JKMU Research Article

Journal of Kerman University of Medical Sciences, 2017; 24(1): 1-15

Modeling Leukemia in Children Using Phase-type Distribution

Marzieh Mahmoudimanesh, M.Sc.[1], Abas Bahrampour, Ph.D.[2], Zahra Farahmandinia, M.D.[3]

1- Department of Epidemiology and Biostatistics, School of Health, Kerman University of Medical Sciences, Kerman, Iran

2- Professor of Biostatistics, Research Center for Health Modelling, Kerman University of Medical Sciences, Kerman, Iran (Corresponding author; e-mail: abahrampour@yahoo.com)

3- Assistant Professor of Pediatrics, Afzalipour School of Medicine, Kerman University of Medical Sciences, Kerman, Iran

Received: 26 April, 2015 Accepted: 14 July 2015

Abstract

Background: In this study, with the aim of modeling Leukemia in children using Phase-type distribution, three transitional phases including diagnosis, brain metastasis and testis/ovary metastasis, and one absorotion phase of recovery/death have been considered. The distribution was fitted and the probabilities of death or recovery were determined based on the independent variables including age, sex, blood group, etc.

Methods: In this modeling study, necessary information was extracted from patients' medical records (recorded during 2006-2013) available in the Medical Records Department of Afzalipour Hospital of Kerman/ Iran. After excluding the unrelated cases, Phase-type distribution was fitted in which four phases including three transitional phases (cancer diagnosis, brain metastasis, and testis/ ovary metastasis) and one absorption phase (death or recovery) have been considered. For this purpose, different modeling methods were used for patients who had died and recovered. EM algorithm was used for modeling and fitting Phase-type distribution. Data were analyzed using SPSS22 and R. After fitting Phase-type distribution and determining the probabilities of absorption, the effect of each independent variable on these probabilities was evaluated, and t-test, Pearson's correlation coefficient, and One-way analysis of variance (ANOVA) were used for the analysis of different variables.

Results: The variables of sex and the presence or absence of splenomegaly and hepatomegaly had no effect on the probability of death and recovery. However, the probability of death showed significant relationship ($p<0.0001$) with the diagnosis of cancer type (ALL or AML) and it was more in patients diagnosed with ALL. Death probability had also significant relationship with brain and testis/ovary metastasis ($P=0.002$). As expected, the probability of death in patients with brain or testis/ovary metastasis compared to those without matastasis was more. In addition, the p-value of the test used to assess the association between the probability of death and blood groups was 0.025; therefore, there is a significant difference in the probability of death between at least two blood groups.

Conclusion: The results show that the diagnosis of cancer type and treatment method can affect the probabilities of death and recovery. Further studies on other variables can help physicians to predict the probabilities of death or recovery during the development of cancer and choose the best treatment method to enhance the probability of recovery in these patients.

ARTICLE INFO

Article type:
Original article

Keywords:
Phase-type Distribution
EM Algorithm
Transitional Phase
Absorption Phase
Death or Recovery probability
Leukemia
Acute Myeloid Leukemia
Acute Lymphoblastic Leukemia
Metastasis

Introduction

The causes of death have been recently shifted from infectious diseases to noninfectious diseases and cancer after cardiovascular diseases (CVD) and accidents has been considered as the third main cause of death in Iran (1).

Leukemia is the fifth common type of cancer worldwide and accounts for about 8% of all cancers. Leukemia is known as children cancer and its occurrence is affected by many factors such as age, gender, race, blood groups, radiation exposure, etc. (2).

Leukemia can occur in every race and at every age. There are two main types of Leukemia; acute Leukemia including acute myeloid leukemia (AML) and acute lymphoblastic Leukemia (ALL) and chronic Leukemia including chronic Myelocytic Leukemia (CML) and chronic Lymphocytic Leukemia (CLL). The frequency of Leukemia is 10 times more in adults and 2 times more in men. The prevalence rates of different types of leukemia are as follows: ALL=11%, CLL= 29%, AML= 46% and CML= 14%; among which ALL and AML are the most common types of cancer in children (3).

In most diseases, the disease as a dynamic process, develops from early to advanced stages. The process of cancer progress is determined by measuring tumor size and metastasis (4).

The progress of disease is determined by some tests during the follow-up period and since there might be right censoring or a gap during different stages of disease, using statistical models, like regression, is not easily possible for modeling (5).

For this purpose, Markov and Phase-type distribution can be useful. In fact, this distribution indicates distribution of absorption time for each continuous-time Markov chain on a finite discrete state space (6).

In a study performed by Sali I. MacKlein et al (2012), AIDS development has been modeled using Phase-type distributions. Data were obtained from 2092 HIV-infected patients. The variables evaluated in the mentioned study included: patients' history, treatment, dietary history, environmental factors, and HIV diagnostic stage. In this study, the Coxian Phase-type distribution, an intuitive and new method, was used for modeling HIV progress. EM algorithm was used to estimate the parameters of this distribution. Moreover, data obtained from all patients were first fitted to the distribution and four HD stages were considered for the patients. Then, data were clustered and four HI stages were considered for all patients, and in each HI stage, there were four HD stages to which the patients were assigned. Finally, it was observed that for every HI stage, AIC value, after being clustered by HD stage, significantly improved (5).

In the study of Ali Zare et al (2014) on 330 patients with gastric cancer who had undergone surgery at Iran Cancer Institute during 1995– 1999, Markov models and time homogeneity assumption were evaluated in multistate models. In this study, there were three transmission states for the patients: 1) risk of death without recurrence; 2) risk of recurrence; and 3) risk of death with recurrence.

One of the methods for testing Markov hypothesis is determination of distribution of sojoum time in a steady state.

A wide range of statistical distributions can be considered for this time: Log-normal, Log-logistic, etc.

In this study, Akaike information criteria (AIC) and Cox-Snell residuals were used to determine which distribution is the most suitable one. Eventually, it was concluded that although the multistate model is a suitable model for studying cancers and it can provide a better analysis of variables' behavior, for this purpose, assumptions like Markov and time homogeneity are also required. These assumptions can simplify multistate models (7).

The aim of this study was modeling Leukemia in children using Phase-type distribution and since there are no similar studies on modeling Leukemia in children using Phase-type distribution and in most medical studies, the Phase-type distribution have been used to study the patients' length of stay and treatment/post-treatment costs or to predict or reduce the number of re-hospitalizations, only the two above-mentioned studies were introduced in this study to explain the applicability of this test in medicine.

Methods

In this modeling study, all patients with ALL and AML (<15 years old) referred to Afzalipour hospital of Kerman during 2006-2013 were included. Data were obtained from the medical records available in medical record archives of the hospital. The database used in this study included 538 cases that after excluding the irrelevant cases (patients over 15 years old as well as patients with other types of Leukemia rather than ALL or AML), 177 cases (132 with ALL and 45 with AML) were investigated among which 32 patients with ALL and 13 patients with AML had died.

As mentioned before, Phase-type distribution has been used for modeling Leukemia in this study. Phase-type distribution shows the absorption time of a finite-state Markov chain that is a generalized exponential distribution and a useful and flexible tool for modeling.

There are two parallel descriptions of the Phase-type distributions. One corresponds to the obtained distribution from the absorption times in Continuous-time Markov chains (CTMC) over the interval$(0\ ,\infty)$, and the other one corresponds to the distribution on the nonnegative integers resulted from Discrete-time Markov chains (8). In this study, Discrete Phase-type distribution (DPTD) was used.

Phase-type Distribution

If the Markov process$\{X_t\}_{t\geq 0}$be considered on the state space S={1,2,...,p,p+1}, in a way that 1,2,...,p and P+1 determine the transitional and absorption phases, respectively; then, for this process, a matrix will be defined as the intensity matrix ($d = \begin{pmatrix} T & t \\ 0 & 0 \end{pmatrix}$), where T, that is defined as the sub-intensity matrix, determines the inter-phase transition rate, and ‘t’ is the exit rate vector from each phase to the absorption phase and it is p × 1 that is defined as t= e–T.e, where e is a p-dimensional column vector.

If $\pi_i = p(X_0 = i)$, the probability that it be in stage i when time=0 and $\pi = (\pi_1, \dots, \pi_p)$ shows the initial distribution of $\{X_t\}_{t\geq 0}$on transition phase in a way that $\sum_{i=1}^{p} \pi_i = 1$, and $\tau = \inf\{t > 0 | X_t = p + 1\}$ shows the time of absorption phase; therefore, distribution only depends on π and T, and as a result, τ would have discrete Phase-type distribution which is shown as $\tau \sim PH(\pi, T)$ and its probability density function is as follows (9,10):

$$f(x) = \pi T^{x-1} t$$

Variables

Independent variables defined in this study include gender, age, splenomegaly (presence or absence), hepatomegaly (presence or absence), CBC tests in each phase, treatment type (chemotherapy, radiotherapy, chemoradiotherapy), the interval between admission and discharge, brain metastasis (yes or no), testis or ovary metastasis (yes or no).

Response variable: In this study, for modeling the intended response variable is actually the treatment outcomes (recovery or death).

In order to record the variables, a check list including the patient’s file number, cancer diagnosis date, discharge date, date of brain metastasis diagnosis, date of testis/ovary metastasis diagnosis, gender, age, blood group, RH, diagnosis of cancer type (AML or ALL), splenomegaly (presence or absence), hepatomegaly (presence or absence), treatment type, CBC tests in each phase, and treatment result was used.

Fitting Phase-type Distribution

In Phase-type distribution, observation y that determines the time to absorption, can be considered as an incomplete observation of Markov process$\{X_t\}_{t\geq 0}$; because it merely defines the time to absorption, and does not provide information about which phase was the starting point, to which phases it had reached during the process up to absorption, and how long it had stayed in each phase. Therefore, to estimate the parameters of this distribution, EM algorithm is used (11).

EM algorithm is a broadly applicable approach for iterative computing maximum likelihood estimates that can be used for solving the problems of the studies with incomplete data and it is simpler than Newton-Raphson method algorithm.

Each iteration of the EM algorithm is composed of two phases, Expectation (E) and Maximization (M).

This algorithm is not only used when data seem to be incomplete or missing, but also can be used where there are truncated distributions or censored observations and the data do not seem to be obviously incomplete. Therefore, application of EM algorithm has been developed in almost all fields where statistical techniques are applied.

At phase E, for an equation with incomplete data, using a set of observed data, data are produced for completing data and a value is also defined for parameters of the equation so that, the next stage computation is made easy. At phase M, by starting from the initial values of parameter, the next values would be defined for the parameters of distribution, and the phases are repeated until convergence to a fixed value (12).

Suppose there is M observation of $\mathrm{y}_1, \dots, \mathrm{y}_\mathrm{M} \in \mathbb{N}$ from distribution of $\mathrm{DPH}_\mathrm{p}(\pi, \mathrm{T})$ which are independent from one another and for everyy_k, there is $\mathrm{X}^{(\mathrm{k})} = \left(\mathrm{x}_0^{(\mathrm{k})}, \mathrm{x}_1^{(\mathrm{k})}, \dots, \mathrm{x}_{\mathrm{y}_\mathrm{k}}^{(\mathrm{k})}\right)$. If $\mathrm{X}^{(\mathrm{k})} = \left(\mathrm{x}_0^{(\mathrm{k})}, \mathrm{x}_1^{(\mathrm{k})}, \dots, \mathrm{x}_{\mathrm{y}_\mathrm{k}}^{(\mathrm{k})}\right)$ is the complete set of data and $\mathrm{y} = (\mathrm{y}_1, \dots, \mathrm{y}_\mathrm{M})$ defines incompletely observed data sets; then for $\theta = (\pi, \mathrm{T}, \mathrm{t})$, the likelihood function is:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{k=1}^{M} \boldsymbol{\pi}\mathbf{T}^{y_k - 1}\mathbf{t}$$

Therefore, the logarithm of relation (2) is demonstrated as:

$$l(\theta; y) = \sum\nolimits_{k=1}^{M} \log f(y_k)$$

Substituting

$$\pi = \textstyle\sum_{s=1}^{p-1} \pi_s e_s' + \left(1 - \sum_{s=1}^{p-1} \pi_s\right) e_p'$$

into the density function $[\mathrm{f}(\mathrm{y}_\mathrm{k}) = \pi \mathrm{T}^{\mathrm{y}_\mathrm{k} - 1}\mathrm{t}]$, the following relation is obtained:

$$\mathrm{f}(\mathrm{y}_\mathrm{k}) = \sum\nolimits_{s=1}^{p-1} \pi_s \mathbf{e}_s' \mathbf{T}^{y_k - 1}\mathbf{t} - \left(1 - \sum\nolimits_{s=1}^{p-1} \pi_s\right) \mathbf{e}_p' \mathbf{T}^{y_k - 1}\mathbf{t}$$

If Markov chains data be obtained from $\mathrm{X}^* \in \left\{\mathrm{X}^{(\mathrm{k})}\right\}_{\mathrm{k}=1,\dots,\mathrm{M}}$, by putting relation 3 in the likelihood function (multiplying the density functions) and simplifying it, the likelihood function for discrete Phase-type distribution will be as follows:

$$L_f(\boldsymbol{\theta}; \mathrm{X}^*) = \prod_{i=1}^{p} \pi_i^{B_i} \prod_{i=1}^{p} \prod_{j=1}^{p} t_{ij}^{N_{ij}} \prod_{i=1}^{p} t_i^{N_i}$$

In this relation, Bi refers to the number of initiating processes in state i, Ni refers to the number of transmission processes from mode i to absorption mode, and Nij refers to the number of jumps from state i to state j between all processes.

The logarithm of likelihood function which is obtained from relation 4 will be as follows:

$$\boldsymbol{l_f(\theta;X^*)=\sum_{i=1}^{p}B_i\log(\pi_i)+\sum_{i=1}^{p}\prod_{j=1}^{p}N_{ij}\log(t_{ij})+\sum_{i=1}^{p}N_i\log(t_i)}$$

As there are M independent series from the observations, therefore:

$$B_i=\sum_{k=1}^{M}B_i^k \quad , \quad N_i=\sum_{k=1}^{M}N_i^k \quad , \quad N_{ij}=\sum_{k=1}^{M}N_{ij}^k$$

where $\mathrm{B_i^k}$,$\mathrm{N_i^k}$, and $\mathrm{N_{ij}^k}$are the corresponding components of K observation (13):

$$B_i=\sum_{k=1}^{M}B_i^k \quad , \quad N_i=\sum_{k=1}^{M}N_i^k \quad , \quad N_{ij}=\sum_{k=1}^{M}N_{ij}^k$$

Relation 5 is simply maximized by using Lagrange coefficients.

$$\mathbf{LA}\left(\pi_i,t_{ij},t_i,\lambda_1,\lambda_2\right)=l_f(\theta;X^*)+\lambda_1(1-\sum_{j=1}^{p}t_{ij}-t_i)+\lambda_2(1-\sum_{i=1}^{p}\pi_i)$$

Finally, by differentiating of the aforementioned relation and solving it, the following formulas would be obtained to determine the initial values of the parameters.

$$\widehat{\mathbf{t_\imath}}=\frac{\mathbf{N_i}}{\sum_{j=1}^{p}\mathbf{N_{ij}}+\mathbf{N_i}},\widehat{\boldsymbol{t_{\imath\jmath}}}=\frac{\boldsymbol{N_{ij}}}{\sum_{s=1}^{p}\boldsymbol{N_{is}}+\boldsymbol{N_i}},\widehat{\boldsymbol{\pi_\imath}}=\frac{\boldsymbol{B_i}}{\boldsymbol{\lambda_2}}=\frac{\boldsymbol{B_i}}{\boldsymbol{M}}$$

Stage E in EM algorithm, is calculation of expectation. Accordingly, conditional expectations are calculated as follows:

$$E_{\boldsymbol{\theta}}\left(\mathrm{B_i^k}\middle|\mathrm{Y_k}=\mathrm{y_k}\right)=\frac{\mathbf{e}_{\mathrm{i}}^{'}\mathbf{T}^{\mathrm{y_k}-1}\mathbf{t}}{\pi\mathbf{T}^{\mathrm{y_k}-1}\mathbf{t}}\pi_{\mathrm{i}}$$

$$E_{\boldsymbol{\theta}}\left(N_{ij}^k\middle|\mathrm{Y_k}=\mathrm{y_k}\right)=\ \mathbf{1}_{\{y_k\geq2\}}\sum\nolimits_{l=0}^{y_k-2}\frac{\mathbf{e}_j^{'}\mathbf{T}^{(y_k-(l+1)-1)}\mathbf{t}\boldsymbol{\pi}\mathbf{T}^{l}\mathbf{e}_i}{\boldsymbol{\pi}\mathbf{T}^{y_k-1}\mathbf{t}}t_{ij}$$

$$\mathrm{E}_{\boldsymbol{\theta}}\left(\mathrm{N_i^k}\middle|\mathrm{Y_k}=\mathrm{y_k}\right)=\frac{\boldsymbol{\pi}\mathbf{T}^{\mathrm{y_k}-1}\mathbf{e}_{\mathrm{i}}}{\boldsymbol{\pi}\mathbf{T}^{\mathrm{y_k}-1}\mathbf{t}}\mathrm{t_i}$$

The next stage is maximization in which new values of the parameters should be determined so that the values eventually reach a convergent value as a final value of the parameter.

The expectation values for all observations$(\mathrm{y_k}\,;\mathrm{k}=1,\dots,\mathrm{M})$ are calculated based on the relation 7 and the results of their sum are considered as new values for $\mathrm{B_i}$ ،$\mathrm{N_{ij}}$ و $\mathrm{N_i}$ and then put in relation 6 and subsequently a new value for vector t, matrix T , and vector π is obtained. This process is carried on until a convergent value for parameters t, T, and π, which is used in Phase-type distribution, be obtained.

In summary, for fitting Phase-type distribution, EM algorithm functions as following stages:

1- Initializes the parameters of $\boldsymbol{\theta}_0=(\boldsymbol{\pi}_0,\mathbf{T}_0,\mathbf{t}_0)$
2- Finds $\sum_{k=1}^{M}E_{\boldsymbol{\theta}_0}\left(B_i^k\middle|Y_k=y_k\right)$, $\sum_{k=1}^{M}E_{\boldsymbol{\theta}_0}(N_{ij}^k\,|Y_k=y_k)$, and $\sum_{k=1}^{M}E_{\boldsymbol{\theta}_0}\left(N_i^k\middle|Y_k=y_k\right)$
3- Calculates $\widehat{\boldsymbol{\theta}}=\left(\widehat{\boldsymbol{\pi}},\widehat{\mathbf{T}},\widehat{\mathbf{t}}\right)$

4. $\widehat{\theta}$ is placed for θ_0 and goes to the second stage (11, 13)

In this study, Phase-type distribution was fitted and data analysis performed using SPSS 22 and R. Since, there was no programmed instruction for fitting Phase-type distribution and determining density function in R software, therefore, all instructions were programmed.

Results

To collect data, firstly it was necessary to specify the processes and then determine Bi, Ni, and Nij based on the processes. Therefore, 9 processes can be considered for this study based on the Figure1.

Process 1: A → D

Process 2: B → D

Process 3: C → D

Process 4: A → B → D

Process 5: A → B → C → D

Process 6: A → C → D

Process 7: B → C → D

Process 8: A → C → B → D

Process 9: C → B → D

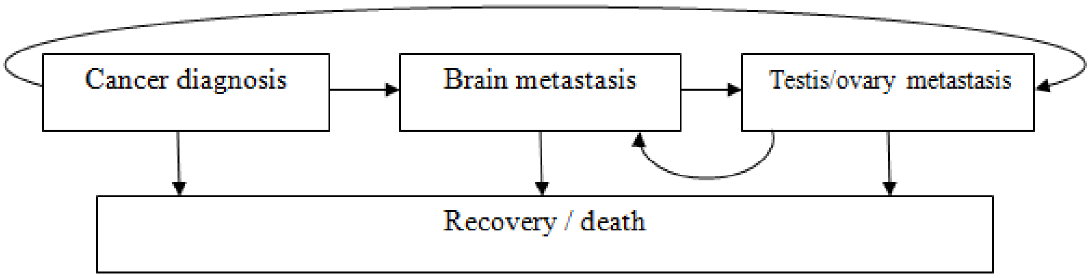

Figure 1. The phases and transmissions considered for the Phase-type distribution of data

Then, data related to the patients who had died were differentiated from data related to those who had recovered, and Phase-type distribution was separately fitted for these two groups by using the EM algorithm.

The results showed that 45 patients had died (M=45), among them 35 patients had passed the process 1 to reach absorption phase (death) and the number of patients who had reached to the 2nd to 9th processes were respectively 0, 1, 5, 1, 1, 0, 2, 0. Therefore, the values of B_i ،N_i, and N_{ij} were determined.

B1 refers to the number of processes initiating from phase 1 (A). It is clear that processes 1,4,5,6,8 are initiated from phase 1. As a result, B1= 35+5+1+1+2=44, B2= 0, and B3= 1+0= 1.

When B_is was determined, N_1, N_2 and N_3 showing respectively the number of transmissions from phase 1,2, and 3 to the absorption phase, were determined.

Considering all the processes, only the first process indicates transmission from process 1 to the absorption phase and in the died group, 35 patients after diagnosis phase had died (N_1=35). The processes 2,4,8,9 show the transmission from phase 2 to the absorption phase. Accordingly, 7 patients had died after brain metastasis (N_2=7). The processes 3,5,6,7 show transmission from phase 3 to the absorption phase (death), and 3 patients passed these processes and had died after metastasis of testis/ovary (N_3=3).

N_{12}, indicating the number of jumps or transmission from phase 1 to 2, is obtained by adding the number of patients passed the processes 4 and 5 (N_{12}=6).

N_{13} is obtained by adding the number of patients who passed the processes 6 and 8 (N_{13}=3).

As shown in the Figure1, there is no transmission from phase 2 to 1 and also from phase 3 to 1; therefore, N_{21} and N_{31} =zero. Accordingly, there is also no interphase transmission (N_{11}, N_{22}, N_{33}=0). N_{32} shows transmission from phase 3 to 2 and is determined according to the processes 8 and 9 (N_{32}=2 and N_{23}=1).

After determination of N_i, N_{ij}, and B_i, initial values of the parameters were determined by using relation 6. After four times repetition of the phases, the algorithm reached a convergent value and after higher repetitions, the results obtained for *π*, *t*, and *T* were the same. And finally:

$$\pi = (0.97777778, 0.00000000, 0.02222222)$$

$$t = \begin{bmatrix} 6.949164e-05 \\ 1.250780e-04 \\ 2.680506e-05 \end{bmatrix} \qquad T = \begin{bmatrix} 0 & 0.6666203 & 0.3333102 \\ 0 & 0.0000000 & 0.9998749 \\ 0 & 0.9999732 & 0.0000000 \end{bmatrix}$$

Then, by putting these parameters in the formula of Phase-type distribution, the values of the probability density function for all the 45 observations were determined. It is worth mentioning that x in this function, shows the interval between admission and absorption phase that is shown in terms of years (Table 1).

Table 1. Observations of time to absorption (death) and their correspondent values of the probability density function

x	f(x)	x	f(x)	x	f(x)
6.333	1.22E-05	0.167	0.000139	0.0833	0.000145
4.917	1.96E-05	0.167	0.000139	0.0833	0.000145
5.667	1.52E-05	0.003	0.000148	0.75	0.000105
1.417	7.74E-05	1.083	8.97E-05	0.5833	0.000113
4.583	0.000022	0.667	0.000109	0.0008	0.000151
0.25	0.000133	0.333	0.000128	0.3333	0.000128
1	0.000093	3.583	3.19E-05	0.0004	0.000151
0.833	0.000101	0.25	0.000133	0.0833	0.000145
0.5	0.000118	1.75	6.68E-05	0.0833	0.000145
0.167	0.000139	0.333	0.000128	0.1667	0.000139
0.083	0.000145	1.667	6.94E-05	0.0002	0.000151
2.583	4.73E-05	0.006	0.000145	0.0833	0.000145
0.5	0.000118	0.583	0.000113	0.0004	0.000151
0.917	0.000097	0.25	0.000133	0.0008	0.000151
0.5	0.000118	1.083	8.97E-05	0.0002	0.000151

The results show that increase in intervals between admission and absorption phase (death), decreases the probability of death. In other words, a patient who undergoes a long treatment period is less likely to die.

Figure 1 shows the probability density function for observations of time to absorption for these patients.

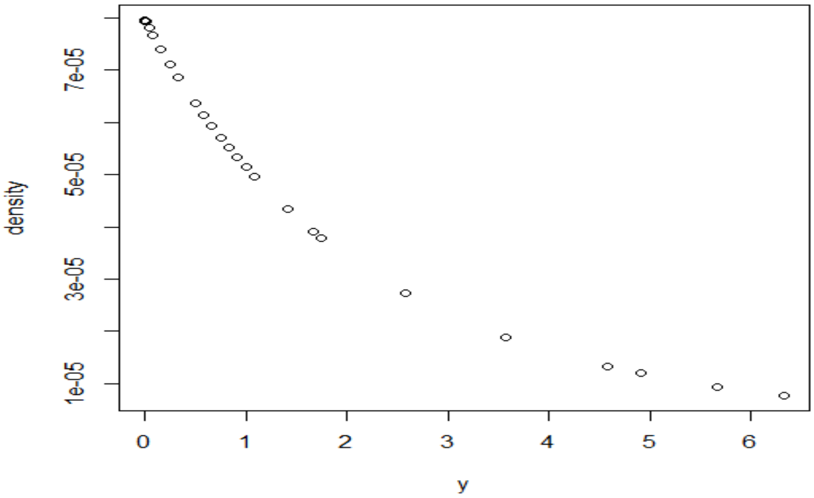

Figure 1. The values of Phase-type probability density function for time observations to absorption for the dead group

Then, the relationship between the absorption probabilities and other independent variables was measured and based on the type of the variable, t-test, Pearson correlation coefficient and one way analysis of variance (ANOVA) were used for data analysis.

For the variables including gender, diagnosis of cancer type (AML or ALL), splenomegaly (presence or absence) and hepatomegaly (presence or absence), brain/testis/ovary metastasis (yes or no), which are dual-mode qualitative variables, t-test was used, and p-values of each test are shown in Table 2.

Table 2. The relationship between dual-mode qualitative variables and the probability of death

Variables	Testis/ovary Metastasis	Brain Metastasis	Hepatomegaly	Splenomegaly	Diagnosis	Sex	RH
P-value	0.002	0.002	0.84	0.977	<0.0001	0.565	0.43

The results show that there is a significant difference in the probability of death between AML and ALL groups. In addition, there is a meaningful relationship between the probability of death and brain/testis/ovary metastasis.

The relationship between death probability and patient's age and also WBC, RBC, HGB, HCT, MCV, MCH, MCHC, and PLT defined in CBC test taken during the diagnosis of cancer, was analyzed using Pierson's corelation coefficinet. The results are shown in Table 3.

Table 3. The relationship between probability of death and independent quantitative variables

	age	MCH	MCV	HCT	HGB	RBC	WBC	MCHC	PLT
Regression Coefficientf(x)	-0.191	0.05-	0.16-	0.14-	0.14-	0.11-	0.099	-0.038	-0.28
P-value	0.208	0.725	0.917	0.335	0.355	0.448	0.517	0.8	0.06

As it is seen, the probability of death has a direct relationship with the number of white blood cells (WBCs) determined by CBC test, but it has a reverse relationship with age and other variables determined in this test. Considering p-values, there is no linear relationship between the probability of death and other quantitative variables.

The relationship between the probability of death and the blood group and also the type of treatment was analyzed using One-way analysis of variance (ANOVA). The results show that there is no significant difference between the probability of death and different treatments (p= 0.145), but there is significant difference between the probability of death and blood group (at least for two blood groups, p=0.025).

The same processes were performed for the second group and the values of probability density function for 132 recovered patients were determined (Table 4).

Figure 2 shows the probability density function for observations of time to absorption for these patients.

Table 4. Observations of time to absorption (recovery) and their correspondent values of the probability density function

x	f(x)	x	f(x)	x	f(x)	x	f(x)	x	f(x)	x	f(x)
3.42	0.09051	3.42	0.09051	3.3	0.10113	3	0.12027	3.5	0.08504	2.58	0.16091
3.33	0.09568	2.83	0.13531	3.3	0.09568	2.5	0.17008	3.67	0.07611	2.58	0.16091
3.17	0.10764	2.75	0.14302	3.3	0.09568	3.3	0.09568	3.17	0.10764	2	0.24053
10	0.00094	5.17	0.02691	3.3	0.10113	3.8	0.07151	3.42	0.09051	2.42	0.18103
3.08	0.11378	3.08	0.11378	3.3	0.09568	2.2	0.21528	3.08	0.11378	1.83	0.27061
3	0.12027	3.25	0.10113	3.3	0.09568	1.8	0.27061	3.25	0.10113	2.17	0.21528
2.75	0.14302	2.33	0.19135	2.3	0.19135	4.5	0.04252	3.25	0.10113	2	0.24053
2.5	0.17008	5.33	0.02392	3.1	0.11378	4.5	0.04252	3.08	0.11378	2	0.24053
2.75	0.14302	2	0.24053	4.3	0.05057	3	0.12027	3.17	0.10764	2	0.24053
3.33	0.09568	4.08	0.05689	3.3	0.10113	3.3	0.10113	2.5	0.17008	1.83	0.27061
2.08	0.22756	3.58	0.08045	3.2	0.10764	2.5	0.17008	3.17	0.10764	1.75	0.28604
3.33	0.09568	5	0.03007	3.1	0.11378	3.7	0.07611	3.17	0.10764	3.33	0.09568
3.17	0.10764	1.92	0.25601	2.8	0.13531	2.6	0.16091	2.25	0.20226	1.42	0.36206
2.83	0.13531	3.33	0.09568	3	0.12027	3.2	0.10764	2.92	0.12801	1.25	0.40452
2.83	0.13531	3.33	0.0957	3	0.12027	3.17	0.1076	2.92	0.128	1.25	0.4045
8.08	0.00356	3.33	0.0957	2.08	0.22756	3.33	0.0957	3	0.1203	2.33	0.1914
6.08	0.01422	3.33	0.0957	3.33	0.09568	2.17	0.2153	3	0.1203	2.58	0.1609
4	0.06013	3.42	0.0905	4.92	0.032	2.08	0.2276	2.92	0.128	6.25	0.0126
3.33	0.09568	2.58	0.1609	3.25	0.10113	2.25	0.2023	2.83	0.1353	3.33	0.0957
2.5	0.17008	6	0.015	2.33	0.19135	2.67	0.1522	2.83	0.1353	4.5	0.0425
5.58	0.02011	2.25	0.2023	3.33	0.09568	2.33	0.1914	2.5	0.1701	2.5	0.1701
3.33	0.09568	2.67	0.1522	3.17	0.10764	3.58	0.0805	2.58	0.1609	2.08	0.2276

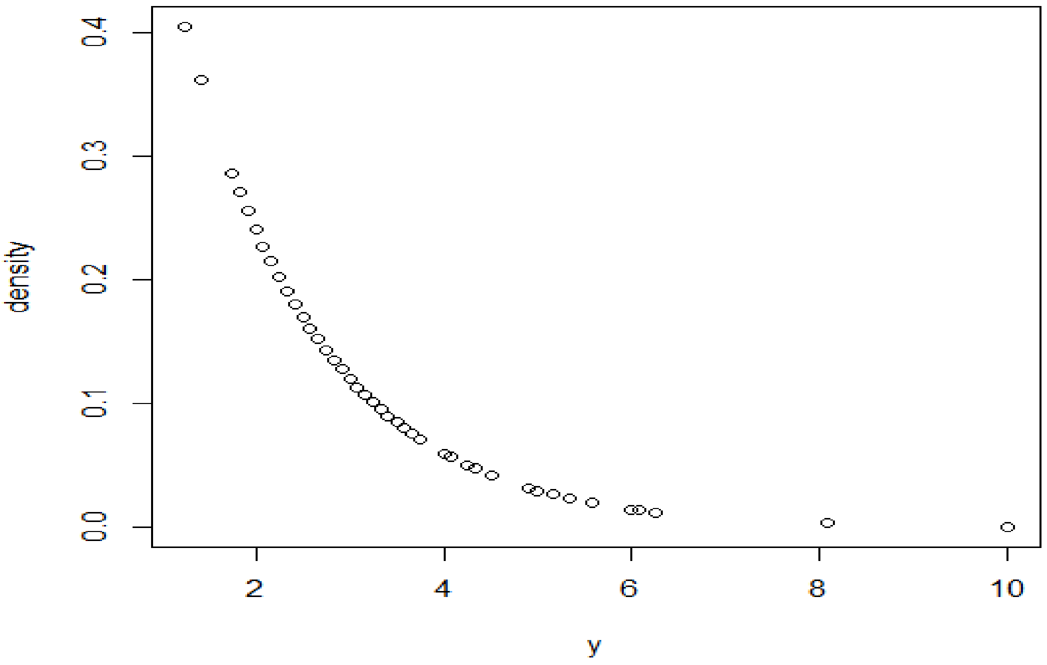

Figure 2. The values of Phase-type probability density function for time observations to absorption for the recovered group

Analysis of these probabilities shows that decrease in treatment period associates with increase in the probability of recovery and vice versa.

The relatinship between the probability of absorption (recovery) and independant dual-mode qualitative variables such as gender, diagnosis of the cancer type, splenomegaly (presence or absence) and hepanomegaly (presence or absence) was analyzed using t-test (Table 5).

Table 5. The relationship between dual-mode qualitative variables and the probability of recovery

Variables	RH	Metastasis of testis/ovary	Brain Metastasis	Hepatomegaly	Splenomegaly	Diagnosis	Sex
P-value	0.271	0.203	0.111	0.45	0.115	0.669	0.525

As shown inTable 5, there is no meaningful relationship between the probability of recovery and these independant dual-mode qualitative variables.

Pearson’s corelation coefficinet was used to determine the relationship between recovery probability and independant quantitative variables such as patients’ age and also WBC, RBC, HGB, HCT, MCV, MCH, MCHC, and PLT defined in CBC test that were taken during the cancer diagnosis (Table 6).

Table 6. The relationship between the probability of recovery and independent quantitative variables

	age	MCH	MCV	HCT	HGB	RBC	WBC	MCHC	PLT
Regression Coeficient f(x)	0.073	-0.058	0.025-	-0.067	-0.071	0.017-	0.106	-0.057	-0.012
P-value	0.405	0.511	0.776	0.448	0.418	0.849	0.225	0.513	0.894

As shown in Table 6, there is a direct relationship between the probability of recovery and variables of age and WBC. Considering p-values, there is no linear relationship between the probability of recovery and the mentioned variables.

The relationship of the probability of recovery with the blood group and treatment was analyzed using One-way analysis of variance (ANOVA) and according to the obtained results, the null hypothesis (H0) is not rejected and there is no significant relationship between recovery and variables of blood group and type of treatment (p=0.708 and p=0.278 respectively).

Discussion and Conclusion

Phase-type distribution can be used in modelling diseases like cancer, diabetes, AIDS, renal diseases and cardiovascular diseases which develop from the dignosis and initial stages to the advanced stages and also to study the the effect of different factors on the development and progress of disease.

Since there might be a censorship or an interval in different stages of the diseases, using statistical models such as regression for modeling is difficult and Phase-type distribution can be an effective method. In such situations in which there is many censorship and missed data are not clear, EM algorithm that is used for fitting Phase-type distribution can be used.

In addition, in such diseases, using Cox proportional hazards model, only one outcome can be studied but using multi-mode models, several outcomes can be simultaneously investigated and also the effects of the intermediate variables on the outcomes can be measured.

The aim of this study was to model Leukemia in children by using Phase-type distribution. In this study, medical records of 177 patients with Leukemia were investigated (ALL=149 and AML=28). From these patients, 132 ones had been recovered and 45 patients had died.

For fitting Phase-type distribution, the required data were extracted from the designed check list; for example, the interval between the diagnosis of disease and the end of treatment or death, as the time to absorption (x), was put in the probability density function [f(x)] to find the absorption probabilities.

In the dead group, increase of the interval between admission and absorption phase (death) was associated with decrease of the probability of death, which is cosistent with the expected results and indicates that Phase-type distribution has been properly fitted. The analysis of the recovered group shows that decrease in treatment period increases the probability of recovery.

There was a significant difference between two different diagnostic groups (ALL and AML) in regard to the probability of death and in patients with ALL the probability of death was more than those with AML; while according to the previous clinical trials and intuitively, the probability of death in patients with AML is more than that in ALL group.

It was also revealed that there is significant difference between the probability of death in patients with brain/testis/ovary metastasis compared to the patients without metastasis that is clinically justifiable.

There was no significant relationship between the probability of death and other variables such as gender, spelnomegaly or hepatomegaly (presence or absence), RH, CBC test, age, and blood group.

However, in the study of Ali Mohammad Zand et al (2009), it was reported that AML and ALL are more common in blood group O and some how in blood group A and also the risk of cancer in men is 1.5 times higher than that in women (2).

In the recovered group, there was no significant relationship between the probability of recovery and the other independant variables. After fitting Phase-type distribution and analyzing the values of absorption probabilities, it was revealed that this distribution presents considerable results and accordingly, it can be a useful method for modeling cancers.

Using this method, from primary estimation of treatment period, the probability of recovery can be estimated. By analyzing some variables such as the type of treatment, it can be predicted which treatment would have positive effect on the probability of recovery and consequently, according to the probabilities of recovery or death, the best decision can be made.

However, further studies are needed to investigate the effects of other variables and factors on the probabilities of absorption. In addition,besides the considered metastases, other possible problems in other parts of the body can be considered as different phases.

In this study, only EM algorithm was used for fitting Phase-type distribution, but it is better to analyze and compare other methods of parameter estimation for this distribution and specify which one can present more acceptable results at the end.

As there are many diseases for which different phases and stages can be considered for controlling their progress, it is better to use Phase-type distribution for modeling such diseases in order to obtain more exact understanding of their progress.

References

1. Mashhadi MA, Zakeri Z, Abdollahinejad M. Cancer incidence in South East of Iran: results of a population-based cancer registry. *Shiraz E Medical Journal* 2010; 11(3): PP 148-55.
2. Zand A, Sa'adati M, Borna H, Ziaei R, Honari H. Effect of age, gender and blood group on blood cancer types. *Kowsar Medical Journal* 2010; 15: 111-4 [In Persian].
3. Lafzi A, Asvadi Kermani I, Ghanbari HA, Raadi E. Clinical comparison of the prevalence and type of the periodontal diseases in leukemic patients hospitalized in Tabriz through 2004-2005. *Journal of Mashhad Dental School* 2005; 29(1-2): 123-30.
4. Khalkhali H, Kazemnejad A, Ghafari Moghadam A. Prediction of chronic renal failure in patients with impaired renal function. *Journal of Epidemiology* 2005; 6: 25-31.
5. Garg L, Masala G, McClean SI, Micocci M, Cannas G. Using phase type distributions for modelling HIV disease progression. 25th International symposium on Computer-Based Medical Systems (CBMS), 2012;.
6. Bladt M, Nielsen B.F. Lecture notes on phase type distributions. 2008.
7. Zare A, Mahmoudi M, Mohammad K, Zeraati H, Hosseini H, Naeini KH. Assessing Misdiagnosis of Relapse in Patients with Gastric Cancer in Iran

Cancer Institute Based on a Hidden Markov Multi-state Model. *Asian Pac j cancer prev: APJCP*, 2014; 15(9): 4109-15.

8. Neuts M.F, Pagano ME. Generating random variates from a distribution of phase type. in Proceedings of the 13th conference on Winter simulation-Volume 2, NJ,USA 1981. IEEE Press, 1981.

9. McGrory C.A, Pettitt AN, Faddy MJ. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis* 2009; 53(12): 4311-21.

10. Bobbio A, Horvath A, Scarpa M, Telek M. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. *Performance evaluation* 2003; 54(1): 1-32.

11. Asmussen S, Nerman O, Olsson M. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* 1996 23(4): 419-41.

12. McLachlan G. Krishnan T. The EM algorithm and extensions. 2nd ed., John Wiley & Sons, 2007; P 382.

13. Esparza L.J.R., Nielsen B.F, Bladt M. Maximum likelihood estimation of phase-type distributions. Technical University of Denmark. 2011.