Modeling Survival Time in Children Undergoing Bone Marrow Transplantation

Sathvik Kommireddy
Rancho Cucamonga High School,
Rancho Cucamonga, CA

Abstract

This study models survival times in pediatric patients following bone marrow transplantation by comparing several modeling approaches. We analyze time-to-event data using the semi-parametric Cox proportional hazards model, as well as fully parametric exponential, log-logistic, and generalized gamma models. The model performance is assessed using the Akaike information criterion and the concordance index.

Keywords: bone marrow transplant, survival analysis, Kaplan-Meier survival curve, Cox proportional hazards model, exponential survival regression model, log-logistic regression model, generalized gamma regression model, Akaike information criterion, concordance index

1 Introduction

Bone marrow transplantation is a versatile and often life-saving treatment for children with a wide range of serious diseases. Bone marrow cells, also known as stem cells, are immature blood cells capable of developing into any of the major blood cell types: red blood cells, white blood cells, and platelets. The primary types of transplants include autologous transplantation, in which patients receive stem cells collected from their own bodies; allogeneic transplantation, where stem cells are obtained from a donor; and umbilical cord blood transplantation, which uses stem cells harvested from the umbilical cord immediately after birth. The overarching goal of bone marrow transplantation is to replace diseased or damaged marrow, thereby treating or curing otherwise life-threatening conditions, including certain cancers. Although the procedure carries risks—such as infection, low platelet counts, pain, diarrhea, nausea, and vomiting—these risks are often outweighed by the severity of the underlying diseases for which the transplant is indicated.

Research on bone marrow transplantation began in the late 1940s, motivated in part by the urgent need to understand the effects of high levels of radiation exposure in humans. Early experiments in mice revealed that injections of bone marrow could restore blood cell production and protect against radiation, a groundbreaking discovery at the time. Translating these findings to humans, however, proved far more challenging, and initial attempts at human transplants frequently failed, causing many researchers to abandon the field. Progress resumed with successful transplant experiments in dogs, paving the way for controlled human trials. Early clinical results were promising: some patients achieved long-term survival rates of roughly 45%, a remarkable outcome given the complexity of the procedure. Today, advances in conditioning regimens, donor matching, and supportive care have dramatically improved outcomes, with long-term survival rates reaching around 80% for many pediatric conditions.[1,2] This history reflects the remarkable evolution of bone marrow transplantation from experimental animal studies to a life-saving therapy for children worldwide.

1.1 Literature Review

In the context of bone marrow transplantation, several studies have demonstrated the effectiveness of survival modeling techniques, both traditional and machine learning-based, for predicting patient outcomes.

The Cox proportional hazards model remains the standard approach in survival analysis. For example, Bhatia et al. (2021) employed the Cox model to predict survival over four decades for patients who received allogeneic blood or marrow transplants.[3] Such studies highlight the robustness and interpretability of the Cox model in analyzing long-term survival outcomes.

In addition to the semi-parametric Cox model, fully parametric regression models are commonly used in survival analysis. These models have shown particular promise in transplant research. Sayemiri et al. (2009) applied parametric models to investigate predictor variables influencing survival time after stem cell transplantation and demonstrated improved estimation through maximum likelihood methods.[4] In the present study, we use the Cox model along with exponential, log-logistic, and generalized gamma parametric regressions to compare their performance and goodness-of-fit within our dataset of pediatric bone marrow transplant patients.

Taken together, the literature demonstrates that both parametric and semi-parametric models provide valuable insights into post-transplant outcomes. This study builds on prior work by comparing the performance of Cox and parametric models within a unified framework, thereby providing a comprehensive assessment of predictive methods for pediatric bone marrow transplant survival.

1.2 Data Description and Preprocessing

The dataset used in this study, titled *Bone Marrow Transplant: Children*, was obtained from Kaggle. It includes clinical and demographic data for 187 pediatric patients who underwent unmanipulated allogeneic unrelated donor hematopoietic stem cell transplantation. The patients were treated for various hematologic disorders, including leukemias, myelodysplastic syndrome, aplastic and Fanconi anemia, and X-linked adrenoleukodystrophy.

The original motivation for collecting this dataset, as described by Kałwak et al. (2010), was to identify the key factors that influence the success or failure of pediatric stem cell transplantation. In particular, their study tested the hypothesis that a higher dose of $CD34^+$ cells per kilogram of recipient body weight improves long-term survival without increasing the risk of severe graft-versus-host disease or other adverse outcomes.[5] This dataset has since been used in several machine learning and rule-based modeling studies on survival rule induction (Wróbel et al., 2017; Sikora et al., 2019; Gudyś et al., 2020).[6, 7, 8]

Overall, the dataset contains 37 predictor variables encompassing donor and recipient characteristics, transplant parameters, immunologic compatibility, and post-transplant outcomes. A full descriptor file, the original dataset, and the modified dataset can be found in this project's GitHub. In preparing the data for analysis, we removed all columns and rows with missing values, as well as derived variables, reducing the original 37 variables to 22. Based on the computed Spearman correlation values, we selected five variables for detailed analysis due to their known or hypothesized relevance to post-transplant survival outcomes. The primary outcome variables are survival time and survival status. A detailed table (Table A) of variable descriptions and distributions is provided in the appendix.

2 Methodology: Modeling Survival Functions with Predictors

2.1 The Kaplan-Meier Survival Curve

Let T be a non-negative random variable representing the time until an event of interest occurs. The survival function $S(t)$ is defined as

$$S(t) = P(T > t).$$

Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ be the ordered observed survival times, d_i the number of deaths at time $t_{(i)}$, and n_i the number at risk just before $t_{(i)}$. The Kaplan–Meier estimator of the survival function is

$$\hat{S}(t) = \prod_{t_{(i)} \le t} \left(1 - \frac{d_i}{n_i}\right), \quad t > 0.$$

The Kaplan–Meier survival curve is a plot of this estimator, represented as a step-wise decreasing function with jumps at observed event times.[9]

2.2 Cox Proportional Hazards Model

Let T be a survival time, and let $x_1, \ldots, x_p$ be the predictors. The Cox model specifies the survival function in the form

$$S(t, x_1, \ldots, x_p) \;=\; \Big[S_0(t)\Big]^r, \;\; t > 0,$$

where $S_0(t)$ is the *baseline survival function* corresponding to an often hypothetical baseline individual with zero predictors, and $r = \exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}$ is the *relative risk.*[9] Alternatively, the model may be formulated in terms of the hazard function

$$h(t) = \frac{f(t)}{S(t)}, \quad t > 0,$$

which represents the instantaneous event rate given survival up to time t. In the Cox model,

$$h(t) = h_0(t)\, r, \;\; t > 0,$$

where $h_0(t)$ is the baseline hazard function. For any two individuals, the ratio of their hazard functions depends only on their covariates and not on time. Hence the name *proportional hazards.*

The estimated regression coefficients admit the following interpretation. When x_1 is continuous and increases by one unit, the estimated hazard changes by $\big(\exp\{\hat{\beta}_1\} - 1\big) \cdot 100\%$, holding all other predictors fixed. If x_1 is a binary (0–1) variable, then $\exp\{\hat{\beta}_1\} \cdot 100\%$ represents the estimated percent hazard ratio comparing individuals with $x_1 = 1$ to those with $x_1 = 0$, controlling for the other predictors.

2.3 Exponential Model

The exponential survival model provides a parametric estimator of the survival function. It is assumed that the time to event T follows an exponential distribution with density

$$f(t) = \lambda \exp(-\lambda t), \;\; t > 0. \quad [9]$$

Thus, the survival function of T is

$$S(t) = \exp(-\lambda t), \;\; t > 0,$$

where

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p).$$

In this model, the hazard function is independent of time, since

$$h(t) = \frac{f(t)}{S(t)} = \lambda.$$

The interpretation of the estimated coefficients is analogous to that in the Cox model, expressed in terms of the hazard function.

2.4 Log-Logistic Survival Regression Model

The log-logistic survival regression model is a parametric model for time-to-event data in which the logarithm of the survival time T follows a logistic distribution. The survival function of T is

$$S(t) = \frac{1}{1 + \exp\left(\frac{\log t - \mu}{\sigma}\right)}, \quad t > 0,$$

where the location parameter

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

is modeled as a linear function of the predictors, and the scale parameter $\sigma > 0$.[9]
In contrast to the Cox and exponential models, which satisfy the proportional hazards property, the log-logistic model satisfies the *proportional odds* property. The odds of failing by time t are

$$\frac{1 - S(t)}{S(t)} = \exp\left(\frac{\log t - \mu}{\sigma}\right) = t^{1/\sigma}\, e^{-\mu/\sigma},$$

which are proportional across individuals. That is, the ratio of these odds for any two individuals is independent of time.

The interpretation of the estimated regression coefficients is as follows. For a continuous predictor x_1, the quantity

$$\left(\exp\{-\hat{\beta}_1/\hat{\sigma}\} - 1\right) \cdot 100\%$$

represents the percent change in the estimated odds of failing by time t when x_1 increases by one unit, holding all other predictors fixed. If x_1 is an indicator variable, then $\exp(-\hat{\beta}_1/\hat{\sigma}) \cdot 100\%$ gives the percent factor of the estimated odds comparing individuals with $x_1 = 1$ to those with $x_1 = 0$ (the reference group), again holding all other predictors constant.

2.5 Generalized Gamma Survival Regression Model

In the generalized gamma survival regression model, the probability density function of T can be written as

$$f(t) = \frac{\lambda\,\alpha\,(\lambda\,t)^{p\,\alpha-1}}{\Gamma(k)}\,e^{-(\lambda\,t)^{\alpha}},\ \ t>0,$$

where the scale parameter λ is modeled as a function of predictors as

$$\lambda = \exp\{\beta_0 + \beta_1\,x_1 + \cdots + \beta_p\,x_p\},$$

and $k>0$ and $\alpha>0$ are shape parameters. This distribution includes the following special cases: gamma when $\alpha=1$, exponential when $\alpha=1$ and $k=1$, and log-normal as a limiting distribution when $k\to\infty$.[9]

The survival function of T is derived as

$$S(t)\ =\ \int_t^{\infty}\ \frac{\lambda\,\alpha\,(\lambda\,u)^{p\,k-1}}{\Gamma(k)}\,e^{-(\lambda\,u)^{\alpha}}\,du.$$

Substituting $y=(\lambda\,u)^{\alpha}$ with $dy=\lambda\,\alpha\,(\lambda\,u)^{\alpha-1}du$, we arrive at

$$S(t)\ =\ \frac{1}{\Gamma(k)}\int_{(\lambda\,t)^{\alpha}}^{\infty}\ y^{k-1}\,e^{-y}\,dy\ =\ \frac{\Gamma\big(k,(\lambda\,t)^{\alpha}\big)}{\Gamma(k)},\ \ t>0.$$

The generalized gamma model satisfies the *accelerated failure time* assumption, which implies that the times to event for two individuals are proportional. In other words, the survival experience of one individual can be viewed as a time-scaled version of the other's. Formally, there exists a constant θ, independent of time, such that the survival function for one individual, $S(t)$, can be expressed as the survival function for the other individual evaluated at θt. Specifically,

$$S(t,\lambda)\ =\ \frac{1}{\Gamma(k)}\int_{(\lambda_1 t)^{\alpha}}^{\infty}\ y^{k-1}e^{-y}\,dy\ =\ \frac{1}{\Gamma(k)}\int_{(\lambda_2\theta t)^{\alpha}}^{\infty}\ y^{k-1}e^{-y}\,dy\ =\ S(\theta t,\lambda_2)$$

where $\theta=\lambda_1/\lambda_2$, and λ_1 and λ_2 are the scale parameters for the two individuals. The parameter θ is termed the *acceleration factor* since when $\theta>1$, the survival time is stretched or decelerated (longer survival), whereas when $\theta<1$, the survival time is compressed or accelerated (shorter survival).

The interpretation of the estimated regression coefficients is as follows. If x_1 is continuous, then for a one-unit increase in x_1, the estimated acceleration factor changes by $\big(\exp\{-\hat{\beta}_1\}-1\big)\cdot 100\%$, keeping the other predictors constant. If x_1 is a 0-1 variable, then $\exp\big(-\hat{\beta}_1\big)\cdot 100\%$ measures the percent ratio of the estimated acceleration factors for individuals with $x_1=1$ to those with $x_1=0$, again controlling for the other predictors.

2.6 Performance Measures

Model performance in survival analysis can be assessed using both goodness-of-fit and discrimination measures. For parametric survival models, the *Akaike Information Criterion* (AIC) is commonly used to compare model performance. It is defined as

$$\mathrm{AIC} = -2\ln(\hat{L}) + 2p,$$

where $\hat{L}$ is the maximized likelihood function, and p denotes the total number of estimated parameters of the model. Lower AIC values indicate better model fit.[10]

However, AIC values are only meaningful when models are fitted via full likelihood maximization. Because the Cox proportional hazards model uses a partial likelihood rather than a full likelihood, its AIC is computed on a different scale. As a result, the AIC for a Cox model cannot be directly compared with the AIC from fully parametric models, and should only be compared among Cox models themselves.

For models that do not rely on full likelihood maximization—such as the Cox model—model performance is more appropriately evaluated using the concordance index (c-index).[10] The c-index measures predictive discrimination: how well the model distinguishes between individuals with different survival times. It represents the proportion of all usable subject pairs for which the model correctly predicts which individual experiences the event earlier. A c-index of 0.5 indicates performance no better than chance, whereas a value of 1 reflects perfect discrimination. Because the c-index is scale-free and does not depend on how the underlying risk is parameterized, it allows for direct comparison across Cox and parametric survival analysis models.

3 Applications and Results

First, we estimate the survival function of the survival times following bone marrow stem cell surgery using the Kaplan–Meier estimator and plot the corresponding survival curve (see Figure 1). The curve exhibits an approximately exponential decay, suggesting that an exponential survival model may provide a good fit.

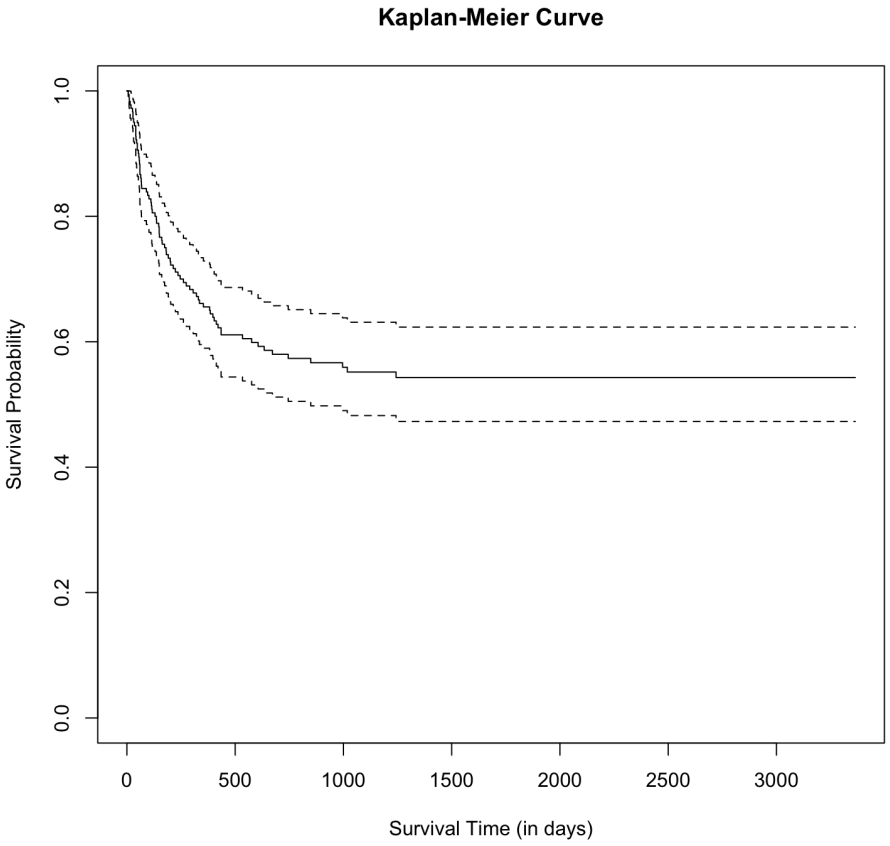


Figure 1: Kaplan–Meier survival curve.

3.1 Comparison of Model Fit

Next, we fit the Cox proportional hazards model, and the three parametric regressions (exponential, log-logistic, and generalized gamma). The corresponding model performance measures are summarized in Table 1.

Table 1. Model Performance Measures.

Model Name	**C-index**	**AIC**
Cox Model	0.3283160	NA
Exponential	0.3293195	1337.411
Log-Logistic Model	0.6722313	1286.734
Generalized Gamma Model	0.6712279	1284.124

We can see that, of the three parametric models, the generalized gamma provides the best fit, as its AIC is the lowest, with almost the highest c-index, suggesting that the data may

exhibit the accelerated time property. It significantly outperforms the Cox and exponential models, suggesting that the data exhibit properties such as non-constant hazard rates and departures from strict proportional hazards.

3.2 Results of the Fitted Generalized Gamma Model

The estimated parameters of the generalized gamma regression model are given in Table 2 along with the corresponding p-values for testing equality to zero, when appropriate.

Table 2. Parameter estimators and p-values for generalized gamma regression.

Parameter	Estimator	p-value
Intercept	8.480181	NA
relapse	- 1.145374	0.0032
acuteGvHD	- 1.354907	0.0096
rh	- 1.058534	0.022
bodymass	- 0.026892	0.956
CD34Dosage	0.049443	0.998
k	1.63	NA
α	0.29	NA

The fitted generalized gamma survival model has the form

$$\hat{S}(t) = \frac{\Gamma\big(\hat{k}, (\hat{\lambda}\, t)^{\hat{\alpha}}\big)}{\Gamma(\hat{k})}, \quad t > 0,$$

with $\hat{\lambda} = \exp\big\{8.480181 - 0.026892\,\texttt{bodymass} - 1.058534\,\texttt{rh} - 1.354907\,\texttt{acuteGvHD} - 1.145374\,\texttt{relapse} + 0.049443\,\texttt{CD34Dosage}\big\}$, and the estimated shape parameters $\hat{k} = 1.63$ and $\hat{\alpha} = 0.29$.

Because the generalized gamma model performed best among all candidates, our results suggest that the data are more consistent with an accelerated failure time structure rather than a proportional hazards or proportional odds framework. Analysis of the generalized gamma model indicates that post-transplant complications, particularly relapse and the development of acute graft-versus-host disease (GvHD), are the strongest predictors of reduced survival time, with p-values of 0.0032 and 0.0096, respectively.

Recipients who experience relapse are estimated to have a survival time equal to $\exp(-1.145374) \cdot 100\% = 31.81\%$ of that of recipients without relapse, underscoring the critical importance of preventing disease recurrence. Similarly, recipients who develop severe

acute GvHD are estimated to have a survival time of $\exp(-1.354907) \cdot 100\% = 25.80\%$ of that of recipients who do not develop acute GvHD, highlighting its profound adverse impact on patient outcomes.

Survival time for rhesus-positive recipients is estimated to be $\exp(-1.058534) \cdot 100\% = 34.70\%$ of that for rhesus-negative recipients. With a p-value of 0.022, this effect is statistically significant, although less pronounced than those associated with relapse and acute GvHD.

In addition, the fitted model suggests that a one-kilogram increase in body mass is associated with an estimated $|(\exp(-0.026892) - 1) \cdot 100\%| = 2.65\%$ decrease in expected survival time. Conversely, a one-unit increase in CD34 dosage is associated with an estimated $(\exp(0.049443) - 1) \cdot 100\% = 5.07\%$ increase in survival time. However, neither body mass nor CD34 dosage is a statistically significant predictor, with corresponding p-values of 0.956 and 0.998, respectively.

3.3 Discussion

These results are consistent with previously published work on parametric and semi-parametric survival models. In a 2010 study, Sayemiri et al. applied similar analytical methods to a comparable dataset of leukemia patients undergoing hematopoietic stem cell transplantation and found that the generalized gamma model provided the best overall fit across multiple data characteristics.[4] Consistent with their findings, our results suggest that the dataset exhibits accelerated failure time properties rather than proportional hazards or proportional odbds behavior.

Moreover, the significance of key covariates aligns across studies, with post-transplant complications such as acute graft-versus-host disease (acute GvHD) and relapse emerging as dominant predictors of survival. Overall, the concordance between our findings and established research underscores the utility of parametric models—particularly the generalized gamma model—in effectively characterizing and predicting survival outcomes in hematopoietic stem cell transplant patients.

4 Conclusion and Future Directions

After eliminating highly correlated or redundant covariates, we applied a Cox proportional hazards model along with three parametric regression models to analyze post–bone marrow transplant survival. Among these approaches, the generalized gamma model demonstrated the best overall fit. Relapse, acute GvHD grades III–IV, and recipient Rh status emerged as the most influential predictors of survival.

Future work may focus on the application of machine learning methods specifically designed for time-to-event data, such as random survival forests, gradient boosting models for survival, and survival support vector machines, to this dataset and to other disease domains. These approaches may further improve predictive performance while allowing for the assessment of robustness, interpretability, and clinical utility in real-world medical decision-making.

Supplemental Materials

The dataset, R code, and all relevant outputs are available in the project's GitHub repository (https://github.com/sathvik-kommireddy/Bone-Marrow-Transplant-Analysis) for reproducibility and further exploration.

Acknowledgments

I would like to sincerely thank Dr. Olga Korosteleva, a professor at California State University, Long Beach, for her invaluable guidance and support throughout this project. I would also like to thank Dr. Oleg Gleiser and the UCLA Math Circle program for the incredible opportunities they have presented. I am also grateful to my dedicated math teachers, Mr. Jared Derksen and Mr. Dave Oberhauser, for their encouragement and assistance. Finally, I wish to express my deepest appreciation to my family members for their wholehearted support and motivation.

References

[1] Granot, N. and R. Storb. (2020). History of hematopoietic cell transplantation: Challenges and progress. *Haematologica*, 105(12), 2716–2729.

[2] Martin, P. J., et al. (2010). Life expectancy in patients surviving more than 5 years after hematopoietic cell transplantation. *Journal of Clinical Oncology*, 28(6), 1011–1016.

[3] Bhatia S., et al. (2021). Trends in Late Mortality and Life Expectancy After Allogeneic Blood or Marrow Transplantation Over 4 Decades: A Blood or Marrow Transplant Survivor Study Report. *JAMA Oncology*, 7(11), 1626–1634.

[4] Sayemiri, K., et al. (2010). Predictive factors of survival time after hematopoietic stem cell transplant in acute myeloid leukemia patients who received allogeneic BMT from

matched sibling donors using generalized gamma models. *International Journal of Hematology-Oncology and Stem Cell Research*, 3(1), 21–26.

[5] Kałwak, K., et al. (2010). Higher CD34+ and CD3+ cell doses in the graft promote long-term survival, and have no impact on the incidence of severe acute or chronic graft-versus-host disease after in vivo T cell-depleted unrelated donor hematopoietic stem cell transplantation in children, *Biology of Blood and Marrow Transplantation*, 16(10): 1388-1401.

[6] Wróbel, Ł., Gudyś, A., and M. Sikora. (2017). Learning rule sets from survival data, *BMC Bioinformatics*, 18(1): 285.

[7] Sikora, M., Wróbel, Ł., and A. Gudyś. (2019). GuideR: a guided separate-and-conquer rule learning in classification, regression, and survival settings, *Knowledge-Based Systems*, 173: 1-14.

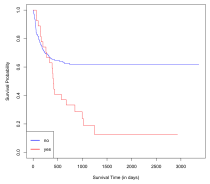
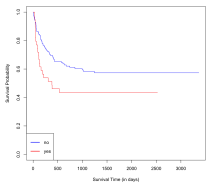
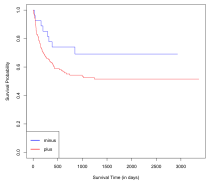
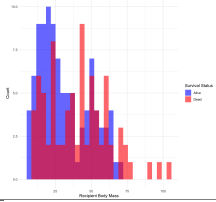
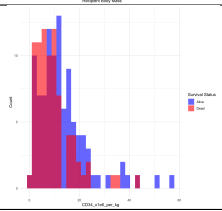
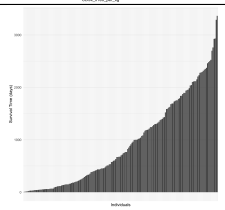
[8] Gudyś, A., Sikora, M., and Ł. Wróbel. (2020). RuleKit: A Comprehensive suite for rule-based learning, *Knowledge-Based Systems*, 194: 105480.

[9] Kleinbaum, D. and M. Klein. (2012). *Survival Analysis: A Self-Learning Text*, Springer Science + Business Media/LLC.

[10] Collett, D. (2023). *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC.

Appendix

Table A. Variable Description

Variable	Description	Distribution
relapse	Binary variable indicating whether the patient's underlying disease recurred after transplantation.	
acuteGvHD	Acute Graft-versus-Host Disease (GvHD) grades III–IV. Binary variable indicating whether the patient developed severe acute GvHD.	
rh	Presence or absence of the Rh (Rhesus) factor, a protein found on the surface of red blood cells, in the recipient's blood (Rh-positive or Rh-negative).	
bodymass	Recipient's body mass of the pediatric patient at the time of transplantation, measured in kg.	
CD34 Dosage	CD34 $\times 10^6$/kg: dose of hematopoietic stem cells marked by the CD34 protein.	
Survival Time	Time-to-event variable representing time to death or follow-up duration for surviving patients.	
Survival Status	Binary indicator of patient status at the end of observation (0 = alive, 1 = deceased).