МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2013
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2013
Proceedings of the Forty Second Spring Conference
of the Union of Bulgarian Mathematicians
Borovetz, April 2–6, 2013

MODELLING SURVIVAL IN CHILDHOOD ACUTE LYMPHOBLAST LEUKEMIA*

Krasimira Prodanova, Nadejda Yurukova

Acute lymphoblastic leukemia (ALL) is the most common malignancy diagnosed in children, representing nearly one third of all pediatric cancers. About 30% of the children with ALL have a gene marker. The most frequent abnormality is in TEL-AML1 gene rearrangement and this marker can be detected in 20-30% of the cases with ALL. In this paper the survival analysis is used to determine the prognostic significance of TEL-AML1 and to models the time it takes for relapse or death. The data are from 160 patients, observed in Specialized Children's Oncohematology Hospital – Sofia, Bulgaria for a time of 8 years. The gene marker TEL-AML1 is detected in 33 of the patients. For estimating event (relapse or death) free survival rate the Kaplan-Meier method is used. Time to event (in months) is calculated as the time from study entry to first event or data of last contact. The log-rank test is used for comparison of survival curves between two groups (with and without TEL-AML1). Multivariate analysis is conducted using Cox proportional hazards regression. Cox regression gives a simple model to describe the survival between two groups The prognostic significance of TEL-AML1 is investigated with Proportional Hazards Model for the first time in Bulgaria.

1. Introduction. Annually, around 13000 children in the world are diagnosed with acute lymphoblastic leukemia. White children are more frequently affected than black children, and there is a slight male preponderance [9]. The two most important factors are reported to be predictive of greater risk of relapse are age and presenting white blood cell (WBC) count at diagnosis [9,10]. The National Cancer Institute (NCI)/ Rome criteria stratify patients into subsets based on age 1 to 9.99 years and/or WBC $< 0,5.10^9$/L (standard risk), or age ≥ 10 years and/or WBC $\geq 50,000/\mu$L (higher risk) [13, 14]. These variables have consistently emerged as independent predictors of outcome in almost all therapeutic studies. The obvious advantage of this system is that these variables can be measured reliably in almost all circumstances, and therefore these criteria can be applied to children worldwide The purpose of this article is to prospectively determine the prognostic significance of marker TEL-AML1 in Bulgarien children with acute lymphoblastic leukemia, to estimate event (relapse and death) survival time. Clinical prognostic risk factors as age, gender, WBC count at diagnosis, risk-type group are taken as a predictors (covariates) of multivariate analysis by using Cox proportional hazards regression. The

*2000 **Mathematics Subject Classification:** 92B15, 62P10.
Key words: survival analysis, Kaplan-Meier estimator, Cox Proportional Hazards Regression.

comparison of the survival characteristics between two groups of patients: with TEL-AML1 and without this marker using the Kaplan-Meier estimator and Cox model is made. The software package STATISTICA 10.0 is used [12].

2. Measurements. We describe the models in this paper using a cancer survival data set on 160 individuals from two groups: with Marker TEL-AML1 (33 patients) and without this marker (127 patients). The explanatory variables (covariates) are:

- Sex – takes the values "1" for male or "0" for female;
- Risk_Group – coded : "0" for standard risk, "1" for middle risk, and "2" for higher risk;
- AGE in years;
- WBC – white blood cell count at diagnosis;
- M_Event – the survival time (in months);
- M_Letal – the survival time (in months) for death ;
- Event and Lethal – coded by "1" if event(s) happened or "0" if no event at the end of the study, i.e. "right censored".
- Marker – coded by "1" if the patient have TEL-AML1 and "0" if not.

For the group without marker TEL-AML1 the minimal value of variable AGE is 3 months (0.25 years) and the maximum is 18 years, the median of the AGE is 7 years. 62.2% of the patients are male, 78.74% of the patient have no marker at all, an relapse occur for 25.98% of the patients, the median for M_Event is 37 monts and for M_Letal is 37.4 months. The minimum value of Leukocytes is $0.5.10^9$ /L. The censored cases for relapse are 75.21% and for death 71.65%.

For the group with marker TEL-AML1 the minimal value of variable AGE is 2 years and the maximum is 16 years, the median of the variable AGE is 4 years. 60.6% of the patients are male, an relapse ocur for 6.06% of the patients, the median for M_Event is 37.70 months and for M_Letal is 37.67 months. The censored cases for events are 93.75%, and for death 90.91%.

3. Statistical methods. One of the oldest and most straightforward methods for analyzing survival data is to compute the Life Table [1] The Life Table method worked well for a homogeneous sample, but it did not address a primary goal of cancer research, namely to determine whether or not certain continuous and/or categorical variables are related to the survival times [7]. This need led to the application of regression methods for analyzing survival data. Two important developments that have greatly enhanced survival analysis methods are the derivation of a nonparametric method for constructing a survival curve from censored data by Kaplan and Meier [3], and the Proportional Hazards model proposed by Cox [4,5]. The Cox model, a multivariate semi parametric regression model, is the most widely used in clinical studies to characterize disease progression on existing cases by revealing the importance of covariates [1–10].

Kaplan-Meier estimator of survival function. For real data the survival function $S(t)$ is unknown but can be estimated from a sample. The estimation of the survival function can be realised nonparametrically by Kaplan-Meier estimator. If the data were not censored, the obvious estimate would be the empirical survival function [3]: $\widehat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(t_i > t\right)$, where I is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. The estimator is simply the proportion alive at t. Kaplan

and Meier [3] extended the estimate to censored data. An observation is right-censored if the subject leaves the study or is alive when the study ends. Let $t_1 < t < \cdots < t_m$ denote the distinct ordered times of death (not counting censoring times). Let d_i be the number of deaths at t_i, and let n_i be the number alive just before t_i. The intervals between each time typically will not be uniform. The Kaplan-Meier or product limit estimate of the survivor function is$S = \prod_{i:t_i>t} (1 - d_i/n_i)$. In case of discrete covariates, survival curves for different factor levels can be calculated and compared by log-rank test [4, 5] which gives an indication of the relevance of the factor, i.e. whether survival in the groups significantly differs. If the two groups have the same survival function, the logrank statistic is approximately standard normal.

Proportional Hazards Model. Let T represent survival time. T is regarded as a random variable with cumulative distribution function $P(t) = P(T \leq t)$ and probability density function $p(t) = dP(t)/dt$. The survival function $S(t)$ is the compliment of the distribution function $S(t) = P(T > t) = 1 - P(t)$. A fourth representation of the distribution of survival times is the hazard function, which assesses the instantaneous risk of demise at time t, conditional on survival to that time [4], [5], [7]:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[(t \leq T < t + \Delta t) \,|T \geq t\right]}{\Delta t}.$$

The hazard is sometimes referred to as the "force of mortality" or the "conditional failure rate". Survival analysis typically estimates the relationship of the survival distributions to covariates. Most commonly, this estimation entails the specification of a linear-like model for the log hazard. A parametric model based on the exponential distribution may be written as $\log h_i(t) = \lambda + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, where i is subscript for observations, and x's are the covariates and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector of model parameters. The Cox Proportional-Hazards Model leaves the baseline hazard function $\lambda(t) = \log h_0(t)$ unspecified. This model relates the hazard function of an individual at time t, with a vector $X = (X_1, X_2, \ldots, X_p)^T$ of p covariates, to a baseline hazard function $h_0(t)$ via a log-linear function: $h(t; X) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X_j\right)$. An important consequence of this formulation, and the reason for the name "Proportional Hazards Model", is that the hazard ratio-*HR* (also called relative risk – RR) for two individuals does not depend on time. The proportional hazards model [4] is a simple model to describe the survival between two groups having hazard functions $h_1(t)$ and $h_2(t)$. The proportional hazards assumption is that ratio of these hazard functions does not depend on time: $\psi = \dfrac{h_2(t)}{h_1(t)}$, where ψ is called the proportionality constant [7]. While ψ does not depend on time, it does depend on the explanatory variables in the model. A convenient and readily interpretable linear expression for the logarithm of ψ is obtained:

$$\log \psi = \beta_1 * (\text{Sex}) + \beta_2 * \text{Age} + \beta_3 * \text{MARKER} + \beta_4 * WBC + \beta_5 * \text{Risk\_Group},$$

where the β's have been subscripted to reflect there role in the hypothesized causal mechanism. The hypothesis of no association of one or more of the p independent variables with survival can be tested by the likelihood ratio test [4] or Wald test [11, 12].

4. Models derived from the experimental data. For estimating event free survival rate the Kaplan–Meier method is used. Time to event (relapse or death) in

months is calculated as the time from study entry to first event or data of last contact. For the both groups (with and without TEL-AML1) resulting Kaplan–Meier estimators built from the data of relapse in M_Event_33, Event_33 and M_Event_127, Event_127 are plotted in Figure 1(A). Log-rang test is used for the test of the null hypothesis $H_0 : S_1(t) = S_2(t)$ at the level of significance $\alpha = 0,05$. The obtained test statistic for survival function for relapse is $Z = 2,31588 > 1,645$ and H_0 is rejected. The same type of analysis is made for survival functions if the event is a death. For the both groups resulting Kaplan–Meier estimators built from the data of relapse in M_Letal_33, Letal_33

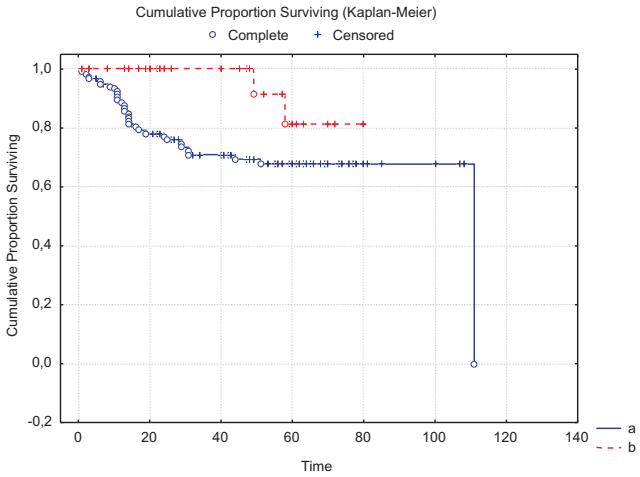

Fig. 1(A). Comparison of the survival functions of Kaplan-Meier estimate using data for relapse in the groups of patients without marker TEL-AML1 (solid line) and with TEL-AML1 (dotted line)

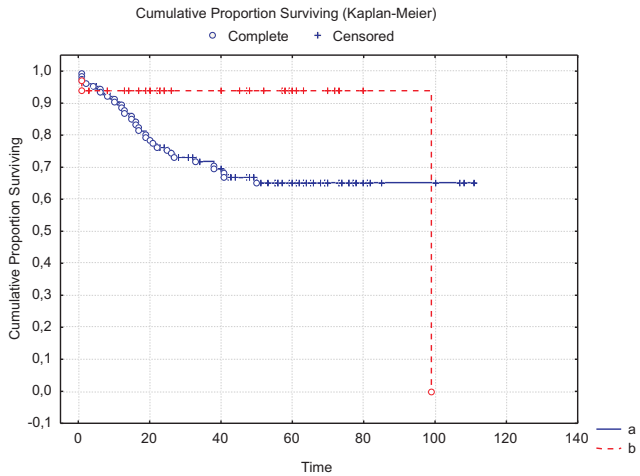

Fig. 1(B). Comparison of Kaplan-Meier estimate for the survival function using data for death in the groups of patients without TEL-AML1 (solid line) and with TEL-AML1 (dotted line)

and M_Letal_127, Letal_127 are plotted in Figure 1(B). The obtained test statistic for survival function for relapse is $Z = 2,09335 > 1,645$ and H_0 is rejected.

To describe the effect of other variables, such as age, sex, white blood cell, risk-type group (Risk-Group), on survival it is convenient to use the hazard function. The coefficients in a Cox regression relate to hazard – a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated.

The fit of the data to the proportional hazard (Cox) regression model for data of relapse is given in the Table 1:

Table 1. Estimated parameters of Cox model using data for relapse

Dependent Variable: M_event (TEL_AMI_sta) Censoring var.: Event = 0 Chi2 = 15.7582df = 5 p = 0.00758						
	Beta	Standard Error Beta	t-value	Exponent Beta or HR	Wald – Statist.	p
Sex	0.08202	0.370231	0.221537	1.085477	0.04907	0.824676
Age	−0.1170	0.157110	−0.42523	0.19166	−0.742	0.5507
Marker	0.86504	0.326319	2.650910	2.375108	7.02732	0.008031
WBC	0.00359	0.001468	2.448091	1.003600	5.99315	0.014367
RISK_Group	1.01665	0.330834	3.073019	2.763943	9.44344	0.002121

Table 2. Estimated parameters of Cox model using data for death

Dependent Variable: M_letal (TEL_AMI_sta) Censoring var.: Letal-0 Chi2 = 12.5435 df = 5 p = 0.02807						
	Beta	Standard Error Beta	t-value	Exponent Beta or HR	Wald – Statist.	p
Sex	0.03964	0.343430	−0.11542	0.961135	0.013323	0.908109
Age	−0.01337	0.035184	−0.37999	0.986719	0.144395	0.703953
Marker	0.79678	0.319194	2.49623	2.218387	6.231165	0.012557
WBC	0.00392	0.001339	2.93284	1.003935	8.601565	0.003361
Risk-Group	0.91166	0.299497	3.04398	2.488457	9.265851	0.002337

The results of this multivariate proportional hazard regression are the following. The variable Sex is not significant prognostic factor for the relapse. The other covariates are prognostic factors for the relapse ($p < 0.05$). Expressed in terms of HR, $\exp(0.865) = 2.375$ indicates that the rate of relapse for the group without TEL-AML1 is 2,375 times the risk of relapse for those with TEL-AML1. From looking at the hazard ratios the model indicates that if the Risk-Group ($\beta_4 = 1.0166 \approx 1.017$) is altered from standard to higher risk, while holding all other variables constant, the rate of relapse increases by 1.7%.

Resulting proportional hazard (Cox) regression model for the data of death is given in the Table II. The results of this regression are the following. The variables Sex and Age are not significant prognostic factors for the death ($p > 0.05$). The other covariates are prognostic factors for the death ($p < 0.05$). In the terms of HR, $\exp(0.796781) = 2.218387$

indicates that the rate of death for the group without TEL-AML1 is approximately 2.2 times the risk of death for those with TEL-AML1.

5. Conclusions. These data support that the presence of TEL-AML1 at diagnosis is an independent favorable prognostic indicator in patients with ALL in Bulgaria. The conclusions derived using Kaplan-Meier estimator for the group of patients with TEL-AML1 are the following: the most critical of relapse is the month's interval $[0, 48)$ and in the month's interval $[48, 58)$ the probability decreases. For the group of patients without TEL-AML1are the following: The month's interval $[0, 30)$ is the most critical of relapse and after 50-th months the probability for relapse free time is a relatively high (0.68). For the lethal outcome in the group of patients with TEL-AML1 the derived conclusion is that the most critical period is till the end of the first month. In the group of patients without TEL-AML1 the month's interval $[1, 50]$ is the most critical of death. After that the probability for a live time is a relatively high (0.65). Similar results, using Kaplan-Meier estimator, are published in [9], [10] and [13].

The 3-year event-free survival (EFS) varies considerably dependin on risk category, from 95% (low risk) to 30% (very high risk), with infant leukemia having the worst outcomes: 20% for patients younger than 90 days. Overall, the cure rate for childhood acute lymphoblastic leukemia (ALL) is more than 80%.

Acknowledgement. The authors are very grateful to the referee for the helpful comments on the paper.

REFERENCES

[1] J. Berkson, R. Gage. The calculation of survival rates for cancer. *Proc. of Staff Meetings, Mayo Clinic*, **25** (1950), 23–35.

[2] B. Everitt, S. Rabe-Hesketh. Analyzing Medical Data Using S-PLUS. New York, NY, Springer-Verlag, 2001.

[3] E. Kaplan, P. Meier. Nonparametric Estimation from Incomplete Observations. *J. Am. Statistical Association*, **53** (1958), 457–481.

[4] D. Cox. Regression models and life tables. *J. Roy. Stat. Soc. B*, **4** (1972), 187–220.

[5] D. R. Cox, D. Oakes. Analysis of survival data. London: Chapman and Hall, 1984.

[6] T. Therneau, P. Grambsch. Modelling Survival Data. New York, Springer, 2000.

[7] F. Ahmed, P. Vos. D. Holbert. Modelling survival prediction in colon cancer. *Molec. Cancer*, **6** (2007), 15–25.

[8] T. Moreau, M. Le Minor, P. Miguel, J. Lellouch. Estimation of the hazard ratio in two grouped samples. *Biometrics*, **41** (1985), 245–252.

[9] S. K. Vikramjit. Pediatric Acute Lymphoblastic Leukemia. *Hematology*, **1** (2003), 102–131.

[10] J. Rubnitz, D. Wichlan, M. Devidas. Prospective Analysis of TEL Gene Rearrangements in Childhood lymphoblastic Leukemia-Children's Oncology group Study. *J. Clin. Oncol*, **26** (2008) No 13, 2186–2191.

[11] J. Fox. Appendix to an R and S-PLUS Companion to Applied Regression. New York, Academic Press, 2002.

[12] STATISTICA10. Statistical Analysis Software, 2009.

[13] U. Ozbek, S. Sirma, L. Agaoglu, L. Yuksel. Prognostic Significance of the TEL-AML1 Fusion Gene in Pediatric Acute Lymphoblastic Leukemia in Turkey. *J. of Pediatric Hematology/Oncology*, **25** (2003) No 3, 204–208.

Krasimira Stoyanova Prodanova
Technical University of Sofia
1000 Sofia, Bulgaria
e-mail: kprod@tu-sofia.bg

Nadejda Nikolova Yurukova
Specialized Children's Oncohaematology Hospital – Sofia
1000 Sofia, Bulgaria

МОДЕЛИРАНЕ НА ПРЕЖИВЯЕМОСТ ПРИ ОСТРА ЛИМФОБЛАСТНА ЛЕВКЕМИЯ В ДЕТСКА ВЪЗРАСТ

Красимира Проданова, Надежда Юрукова

Острата лимфобластна левкемия (ОЛЛ) е най-честото злокачествено заболяване при децата, което се диагностицира в една трета от случаите на ракови заболявания. Около 30% от децата с ОЛЛ имат генетичен маркер. Най-често срещаното отклонение от нормалното прегрупиране е в генетичния маркер TEL-AML1 и то може да се открие при 20–35% от случаите на ОЛЛ. В тази работа се използва анализ на преживяемостта, за да се определи значимостта на TEL-AML1 като прогностичен фактор и да се моделира времето за настъпване на рецидив на болестта или за смърт. Данните са от 160 пациенти наблюдавани през последните 8 години в Специализираната онко-хематологична клиника в София, България. Генетичният маркер TEL-AML1е открит у 33 от пациентите. За оценяване на степента на преживяемост без събитие (рецидив или смърт) е използван метода на Каплан и Майер. Времето до събитие (в месеци) се пресмята като време от началото на изследването до първото появяване или до датата на последен контакт с пациента. За сравнение на кривите на преживяемост за двете групи (с и без TEL-AML1) е използван лог-ранг тестът. Многомерният статистически анализ е осъществен с построяване на регресионенни модели на Кокс, които описват преживяемостта в две групи. Значимостта на TEL-AML1 с използване на пропорционалния рисков модел се изследва за пръв път в България.