased levels are associated with disease progression and reduced outcome. Bajaj et al. (2020) showed that organ dysfunction is one of the factors that influence liver-related survival in advanced liver disease with including predictors for the metabolism of bilirubin. Albumin is widely accepted as an indicator of liver synthetic function as well as nutritional and systemic status. Caraceni et al. (2018) showed a survival benefit in patients with decompensated cirrhosis using sustained albumin infusions and the prognostic role of albumin was reinforced and explained by Bernardi et al. (2022) in which new concepts regarding the clinical significance of albumin in chronic liver disease have been demonstrated.

Prothrombin time is also a relevant predictor because it reflects the liver's ability to synthesize clotting factors. Lisman and Porte (2022) showed that clotting factor disorders are linked with hepatic disease and retain their clinical significance as outcome predictors. The predictor 'age' was added because it could affect outcome through reduced physiological reserve, increased comorbidity index and diminished tolerance of liver injury. These predictors therefore provide a clinically meaningful and practical basis for developing a prognostic model.

II. METHODS

Study Design

This research was conducted as a retrospective cohort study aimed at developing and internally validating a multivariable prediction model for mortality among patients with hepatitis. The study relied entirely on secondary data from a publicly available repository, and model development was followed by internal validation using cross-validation techniques.

Data Source and Collection

The dataset was obtained from the Kaggle Hepatitis Survival Dataset, a publicly validated clinical repository used for prognostic modelling in liver disease. A total of 155 patients were included, comprising 123 survivors and 32 non-survivors. Only records with complete information on age, bilirubin, albumin, prothrombin time, and survival status were retained for analysis, following a complete-case approach.

The study used secondary data from a publicly available clinical repository for analysis. The underlying data source reflects patient-level hepatitis data and was provided for purposes of research and prediction modelling. The exact number and location of the participating locations were not identified within the accessible data set. Hence, this study could be characterized as a retrospective analysis using anonymised clinical data from an electronic health data source.

The included participants were patients recorded in the dataset as having hepatitis and their study variables were available for the analysis. Only data with values for all predictor variables and the outcome measure(s) were included in the prediction modelling part. Thus, complete-case analysis was employed for regression and for baseline characteristics comparison. No specific inclusion or exclusion criteria other than presence in the data set and availability of necessary study variables were employed for this analysis.

No treatment-related variables such as antiviral therapy, nutritional support, or other clinical

interventions were available in the dataset. Each record represented a single patient, and the dataset was anonymised, with no identifiers for hospitals, centres, or geographic location.

Data Preparation

The dataset was examined and validated for completeness, variable format, accuracy of coding and range of the variables prior to the analysis. Data cleaning involved checking the coding of the binary outcome variable so that survival status was consistently represented as 0 = survived and 1 = died, and verifying that predictor variables were in the format for analysis. Age, bilirubin, albumin, and prothrombin time were retained as recorded continuous variables, while sex was treated as a categorical variable. No feature engineering, transformation or scaling of any predictor was applied and no variable selection or dimensionality reduction was implemented. All statistical analyses were conducted using R software. A two-sided p-value < 0.05 was considered statistically significant.

The checks performed for data quality were minimal as the quality of data can only be checked by missing data and variable structure with the provided dataset. A complete-case approach was used, such that missing values within the required variables for each analysis excluded the participant from the given analysis. Each participant represented a single patient record; therefore, no repeated records from the same individual were present and no leakage of individuals across validation folds occurred

All data quality checks were performed after downloading the data set as information available for dataset audit, standardization of laboratory assays, external data quality checks was minimal.

No specific or different data cleaning methods were applied according to sociodemographic variables; all data preprocessing steps and the rule of complete cases were the same among subjects, and there were no age and gender specific steps. Due to the use of a secondary data set that provided very limited sociodemographic data it was impossible to perform additional procedures on these subgroups, like harmonization.

Outcomes

The outcome of interest was survival status, coded as a binary variable: 0 = survived and 1 = died. Survival status was assessed at the last recorded follow-up available in the dataset rather than at a pre-specified time point such as 30-day or 1-year mortality. Because detailed information on follow-up duration and censoring was unavailable, the study modelled mortality status at last observation rather than time to death. No information was available on whether outcome assessment was blinded to predictor information.

Predictors

The predictors assessed in this study were Age, Sex, Bilirubin, Albumin and Prothrombin Time (Protime). These predictors were selected because they are available and clinically relevant indicators in the assessment of liver disease severity and prognosis.

Age was recorded in years. Sex was treated as a categorical demographic variable. Bilirubin, albumin, and prothrombin time were included as biochemical or clinical laboratory predictors and they were assumed to be baseline characteristics in the used database as measurements before or simultaneously as outcome is later classified. Sex was examined in baseline descriptive analysis but was not included in the final multivariable model. It was excluded from the final model due to lack of statistical significance and limited sample size, which restricted power to detect subgroup differences.

There was no information in the documentation that was used about the particular laboratory technique and the time of blood drawal and test system, also no information was available in the dataset documentation on whether predictor assessment was performed without knowledge of the outcome or other predictors. Therefore, no formal statement on blinding of predictor measurement can be made.

Sample Size

The sample size was determined by the total number of participants available in the dataset. The initial dataset contained 155 patients with hepatitis, and all available observations were considered for analysis subject to completeness of the variables required for modelling. The dataset included 123 survivors and 32 non-survivors. These numbers formed the basis for descriptive analysis and model estimation. Because the number of deaths was relatively small, it is important that it is considered when interpreting model stability and statistical precision. The choice to utilize

cross-validation instead of a simple train-test split was guided by the small sample size and the limited number of events may increase the risk of overfitting and reduce the precision of estimates.

Missing Data

Complete case analysis was employed to deal with the missing data. For baseline comparisons and regression modelling, only observations with complete information on the variables required for a given analysis were included. There was no use of single imputation or multiple imputation. Although this was the more desirable approach as it kept the method uncomplicated and the analysis visible, it could potentially produce less accuracy and a biased result if the missingness mechanism was not completely at random. The effect of missing data is therefore acknowledged as a limitation of the study.

Model Development and Internal Validation

The baseline characteristics of this study are summarised using means and standard deviations for continuous variables and counts and percentages for categorical variables. These statistics were presented for the overall hepatitis patients and divided by the survival outcome.

Continuous variables as mean ± standard deviation were summarized and independent-samples t-tests was used for comparison. Categorical variables were summarized as counts and percentages, with Chi-square tests for comparison. Standardized mean differences (SMDs) for all continuous and binary variables were also calculated to measure imbalance regardless of sample size. SMD ≥ 0.10 indicated a meaningful imbalance. The results of these comparisons are shown in table 1 which also showed the number of observations used for each comparison. The primary modelling approach was multivariable logistic regression, it was chosen because the outcome was binary and also the dataset did not contain sufficient information on follow-up time or censoring to support time-to-event survival analysis. The objective was to estimate the probability of death for a patient given a set of demographic and biochemical characteristics. The main multivariable model included age, bilirubin, albumin, and prothrombin time as predictors. Sex was considered in descriptive analysis but was not included in the final multivariable model reported in the main results. The result is shown in table 3

For transparency, univariate logistic regression analyses were first conducted for each predictor separately to estimate crude associations with mortality. A multivariable logistic regression model was then fitted including the selected predictors simultaneously to obtain adjusted odds ratios with 95% confidence intervals. Model coefficients were estimated using maximum likelihood estimation. The result is shown in table 1b. Internal validation was performed using cross-validation to assess model stability and estimate optimism-adjusted predictive performance within the available dataset.

The full dataset of 155 patients was used for model development and internal validation. Because the sample size, particularly the number of deaths (n = 32), was limited, the data were not partitioned into separate training and test datasets, as such a split would reduce the effective sample available for model development and could produce unstable estimates. Instead, the model was developed using the available dataset and internally validated using k-fold cross-validation. In this procedure, the data were partitioned into k approximately equal folds; the model was trained on $k - 1$ folds and evaluated on the remaining fold, and this process was repeated until each fold had served once as the validation set. This approach allowed efficient use of the full dataset for both model development and performance evaluation while reducing optimism in apparent performance estimates.

The analysis did not explicitly account for clustering because the available dataset did not provide verified cluster identifiers such as hospital, centre, or country for modelling purposes. Therefore, heterogeneity in model parameters or performance across clusters could not be examined.

Model Formulation

The logistic model expresses the conditional probability. $\pi_i = P(Y_i = 1|X_i)$as follows:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}}} \quad (1)$$

where:

- π_i is the probability of the i*i*-th patient dying.
- β_0 is the intercept parameter.

- $\beta_1, \beta_2, \ldots \beta_p$ *are the regression coefficients for the predictor variables* $x_1, x_2, \ldots, x_p$ *(Age, Bilirubin, etc.).*
- x_{1i} *is the Age*
- x_{2i} *is the Bilirubin*
- x_{3i} *is the Albumin*
- x_{4i} *is the Prothrombin time (Protime)*
- *e is the* base of the natural logarithm.

The model is often linearised by using the logit transformation, which is the natural log of the odds of the event:

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad (2)$$

This transformation allows the relationship between the predictors and the outcome to be modelled as a linear combination.

Parameter Estimation Using Maximum Likelihood Estimation (MLE)

The parameters ($\beta_0, \beta_1, \ldots \beta_p$) of the logistic model are estimated using the Method of Maximum Likelihood Estimation (MLE) (Agresti, 2013). Unlike ordinary least squares used in linear regression, MLE seeks to find the parameter values that maximise the likelihood function. The likelihood function expresses the probability of observing the sample data as a function of the unknown parameters.

For a binary outcome, the likelihood function for a sample of *n* independent observations is given by:

$$L(\beta) = \prod_{i=1}^{n} \pi_i{}^{y_i} 1 - \pi_i{}^{1-y_i} \quad (3)$$

The log-likelihood function:

$$\ell(\beta) = ln\,[L(\beta)] = \sum_{i=1}^{n} [y_i ln(\pi_i) + (1-y_i) ln(1-\pi_i)] \quad (4)$$

The MLE process, performed iteratively using computational algorithms like the Newton-Raphson method, finds the values of β that maximise ℓ(β). The resulting estimates are denoted as $\hat{\beta}$. The asymptotic normality of MLE estimators allows for the construction of confidence intervals and hypothesis tests.

Hypothesis Testing and Confidence Intervals

The significance of individual coefficients was tested using the Wald statistic:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (5)$$

which approximately follows a standard normal distribution under the null hypothesis that βj=0. The corresponding p-value was derived from this distribution. Furthermore, 95% Confidence Intervals (CIs) for the odds ratios were constructed as:

$$e^{\hat{\beta}_j \pm Z_{1-\alpha/2}.SE(\hat{\beta}_j)} \quad (6)$$

where $Z_{1-\alpha/2}$ is the 97.5th percentile of the standard normal distribution. An OR whose 95% CI does not include 1.0 is considered statistically significant at the α=0.05 level.

Overall Model Fit

The global null hypothesis (that all slope coefficients are zero, $\beta 1 = \beta 2 = \cdots = \beta p = 0$ was tested using the Likelihood Ratio Test (LRT). The test statistic is:

$$G = -2ln\left[\frac{L(intercept\ only\ model)}{L(full\ model)}\right] \quad (7)$$

This statistic follows a chi-square distribution with p*p* degrees of freedom under the null hypothesis. A significant p-value indicates that the full model provides a better fit than a model with no predictors.

III. DATA ANALYSIS

All analyses were conducted in R. The full dataset of 155 patients was used for model development and internal validation. Because the number of deaths (n = 32) was small, the data were not partitioned into separate training and test sets because the partitioning would have reduced the effective sample available for model development. Internal validation was performed using k-fold cross-validation, the data were partitioned into 10 folds; in each iteration, the model was trained on 9 folds and evaluated on the remaining fold, and performance estimates were summarised across folds. Each participant was treated as representing a single patient, so no leakage of individuals across validation folds was expected.

Age, bilirubin, albumin, and prothrombin time were entered as continuous predictors in their original scales. No standardisation, transformation, categorisation, or feature reduction was performed. The prediction model was developed using multivariable logistic regression because the outcome was binary and the dataset lacked detailed follow-up time and censoring information required for time-to-event modelling. Predictors were selected a priori based on clinical relevance and data availability.

Univariate logistic regression analyses were first fitted for each predictor, followed by a multivariable logistic regression model including age, bilirubin, albumin, and prothrombin time simultaneously. No hyperparameter tuning was required.

Model performance was assessed using discrimination, calibration, and overall prediction error. Discrimination was also assessed by the area under the receiver operating characteristic curve (AUC). Furthermore, calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test and a calibration plot. Overall predictive accuracy was assessed using the Brier score. Predicted probabilities of death were obtained from the fitted logistic regression equation for each individual. No clustering adjustment was performed because the dataset did not provide usable cluster identifiers such as hospital or health centre.

IV. RESULTS

Participants flow and Characteristics

A total of 155 patients with hepatitis were included in the dataset, comprising 123 survivors and 32 non-survivors. The study analysed all available observations subject to completeness of the variables required for each analysis. Because the dataset did not include detailed follow-up duration, no summary of follow-up time could be reported.

Baseline characteristics stratified by survival status are presented in Table 1. Compared with survivors, non-survivors were older and had worse biochemical profiles. Specifically, non-survivors had higher mean bilirubin levels, lower mean albumin levels, and longer mean prothrombin time. The largest standardized mean differences were observed for albumin and prothrombin time, indicating clinically meaningful imbalance between the two groups.

Unadjusted associations between candidate predictors and mortality

Univariate logistic regression results are shown in Table 2. When considered individually, age, bilirubin, albumin, and prothrombin time were all significantly associated with mortality. Higher age and higher bilirubin were associated with increased odds of death, whereas higher albumin was strongly protective. Prothrombin time also showed a significant association with mortality in the univariate analysis.

Table 2 represented unadjusted associations with mortality, controlling for each predictor considered separately. Undjusted odds ratios greater than 1 indicate increased odds of mortality, while unadjusted odds ratios less than 1 indicate protective effects. Confidence intervals (CIs) are presented at the 95% level and statistical significance was defined as $p < 0.05$.

The result in table 2 showed how each individual predictor relates to mortality when considered one at a time. Age has an odds ratio of about 1.04, meaning that with each additional year, the odds of death increase by roughly 4%, and this effect is statistically very strong. Bilirubin showed odds ratios greater than 1, which suggested that higher levels of these liver function markers are associated with higher mortality, while albumin has a very low odds ratio (around 0.14), it showed a strong protective effect where higher albumin is linked to much lower odds of death. Prothrombin time (Protime) also appears associated with mortality. All of these results are based on separate unadjusted models, so they do not account for confounding between predictors.

Model development

The multivariable logistic regression model included age, bilirubin, albumin, and prothrombin time as prespecified predictors. Sex was not retained in the final multivariable model. A total of 155 observations contributed to the model, including 32 deaths.

The multivariable model results are presented in Table 3. After adjustment for the other predictors, albumin remained the only statistically significant independent predictor of mortality. Higher albumin was associated with substantially lower odds of death. Bilirubin and prothrombin time showed borderline associations, while age was no longer statistically significant after adjustment.

Table 3 represented adjusted associations with mortality, controlling for age, bilirubin, albumin, and protime. Adjusted odds ratios greater than 1 indicate increased odds of mortality, while adjusted odds ratios less than 1 indicate protective effects. Confidence intervals (CIs) are presented at the 95% level and statistical significance was defined as $p < 0.05$.

The result in table 1a showed the adjusted logistic regression model, age, bilirubin, albumin, and prothrombin time (protime) were evaluated together as predictors of mortality. Age has an adjusted odds ratio

of 1.039 with a 95% confidence interval from 0.98 to 1.10 and a p-value of 0.193, which means that although the point estimate suggests slightly higher odds of death with increasing age, this association is not statistically significant. Bilirubin has an adjusted odds ratio of 1.771 (95% CI 1.00–3.15, p = 0.052), which indicated that higher bilirubin levels are associated with higher odds of mortality and that this effect showed borderline statistical significance. Albumin showed a strong and clearly significant association, with an adjusted odds ratio of 0.210 (95% CI 0.06–0.75, p = 0.016); which means that higher albumin levels are independently protective, substantially reducing the odds of death even after adjusting for the other factors. Prothrombin time has an adjusted odds ratio of 0.964 (95% CI 0.93–1.00, p = 0.062), which suggested a small potential protective effect as values increase, but the confidence interval just touches 1 and the p-value is slightly above 0.05, so this is best interpreted as a marginal association rather than a clearly statistically significant one. Although the adjusted odds ratio suggested a marginal protective effect, the confidence interval includes 1 and the p-value is borderline, so this should be interpreted with cautious.

Baseline characteristics of the study cohort stratified by survival status presented in Table 1, showed characteristics comparison of survivors (123 patients) and non-survivors (32 patients). The SMD (standardized mean difference) showed how big the difference is in practical terms, and the p-value tells you whether it's statistically significant.

Survivors are on average about 40 years old, while non-survivors are about 47 years old. The SMD of 0.53 and p-value 0.019 mean that non-survivors tend to be meaningfully and significantly older than survivors. So, higher age is associated with a higher chance of dying in the group.

Survivors have a mean bilirubin of about 3.2 mg/dL, while non-survivors have about 5.7 mg/dL. The SMD of 0.60 and p-value 0.020 show that non-survivors have substantially and significantly higher bilirubin levels. Clinically, this means that worse liver function is linked with a higher risk of death.

Survivors have an average albumin of about 3.54 g/dL, whereas non-survivors average 2.87 g/dL. The SMD here is 0.90 with a p-value < 0.001, which is a large and highly statistically significant difference. This showed that non-survivors have much lower albumin levels, and low albumin is very strongly associated with death in this cohort. Albumin stands out as one of the strongest markers of poor outcome.

Survivors have a mean prothrombin time of about 15.8 seconds, compared with 18.9 seconds in non-survivors. The SMD of 0.82 and p-value 0.001 indicate another large and statistically significant difference. Longer prothrombin time reflects worse blood clotting function due to liver failure; it indicated that worse coagulation (longer PT) is strongly linked to mortality.

The final prediction model generated an individual predicted probability of death from the multivariable logistic regression equation including age, bilirubin, albumin, and prothrombin time. The primary purpose of the model was probabilistic risk estimation rather than classification into predefined risk groups. No decision threshold or risk categorisation scheme was applied in the present analysis.

Model Performance

Model performance was evaluated using discrimination, calibration, and overall predictive accuracy. The model showed good discrimination, with an AUC of 0.895, indicating strong ability to distinguish between survivors and non-survivors. Calibration was acceptable, with a Hosmer-Lemeshow test p-value of 0.354, suggesting no strong evidence of lack of fit. The Brier score was 0.087, indicating good overall agreement between predicted probabilities and observed outcomes.

The ROC curve of figure 1showed that the model performed well across a range of sensitivity and specificity values, with the curve lying well above the no-discrimination reference line. The calibration plot showed reasonable agreement between predicted and observed mortality across risk strata, indicating that the model's probability estimates were broadly reliable within the available data.

The ROC (Receiver Operating Characteristic) curve for the multivariable logistic regression model in figure 1 showed how well the model can discriminate between the hepatitis patients who die and those who survive. The x-axis represents the false positive rate (1 − specificity), and the y-axis represents the true positive rate (sensitivity). The curve lies well above the 45-degree diagonal "no discrimination" line, and the corresponding area under the curve (AUC) is about 0.89.

The shape of the curve, bending strongly toward the top-left corner, showed that there are points one can achieve high sensitivity while still maintaining relatively low false positive rates. Moreover, the ROC analysis supports the conclusion that the model is very effective at ranking patients from lower to higher mortality risk.

Figure 2: Calibration curve showing agreement between predicted probabilities of mortality and observed event rates across deciles of risk. The x-axis is the predicted probability of death (from the model), and the y-axis is the actually observed proportion of deaths within groups of patients who have similar predicted risks. The 45-degree diagonal line represents "perfect calibration" and the points representing deciles of predicted risk lie reasonably close to this diagonal line across the spectrum of predicted probabilities. However, Figure 2 above suggested that the model's probability estimates are trustworthy as absolute risks, not just as relative rankings.

A Brier score of about 0.087 is low, which is good which means the predicted probabilities are on average close to the true outcomes (0 or 1). Calibration assessed by the Hosmer–Lemeshow test was acceptable ($p > 0.05$). the points lie reasonably close to the diagonal, indicating that the model is reasonably well calibrated.

In univariate model, age, bilirubin and albumin all showed strong associations with mortality when considered on their own, with albumin being notably protective and bilirubin harmful. When you build a multivariable logistic model with age, bilirubin, albumin, and protime together, the model discriminates very well (AUC ~0.89), has good overall probability accuracy (low Brier score), and appears reasonably well calibrated by the calibration curve. This suggested that these variables together form a strong and fairly reliable risk prediction model for mortality in this hepatitis patients.

Internal validation was performed using cross-validation within the available dataset. This approach allowed the predictive performance of the model to be evaluated across repeated training and validation folds while preserving efficient use of the small dataset. The internal validation findings supported the overall stability of the model, although external validation in an independent cohort remains necessary before clinical implementation.

In summary, the study identified albumin as the most robust independent predictor of mortality among the variables assessed. Bilirubin and prothrombin time showed clinically relevant but borderline adjusted associations, whereas age was not independently associated with mortality after multivariable adjustment. The final model showed good discrimination, acceptable calibration, and low overall prediction error, supporting its potential value as a simple prognostic tool based on routinely available variables.

V. DISCUSSION

This study developed and internally validated a multivariable logistic regression model for mortality prediction among patients with hepatitis using routinely available demographic and biochemical predictors. The finding showed that albumin was the strongest independent predictor of mortality after adjustment for age, bilirubin, and prothrombin time. Higher albumin levels were associated with substantially lower odds of death, while bilirubin and prothrombin time showed borderline adjusted associations. Age was associated with mortality in univariate analysis but did not remain statistically significant in the multivariable model. Overall, the model showed excellent discrimination (AUC = 0.895), strong calibration, and high overall accuracy.

The findings highlighted the role of albumin in predicting survival outcomes. Lower albumin levels were strongly associated with mortality, reflecting impaired liver synthetic function and poor nutritional status. This reinforces the clinical utility of albumin as a practical and inexpensive predictor for risk stratification in hepatitis patients. The observed trends for bilirubin and prothrombin time are consistent with their known roles in hepatic excretory and synthetic function, though their lack of statistical significance in the adjusted model suggested overlapping contributions with albumin.

These results aligned with previous studies that have emphasized albumin as a key prognostic predictor in chronic liver disease.Caraceni et al. (2018), who demonstrated the clinical importance of albumin in decompensated cirrhosis, and with Bernardi et al. (2022), who further highlighted albumin as a central prognostic marker in advanced liver disease. The finding that bilirubin showed a positive but borderline

adjusted association with mortality is also consistent with its established role as a marker of hepatic dysfunction and disease severity, as discussed by Bajaj et al. (2020). Similarly, the borderline association for prothrombin time is biologically reasonable because worsening coagulation abnormalities often reflect progressive hepatic impairment, as noted by Lisman and Porte (2022).

The results also fit within the wider literature on prognostic modelling in hepatology. Kamath et al. (2001) showed that combinations of routine clinical and laboratory variables can provide meaningful prognostic information in liver disease. More recent work by Strandberg et al. (2024) has emphasised that prediction modelling in hepatology should prioritise both methodological rigour and practical clinical usefulness. Arupzhanov et al. (2024) likewise demonstrated that prediction approaches can perform well in hepatitis mortality prediction, although their approach used different data sources and modelling techniques. In that context, the present study contributes by showing that a simple regression-based model using accessible biomarkers can still provide clinically useful prognostic information.

At the same time, the interpretation of these findings should remain cautious. This study supports the prognostic value of albumin and suggests possible added contributions from bilirubin and prothrombin time, but it does not establish causal effects. In addition, because the model was developed and internally validated within a single retrospective dataset, the reported performance should not be interpreted as proof of transportability to other clinical settings. The findings therefore support the model as a promising prognostic tool rather than a model ready for routine clinical deployment.

Although bilirubin and prothrombin time did not reach statistical significance in the adjusted model, their clinical relevance suggests they may contribute to prognosis in larger or more diverse cohorts.

VI. CLINICAL IMPLICATIONS

The results of this study have meaningful implications for clinical practice and management of hepatitis patients. Among the predictors evaluated, albumin showed the most reliable and independent marker of survival, with higher levels strongly protective against mortality. This underscores the value of routine monitoring of albumin in hepatitis patients, as low albumin may serve as an early warning sign of poor prognosis and guide timely interventions such as nutritional support or closer clinical surveillance.

Bilirubin and prothrombin time (protime) also showed associations with mortality. Bilirubin reflected impaired liver function and may indicate progression toward hepatic failure, while prolonged prothrombin time highlighted compromised coagulation capacity. Clinically, these markers remain relevant for risk stratification, and their inclusion in predictive models improves the ability to identify patients at higher risk of adverse outcomes.

Although age showed a association with mortality in univariate analysis, it did not remain significant in the multivariable model, suggesting that its effect may be influenced by liver function parameters rather than acting as an independent predictor.

In summary, these results suggested that albumin, bilirubin, and prothrombin time can provide meaningful prognostic information in hepatitis patients. Including these variables into clinical risk assessment tools may improve early identification of high-risk individuals, support more personalized management strategies, and ultimately enhance patient outcomes.

VII. LIMITATIONS

This study has several important limitations. First, the dataset lacked temporal information such as enrolment dates, follow-up duration, and censoring, which prevented the use of time-to-event survival analysis and restricted the model to logistic regression. Second, the relatively small number of deaths (n = 32) limited statistical precision and increased the risk of overfitting, which may affect the stability of the estimates. Third, treatment-related variables (e.g., antiviral therapy, nutritional support, or clinical interventions) were unavailable, meaning the model could not account for therapeutic effects on survival. Fourth, complete-case analysis was used to handle missing data, which may introduce bias if the missingness was not completely at random. Finally, the study relied on a single, secondary dataset without external validation, restricting generalizability across different populations and healthcare settings.

VIII. CONCLUSION

This study developed and internally validated a multivariable logistic regression model for survival prediction in hepatitis patients using routinely available clinical variables. The analysis identified albumin as the most significant independent predictor of survival, while bilirubin and prothrombin time showed clinical trends toward mortality but did not reach statistical significance in the adjusted model. The multivariable logistic regression model showed excellent discrimination (AUC = 0.895), strong calibration, and high overall accuracy, underscoring its potential utility in clinical practice, but several limitations must be acknowledged, including the small number of deaths, absence of time-to-event data, and lack of external validation.

The findings emphasize the value of albumin as an accurate and widely accessible predictor for risk stratification, particularly in resource-limited settings where advanced prognostic tools may not be available. Furthermore, this internally validated model offers a practical framework for early risk stratification in resource-limited settings. Nevertheless, external validation in prospective African cohorts is essential to confirm its generalizability and clinical utility before widespread adoption.

IX. FUTURE RESEARCH

Further work is needed before this model can be recommended for implementation. The most important next step is external validation in independent and more representative hepatitis cohorts, particularly in African clinical settings and in datasets with clearer documentation of healthcare setting, recruitment, and follow-up. Such studies should assess not only discrimination and calibration, but also subgroup performance and fairness across relevant demographic and clinical groups.

Future research should also evaluate whether adding other clinically relevant predictors, such as treatment variables, virological markers, fibrosis measures, or comorbidities, improves predictive performance without sacrificing interpretability. If larger datasets become available, more robust modelling and validation strategies, including bootstrap validation, temporal validation, and head-to-head comparison with existing prognostic tools, should be considered. In addition, prospective studies with complete time-to-event data would allow proper survival modelling and estimation of risk over clinically meaningful time horizons.

Overall, the present model represents a useful step toward simple and accessible mortality risk prediction in hepatitis, but it should be considered an early-stage prognostic model requiring further validation, refinement, and implementation-focused evaluation before adoption in routine care.

ACKNOWLEDGMENT

The author declares that there are no acknowledgments to report for this study. No sponsorship, or institutional support was received.

REFERENCES

[1] Arupzhanov *et al.*, “One-year mortality prediction of patients with hepatitis in Kazakhstan based on administrative health data: A machine learning approach,” *Electron. J. Gen. Med.*, vol. 21, no. 6, 2024, doi: 10.29333/ejgm/15747.

[2] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, “Burden of liver diseases in the world,” *J. Hepatol.*, vol. 70, no. 1, pp. 151–171, 2019.

[3] J. S. Bajaj *et al.*, “Survival in infection-related acute-on-chronic liver failure is defined by extrahepatic organ failures,” *Hepatology*, vol. 72, no. 3, pp. 977–984, 2020.

[4] M. Bernardi *et al.*, “Albumin in decompensated cirrhosis: New concepts and perspectives,” *Gut*, vol. 71, no. 2, pp. 250–262, 2022.

[5] P. Caraceni *et al.*, “Long-term albumin administration in decompensated cirrhosis (ANSWER): An open-label randomised trial,” *The Lancet*, vol. 391, no. 10138, pp. 2417–2429, 2018.

[6] G. D’Amico *et al.*, “Clinical states of cirrhosis and competing risks,” *J. Hepatol.*, vol. 78, no. 1, pp. 134–144, 2023.

[7] G. N. Ioannou, P. K. Green, and K. Berry, “HCV eradication induced by direct-acting antiviral agents and the risk of hepatocellular carcinoma,” *J. Hepatol.*, vol. 73, no. 4, pp. 1392–1394, 2020.

[8] P. S. Kamath *et al.*, “A model to predict survival in patients with end-stage liver disease,” *Hepatology*, vol. 33, no. 2, pp. 464–470, 2001.

[9] T. Lisman and R. J. Porte, "Mechanisms of haemostasis and coagulation in the liver," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 19, no. 1, pp. 45–59, 2022.

[10] K. G. Moons *et al.*, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration," *Ann. Intern. Med.*, vol. 162, no. 1, pp. W1–W73, 2015.

[11] M. E. Rinella *et al.*, "A multisociety Delphi consensus statement on new fatty liver disease nomenclature," *J. Hepatol.*, vol. 79, no. 6, pp. 1542–1556, 2023.

[12] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: Seven steps for development and an ABCD for validation," *Eur. Heart J.*, vol. 35, no. 29, pp. 1925–1931, 2014.

[13] R. Strandberg, P. Jepsen, and H. Hagström, "Developing and validating clinical prediction models in hepatology – An overview for clinicians," *J. Hepatol.*, vol. 81, no. 1, pp. 149–162, 2024, doi: 10.1016/j.jhep.2024.03.030.

[14] E. B. Tapper and F. Kanwal, "The use of big data in the management of cirrhosis and hepatocellular carcinoma," *Clin. Gastroenterol. Hepatol.*, vol. 19, no. 6, pp. 1109–1117, 2021.

[15] World Health Organization, *Global hepatitis report 2024: Action for access in low- and middle-income countries*. Geneva: WHO, 2024.

[16] J. F. Wu *et al.*, "Clinical predictors of liver fibrosis in patients with chronic hepatitis B virus infection from children to adults," *J. Infect. Dis.*, vol. 217, no. 9, pp. 1408–1416, 2018, doi: 10.1093/infdis/jiy048.

Table 1. Baseline characteristics of hepatitis patients stratified by survival status

Variable	Survivors (n = 123)	Non-survivors (n = 32)	SMDs	*p*-value
Age (years, mean ± SD)	39.6 ± 12.4	46.6 ± 15.8	0.53	0.019
Sex (Male %)	68.3%	78.1%	0.21	0.309
Bilirubin (mg/dL, mean ± SD)	3.17 ± 3.22	5.66 ± 6.68	0.60	0.020
Albumin (g/dL, mean ± SD)	3.54 ± 0.74	2.87 ± 0.76	0.90	< 0.001
Prothrombin time (s, mean ± SD)	15.8 ± 3.5	18.9 ± 4.8	0.82	0.001

Overall, patients who died were generally older and showed evidence of poorer liver function at baseline.

Table 2: Univariate Logistic Regression Results for Predictors of Mortality in Hepatitis Patients

Variable	Unadjusted Odds Ratio	95% CI (Lower–Upper)	*p*-value
Age	1.044	1.03 – 1.05	0
Bilirubin	2.413	1.92 – 3.03	4.1e-14
Albumin	0.137	0.12 – 0.16	0
Protime	0.944	0.93 – 0.95	0

These unadjusted findings indicate that each predictor had prognostic relevance when examined separately, but they do not account for confounding or overlap between predictors.

Table 3. Multivariable logistic regression model for mortality in hepatitis patients

Variable	Adjusted Odds Ratio	95% CI (Lower–Upper)	*p*-value
Age	1.039	0.98 – 1.10	0.193
Bilirubin	1.771	1.00 – 3.15	0.052
Albumin	0.210	0.06 – 0.75	0.016
Protime	0.964	0.93 – 1.00	0.062

Albumin therefore emerged as the strongest independent predictor in the final model.

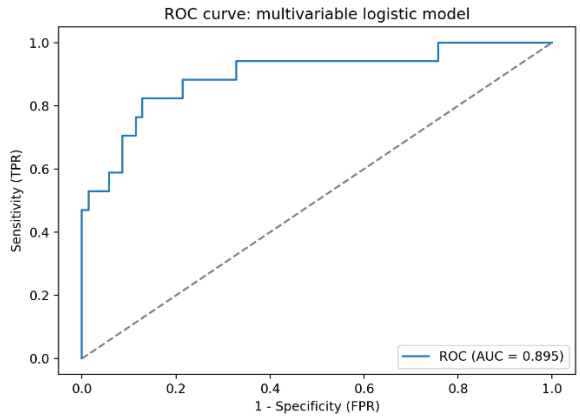

Fig. 1: Receiver Operating Characteristic (ROC) Curve for the Final Prediction Model

The ROC curve demonstrates the discriminatory ability of the logistic regression model. The Area Under the Curve (AUC) was 0.895, indicating excellent discrimination between survivors and non-survivors.

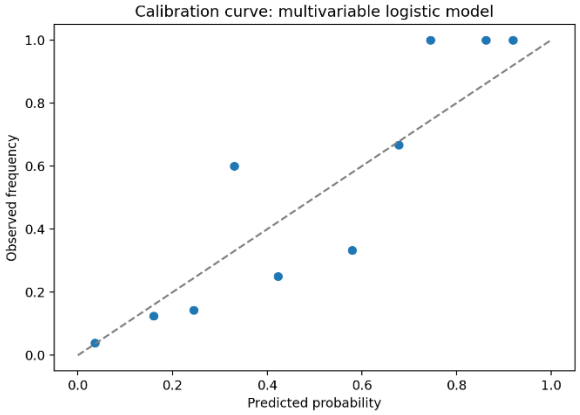

Fig. 2: Calibration Plot of Predicted vs. Observed Mortality

The calibration plot compares predicted probabilities of death with observed outcomes. The Hosmer-Lemeshow goodness-of-fit test yielded p = 0.354, confirming strong calibration of the model

Ethical Considerations

Plagiarism:

The author confirms that this manuscript is original and free from plagiarism. All sources have been properly acknowledged and referenced.

Conflicts of Interest:

The author declares that there are no conflicts of interest related to this study.

Ethical Approvals:

This research was conducted as a secondary analysis of anonymised data obtained from the publicly available Kaggle Hepatitis Survival Dataset. No direct patient contact occurred, and no personally identifiable information was included. Therefore, formal institutional ethical approval and informed consent were not required.