Reduction of Number of Predictors using Correlation Techniques for Estimation of Survival Time: An Application on Acute Lymphoblastic Leukemia Patients

Anurag Sharma[1] and Deepak Kumar[2]*

Abstract

In this research paper, we have undertaken an effort to minimize the quantity of predictors. This was achieved by identifying the predictors exhibiting strong positive correlations and subsequently assessing the impact of this correlation or association on the follow-up time of patients with acute lymphoblastic leukemia, utilizing Accelerated Failure Time (AFT) models. We conducted a one-way ANOVA to examine the connection between a quantitative and a categorical variable, as well as a Chi-square test to evaluate the association between two categorical variables. To choose between two highly correlated or associated predictors, we employed a method where one predictor was treated as independent while the other was considered dependent in separate model fits. The model with the smaller Akaike Information Criterion (AIC) was chosen, and the independent variable from that model was included in the reduced model. Subsequently, we compared the complete model (containing all predictors) and the reduced model (with fewer predictors) based on their AIC values, selecting the model with the lowest AIC. Through this approach, we were able to reduce the number of predictors by nearly 50% without impacting the estimated follow-up time and maintaining a reduced standard error. This pioneering study marks the first instance of such an analysis using data from acute lymphoblastic leukemia patients.

Key words: Acute lymphoblastic leukemia; AFT, Correlation; Chi- Square test; One-Way ANOVA.

Introduction

Acute lymphoblastic leukemia is a cancer of the lymphoid line of the blood cells and is caused due to development of the large number of immature lymphocytes also known as "Lymphoblasts" (Board, 2023). Normally lymphoblasts are found in the bone marrow, but in acute lymphoblastic leukemia (ALL), these lymphoblasts increase rapidly and as a result are found in the peripheral blood. Their size lies between 10 and 20 µm (Rozenberg, 2011). These excessive immature lymphocytes in the bone marrow interfere with the production of new RBCs, WBCs and platelets. Common symptoms of ALL include feeling tired, pale skin color, fever, easy bleeding or bruising, enlarged lymph nodes, or bone pain. ALL spreads rapidly and is lethal if not treated properly or left untreated within weeks or months (Vos et al., 2016). Reasons for ALL are very difficult to find in most of the cases. There may be a number of risk factors associated with survival of ALL. Some of the risk factors namely, Down syndrome, Li-Fraumeni syndrome, or neurofibromatosis type 1 may be included in the genetic risk factors and significant radiation exposure may be considered as an environmental risk factor. Around 8.76 million people were found to be diagnosed with ALL in 2015 globally. Also, it results in over 1.11 million deaths all around the world (Vos et al., 2016; Wang et al., 2016). It occurs most commonly in children, particularly in children between the ages of two and five (SEER, 2023; Heroto et al., 2013).

In India only, every year nearly 25,000 children are diagnosed with this cancer (SEER 2023; Berger and Fuhrman, 2006).

* *Corresponding Author*
[1] *Assistant Professor, Ram Lal Anand College, University of Delhi, Delhi, India*
[2] *Research Scholar, Singhania University, Rajasthan, India. Email Id: deepak.kumar.stats@gmail.com*

Diagnosis of ALL is generally based on the blood tests and bone marrow examination. Lumbar puncture is used to determine whether the spinal column and brain have been invaded. Central nervous system (CNS) directed therapy is a fundamental component in the management of pediatric acute lymphoblastic leukemia (ALL). Immunophenotyping (IPT) is a laboratory test which is used to identify proteins that are expressed on their cell surface. It is a key component in the diagnosis of ALL (Brown, 2013). The initial choice of treatment for ALL is chemotherapy. Most of ALL patients receive a combination of medications.

Several studies have been performed to find the prognostic factors by using nonparametric survival methods. Le QH et al. (2006) used initial (during the induction period) and late (consolidation period during the post induction treatment) prognostic factors to predict survival in adult acute lymphoblastic leukemia. They assessed factors which were able to predict overall survival in adult patients with acute lymphoblastic leukemia according to the period since initiation of the treatment using a Cox proportional hazards model (Le et al. 2006). Locatelli et al. (1995) investigated the role of other variables on the probabilities of relapse, TRM and event-free survival (EFS). They compared the results obtained in 26 children given HSCT (Hematopoietic stem cell transplantation) before January 1998 with those of 37 patients transplanted beyond that date (Uderzo et al. 1995). Bonnet M et al. (1980) studied the prognostic factors to the treatment of childhood (less than 20 years old) acute lymphoblastic leukemia (Jacquillat et al., 1980). Sayehmiri et al. (2008) evaluated the impact of prognostic factors of overall survival (OS) after hematopoietic stem cell transplant (HSCT) in acute lymphoblastic leukemia (ALL) patients using accelerated failure time (AFTM), Cox proportional hazard (PH), and Cox time-varying coefficient models (Sayehmiri et al. 2008). Howard et al. (2002) determined the risk factors for traumatic and bloody Lumbar Puncture in Children with Acute Lymphoblastic Leukemia. In this study all patients underwent a diagnostic LP followed by a median of 4 LPs to instill intrathecal chemotherapy (Howard et al., 2002). Sung et al. (2014) evaluated the risk factors not examined in the study by Howard et al. including the effects of older age, body mass index (BMI) percentile, treatment with anticoagulation and the use of image-guidance (Shaikh et al., 2014). Totadri et al. (2015) evaluated the impact of traumatic lumbar puncture (TLP) at diagnosis of relapse in childhood acute lymphoblastic leukemia (ALL). Risk factors associated with TLP were also assessed (Totadri et al., 2015).

In all the studies that have been conducted till now, all the available risk factors or the predictors have been considered which may result in the larger standard error of the model. In this paper, we have proposed a novel correlation screening procedure for reducing the number of predictors/ risk factors to be included in the model without effecting the estimated survival time with lower standard error.

Methods

Data

Data on 107 patients suffering from acute lymphoblastic leukemia patients were obtained from a hospital in Delhi, India. Predictors such as *Age, Sex, BMI, Obese, bulky and duration of symptoms*, platelets at the time of *lumbar puncture* and days in which patients were in lumbar puncture to find a cause for symptoms were recorded. Also, total lymphocyte count was recorded for each patient. The TLC level (0- less than

50000 lac/cumm TLCs in a patient, 1-greater than 50,000 but less than 99000 lac /cumm, ,2- for TLCs greater than 1 lac/cumm) was also recorded. Days of *delay/ interruptions* in chemotherapy due to toxicity such as neutropenia or severe infections were also noted. Also, whether patient skipped the therapy or not were also recorded. *Immunophenotyping* is used as a diagnostic tool. It allows a proper definition of hematological malignancies' lineage and differentiation. The following entities were identified through immunophenotyping (0-B lineage acute lymphoblastic leukemia, 1-T lineage acute lymphoblastic leukemia). Then blasts/excessive immature lymphocytes growth occurred at the time of LP were recorded and categorized (0-No blasts, 1-Excessive growth of immature lymphocytes growth). Patients were then given combination of steroids or leunase or both depending on the condition of patients. Augmented and standard Berlin-Frankfurt-Munster chemotherapy were used for treatment. CNS status is also recorded. It is defined as :- CNS 1 – puncture not traumatic (< 10 red blood cells per µL) and no identifiable leukemic blast cells after cytocentrifugation, CNS 2 – puncture not traumatic (< 10 red blood cells per µL), < 5 white blood cells (WBCs) per µL with leukemic blast cells after cytocentrifugation; Negative traumatic lumbar puncture (TLP) – puncture traumatic (≥ 10 red blood cells per µL) with no leukemic blast cells after cytocentrifugation; Positive TLP – puncture traumatic (≥ 10 red blood cells per µL) with leukemic blast cells after cytocentrifugation. Follow time up from diagnosis to the time of last contact is considered as the survival time for patients.

Methods

Suppose one aim to predict the follow-up time of Acute Lymphoblastic Leukemia (ALL) patients using an Accelerated Failure Time (AFT) model while considering all potential predictors. We can denote the survival time of the i^{th} patient as a random variable, T_i, then

$$\log(T_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots \beta_n x_{ni} + \sigma\varepsilon_i \qquad (1)$$

Where, β_0 is the intercept, $\beta_i's$ are the coefficients of "n" explanatory variables for i^{th} patient., σ is the scale parameter ε_i is a random variable used to model the deviation of values of $log_e(T_i)$ from the linear part of model.

The parameters of AFT model are estimated by the maximum likelihood estimation method and by using Newton- Raphson procedure.

Correlation between two variables measures the extent of their linear relationship. When a significant correlation exists between two variable pairs, it implies that instead of including both variables in a model, including just one of them would suffice. For instance, in Model (1), if X_1 *and* X_2 exhibit significant correlation or association, retaining only one of them in the model will serve the purpose of representing both variables. So, if we assume that among the "n" independent variables, "m" pairs of predictors are correlated or associated, then, from these "m" pairs, only one variable from each pair can be included in the model. Consequently, this reduces the number of predictors in the model to "n-m". The reduced model will then be given by;

$$\log(T_i) = \beta_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \ldots \gamma_{n-m} x_{(n-m)i} + \sigma\varepsilon_i \qquad (2)$$

Where the notations have their usual meaning.

To examine the independence of two categorical variables, we employ the chi-square test of independence. When the p-

value is less than the predetermined significance level (usually 0.05 in our case), null hypothesis is rejected and it is concluded that the variables are not independent.

The reduced model is expected to have a decreased standard error because it eliminates correlated predictors from the full model, consequently improving the fit of the general model. We can then compare these two models, namely, the general model and the reduced model, by evaluating their Akaike Information Criteria (AICs) where AIC is;

$$AIC = -2LL + 2(a + c) \qquad (3)$$

Where LL = Log-likelihood of the model, a = number of parameters of the assumed probability distribution (for example; $a = 2$ for Log-Logistic AFT model as there are two parameters involved) and c, the number of coefficients (excluding constant) in the final model. The model with smaller value of AIC can be considered as a better model compared to other models under consideration.

Results and discussion

The follow up time for each patient is recorded in months. An accelerated failure time model having all the predictors is fitted and the results are shown in the Table 1. All these models are assessed using statistical criteria, specifically the log-likelihood ratio test and AIC. The model with the lowest AIC value is regarded as the most suitable model. According to the information presented in Table 1, the Weibull model has the lowest AIC value, indicating that it appears to be the most appropriate Accelerated Failure Time (AFT) model for the analysis.

The results of the Weibull AFT model are presented in the Table 2. It can be observed that the TR for Age is greater than 1 which signifies that patients who are old (> 18 years) are expected to have larger follow up time as compared to the patients who are young (<=18 years). Similarly, patients whose BMI are higher are expected to have smaller follow up time as compared to the patients whose BMI are lower (as TR<1). Also, patients with smaller mean duration of symptoms have greater follow up time as compared to the patients whose mean duration is more (as TR< 1). Similarly, patients having higher total lymphocyte count have lower follow up time as compared to the patients who had lower total lymphocyte count. Females are expected to have larger follow up time as compared to the males (TR>1) (Sousa et al. (2015)). Also, patients who have obesity had less follow up time (TR<1). The T- cell phenotyping has been shown to be an adverse clinical prognostic factor in ALL by Gajjar (2000). In our study also, Patients suffering from T- lineage ALL have lower follow time as compared to the patients suffering from B- lineage ALL.

Patients who experienced blasts have lower follow up time when compared to the patients who didn't experience blasts (TR<1). Patients who have CNS 2, negative TLP, positive TLP have lower follow up time as compared to the reference group (TR<1). ABFM was reducing the lower follow up time of the patients as compared to SBFM (TR<1). Also, patients who skipped therapy had lower follow up time in comparison to those who didn't skip the therapy (TR >1).

We employed several statistical tests, including pairwise correlation, One-Way ANOVA, and the Chi-square test, to examine the independence or association between predictors.

Table 1 Akaike Information Criterion (AIC) values for AFT models

Distribution	Log-Likelihood	DF	C	AIC
Exponential	-111.14	24	1	260.28
Weibull	-88.19	24	2	216.38
Log- Normal	-91.42	24	2	222.84
Log- Logistic	-91.52	24	2	223.04

Table 2 Results of Weibull AFT model for Acute Lymphoblastic Leukemia patients

Predictors	Accelerated factor	Std. Err.	Z	P>|z|	[95% Conf.	Interval]
Age	1.016211	0.008474	1.93	0.054	0.999736	1.032956
BMI	0.965953	0.027697	-1.21	0.227	0.913166	1.021791
DOS	0.998117	0.001617	-1.16	0.245	0.994952	1.001292
Delay in days	0.997155	0.004662	-0.61	0.542	0.988059	1.006335
Pre-TLC	0.999998	1.24E-06	-1.28	0.199	0.999996	1.000001
Platelets at LP	1.000013	1.18E-05	1.12	0.264	0.99999	1.000036
Days of LP	0.980251	0.01427	-1.37	0.171	0.952678	1.008623
Sex						
Male	1					
Female	1.117255	0.169303	0.73	0.464	0.830167	1.503624
Obese						
No	1					
Yes	0.902855	0.281864	-0.33	0.743	0.489643	1.664778
Bulky						
Yes	1					
No	1.916426	0.367978	3.39	0.001	1.315378	2.792116
Study type						
Retro	1					
Prospective	5.230369	0.946507	9.14	0	3.668561	7.457083
TLC Level						
< 50K	1					
50K to 99K	1.111534	0.249987	0.47	0.638	0.715296	1.727269
>99K	1.76574	0.607666	1.65	0.099	0.899482	3.466261
IPT						
B lineage	1					
T lineage	0.704082	0.135606	-1.82	0.069	0.482705	1.026987
Blast						
No	1					
Yes	0.998755	0.16185	-0.01	0.994	0.726977	1.372136
CNS type						
CNS 1	1					
CNS 2	0.601443	0.233873	-1.31	0.191	0.280674	1.288805
Negative traumatic lumbar puncture (TLP)	0.706431	0.117698	-2.09	0.037	0.509627	0.979237
Positive TLP	0.600385	0.369817	-0.83	0.408	0.179522	2.0079
Steroid plus Leunase						
No	1					
Only Steroid	0.692845	0.136785	-1.86	0.063	0.47053	1.020198
Both	0.759495	0.185789	-1.12	0.261	0.470222	1.226723
Protocol						
SBFM	1					
ABFM	0.623149	0.124648	-2.36	0.018	0.421044	0.922267
Therapy omissions						
Yes	1					
No	1.42172	0.245483	2.04	0.042	1.013539	1.994288

One-Way ANOVA assesses the independence between a categorical predictor and a continuous predictor, while the Chi-square test investigates the independence between two categorical predictors. Pairwise correlation is utilized to evaluate the correlation between two continuous predictors. The results of these tests are displayed in the table below. In Table 3, the values that are bold represent either the correlation coefficients (for continuous predictors) or significant p-values (indicating at least one predictor in the pair is significant). These bold values indicate that the pairs are significantly correlated or associated. Table 4 presents the pairs of predictors that have been identified as significantly correlated or associated. The predictors listed in the first column exhibit pairwise correlation or association with each predictor in the second column.

Once we have identified the significantly correlated or associated pairs of predictors, we proceed to determine which predictor is dependent on the other within these pairs. This determination is made by fitting two models:

1) The first model takes one predictor as the dependent variable and the other predictor as independent from the significantly correlated or associated predictor.
2) In the second model, we swap the roles of the dependent and independent variables.

These two models are then compared, and the one with the lower AIC is selected. The corresponding independent variable is chosen, while the dependent variable is omitted. Depending on the nature of the dependent variable, three types of models are fitted: Binomial Logistic, Linear, and Multinomial Logistic models. Logistic model is fitted when the independent variable is a categorical variable having two categories. Linear regression model is applied when the independent variable is a continuous variable. Multinomial Logistic regression is used to fit the model when the independent variable is a categorical variable having more than two categories. The fitted models and their corresponding AIC values are presented in Table 5.

In the next step, we retain only those independent predictors whose models have lower AIC values compared to the models in which they are taken as the dependent variable, with the other predictor of the significantly correlated or associated pair as the independent variable. As a result, out of the initial eighteen predictors, eight have been selected for inclusion in the final model. These selected predictors are as follows: Age, Duration of symptoms, Delay in therapy, Presenting Total Lymphocyte count, Days of LP, Platelet count at LP, CNS status, and Protocol. Subsequently, we estimate the follow-up time once again, this time using only these eight predictors.

In this reduced model as well, the Weibull Accelerated Failure Time Model (AFTM) has been found to be the most appropriate choice. The results of this final model have been presented in Table 6. Also, standard errors of the coefficients of the reduced model are found to be less than the standard errors of the coefficients of the General model shown in the Table 7. As the AIC of the reduced model is small as compared to the AIC of the general model, the reduced model fits better than the general model as shown in the Table 8.

Table 3 Result of the tests of dependence between predictors (p- values)

	Age	BMI	DOS	Delay (days)	Pre-TLC	Plt at LP	Day of LP	Sex	Obese	Bulky	P/R	Level of TLC	IPT	Blast	S+L	CNS status	Protocol	Therapy omissions
Age		0.729	-0.133	0.016	0.042	-0.004	0.384	0.682	0	0.548	0.264	0.095	0.936	0.011	0.001	0.548	0.196	0
BMI			-0.181	-0.067	0.041	0.016	0.353	0.561	0	0.506	0.43	0.007	0.825	0.133	0.014	0.468	0.579	0.002
DOS				0.28	-0.208	0.057	-0.087	0.425	0.141	0.425	0.44	0.039	0.433	0.914	0.718	0.611	0.527	0.995
Delay (days)					-0.153	-0.003	-0.209	0.082	0.423	0.271	0.28	0.571	0.657	0.619	0.019	0.857	0.584	0.711
Pre-TLC						0.077	0.057	0.321	0.762	0.03	0.181	0	0.001	0.26	0.22	0.191	0.66	0.41
Plt at LP							0.108	0.579	0.946	0.542	0.34	0.168	0.053	0.393	0.485	0.358	0.791	0.141
Day of LP								0.854	0.015	0.047	0.342	0.227	0.731	0	0	0.827	0.14	0.011
Sex									0.298	0.884	0.049	0.452	0.584	0.29	0.746	0.165	0.841	0.4
Obese										0.35	0.453	0.009	0.699	0.71	0.011	0.819	0.072	0.064
Bulky											0.803	0.006	0	0.006	0.28	0.91	0.974	0.805
P/R												0.858	0.305	0.184	0.048	0.453	0.816	0.023
Level of TLC													0	0.499	0.041	0.539	0.753	0.69
IPT														0.033	0.667	0.643	0.515	0.243
Blast															0	0.775	0.049	0.093
S+L																0.27	0.007	0.154
CNS status																	0.22	0.481
Protocol																		0.976

Table 4 Significantly Correlated/ associated predictors

Predictor	Significantly correlated/associated predictors
Age	BMI, Obese, Blast at LP, Therapy Omissions
BMI	Obese, TLC level, Steroid plus Leunase, Therapy Omissions
DOS	Obese
Present TLC	Bulky, TLC level, IPT
Day of LP	Obese, Bulky, Blast At LP, Steroid plus Leunase, Therapy Omissions
Sex	P/R,
Obese	TLC level, Steroid plus Leunase
Bulky	TLC level, IPT, Blast
P/R	TLC level
TLC level	IPT, Blast
IPT	Blast
Blast	Steroid plus Leunase
Steroid plus Leunase	Protocol

Table 5 AIC values for choosing suitable predictors in the reduced model

Model	Dependent	Independent	AIC
Linear	BMI	Age	566.45
Linear	Age	BMI	777.66
Logistic	Obese	Age	67.67
Linear	Age	Obese	810.84
Logistic	Blast at LP	Age	141.91
Linear	Age	Blast at LP	852.18
Linear	Age	Steroid plus Leunase	845.50
Multinomial Logistic	Steroid plus Leunase	Age	221.90
Linear	Age	Therapy Omissions	844.16
Logistic	Therapy Omissions	Age	123.97
Logistic	Obese	BMI	17.75
Linear	BMI	Obese	558.37
Linear	BMI	TLC level	645.37
Multinomial Logistic	TLC level	BMI	186.53
Multinomial Logistic	Steroid plus Leunase	BMI	228.16
Linear	BMI	Steroid plus Leunase	640.84
Linear	BMI	Therapy Omissions	638.19
Logistic	Therapy Omissions	BMI	128.86
Linear	Duration of symptoms	TLC level	1134.20
Multinomial Logistic	TLC level	Duration of symptoms	187.52
Multinomial Logistic	Steroid plus Leunase	Delay in therapy	227.69
Linear	Delay in therapy	Steroid plus Leunase	890.44
Linear	Present TLC	Bulky	2785.48
Logistic	Bulky	Present TLC	126.66
Linear	Present TLC	TLC level	2673.68
Multinomial Logistic	TLC level	Present TLC	71.36
Linear	Present TLC	IPT	2777.84
Logistic	IPT	Present TLC	105.42
Logistic	Obese	Days of LP	99.04
Linear	Days of LP	Obese	687.85
Linear	Days of LP	Bulky	690.41
Logistic	Bulky	Days of LP	127.25
Logistic	Blast at LP	Days of LP	131.00
Linear	Days of LP	Blast at LP	679.55
Linear	Days of LP	Steroid plus Leunase	675.67
Multinomial Logistic	Steroid plus Leunase	Days of LP	203.20
Linear	Days of LP	Therapy Omissions	687.85
Logistic	Therapy Omissions	Days of LP	131.60
Logistic	Sex	P/R	118.75
Logistic	P/R	Sex	148.33
Logistic	Obese	TLC level	105.66
Multinomial Logistic	TLC level	Obese	185.45
Logistic	Bulky	TLC level	121.31
Multinomial Logistic	TLC level	Bulky	185.39
Logistic	Bulky	IPT	116.59
Logistic	IPT	Bulky	101.00
Logistic	Bulky	Blast at LP	122.78
Logistic	Blast at LP	Bulky	140.76
Multinomial Logistic	Steroid plus Leunase	Obese	220.25
Logistic	Obese	Steroid plus Leunase	100.80
Logistic	P/R	Steroid plus Leunase	146.14
Multinomial Logistic	Steroid plus Leunase	P/R	230.65
Logistic	P/R	Therapy Omissions	147.02
Logistic	Therapy Omissions	P/R	132.56
Multinomial Logistic	TLC level	IPT	175.61
Logistic	IPT	TLC level	98.15
Multinomial Logistic	Steroid plus Leunase	TLC level	229.78
Multinomial Logistic	TLC level	Steroid plus Leunase	192.49
Logistic	IPT	Blast at LP	110.54
Logistic	Blast at LP	IPT	144.10
Multinomial Logistic	Steroid plus Leunase	Blast at LP	216.03
Logistic	Blast at LP	Steroid plus Leunase	129.84
Logistic	Blast at LP	Protocol	142.71
Logistic	Protocol	Blast at LP	159.05
Multinomial Logistic	Steroid plus Leunase	Protocol	154.49
Logistic	Protocol	Steroid plus Leunase	224.41

Table 6 Results of reduced Logistic AFTM model for ALL patients

Predictors	Time ratio	Std. Err.	z	95% Conf. Interval	
Age	1.00656	0.00772	0.61	-0.01459	0.02763
DOS	0.99781	0.00139	-1.07	-0.00726	0.00211
Delay in days	0.99169	0.00200	-0.77	-0.01909	0.00836
Pre-TLC	1.125	1.12E-6	0.34	-1.80E-06	2.57E-06
Days of LP	1.00001	0.01722	-1.25	-0.05526	0.01226
Platelets at LP	0.98011	1.06E-1	1.02	-1.50E-05	0.00004
Protocol					
SBFM	1				
ABFM	0.91137	0.10153	-0.25	-0.53367	0.41313
CNS type					
CNS 1	1				
CNS 2	0.78181	0.15819	-0.47	-1.5742	0.96668
Negative traumatic lumbar puncture (TLP)	0.92420	0.10578	-0.47	-0.59727	0.36618
Positive TLP	1.06141	0.30593	0.02	-1.81919	1.85352

Table 7 Comparison of Standard errors of Coefficients of general and reduced model

Predictors	Std. Err. (General model)	Std. Err. (Reduced model)
Age	0.008474	0.007721
BMI	0.027697	
DOS	0.001617	0.001392
Delay in days	0.004662	0.002003
Pre-TLC	1.24E-06	1.12E-06
Platelets at LP	1.18E-05	1.06E-05
Days of LP	0.01427	0.0017226
Male		
Female	0.169303	
Obese- Yes		
Obese- No	0.281864	
Obese- NA	0.739653	
Bulky- Yes		
Bulky- NO	0.367978	
Study- Retrospective		
Study- Prospective	0.946507	
TLC Level		
< 50K		
50K to 99K	0.249987	
>99K	0.607666	
IPT- B lineage		
IPT- T lineage	0.135606	
Blast- No		
Blast- Yes	0.16185	
CNS 1		
CNS 2	0.233873	0.158196
Negative traumatic lumbar puncture (TLP)	0.117698	0.105784
Positive TLP	0.369817	0.305934
Steroid plus Leunase- No		
Only Steroid	0.136785	
Both	0.185789	
Protocol- ABFM		
Protocol- SBFM	0.124648	0.101535
Therapy omissions- No		
Therapy omissions- Yes	0.245483	

Table 8 Comparison of general and reduced model

Model	**No. of predictors**	**AIC**
General	18	216.38
Reduced	8	210.40

Discussion

The research begins by aiming to estimate the follow-up time of patients with Acute Lymphoblastic Leukemia (ALL) from the time of diagnosis until their last contact. To achieve this, the study initially employs 18 different predictors. These predictors likely encompass a range of demographic, clinical, and biological variables that could potentially influence the follow-up time. To identify the most suitable model for this estimation, the study considers various survival models. Survival analysis is particularly useful when dealing with time-to-event data, such as the time until a patient's last contact. Among the tested models, the Weibull Accelerated Failure Time Model (AFTM) is selected as the best fit based on the Log-Likelihood and Akaike Information Criterion (AIC) values. The choice of the Weibull AFTM indicates that the researchers believe that the hazards (the risk of experiencing the event of interest) change over time in a specific way, as described by the Weibull distribution. This model is flexible and can capture different shapes of hazard functions. After selecting the Weibull AFTM, the study progresses to assess the independence between each pair of the 18 prognostic factors (predictors). This is a crucial step in survival analysis because the assumptions of many survival models, including the Weibull AFTM, often assume that predictors are independent. If predictors are highly correlated or associated with each other, it can lead to multicollinearity issues and make the model less interpretable.

The results of the independence testing lead to the identification of significantly correlated or associated pairs of predictors. This step is essential for reducing the dimensionality of the model and selecting a more manageable set of variables. From the initial 18 predictors, the study narrows it down to 8 predictors that are deemed most relevant based on correlation analysis. These 8 predictors represent a subset of variables that have been shown to have a meaningful relationship with the follow-up time. The selected predictors are: Age, duration of symptoms, platelets at the time of lumbar puncture, days in which patients were in lumbar puncture, total lymphocyte count, days of delay/interruptions in chemotherapy due to toxicity, chemotherapy type, and CNS (Central Nervous System) status.

With the reduced set of 8 independent predictors, the study revisits the task of estimating the follow-up time using various survival models. Once again, the Weibull AFTM is found to be the best fit for the data. This reaffirms the choice of the Weibull AFTM as the most appropriate model for this specific research question and dataset. One of the key findings of this research is the comparison between the full model (with 18 predictors) and the reduced model (with 8 independent predictors). This comparison is based on the AIC values. The AIC is a measure of model goodness-of-fit that also considers model complexity. Smaller AIC values indicate better-fitting models.

The crucial insight here is that the reduced model (8 independent predictors) has a smaller AIC value than the full model (18 correlated/associated predictors). This suggests that the simpler model with fewer predictors is a better fit for the data. Furthermore, the reduced model has a smaller standard error, indicating more precise parameter estimates.

The research's main takeaway is that, for the specific task of estimating follow-up times for ALL patients, a simpler model with fewer, independent predictors is not only more interpretable but also provides a better fit to the data. This finding can have important clinical implications by identifying which variables are most relevant for predicting follow-up times, thus aiding in treatment planning and prognosis. Additionally, this research illustrates the importance of model selection and dimensionality reduction in survival analysis. It showcases how rigorous statistical methods can enhance the accuracy and interpretability of models, ultimately leading to more meaningful insights in the

medical field. Furthermore, the Weibull AFTM's selection underscores its applicability in modeling time-to-event data with changing hazard rates.

Conclusion

This aim of this research was to estimate the follow up time of ALL patients from the time of diagnosis till the time of last contact. Initially, the follow up time was estimated using 18 predictors. Among a number of survival models, Weibull AFTM is found to be the best model on the basis of the Log-Likelihood and AIC values. Then the independence between each pair of prognostic factors is tested and significantly correlated/associated pairs are selected. Then among these 18 predictors, only 8 predictors namely, Age, duration of symptoms, platelets at the time of lumbar puncture, days in which patients was in lumbar puncture, total lymphocyte count, days of delay/ interruptions in chemotherapy due to toxicity, chemotherapy type and CNS status are selected using correlation analysis. Then follow up time is estimated in the presence of 8 selected predictors using a number of survival models. Again, Weibull AFTM is found to be the best fit. Also, the AIC value of this reduced model is found to be small as compared to the true fitted model with a reduced standard error which shows that the reduced model is a good fit as compared to the true model. So, the model which has 8 independent predictors is a good fit as compared to the model which has 18 correlated/ associated predictors.

References

Board, P D Q Pediatric Treatment Editorial. 2023. "Childhood Acute Lymphoblastic Leukemia Treatment (PDQ®)." In *PDQ Cancer Information Summaries [Internet]*. National Cancer Institute (US).

Brown, Patrick. 2013. "Treatment of Infant Leukemias: Challenge and Promise." *Hematology 2013, the American Society of Hematology Education Program Book* 2013 (1): 596–600.

Feig, Barry W, David H Berger, and George M Fuhrman. 2006. *The MD Anderson Surgical Oncology Handbook*. Lippincott Williams & Wilkins.

Howard, Scott C, Amar J Gajjar, Cheng Cheng, Stephen B Kritchevsky, Grant W Somes, Patricia L Harrison, Raul C Ribeiro, Gaston K Rivera, Jeffrey E Rubnitz, and John T Sandlund. 2002. "Risk Factors for Traumatic and Bloody Lumbar Puncture in Children with Acute Lymphoblastic Leukemia." *Jama* 288 (16): 2001–7.

Inaba, Hiroto, Mel Greaves, and Charles G Mullighan. 2013. "Acute Lymphoblastic Leukaemia." *The Lancet* 381 (9881): 1943–55.

Jacquillat, C, M Weil, M F Auclerc, G Schaison, C Chastang, J L Harousseau, F Bauters, D Olive, C Griscelli, and M Bonnet. 1980. "Application of the Study of Prognostic Factors to the Treatment of Childhood (Less than 20 Years Old) Acute Lymphoblastic Leukemia." *Bulletin Du Cancer* 67 (4): 458–69.

Le, Quoc-Hung, Xavier Thomas, René Ecochard, Jean Iwaz, Véronique Lhéritier, Mauricette Michallet, and Denis Fiere. 2006. "Initial and Late Prognostic Factors to Predict Survival in Adult Acute Lymphoblastic Leukaemia." *European Journal of Haematology* 77 (6): 471–79.

Rozenberg, Gillian. 2011. *Microscopic Haematology: A Practical Guide for the Laboratory*. Elsevier Australia.

Sayehmiri, Kourosh, Mohammad R Eshraghian, Kazem Mohammad, Kamran Alimoghaddam, Abbas Rahimi Foroushani, Hojjat Zeraati, Banafsheh Golestan, and Ardeshir Ghavamzadeh. 2008. "Prognostic Factors of Survival Time after Hematopoietic Stem Cell Transplant in Acute Lymphoblastic Leukemia Patients: Cox Proportional Hazard versus Accelerated Failure Time Models." *Journal of Experimental & Clinical Cancer Research* 27 (1): 74.

SEER. 2023. "Acute Lymphocytic Leukemia-Cancer Stat Facts." *SEER*.

Shaikh, Furqan, Laura Voicu, Soumitra Tole, Teresa To, Andrea S Doria, Lillian Sung, and Sarah Alexander. 2014. "The Risk of Traumatic Lumbar Punctures in Children with Acute Lymphoblastic Leukaemia." *European Journal of Cancer* 50 (8): 1482–89.

Totadri, S, A Trehan, R Srinivasan, D Bansal, and P Bhatia. 2015. "Do Traumatic Lumbar Punctures Lead to Greater Relapses in Acute Lymphoblastic Leukemia? Experience at a University Hospital in India." *Indian Journal of Cancer* 52 (3): 300.

Uderzo, Cornelio, Maria Grazia Valsecchi, Andrea Bacigalupo, Giovanna Meloni, Chiara Messina, Paola Polchi, Gabriele Di Girolamo, Giorgio Dini, Roberto Miniero, and Franco Locatelli. 1995. "Treatment of Childhood Acute Lymphoblastic Leukemia in Second Remission with Allogeneic Bone Marrow Transplantation and Chemotherapy: Ten-Year Experience of the Italian Bone Marrow Transplantation Group and the Italian Pediatric Hematology Oncology Association." *Journal of Clinical Oncology* 13 (2): 352–58.

Vos, Theo, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, and Alan Z Chen. 2016. "Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 310 Diseases and Injuries, 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015." *The Lancet* 388 (10053): 1545–1602.

Wang, Haidong, Mohsen Naghavi, Christine Allen, Ryan M Barber, Zulfiqar A Bhutta, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Zian Chen, and Matthew M Coates. 2016. "Global, Regional, and National Life Expectancy, All-Cause Mortality, and Cause-Specific Mortality for 249 Causes of Death, 1980–2015: A Systematic Analysis for the Global Burden of Disease Study 2015." *The Lancet* 388 (10053): 1459–1544.