Published online on the journal's webpage: **https://journals.telkomuniversity.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 6 No. 2 (2025) 85 - 94 E-ISSN: 2809-8064

Application of Cox Proportional Hazards Modeling for Length of Stay in Anemia Patients

Amandasari Dinda Rabbani[*], Kartika Maulida Hindrayani[2*], Muhammad Nasrudin[3]

[1*,2*,3]Data Science, Computer Science, Universitas Pembangunan Veteran Jawa Timur

[1*]22083010093@student.upnjatim.ac.id, [2*]kartika.maulida.ds@upnjatim.ac.id , [3]nasrudin.fasilkom@upnjatim.ac.id

Abstract

Anemia is a major public health problem that may increase the risk of clinical complications and prolong the length of hospital stay (LOS). Identifying factors associated with LOS is important for understanding hospitalization patterns and supporting hospital resource management. This study applied the Cox Proportional Hazards model to identify demographic and hematological factors associated with the length of hospital stay among hospitalized anemia patients. Secondary data consisting of 1,098 inpatient hospitalization episodes recorded in 2025 at Haji Regional General Hospital, Surabaya, Indonesia, were analyzed using anonymized electronic medical records and clinical pathology laboratory data. The analysis included data preprocessing, exploratory survival analysis using Kaplan–Meier curves and the log-rank test, iterative variable selection, proportional hazards assumption testing, and model evaluation. The final Cox Proportional Hazards model identified age (HR = 0.993, $p < 0.05$), gender (HR = 0.815, $p < 0.05$), and hemoglobin level (HR = 0.947, $p < 0.05$) as significant factors associated with the hazard of hospital discharge. The final model satisfied the proportional hazards assumption and produced a log-likelihood value of −6022.375, an AIC of 12050.749, and a C-index of 0.565. Although the model satisfied the statistical assumptions, its discriminative ability remained limited. Therefore, the findings should be interpreted primarily as identifying factors associated with LOS rather than as a clinically validated predictive model. Future studies should incorporate additional clinically relevant variables and perform internal and external validation to improve model robustness, generalizability, and predictive performance.

Keywords: *Anemia, Survival Analysis, Cox Proportional Hazards, Length of Stay (LOS)*

1. Introduction

Anemia is a condition caused by low hemoglobin levels in the blood, significantly reducing the blood's ability to distribute oxygen to body tissues [1]. Anemia remains a global public health problem with a high prevalence, particularly among women of reproductive age, pregnant women, and children. The greatest burden occurs in Africa and Southeast Asia [2]. In Indonesia, the prevalence of anemia among adolescents aged 15-24 years reaches 32% and among pregnant women reaches 48.9%, indicating that anemia remains a serious health problem [3]. This condition can reduce patients' quality of life, increase the risk of clinical complications, prolong hospitalizations, and increase the burden on healthcare services. Therefore, effective, data-driven clinical evaluation is essential [1].

Length of Stay (LOS) is an important indicator in evaluating the quality of hospital services and the efficiency of healthcare resource utilization [27]. In anemic patients, LOS is influenced by various clinical and demographic factors such as hemoglobin (Hb), hematocrit (Hct), leukocytes, platelets, erythrocytes, Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), age, gender, and other medical conditions. However, conventional analysis methods are generally unable to optimally capture the complex risk relationships between these factors [4]-[6]. To address this problem, survival analysis is used because it is able to model time-to-event data and handle censored data effectively. One widely used method is the Cox Proportional Hazards model, which can identify factors that influence patient length of stay through a multivariate approach [7]-[10]. In addition, this model requires testing the proportional hazards assumption to ensure that the hazard ratio is constant throughout the observation period [9]. Survival analysis in this study is also supported by Kaplan-Meier curves and log-rank

Received: 24-06-2026 | Revised: 22-07-2026 | Published: 21-08-2026

tests to explore survival patterns between patient groups [8], [11].

Previous studies have applied survival analysis using the Cox Proportional Hazards model to analyze hospital length of stay among COVID-19 patients and general inpatient populations. Wang et al. reported that age, health conditions, and laboratory findings significantly influenced the length of stay among COVID-19 patients [5]. Similarly, Kim et al. demonstrated that demographic and clinical factors significantly affected hospital length of stay using the Cox Proportional Hazards model [4], while Momeni et al. applied survival analysis to predict hospital length of stay using large-scale inpatient data [6]. However, studies specifically investigating the length of stay among anemia patients remain limited, particularly those based on integrated hospital electronic medical records in Indonesia. Moreover, previous studies have mainly emphasized prediction performance or methodological comparisons rather than providing a comprehensive evaluation of the clinical, hematological, and demographic factors associated with hospitalization duration in anemia patients. Therefore, this study applies the Cox Proportional Hazards model to integrated electronic medical records from RSUD Haji Surabaya to identify the significant factors associated with the length of stay among anemia patients. By providing context specific empirical evidence from an Indonesian hospital setting, this study contributes to a better understanding of the determinants of hospitalization duration in anemia patients and provides evidence that may support clinical decision making and hospital resource planning for more effective anemia patient management.

Advances in health information technology support comprehensive patient data analysis through integrated electronic medical records and data-driven healthcare decision-making [12], [28]. The SATUSEHAT platform improves data interoperability and clinical data quality, which are important for accurate survival analysis and Cox Proportional Hazards model estimation [9]-[10], [13]-[14]. RSUD Haji Surabaya was selected because it has integrated SATUSEHAT-based medical records and a diverse patient population, making it suitable for developing a Cox Proportional Hazards model based on clinical, hematological, and demographic variables [13]-[15]. This study aims to identify factors influencing the length of stay of anemia patients and develop a Cox Proportional Hazards model to support data driven clinical decision making.

2. Research Methods

The research included data collection and preprocessing, exploratory data analysis, and Cox Proportional Hazards modeling, which included variable selection, assumption testing, evaluation, and model finalization. The research flowchart is presented in the Figure 1.

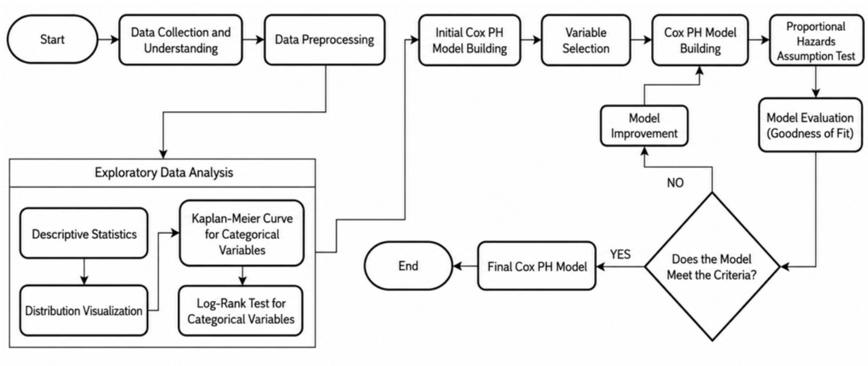

Figure 1. Research flowchart

2.1. *Data Collection and Interpretation*

This study began with the collection and interpretation of medical records and clinical pathology lab data from inpatients with anemia at Haji Regional General Hospital in Surabaya, including demographic, clinical, and hematological variables. This phase aimed to identify the characteristics and structure of the data before proceeding to the next stage of analysis. The study included inpatient medical records of anemia patients hospitalized during 2025. Patients were identified as anemia patients based on the hospital inpatient registry and electronic medical records provided by the hospital.

The inclusion criteria consisted of inpatient records with complete demographic, laboratory, discharge status, and length of stay information required for survival analysis. No records were excluded because all retrieved medical records contained complete information for the variables analyzed. Each hospitalization episode was treated as an independent observation. Therefore, if a patient was hospitalized more than once during the study period, each admission was included separately because it represented a different hospitalization episode with potentially different clinical characteristics and outcomes. In the survival analysis, hospital discharge (recovered/healthy or outpatient follow-up) was defined as the event (Status = 1), whereas referral to a higher-level hospital and in-hospital death were treated as censored observations (Status = 0).

This study received ethical approval from the Health Research Ethics Committee of Haji Regional General Hospital, East Java Province (Ethical Clearance No. 445/15/KOM.ETIK/2026). Electronic medical records were obtained with permission from the hospital and used exclusively for research purposes. To protect patient privacy and confidentiality, all personal identifiers were removed before analysis. Original hospital patient identifiers were replaced with new unique identification codes generated specifically for this study to ensure complete anonymization. All analyses were performed using anonymized data, and no information that could directly identify individual patients was retained.

2.2. *Data Preprocessing*

This stage prepares the data for survival analysis. This process includes data type conversion, adjusting variable column names, encoding categorical variables, establishing survival status, handling missing values and outliers, and creating categorical variables for Kaplan-Meier curve analysis and the log-rank test. Gender was encoded as 1 = Male and 0 = Female. This encoding was used consistently throughout the Cox Proportional Hazards analysis. Outliers in continuous laboratory variables (hemoglobin, leukocyte count, hematocrit, platelet count, erythrocyte count, MCV, MCH, and MCHC) were handled using an interquartile range (IQR) based winsorization approach. For each variable, the lower and upper thresholds were defined as $Q_1 - 1.5 \times IQR$and $Q_3 + 1.5 \times IQR$, respectively, where $IQR = Q_3 - Q_1$. Values falling outside these thresholds were capped at the corresponding boundary values rather than removed from the dataset. This approach was selected to reduce the influence of extreme observations on model estimation while preserving all patient records. Since extreme laboratory values may represent clinically meaningful severe conditions rather than measurement errors, winsorization was preferred over case deletion to retain potentially important clinical information while minimizing the disproportionate influence of outliers on the Cox Proportional Hazards model.

2.3. *Exploratory Data Analysis*

This stage aims to understand data characteristics before modeling through descriptive statistics, data visualization, Kaplan Meier curve analysis, and the log-rank test. Kaplan Meier curves were used to estimate survival probabilities based on gender, hemoglobin, age, leukocyte, and platelet groups [16]-[18], while the log-rank test assessed differences in survival between groups [11]. For the Kaplan Meier and log-rank analyses, continuous variables were categorized using clinically meaningful thresholds to facilitate survival curve visualization and group comparisons. Hemoglobin levels were classified into severe, moderate, mild, and normal categories according to the World Health Organization (WHO) guideline for anemia classification [2]. Age was categorized into infants/toddlers (0-5 years), children (6-12 years), adolescents (13-25 years), adults (26-59 years), and older adults (≥60 years) for exploratory survival analysis. Leukocyte categories (low, normal, and high) were determined using age-specific clinical laboratory reference ranges because normal leukocyte counts vary across age groups. Platelet categories were classified as low (<150,000/µL), normal (150,000-450,000/µL), and high (>450,000/µL) according to commonly accepted clinical laboratory reference ranges. These categorized variables were used exclusively for Kaplan Meier curve and log-rank analyses, whereas the Cox Proportional Hazards model retained the original continuous variables to preserve statistical information and avoid information loss due to categorization.

2.4. *Survival Analysis*

Survival analysis is a statistical technique used to analyze time to event data by considering the time until a particular event occurs while appropriately handling censored observations [19]-[20]. In this study, survival time was defined as the Length of Stay (LOS), measured as the number of days from hospital admission until hospital discharge. The event of interest was hospital discharge. Patients who did not experience the event during the observation period, such as those who died or were referred to another hospital, were treated as right censored observations. Survival analysis was selected because it can estimate the probability of discharge over time while accounting for censored data [8]. Because the event was defined as hospital discharge, the hazard ratio (HR) represents the relative rate of discharge. Therefore, an HR greater than 1 indicates a higher probability of earlier discharge (shorter LOS), whereas an HR less than 1 indicates a lower probability of discharge at any given time, corresponding to a longer LOS. The stages of the survival analysis are described as follows:

2.4.1. *Establishment of the Initial PH Cox Model*

At this stage, all preprocessed numerical variables were entered into the initial Cox Proportional Hazards model, including age, gender, hemoglobin, leukocyte count, hematocrit, platelet count, erythrocyte count, MCV, MCH, and MCHC. The initial model was fitted to estimate the regression coefficients, hazard ratios, and statistical significance of each predictor before the variable selection process.

2.4.2. *Variable Selection*

At this stage, variable selection was performed iteratively based on the Hazard Ratio (HR), p-value, Variance Inflation Factor (VIF), and proportional hazards assumption testing. The HR and p-value were used to evaluate the magnitude and statistical significance of the association between each predictor and the length of stay in the initial Cox Proportional Hazards model [18], [20]. Variables with a p-value < 0.05 were considered statistically significant and selected as candidate predictors for subsequent model development. Multicollinearity was assessed using the Variance Inflation Factor (VIF), which was used as a diagnostic indicator rather than a strict exclusion criterion [21]. Variables with elevated VIF values were further evaluated based on their statistical significance, clinical relevance, and contribution to the overall model before determining their inclusion. Variables with elevated VIF values were retained only when they remained statistically significant, clinically relevant, and satisfied the proportional hazards assumption. Variable selection was subsequently refined through proportional hazards assumption testing. Variables that violated the proportional hazards assumption were excluded, and the model was refitted iteratively until all retained variables satisfied the model assumptions.

2.4.3. *Establishing the Cox Proportional Hazards Model*

The Cox Proportional Hazards (Cox PH) model is a semiparametric regression method used to analyze the influence of covariate variables on the time to an event in survival data [9]-[10],[22]. The Cox Proportional Hazards model is formulated as follows:

$$\lambda(t \mid X) = \lambda_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) \quad (1)$$

$\lambda_0(t)$ is the baseline hazard or basic risk of an event at time (t) when all covariates are zero. The coefficients β_1 , β_2, …, β_p indicate the magnitude of the influence of each covariate X_1, X_2, ..., X_p on the risk of an event. Meanwhile, the exponential component $\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$ functions to modify the baseline hazard based on the covariate value and becomes the basis for calculating the hazard ratio.

This model produces a Hazard Ratio (HR) value to interpret the influence of each variable on the hazard of an event and assumes a constant hazard ratio during the observation period (proportional hazards) [10]. At this stage, the Cox Proportional Hazards model is formed using variables that have passed the selection based on the HR value, p-value, and VIF. The model is then evaluated using proportional hazards assumption testing and goodness of fit measures. The resulting model was subsequently evaluated using proportional hazards assumption testing and goodness of fit measures. If the evaluation criteria were not satisfied, the variable selection and model fitting procedures were repeated until a satisfactory model was obtained.

2.4.4. *Proportional Hazards Assumption Test*

The Proportional Hazards assumption test was conducted to ensure that the hazard ratio between groups remained constant throughout the observation period [9], [23]-[24]. In this study, the test was conducted using the Schoenfeld Residual, which is formulated as follows:

$$r_{ik} = x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{\beta x_j}}{\sum_{j \in R(t_i)} e^{\beta x_j}} \quad (2)$$

β is the regression coefficient in the Cox Proportional Hazards model that shows the influence of covariates on the risk of an event. The values x_{ik} and x_{jk} epresent the value of the (k)th covariate in the (i)th and (j)th patients, respectively. Meanwhile $R(t_i)$ is the set of patients who were still in the risk group immediately before the event time (t_i).

The test hypotheses are as follows:
H_0: The data meets the Proportional Hazards assumption
H_1: The data does not meet the Proportional Hazards assumption
The test decision is based on a significance level of $\alpha = 0.05$, where the assumption is declared met if the p-value is > 0.05. At this stage, testing is performed on the Cox Proportional Hazards model to ensure model validity. Variables that do not meet the Proportional Hazards assumption are considered for elimination or model refinement before developing the final model

2.4.5. *Model Evaluation (Goodness of Fit)*

This stage assesses the suitability of the Cox Proportional Hazards model to the survival data being analyzed. In this study, the evaluation was conducted using the log-likelihood value, the Akaike Information Criterion (AIC), and the Concordance Index (C-index). The log-likelihood value is used to measure the model's fit to the data, the AIC to assess the balance between fit and model complexity, and the C-index to evaluate the model's ability to predict the sequence of risk events. A better model is indicated by higher log-likelihood and C-index values and lower AIC values. Evaluation results are used to determine whether model refinement is required [23], [25]-[26]

2.5. *Refinement and Finalization of the Cox Proportional Hazards Model*

Model refinement was performed based on the proportional hazards assumption test and goodness-of-fit evaluation. Variables that violated the assumption or reduced model performance were considered for elimination. The Cox Proportional Hazards model was then rebuilt, retested, and re-evaluated iteratively until the established criteria were met. The model that satisfied the proportional hazards assumption and showed good evaluation results was designated as the final Cox Proportional Hazards model.

3. Results and Discussion

3.1. *Data Collection and Interpretation*

This study analyzed secondary data consisting of 1,098 medical records and clinical pathology lab reports of anemia inpatients collected at Haji Regional General Hospital, Surabaya, in 2025. The dataset comprised 13 demographic, clinical, and hematological variables, including age, sex, hemoglobin, leukocyte count, hematocrit, platelet count, erythrocyte count, MCV, MCH, MCHC, hospital discharge status, and length of stay (LOS). The availability of complete demographic, laboratory, and outcome information allowed all medical records to be retained for preprocessing and subsequent survival analysis without excluding any observations due to missing data. The dataset contained both event observations and censored observations, providing a sound basis for survival analysis using the Cox Proportional Hazards model. The inclusion of both hospital discharges and censored patients allowed survival analysis to estimate the probability of hospital discharge while appropriately accounting for missing observations. Furthermore, the absence of missing values minimizes the need for additional data imputation or case exclusion, thus maintaining the full sample size for model development.

3.2. *Data Preprocessing*

This stage was successfully completed before the survival analysis. No missing values were identified among the 1,098 medical records and clinical laboratory records of anemia patients. Therefore, all records were retained for subsequent analysis without requiring data imputation or case exclusion. After winsorization, the distribution of continuous laboratory variables became more stable while all observations were retained. Furthermore, the categorical variables required for Kaplan Meier and log-rank analyses, including anemia severity, age group, leukocyte category, and platelet category, were successfully generated, while the original continuous variables were retained for Cox Proportional Hazards modeling.

Ukuran data : (1098, 19)

	Kode	Jenis Kelamin	Usia	Kondisi Pulang	Hb	Leukosit	Hematokrit	Trombosit	Eritrosit	MCV	MCH	MCHC	LOS	Jenis Kelamin Encode	Status	Kategori Hb	Kategori Usia	Kategori Leukosit	Kategori Trombosit
0	A001	Laki-Laki	46.0	Meninggal > 48 Jam	9.2	15280.0	24.4	292000.0	3.55	68.7	25.9	36.5	25	1	0	Sedang	Dewasa	Tinggi	Normal
1	A002	Perempuan	55.0	Sembuh / Sehat	5.5	7230.0	18.0	106000.0	2.39	75.3	23.0	30.6	8	0	1	Berat	Lansia	Normal	Rendah
2	A003	Laki-Laki	53.0	Berobat Jalan (Kontrol RS)	9.9	17400.0	32.8	250000.0	4.18	78.5	23.7	30.2	8	1	1	Sedang	Lansia	Tinggi	Normal
3	A004	Laki-Laki	50.0	Berobat Jalan (Kontrol RS)	8.8	9390.0	25.1	298000.0	2.93	85.7	30.0	35.1	10	1	1	Sedang	Lansia	Normal	Normal
4	A005	Perempuan	25.0	Meninggal > 48 Jam	8.4	9640.0	23.8	301000.0	3.19	74.6	26.3	35.3	6	0	0	Sedang	Remaja	Normal	Normal

Figure 2. Data Preprocessing Results

Figure 2 illustrates the dataset after preprocessing. All numeric variables were successfully converted to the appropriate numeric format, and categorical variables were generated for exploratory survival analysis. Consequently, the dataset was deemed suitable for Kaplan-Meier analysis, log-rank testing, and Cox Proportional Hazards modeling. These categorical variables were created solely to visualize the Kaplan-Meier curve and log-rank testing, while the Cox Proportional Hazards model was developed using the original continuous variables to preserve statistical information.

3.3. *Exploratory Data Analysis*

Based on descriptive statistics, the dataset consisted of 1,098 medical records and clinical laboratory records of anemia patients with a mean age of 52.93 years, a mean hemoglobin level of 8.11 g/dL, and a mean length of stay of 6.57 days. Most continuous variables showed approximately normal distributions after winsorization. However, leukocyte and platelet counts remained positively (right) skewed, indicating the presence of relatively few patients with substantially elevated laboratory values.

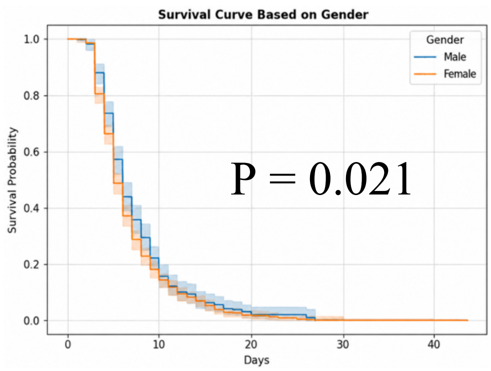

Figure 3. Gender survival curve

Based on Figures 3 to 7, the Kaplan Meier survival curves and log-rank test demonstrated significant differences in survival among gender ($p = 0.021$), hemoglobin category ($p = 0.001$), age category ($p = 0.003$), leukocyte category ($p < 0.001$), and platelet category ($p = 0.001$). These findings indicate that the probability of hospital discharge differed across the categorized groups during hospitalization. However, the Kaplan Meier and log-rank analyses were conducted solely as exploratory analyses to visualize survival patterns and compare survival distributions across categories. These categorized variables were not included in the Cox Proportional Hazards model. Instead, the Cox model was developed using the original continuous variables to preserve statistical information and estimate their associations with the hazard of hospital discharge among anemia patients.

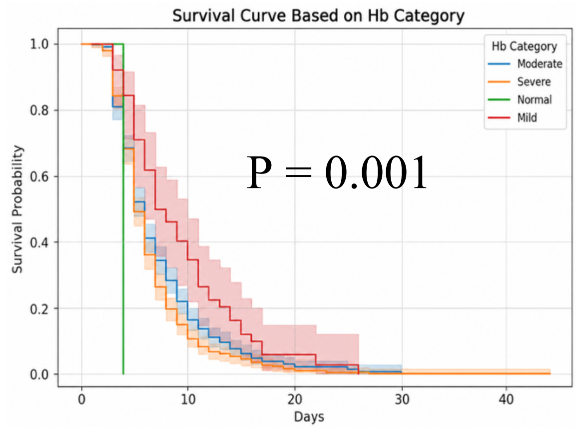

Figure 4. Hb category survival curve

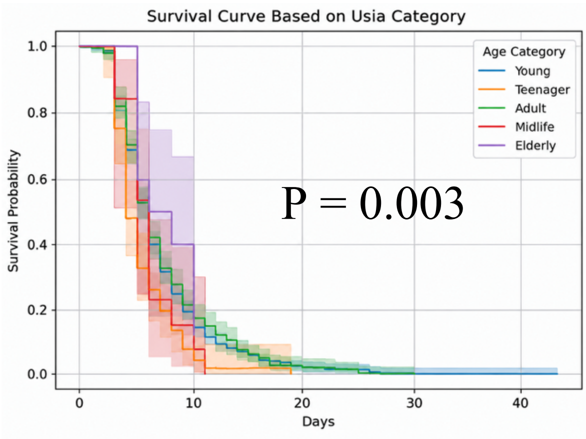

Figure 5. Age survival curve

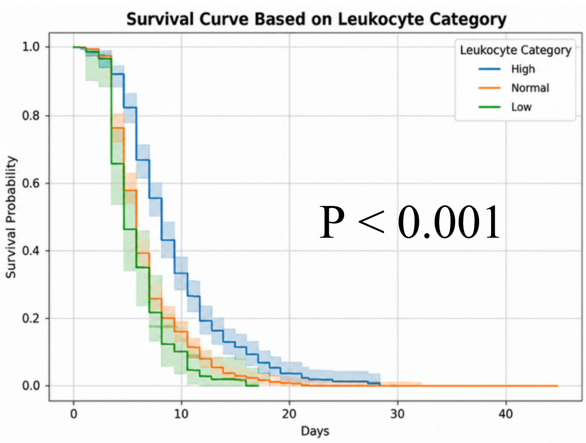

Figure 6. Leukocyte category survival curve

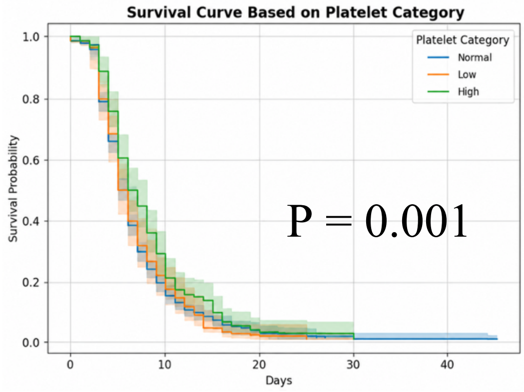

Figure 7. Platelet category survival curve

3.4. *Establishment of the Initial PH Cox Model and Variable Selection*

All numeric variables were entered into the initial Cox Proportional Hazards model. Variable selection was then performed based on the Hazard Ratio (HR), p-value, and Variance Inflation Factor (VIF) to identify suitable variables for use in the Cox PH model. The results of these stages are presented in the table below.

Table 1. Results of the Initial Model and Variable Selection

Variable	HR	P-Value	VIF	Decision
Usia	0.99	< 0.005	12.152	Dipilih
Jenis Kelamin Encode	0.86	0.03	1.616	Dipilih
Hb	0.65	0.03	2091.290	Dipilih
Leukosit	1.00	< 0.005	5.026	Dipilih
Hematokrit	1.14	0.07	3328.788	Tidak Dipilih
Trombosit	1.00	0.14	6.320	Tidak Dipilih
Eritrosit	1.03	0.89	479.327	Tidak Dipilih
MCV	1.01	0.69	1999.153	Tidak Dipilih
MCH	0.98	0.82	1923.011	Tidak Dipilih
MCHC	1.08	0.36	598.865	Tidak Dipilih

Based on Table 1, the variable selection results indicate that age, gender, hemoglobin, and leukocyte count were selected as candidate predictors for the Cox Proportional Hazards model because they showed statistically significant associations with the length of stay among anemia patients ($p < 0.05$). Variable selection also considered multicollinearity using the Variance Inflation Factor (VIF). Several hematological variables, including hemoglobin, hematocrit, erythrocyte count, MCV, MCH, and MCHC, exhibited elevated VIF values because these laboratory parameters represent related hematological characteristics and are biologically correlated in the assessment of anemia. Therefore, the observed multicollinearity among these variables was expected.

Despite the elevated VIF values, Hb was retained because it remained statistically significant throughout the iterative variable selection process and represents the primary laboratory indicator for diagnosing and assessing anemia. Similarly, leukocyte count was retained because of its clinical relevance in reflecting inflammatory responses that may influence recovery during hospitalization. Other hematological variables, including hematocrit, erythrocyte count, MCV, MCH, and MCHC, were excluded because they did not satisfy the predefined variable selection criteria based on statistical significance, overall contribution to the model, and subsequent proportional hazards assumption testing.

Although multicollinearity was observed among several hematological variables, the primary objective of this study was to identify clinically relevant factors associated with the length of hospital stay rather than to develop a purely predictive model. Consequently, statistical significance, multicollinearity diagnostics, clinical relevance, and proportional hazards assumption testing were jointly considered during the iterative variable selection process to obtain the final Cox Proportional Hazards model.

3.5. *Establishing the Cox Proportional Hazards Model, Assumption Testing, Model Evaluation (GOF)*

The variables selected in the previous stage were used to develop a Cox Proportional Hazards model. This model was then evaluated based on the Hazard Ratio (HR) value, proportional hazards assumption testing, and goodness of fit to assess the model's suitability to the data.

Table 2. Cox PH Model Results

Variable	HR	P-Value
Usia	0.99	< 0.005
Jenis Kelamin Encode	0.85	0.02
Hb	0.97	0.12
Leukosit	1.00	< 0.005

Based on Table 2, the variables age, sex, and leukocyte counts showed a significant effect on the length of stay of anemia patients with a p-value < 0.05. Meanwhile, the Hb variable did not show a significant effect, having a p-value of 0.12. HR values for age and sex less than 1 indicate a decreased hazard of an event, while HR values for leukocytes approaching 1 indicate a relatively small effect on the hazard.

Next, the proportional hazards assumption was tested using Schoenfeld residuals. The test results showed that the variables age ($p = 0.651$) and gender ($p = 0.172$) satisfied the proportional hazards assumption because their p-values were greater than 0.05. In contrast, Hb ($p = 0.021$) and leukocyte count ($p = 0.000$) violated the proportional hazards assumption, indicating that further model refinement was required. The goodness-of-fit evaluation showed that this Cox PH model had a log-likelihood value of -5982.938, an AIC value of 11973.875, and a C-index value of 0.637, indicating a relatively good ability to discriminate between patients with different lengths of stay. However, because the proportional hazards assumption was not fully satisfied, variables violating the assumption were iteratively re-evaluated and the model was repeatedly refitted until all retained variables satisfied the proportional hazards assumption. The results of this iterative refinement are presented in the final model.

3.6. *Refinement and Finalization of the Cox Proportional Hazards Model*

This stage was carried out based on the results of the proportional hazards assumption test in the previous model, which indicated that several variables violated the assumption. After iterative refinement and model rebuilding, the final Cox Proportional Hazards model was obtained.

Table 3. Final Model Results

Variable	Coefficient	HR	95% CI	P-Value
Usia	-0.007	0.993	0.989 – 0.996	< 0.001
Jenis Kelamin Encode	-0.205	0.815	0.714 – 0.931	0.003
Hb	-0.054	0.947	0.918 – 0.978	0.001

Based on Table 3, age, gender, and hemoglobin significantly influenced the length of stay among anemia patients ($p < 0.05$). The estimated hazard ratios were accompanied by 95% confidence intervals that did not include 1.00, providing additional support for the statistical significance and precision of the estimated associations. Because hospital discharge was defined as the event of interest, hazard ratios below 1 indicate a lower probability of discharge at any given time and therefore a tendency toward a longer hospital stay.

For age, the hazard ratio of 0.993 indicates that each one-year increase in age was associated with approximately a 0.7% lower hazard of hospital discharge, suggesting that older patients tended to remain hospitalized longer than younger patients. This finding may reflect the greater burden of comorbidities, reduced physiological reserve, impaired immune function and slower recovery commonly observed among older adults. For gender, the hazard ratio of 0.815 indicates that male patients (coded as 1) had an 18.5% lower hazard of hospital discharge than female patients (coded as 0), suggesting a longer length of stay. This difference may be associated with variations in underlying diseases, physiological characteristics, treatment response, or healthcare utilization between male and female patients. However, because information on disease severity, comorbidities, and treatment was unavailable in the present study, the mechanisms underlying this association could not be directly evaluated. For hemoglobin, the hazard ratio of

0.947 indicates that each one-unit increase in hemoglobin level was associated with approximately a 5.3% lower hazard of hospital discharge. Although this direction of association may appear counterintuitive from a clinical perspective, it should be interpreted within the context of the variables included in the model. The observed association may reflect residual confounding because important clinical factors, including anemia severity, comorbidities, transfusion status, treatment procedures, and discharge condition, were unavailable in the electronic medical records used in this study. Therefore, the estimated effect of hemoglobin should be interpreted with caution rather than as evidence of a direct causal relationship.

These findings are generally consistent with previous studies reporting that demographic characteristics and hematological parameters influence hospital length of stay. Kim et al. reported that increasing age was significantly associated with prolonged hospitalization among medical inpatients, while Wang et al. also observed longer hospital stays among older patients with COVID-19. The consistency of these findings suggests that age is an important determinant of recovery across different clinical settings. However, the magnitude of the association observed in the present study was relatively modest, which may be explained by differences in patient populations, disease characteristics, healthcare settings, and the variables included in the Cox Proportional Hazards model. Similarly, previous studies have reported that sex-related differences in hospitalization outcomes may vary according to disease type and patient characteristics, indicating that the observed association between gender and length of stay should be interpreted within the specific clinical context of anemia patients. Furthermore, although hemoglobin is an important indicator of hematological status, its association with length of stay may be influenced by unmeasured clinical factors that were unavailable in this study. The final Cox Proportional Hazards model is expressed as follows.

$$\lambda(t \mid X) = \lambda_0(t)\ \exp(-\,0.007\,Usia - 0.205\,Jenis\ Kelamin\ Encode - 0.054\,Hb) \quad (3)$$

This equation indicates that the hazard of hospital discharge is influenced by age, gender, and hemoglobin level. After the model was iteratively refitted during the refinement process, the estimated effect of hemoglobin became statistically significant, indicating that its association with the length of stay became more stable after accounting for the remaining covariates that satisfied the proportional hazards assumption. The negative regression coefficients indicate that increasing the values of these variables decreases the hazard of discharge during the observation period. After variables that violated the proportional hazards assumption were excluded and the model was iteratively rebuilt, hemoglobin remained statistically significant in the final model. Although hemoglobin exhibited an elevated Variance Inflation Factor (VIF), it was retained because it remained statistically significant and is the primary laboratory parameter used to diagnose and assess anemia severity. The elevated VIF was likely attributable to the strong correlation between hemoglobin and other hematological variables, particularly hematocrit and erythrocyte related indices, which represent related physiological characteristics in anemia assessment. Therefore, variable selection in this study was not based solely on the VIF value but also considered statistical significance, multicollinearity diagnostics, and clinical relevance. These findings indicate that patient characteristics (age and gender), together with hematological status (hemoglobin level), are important factors associated with the length of hospital stay among anemia patients.

The Schoenfeld residual test confirmed that all variables in the final model satisfied the proportional hazards assumption (Hb = 0.255, gender = 0.125, and age = 0.768, $p > 0.05$), indicating that the estimated hazard ratios remained constant throughout the observation period. The model performance evaluation produced a log-likelihood value of -6022.375, an AIC value of 12050.749, and a C-index of 0.565. Although the final model exhibited a lower C-index than the previous model, the earlier model violated the proportional hazards assumption and therefore did not satisfy the fundamental assumption of the Cox Proportional Hazards model. Since the primary objective of this study was to identify factors associated with the length of stay rather than to maximize predictive accuracy, satisfying the proportional hazards assumption was prioritized over achieving a higher C-index. Consequently, the final model was selected because it satisfied the proportional hazards assumption while retaining statistically significant variables, resulting in more reliable hazard ratio estimates for identifying factors associated with the length of stay among anemia patients. Therefore, despite its relatively lower discriminative ability, the final model was considered more appropriate for inferential interpretation than the previous model.

This study was limited to variables available in the hospital electronic medical records. Consequently, several clinically relevant variables, including comorbidities, anemia severity, anemia type, transfusion status, primary diagnosis, treatment procedures, complications, and discharge condition, could not be included in the analysis. The omission of these variables may have introduced potential omitted-variable bias because they are also likely to influence the length of hospital stay. Therefore, the estimated associations in the final model should be interpreted within the context of the available variables. In addition, this study did not perform internal validation using bootstrapping, calibration analysis, or train-test data splitting because the primary objective was to identify factors associated with the length of hospital stay rather than to develop a predictive model. Consequently, the robustness and generalizability of the model should be interpreted with caution. Future studies are encouraged to incorporate additional clinical variables and perform internal and external validation to improve both the comprehensiveness and predictive performance of the survival model.

4. Conclusion

This study applied the Cox Proportional Hazards model to identify factors associated with the length of hospital stay among hospitalized anemia patients using inpatient medical records and clinical pathology laboratory data. The final model identified age, gender, and hemoglobin level as significant factors associated with the hazard of hospital discharge. After iterative model refinement, all retained variables satisfy the proportional hazards assumption. The final model produced a log-likelihood value of -6022.375, an AIC of 12050.749, and a C-index of 0.565. These results indicate that although the model satisfied the proportional hazards assumption, its discriminative ability remained limited. Therefore, the estimated associations should be interpreted within the scope of the available data.

Accordingly, the final model should be interpreted primarily as an explanatory model for identifying factors associated with the length of hospital stay rather than as a clinically validated predictive model for supporting clinical decision-making. This study was also limited by the availability of variables in the hospital electronic medical records and clinical pathology laboratory data, which did not include several clinically relevant factors such as comorbidities, anemia severity, anemia type, transfusion status, treatment procedures, primary diagnosis, complications, and discharge condition. Furthermore, internal validation using bootstrapping, calibration analysis, or train-test data splitting was not performed. Future studies should incorporate more comprehensive clinical variables and perform both internal and external validation to improve the robustness, generalizability, and predictive performance of the model.

References

[1] W. M. Gardner *et al.*, “Prevalence, years lived with disability, and trends in anaemia burden by severity and cause, 1990–2021: findings from the Global Burden of Disease Study 2021,” *Lancet Haematol.*, vol. 10, no. 9, pp. e713–e734, 2023.

[2] W. H. Organization, “Guideline on haemoglobin cutoffs to define anaemia in individuals and populations,” *Geneva: World Health Organization*, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240088542.

[3] C. I. Emanuella Bungasmara Ega Tirta, “19 Juta Warga RI Derita Anemia, Ternyata Dampaknya ke Otak Bikin Ngeri,” *CNBC Indonesia*, 2025. [Online]. Available: https://www.cnbcindonesia.com/research/20250803131307-128-654575/19-juta-warga-ri-derita-anemia-ternyata-dampaknya-ke-otak-bikin-ngeri.

[4] J. Kim, H. Chang, S. Ryu, I. Choi, A. E. Kwon, and H. Ji, “Determinants of Length of Stay for Medical Inpatients Using Survival Analysis,” *Int. J. Environ. Res. Public Health*, vol. 21, no. 11, pp. 1–14, 2024.

[5] Z. Wang *et al.*, “What are the risk factors of hospital length of stay in the novel coronavirus pneumonia (COVID-19) patients ? A survival analysis in southwest China,” *PLoS One*, vol. 17, no. 1, pp. 1–13, 2022.

[6] K. Momeni, M. Raadabadi, J. Sadeghifar, A. Rashidi, Z. Toulideh, and Z. Shoara, “Survival Analysis of Hospital Length of Stay of COVID-19 Patients in Ilam Province , Iran : A Retrospective Cross-Sectional Study,” *J. Clin. Med. Artic.*, vol. 12, no. 20, pp. 1–10, 2023.

[7] Y. Wen *et al.*, “Time-to-event modeling for hospital length of stay prediction for COVID-19 patients,” *Mach. Learn. with Appl.*, vol. 9, no. 100365, pp. 1–8, 2022.

[8] S. W. Lee, “Kaplan-Meier and Cox proportional hazards regression in survival analysis : statistical standard and guideline of Life Cycle Committee,” *Life Cycle*, vol. 3, no. e8, pp. 1–9, 2023.

[9] A. L. P. Hamid, S. Subanti, and Y. Susanti, “Analisis Faktor yang Berpengaruh terhadap Waktu Survival Pasien Penyakit Ginjal Kronis menggunakan Uji Asumsi Proportional Hazard,” *Indones. J. Appl. Stat.*, vol. 5, no. 1, pp. 12–18, 2022.

[10] Z. Nisa, B. Poerwanto, and M. F. Sudding, “Survival Analysis with Cox Proportional Hazard Model for Tuberculosis (TBC) Patients,” *J. Varian*, vol. 7, no. 1, pp. 77–86, 2023.

[11] A. Mufarida, N. Saputra, R. T. Yuriska, and Y. Liaunillah, “Analisis Kurva Survival Kaplan Meier Menggunakan Uji Log Rank Pada Pasien Penyakit Kanker Serviks,” *Indones. Counc. Prem. Stat. Sci.*, vol. 2, no. 1, pp. 14–18, 2023.

[12] S. H. Panjaitan, F. Kedokteran, U. S. Utara, and N. N. Dalimunthe, "Korelasi Skor Get with The Guidelines-Heart Failure dan Rasio Platelet Limfosit pada Pasien Rawat Inap dengan Gagal Jantung Correlation between Get with The Guidelines-Heart Failure Korelasi Skor Get with The Guidelines-Heart Failure dan Rasio Platelet Li," *J. Penyakit Dalam Indones.*, vol. 11, no. 3, pp. 129–133, 2024.

[13] B. K. Kristanto, S. Widyayuningtias, P. Listio, and J. F. Palandi, "Perancangan Integrasi Sistem Informasi Rumah Sakit Kristen Mojowarno Modul Instalasi Gawat Darurat dengan Satu Sehat Kementerian Kesehatan Indonesia Menggunakan Metode Scrum," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 7, no. 6, pp. 2270–2276, 2024.

[14] Kemenkes, "Menkes Apresiasi RSUD Haji Surabaya yang Sukses Integrasikan Layanan ke SatuSehat," 2024. [Online]. Available: https://kemkes.go.id/id/menkes-apresiasi-rsud-haji-surabaya-yang-sukses-integrasikan-layanan-ke-satusehat.

[15] A. A. Putri and D. D. Ananda, "Analysis of Pneumonia Patients Using Kaplan-Meier and Cox Proportional Hazard Regression," *Indones. Counc. Prem. Stat. Sci.*, vol. 4, no. 2, pp. 68–73, 2025.

[16] W. Sanusi, "Analisis Survival Penderita Gagal Ginjal Menggunakan Metode Kaplan Meier (Studi Kasus : Rumah Sakit Labuang Baji Makassar)," *J. Mat. dan Stat. serta Apl.*, vol. 13, no. 1, pp. 15–21, 2025.

[17] M. A. Khikman *et al.*, "Survival Analysis Using Kaplan-Meier and Cox Regression in Hypertension Patients at Kefamenanu Regional Hospital," *Eig. Math. J.*, vol. 8, no. 2, pp. 123–132, 2025.

[18] M. Friska, "Lama Mencari Kerja di Indonesia dengan Menggunakan Analisis Survival," *J. Ekon. dan Stat. Indones.*, vol. 1, no. 3, pp. 276–283, 2021.

[19] T. Qomaria, M. Fatekurohman, and J. Matematika, "Aplikasi Model Cox Proportional Hazard pada Pasien Stroke RSD Balung Kabupaten Jember," *Indones. J. Appl. Stat.*, vol. 2, no. 2, pp. 94–112, 2019.

[20] L. E. Widyawati and R. D. Bekti, "ANALISIS FAKTOR YANG MEMPENGARUHI LAJU KESEMBUHAN PASIEN DEMAM BERDARAH DENGUE MENGGUNAKAN REGRESI COX PROPORTIONAL HAZARD DAN REGRESI WEIBULL," *J. Stat. Ind. dan Komputasi*, vol. 05, no. 2, pp. 26–36, 2020.

[21] N. A. Rabukuntari, B. P. Maulidina, and L. A. Hasanah, "Pengaruh Usia dan Jenis Kelamin Terhadap Selera Menonton Tayangan Berita Menggunakan Metode Regresi Logistik Ordinal," *JAIIT(Journal ofAdvances Inf. Ind. Technol.*, vol. 4, no. 1, pp. 11–22, 2022.

[22] L. Firdausiah, M. Fatekurohman, and I. H. Agustin, "Model CPH untuk Menganalisis Lama Pengobatan dan Tingkat Kesembuhan Pengguna Narkoba," *J. Ris. Stat.*, vol. 5, no. 1, pp. 11–20, 2025.

[23] A. R. Faisal, M. N. Bustan, and S. Annas, "Analisis Survival dengan Pemodelan Regresi Cox Proportional Hazard Menggunakan Pendekatan Bayesian (Studi Kasus : Pasien Rawat Inap Penderita Demam Tifoid di RSUD Haji Makassar)," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 2, no. 2, pp. 62–70, 2020.

[24] I. Z. A. Illah, W. S. J. Sapu, and A. T. Damaliana, "Implementasi Metode Klasifikasi LightGBM dan Analisis Survival dalam Memprediksi Pelanggan Churn," *J. Komtika (Komputasi dan Inform.*, vol. 8, no. 1, pp. 43–53, 2024.

[25] A. K. P. Isyanto, Trimono, and A. T. Damaliana, "Multinomial Logistic Application on Factors Affecting Poor Population in East Java," *Bit-Tec (Binary Digit.*, vol. 8, no. 1, pp. 96–106, 2025.

[26] D. M. Manurung, A. T. Damaliana, and D. A. Prasetya, "Application Of Hybrid ARIMAX-ANN In Forecasting The Price Of Chili Bird ' s Eye," *Bit-Tec (Binary Digit.*, vol. 8, no. 2, pp. 1761–1770, 2025.

[27] T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and M. H. P. Swari, "Cluster Analysis of Hospital Inpatient Service Efficiency Based on BOR , BTO , TOI , AvLOS Indicators using Agglomerative Hierarchical Clustering," *Telemat. J. Inform. dan Teknol. Inf.*, vol. 18, no. 2, pp. 194–210, 2021.

[28] S. S. M. Wara, M. Nasrudin, A. F. Adziima, and A. R. Pratama, "Optimasi Sistem Antrian Pada Medical Center Its Dengan Simulasi Discrete Event Dan Response Surface Methodology," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 99, no. 99, pp. 1–8, 2025.