VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0038

Dipolar Tree Ensemble With and Without Adjustment to Competing Risks: Application to Medical Data

Małgorzata Krętowska[1]

[1] Faculty of Computer Science, Bialystok University of Technology, Poland

Abstract. The analysis of survival data often aims at the prediction of failure time distribution. In cases of competing risk events, the time distributions of more than one failure are under investigation. In this paper, the comparison of two approaches to analyzing survival data with competing risks is presented. The analyses are performed by use of an ensemble of dipolar trees with and without adjustment to competing risks.

Introduction

Collecting survival data, we are very often interested in the prediction of disease-free survival. In such investigations, the observation time for every patient may end because of an accident or the end of follow-up time. The patient may also be lost to follow-up for other reasons. In the first case, the exact time of failure occurrence is known for the patient. In the other two cases, the observation is finished without any event, therefore the exact time of failure occurrence is unknown and the observation is considered as censored. For such patients, we only know that the survival time was no less than the observation time.

In many studies, however, the interest focuses not only on disease-free survival, but also on the time distribution of a specified event occurrence (failure occurrence). If the collected data contain information about one event only, the analysis may be performed by use of 'classical' methods for survival analysis (Kalbfleisch, 1980; Marubini et al., 1995). Competing risks data is a special type of survival data, in which more than one type of failure is under investigation, (Putter et. al., 2007) and, as a result, their analysis requires more sophisticated methods (Pintilie, 2006).

In this paper, I would like to compare the distributions of failure time occurrence obtained by two approaches to analyzing competing risks. According to the first method, all types of failure are considered as separate

ISBN 978–83–7431–392–6 ISSN 0860-150X

failures, treating other cases as censored; according to the other, all investigated failures are taken together. A dipolar tree ensemble, proposed in (Kretowska, 2006), is used as a prediction tool in both cases. The difference is the way the dipolar trees are inducted. The use of data with competing risk events requires additional adjustments (Kretowska, 2012), which make the information of competing risk events possible.

Survival Data with Competing Risk Events

Survival data are represented by a set of observations, which, besides the values of covariates, also contains information about the time of occurrence of a specified event. The event, also called failure, may represent e.g. death or disease relapse. A learning sample L is defined as $L = (\boldsymbol{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, where $\boldsymbol{x}_i$ is an N-dimensional covariate vector describing the ith observation (patient), t_i is the follow-up time and $\delta_i \in \{0, 1\}$ indicates whether the failure has occurred. δ_i equal to 0 represents a censored observation – the observation with unknown failure occurrence and δ_i equal to 1 represents an uncensored observation.

In cases of survival data with competing risk events, a patient is at risk of p ($p > 1$) different types of failure (Figure 1). Assuming that the time of occurrence for the ith type of failure is T_j, we are interested only in the failure with the shortest time: $T = \min(T_1, T_2, \ldots, T_p)$. The learning sample L_{CR} for competing risk data is defined as $L_{CR} = (\boldsymbol{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, but unlike in cases of a single failure, t_i is the time to the first event observed and $\delta_i \in \{0, 1, \ldots, p\}$ indicates the type of failure. δ_i equal to 0 represents a censored observation, which means that for a given patient no failure has occurred.

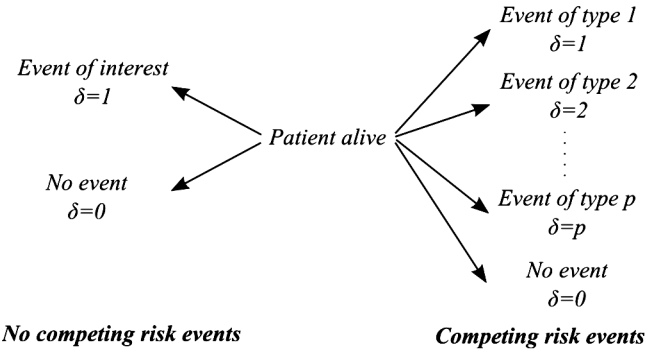

Figure 1. Survival data in cases of absence and presence of competing risk events

The distribution of survival time may be expressed by several functions. The most common approach is the use of the survival function, which presents the probability that the failure does not occur before time t:

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

The Kaplan-Meier estimator of the survival function is given as:

$$KM(t) = \prod_{j/t_{(j)} \leq t} \frac{n_j - d_j}{n_j}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered survival times from the learning sample L, in which the event of interest has occurred, d_j is the number of events at time $t_{(j)}$ and n_j is the number of patients at risk at $t_{(j)}$ (i.e., the number of patients who are alive at $t_{(j)}$ or experience the event of interest at $t_{(j)}$).

A cumulative incidence function (CIF) is used to describe the failure time distribution in cases of competing risks. The CIF for the ith type of event is defined as:

$$F_i(t) = P(T \leq t, \delta = i)$$

and may be interpreted as the probability that an event of type i occurs before or at time t. The estimate of CIF is given as:

$$\tilde{F}_i(t) = \sum_{j/t_{(j)} \leq t} \frac{d_{ij}}{n_j} \tilde{S}(t_{(j-1)})$$

where d_{ij} is the number of events of type i that occur at time $t_{(j)}$ and $\tilde{S}(t_{(j-1)})$ is the Kaplan-Meier estimator of the probability of being free of any event by time $t_{(j-1)}$.

When there are no competing risk events, the value of $(1 - KM(t))$ is equal to $\tilde{F}_1(t)$, in other cases $\tilde{F}_i(t) < 1 - KM_i(t)$, where $KM_i(t)$ is the Kaplan-Meier estimator calculated for the ith type of event. The value of $1 - KM_i(t)$ can be received from the equation (Pintilie, 2006):

$$1 - KM_i(t) = \sum_{j/t_{(j)} \leq t} \frac{d_{ij}}{n_j} KM_i(t_{(j-1)})$$

As we can see, the second part of the formulae, where instead of disease-free survival $\tilde{S}(t_{(j-1)})$ the $KM_i(t_{(j-1)})$ is present, causes the difference between the functions.

Ensemble of Dipolar Tree

An ensemble of dipolar trees (Kretowska, 2006; Kretowska, 2012) is used as a prediction tool. The appropriate construction of single trees enables the analysis of competing risks data as well as the data without competing risk events.

The ensemble is a set of single dipolar trees. Each tree is inducted on the base of a bootstrap sample, drawing with replacement from the learning data L, or in cases of competing risks, from the learning data L_{CR}. The tree induction starts from the root. Next, other internal nodes are created. Each internal node has two children nodes: internal or terminal ones. Internal nodes contain the test, which causes a vector of covariates to go to the appropriate child node. If certain conditions are fulfilled, the node is set as a terminal node, which does not contain any test, but may be viewed as a set of covariate vectors that have reached the node.

In the dipolar tree, the test in the kth internal node has a form of a hyperplane: $H(w_k) = \{\boldsymbol{x} : w_k^T x = 0\}$ and is constructed by minimizing the dipolar criterion function being a sum over some specified criterion functions connected with dipoles. The dipole (Bobrowski et. al., 1997) is a pair of different covariate vectors $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the learning set. Mixed and pure dipoles are distinguished. The pure dipoles are created between pairs that should not be separated, and the mixed ones between pairs that should belong to different groups. Depending on the problem, the dipoles are created in a different manner. In cases of data with no competing risk events, we are interested in dividing the feature space into areas with homogeneous survival times, so the vectors with similar failure times constitute the pure dipoles, and the vectors with distant failure times constitute the mixed dipoles (Kretowska, 2006):

– a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ constitutes the pure dipole, if
$$\delta_i = \delta_j = 1 \wedge |t_i - t_j| < \eta$$
– a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ constitutes the mixed dipole, if
$$\delta_i = \delta_j = 1 \wedge |t_i - t_j| > \zeta$$
$$\delta_i = 0, \delta_j = 1 \wedge t_i - t_j > \zeta \text{ or } \delta_i = 1, \delta_j = 0 \wedge t_j - t_i > \zeta$$

For data with competing risk events, the analysis aims at dividing the feature space into such areas, which would include the patients with the same cases of failure and similar survival times. Taking into account censored cases, the following rules of dipole construction can be formulated (Kretowska, 2012):

– a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ forms the pure dipole, if
$$\delta_i = \delta_j = z \wedge |t_i - t_j| < \eta_z, z = 1, 2, \ldots, p$$

– a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ forms the mixed dipole, if
$\delta_i = \delta_j = z \wedge |t_i - t_j| > \zeta_z, z = 1, 2, \ldots, p$
$(\delta_i = 0, \delta_j = z \wedge t_i - t_j > \zeta_z)$ or $(\delta_i = z, \delta_j = 0 \wedge t_j - t_i > \zeta_z), z = 1, 2, \ldots, p$

Parameters η_z and ζ_z are equal to quartiles of absolute values of differences between uncensored survival times for the zth type of failure, $z = 1, 2, \ldots, p$. Based on earlier experiments, the parameter η_z is fixed as 0.3 quantile and ζ_z – 0.6. Parameters η and ζ are calculated in a similar manner, without taking into account the type of failure.

The general algorithms for generating the ensemble of dipolar trees is as follows:

1. Draw k bootstrap samples $(L_1; L_2; \ldots; L_k)$ of size n with replacement from L or in cases of data with competing risks from L_{CR}.
2. Induction of dipolar survival tree $T(L_i)$ based on each bootstrap sample L_i.
3. For each tree $T(L_i)$, distinguish the set of observations $L_i(\boldsymbol{x}_n)$ which belongs to the same terminal node as $\boldsymbol{x}_n$.
4. Build an aggregated sample $L_A(\boldsymbol{x}_n) = [L_1(\boldsymbol{x}_n), L_2(\boldsymbol{x}_n); \ldots; L_k(\boldsymbol{x}_n)]$.
5. Calculate the Kaplan-Meier aggregated survival function for a new observation $\boldsymbol{x}_n$ as $KM_A(t/\boldsymbol{x}_n)$.
6. In cases of no competing risk events, calculate $1 - KM_A(t/\boldsymbol{x}_n)$; for competing risks calculate the aggregated CIF for all types of failure for a new observation $\boldsymbol{x}_n$: $\tilde{F}_i(t/\boldsymbol{x}_n)$, for $i = 1, 2, \ldots, p$.

The output of the ensemble for a new observation is, depending on the problem, the aggregated Kaplan-Meier function (and then $1 - KM(t)$) or the estimator of CIF. In both cases, some other statistics may also be calculated. Median value or lower and upper quartile belong to the most common statistics, but other quantiles may also be calculated for obtained functions.

Experimental Results

The experiments were performed on two datasets: breast cancer data and follicular type lymphoma data, both with competing risk events. For each dataset, two types of tests were done:

1. For whole data sets, with competing risks events (CR).
2. Without competing risks (nCR). For each basic dataset, p separate datasets were created, each containing the information of one distinguished event, treating other events as censored observations.

All the experiments were conducted using the ensemble of 100 single survival trees.

Breast cancer data (Fyles et al., 2004) contained information about 641 women (50 years old or older) who had undergone breast-conserving surgery for an invasive adenocarcinoma 5 cm or less in diameter. They were randomly assigned to receive breast irradiation plus tamoxifen (321 women) or tamoxifen alone (320 women). The data were collected between 1992 and 2000. The last follow-up was conducted in summer 2002. Table 1 contains a description of the variables (Ibrahim et al., 2008), while Table 2 presents characteristics of each variable – for discrete variables – the number of cases having the same value; for continuous ones – minimum, maximum, lower (Q_1) and upper (Q_3) quartile and median values.

Table 1. Description of variables in breast cancer dataset

Variable name	Description
Tx	Randomized treatment: 1 = tamoxifen, 2 = radiation + tamoxifen
Variables assessed at the time of randomization	
Pathsize	Size of tumor (cm)
Hist	1 = ductal, 2 = lobular, 3 = medullary, 4 = mixed, 5 = other
Hrlevel	0 = negative, 1 = positive
Hgb	Hemoglobin (g/l)
Nodediss	Whether axillary node dissection was done: 0 = no, 1 = yes
Age	Age (years)
Outcome variables	
Time	Time from randomization to event or last follow up (years)
d	Status at last follow up: 0 = censored, 1 = death, 2 = relapse, 3 = malignancy

Table 2. Characteristics of variables in breast cancer dataset

Discrete variables	Value (number of cases)
Tx	1 (321); 2 (320)
Hist	1 (397); 2 (31); 3 (5); 4 (174); 5 (34)
Hrlevel	0 (46); 1 (595)
Nodediss	0 (106); 1 (535)
D	0 (503); 1 (14); 2 (69); 3 (55)
Continuous variables	
Path size	Min = 0.2; Q_1 = 1; Med = 1.5; Q_3 = 2; Max = 4.5
Hgb	Min = 96; Q_1 = 128; Med = 135; Q_3 = 142; Max = 169
Age	Min = 50; Q_1 = 59; Med = 67; Q_3 = 73; Max = 88

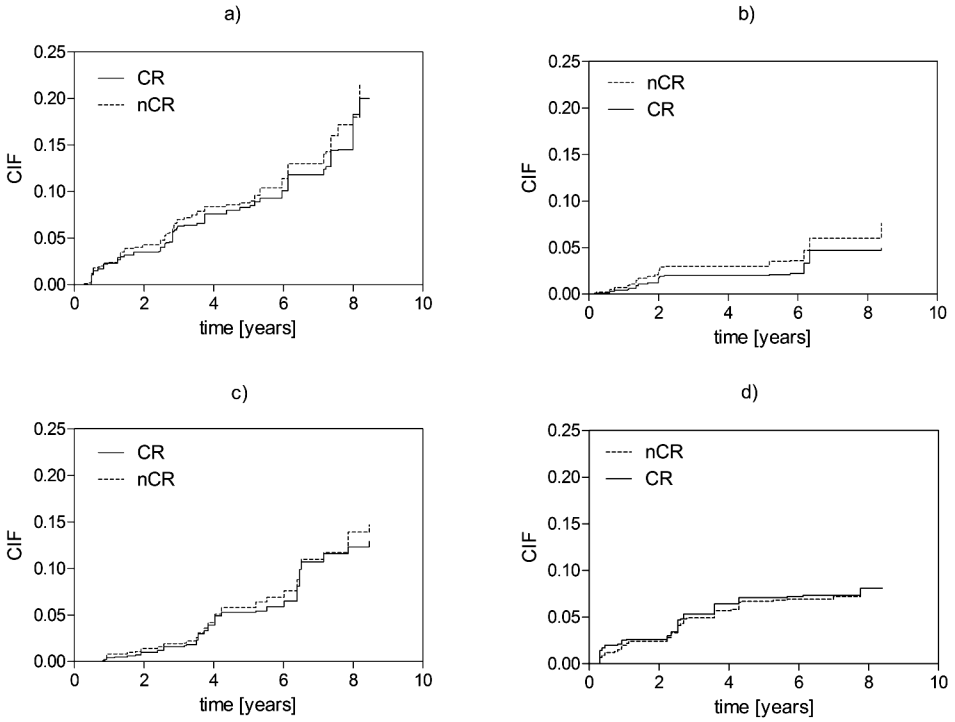

Figure 2. Cumulative incidence functions received for breast cancer data for: a) relapse, Tx = 1; b) relapse, Tx = 2; c) malignancy, Tx = 1; d) malignancy, Tx = 2

Cumulative incidence functions received for breast cancer data are presented in Figure 2. Each graph contains two cumulative incidence functions received for the data with and without competing risks. Figures 2(a) and 2(b) contain CIF's for relapse, while Figures 2(c) and 2(d) – CIF's for malignancy with hist = 1, hrlevel = 1, nodediss = 1, and values of continuous features fixed as their medians.

There are small differences between the two received functions in Figures 2(a) and 2(c) calculated for patients treated with tamoxifen alone, and bigger for the other treatment (Figures 2(c) and 2(d)). The differences may be caused by the previously described relationship between the two functions: $\tilde{F}_i(t) \leq 1 - KM_i(t)$ or by different division of feature space in the two presented approaches, described in "Ensemble of Dipolar Tree" section. The latter is especially visible in Figure 2(d) where $\tilde{F}_i(t)$ is usually greater than $1 - KM_i(t)$.

A Lymphoma patient dataset was created at Princess Margaret Hospital, Toronto (Pintilie, 2006). In the experiments we used the subset

of 541 patients who had follicular type lymphoma, registered for treatment at the hospital between 1967 and 1996, with early stages of disease (I or II) and treated with radiation alone (118 people) or with radiation and chemotherapy (179 people). Each patient was described by four variables, explained in Table 3 and characterized in Table 4. The event of interest was failure from the disease: no response to treatment or relapse. A competing risk type of event was death without failure. There are 272 events of interest and 76 observations of death without relapse.

Table 3. Description of variables in lymphoma patient dataset

Variable name	Description
Variables assessed at the time of diagnosis	
Age	Age (years)
Hgb	Hemoglobin (g/l)
Clinstg	Clinical stage: 1 = stage I; 2 = stage II
Ch	Chemotherapy: 0 = no; 1 = yes
Outcome variables	
Time	Time from diagnosis to event or last follow up (years)
D	Status at last follow up: 0 = censored, 1 = no response to treatment or relapse, 2 = death

Table 4. Characteristics of variables in lymphoma patient dataset

Discrete variables	Value (number of cases)
Clinstg	1 (362); 2 (179)
Ch	0 (118); 1 (179)
D	0 (193); 1 (272); 2 (76)
Continuous variables	
Age	Min = 17; Q_1 = 47; Med = 58; Q_3 = 67; Max = 86
Hgb	Min = 40; Q_1 = 130; Med = 140; Q_3 = 150; Max = 189

In Figure 3, four graphs obtained by an analysis of follicular type lymphoma data are presented. Each of them contains four functions received for death and relapse taking into account CR and nCR approaches and two values of clinical stage (clinstg = 1 and clinstg = 2) and chemotherapy (ch = 0 and ch = 1). The differences between appropriate functions do not always follow the relationship $\tilde{F}_i(t) \leq 1 - KM_i(t)$, which

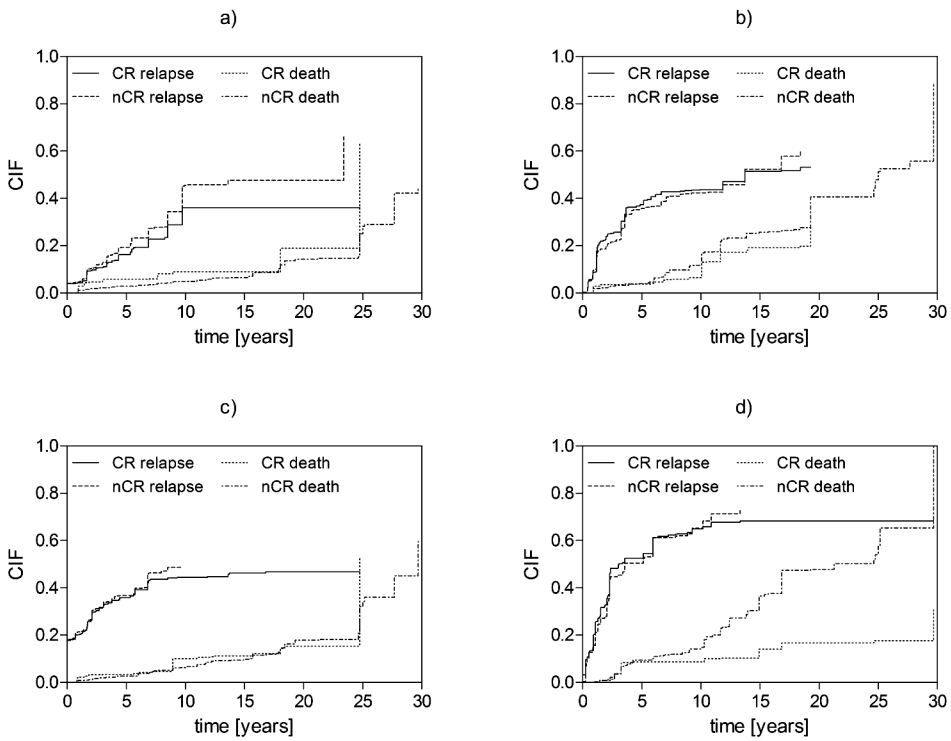

Figure 3. Cumulative incidence functions received for relapse and death in lymphoma data: a) clinstg = 1, ch = 1; b) clinstg = 1, ch = 0; c) clinstg = 2, ch = 1; d) clinstg = 2, ch = 0

may suggest that received groups of patients in the approach adjusted for competing risk events differ from the other approach. This is visible especially in the probabilities received for death, but in Figure 3(a) there are also quite big differences between the two functions calculated for relapse.

In Figures 4 and 5, we can observe the surfaces of lower quartiles obtained from CIF (Figure 4) and $1 - KM$ (Figure 5) functions describing the distribution of no response to treatment or relapse. In cases of the CR approach we can see a better prognosis for patients aged 30–40 with higher values of hemoglobin (150–160 g/l) and for patients aged 50–60 and hemoglobin values of 130–140 g/l. The highest value of the lower quartile received for the CIF function is 3.6 years. In the case of the graph visible in Figure 5, the maximum value of the surface is 7 years. Its shape does not indicate age as a risk factor. The value of the lower quartile depends only on hemoglobin: the higher the value of hemoglobin, the better the prognosis.

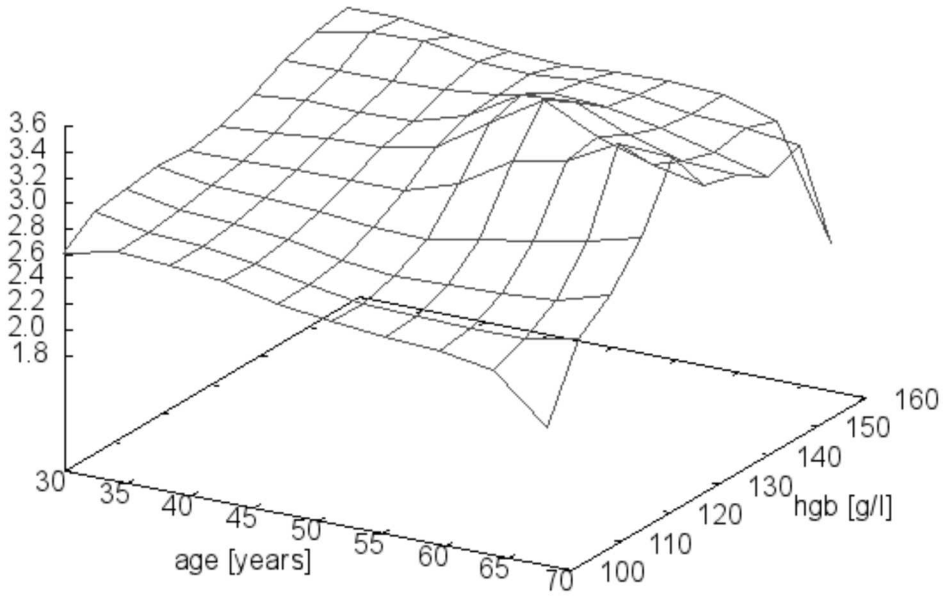

Figure 4. The influence of age and hemoglobin for lower quartiles calculated for CIF functions ($d = 1$) for patients with clinical stage I and chemotherapy

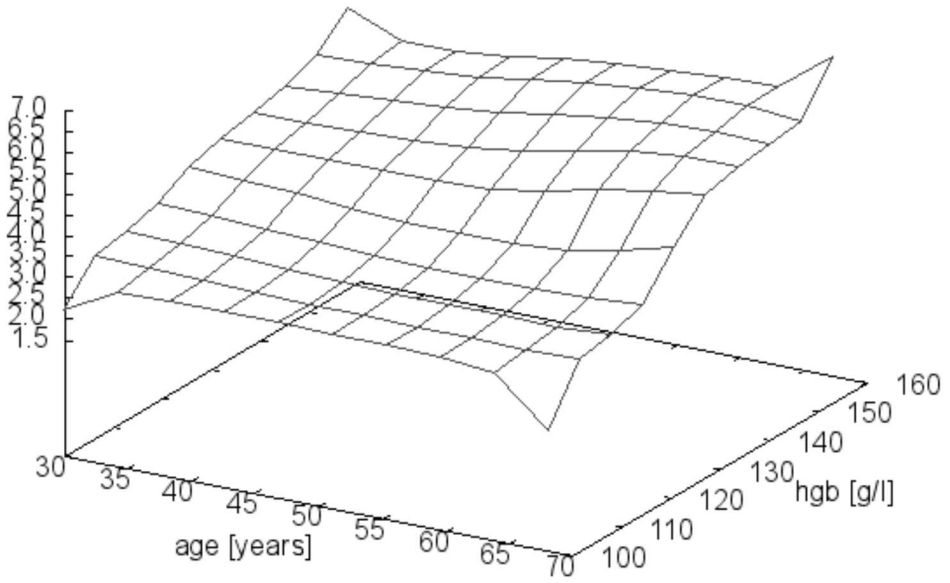

Figure 5. The influence of age and hemoglobin for lower quartiles calculated for $1 - KM$ functions ($d = 1$) for patients with clinical stage I and chemotherapy

Conclusions

In this paper, two approaches to analyzing survival data – with and without adjustment to competing risks – are presented. The experiments were performed on two real datasets: breast cancer data and follicular type lymphoma data, containing a high number of censored observations. To obtain results, cumulative incidence functions or $1 - KM$ estimators were calculated. For follicular type lymphoma data, the surfaces of lower quartiles are also presented. The graph comparisons show how important the use of information about competing risks is. The cumulative incidence functions received for competing risks data differ from the results obtained for single events, treating other events as censored observations. The differences are not only related to the association between the two functions used: $\tilde{F}_i(t) \leq 1 - KM_i(t)$, but also to the way the ensemble of dipolar trees uses the information about competing risks. This leads to dissimilar division of feature space and hence, the established groups of patients with similar survival experience are different for the two approaches. The interpretation of $1 - KM$ estimators obtained for a single event without taking into account competing risks may mislead.

Acknowledgements

This work was supported by the grant S/WI/2/2013 from Bialystok University of Technology.

REFERENCES

Bobrowski, L., Kretowska, M., & Kretowski, M. (1997). Design of neural classifying networks by using dipolar criterions. In proceedings of the Third Conference on Neural Networks and Their Applications, 14–18 October 1997 (pp. 689–694). Kule, Poland.

Fyles, A. W., McCready, D. R., Manchul, L. A., Trudeau, M. E., Merante, P., Pintilie, M., Weir, L. M., & Olivotto, I. A. (2004). Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer. *New England Journal of Medicine*, 351, 963–970.

Ibrahim, N. A., Kudus, A., Daud, I., & Abu Bakar, M. R. (2008). Decision tree for competing risks survival probability in breast cancer study. *World Academy of Science, Engineering and Technology*, 38, 15–19.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley & Sons Ltd.

Kretowska, M. (2006). Random forest of dipolar trees for survival prediction. In L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), ICAISC 2006, LNCS (LNAI) 4029, 909–918.

Kretowska, M. (2012). Competing Risks and Survival Tree Ensemble, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), ICAISC 2012, Part I, LNCS 7267, 387–393.

Marubini, E., & Valsecci, M. G. (1995). *Analysing survival data from clinical trials and observational studies*. England: John Wiley & Sons Ltd.

Pintilie, M. (2006). *Competing Risks: A Practical Perspective*. England: John Wiley & Sons Ltd.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-stage models. *Statistics in Medicine*, 26, 2389–2430.