COMPETING RISK ANALYSIS - GRAPHICAL REPRESENTATION OF VARIABLE INFLUENCE

Małgorzata Krętowska

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: In the paper the possibilities of assessing the variable influence on the failure occurrence is shown. Ensemble of dipolar survival trees is used as a prediction tool. The technique is able to cope with censored data (data with incomplete observations) as well as with competing risks data. The results are presented on the base of two real datasets for which the influence of discrete and continuous variables is examined. To this purpose, the cumulative incidence functions and the quartiles of CIF functions are applied.

Keywords: competing risks, survival analysis, ensemble of survival trees, dipolar criterion

1. Introduction

Survival analysis often aims at discovering risk factors - variables that have great impact on failure occurrence. Failure, according to research field, may have different meanings. In medical domain it usually means death or disease relapse. In case of competing risks data there is not only one event under investigation. For each patient we may observe several events, but only the first one is noticed. Each observation (patient) is described by a set of covariates, the time of the first event occurrence and the failure indicator. The value of failure indicator points the type of the first event. The value equal to 0 means that for a given patients there were no events of interest. We only know its follow-up time. Such incomplete observations are called censored cases.

Discovering the risk factors from survival data may be done by using statistical methods, usually non-parametric or semi-parametric ones. Among the non-parametric methods we may distinguish tests for comparing two CIF functions (e.g. Gray's test, logrank test), the well known Cox model [3] belongs to semi-parametric techniques. The main problem with applying the Cox model to the data is a number

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 8, pp. 19-30, 2011.

of assumptions to fulfill. These requirements are often difficult to obey so alternative techniques are proposed.

Classification or regression trees are ones of the methods successfully used for competing risks. [2] and [6] describe similar approaches, where induction of the proposed between-node tree is based on the difference between cumulative incidence function. Additionally Callahan in [2] presents a within-node tree that use event-specific martingale residuals. The method proposed in [7] is available as an R-package. The method based on ensemble of survival tree for competing risk is presented in [9]. Induction of individual tree is based on minimization of, so called, dipolar criterion function created from dipoles. Dipoles are pairs of feature vectors, formed appropriately for a given problem.

In the paper we use the methodology described in [9]. Based on the results received from the ensemble, we try to test the variable influence on failure occurrence. The examination is done by use of CIF functions as well as the graphical representation of quartiles of a given event time. The experiments are performed on two real datasets described patients with follicular cell lymphoma and the other dataset - patients with breast cancer.

The paper is organized as follows. Section 2. describes the survival data with competing risks and introduces the idea of cumulative incidence function as well as the Kaplan-Meier survival function. In Section 3. short description of ensemble of dipolar survival tree is done. Experimental results are presented in Section 4.. They were carried out on the base of two real datasets describing the patients with breast cancer data and follicular type lymphoma. Section 5. summarizes the results.

2. Competing risks

In case of survival data with competing risks, at the beginning of the follow-up the patient is at risk of p ($p > 1$) different types of failure (Fig 1). Assuming that the time of occurrence for ith failure is T_i, we are interested only in the failure for which the time is the shortest $T = \min(T_1, T_2, \ldots, T_n)$. The learning sample L for competing risk data is defined as $L = (\mathbf{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i is the time to the first event observed and $\delta_i = \{0, 1, \ldots, p\}$ indicates the case of failure. δ_i equals to 0 represents censored observation, which means that for a given patient has not occurred any failure. Variable t_i represents the follow-up time.

The distribution of the random variable T (time), for an event of type i ($i = 1, 2, \ldots, p$) may be represented by several functions. One of the most popular is cumulative incidence function (CIF) defined as the probability that an event of type i

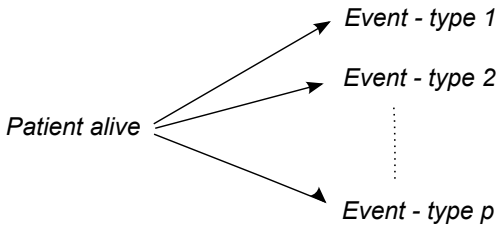

Fig. 1. Competing risks

occurs at or before time t [11]:

$$F_i(t) = P(T \leq t, \delta = i) \tag{1}$$

survival function

$$S_i(t) = P(T > t, \delta = i) \tag{2}$$

or hazard function

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta = i | T \geq t)}{\Delta t} \tag{3}$$

The estimator of the CIF function is calculated as

$$\hat{F}_i(t) = \sum_{j | t_j \leq t} \frac{d_{ij}}{n_j} \hat{S}(t_{j-1}) \tag{4}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered uncensored time points from the learning sample L, d_{ij} is the number of events of type i at time $t_{(j)}$, 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 $t_{(j)}$) and $\hat{S}(t)$ is the Kaplan-Meier estimator of the probability of being free of any event by time t. It is calculated as:

$$\hat{S}(t) = \prod_{j | t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j} \right) \tag{5}$$

where d_j is the number of events at time $t_{(j)}$. Examples of CIF function as well as Kaplan-Meier estimator are given in figure 2.

The "patients specific" cumulative incidence function for the event of type i is given by $\hat{F}_i(t|\mathbf{x}) = P(T \leq t, \delta = i | \mathbf{X} = \mathbf{x})$. The conditional CIF for the new patient with covariate vector $\mathbf{x}_{new}$ is denoted by $\hat{F}_i(t|\mathbf{x}_{new})$.

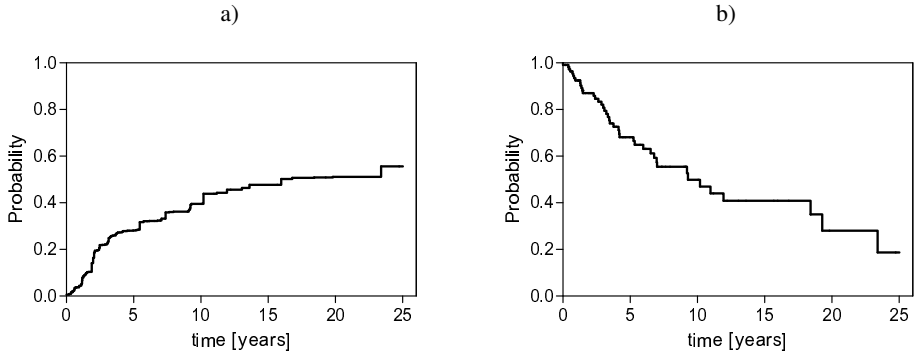

Fig. 2. Examples of a) CIF function; b) Kaplan-Meier estimator.

3. Prediction tool

The analysis of competing risks is performed by the use of ensemble of dipolar survival trees. Detailed description of induction of the ensemble is given in [9]. Below, one can find only the general information about the methodology used.

The general algorithm for generating 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
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 $L_i(\mathbf{x}_n)$ which belongs to the same terminal node as $\mathbf{x}_n$
4. Build aggregated sample $L_A(\mathbf{x}_n) = [L_1(\mathbf{x}_n), L_2(\mathbf{x}_n), \ldots, L_k(\mathbf{x}_n)]$
5. Compute the Kaplan-Meier aggregated survival function for a new observation $\mathbf{x}_n$ as $\hat{S}^A(t|\mathbf{x}_n)$.
6. Compute the aggregated CIF functions for the ith type of failure for a new observation $\mathbf{x}_n$ as $\hat{F}_i^A(t|\mathbf{x}_n)$.

As one can see the main point in presented above algorithm is induction of dipolar survival tree for each generated bootstrap sample L_i. Each internal node contains a split, which tests the value of an expression of the covariates. In the proposed approach the split is equivalent to the hyper-plane $H(\mathbf{w}, \theta) = \{(\mathbf{w}, \mathbf{x}) :< \mathbf{w}, \mathbf{x} >= \theta\}$. The hyper-planes in the internal nodes of a tree are calculated by minimization of dipolar criterion function (detailed description may be found in [8]). This is equivalent to division of possibly high number of mixed dipoles and possibly low number of pure ones constructed for a given dataset.

The dipole [1] is a pair of different covariate vectors $(\mathbf{x}_i, \mathbf{x}_j)$ from the learning set. Mixed and pure dipoles are distinguished. Assuming that the analysis aims at dividing the feature space into such areas, which would include the patients with the same case of failure and similar survival times, pure dipoles are created between pairs of feature vectors with the same failure type, for which the difference of failure times is small, mixed dipoles - between pairs with distant failure times. Taking into account censored cases the following rules of dipole construction can be formulated:

1. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the pure dipole, if
   - $\delta_i \neq 0$ and $\delta_i = \delta_j = z$ and $|t_i - t_j| < \eta_z$, $z = 1, 2, \ldots, p$.
2. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   - $\delta_i \neq 0$ and $\delta_i = \delta_j = z$ and $|t_i - t_j| > \zeta_z$, $z = 1, 2, \ldots, p$
   - $(\delta_i = 0, \delta_j = z$ and $t_i - t_j > \zeta_z)$ or $(\delta_i = z, \delta_j = 0$ and $t_j - t_i > \zeta_z)$, $z = 1, 2, \ldots, p$

Parameters η_z and ζ_z are equal to quartiles of absolute values of differences between uncensored survival times for zth type of failure, $z = 1, 2, \ldots, p$. Basing on the earlier experiments, the parameter η_z is fixed as 0.3 quantile and ζ_z - 0.6.

The straightforward graphical representation of the results is the CIF function calculated for all the analyzed types of failure, for a new patient described by $\mathbf{x}_n$. Studying the influence of single variable for failure occurrence or the interaction of two variables, the tool may results the median value, lower quartile or any other centile of event occurrence for any type of failure. It enables drawing surfaces of a given statistics for different values of examined variables.

4. Experimental results

The experiments were performed on the base of two real datasets: breast cancer data and follicular type lymphoma data. The first analyzed dataset was used to show how to examine the influence of discrete variables for the failure (of any type) occurrence. Here, the cumulative incidence functions were used to model the failure prediction. In case of the other data, we presented the influence of continuous variables for the quartiles of CIF functions. We use here the lower quartile and the median values.

All the experiments were performed using the ensemble of 100 survival trees.

4.1 Breast cancer data

Breast cancer data [4] contain 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 description of the variables [5].

Table 1. Description of variables in breast cancer data

Variable name	Description
tx	Randomized treatment: 1=tamoxifen, 2=radiation + tamoxifen
Variables assessed at the time of randomization	
pathsize	Size of tumor (cm)
hist	Histology: 1=ductal, 2=lobular, 3=medullary, 4=mixed, 5=other
hrlevel	Hormone receptor level: 0=negative, 1=positive
hgb	Haemoglobin (g/l)
nodedis	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,

In figure 3 we can observe the differences between CIF functions calculated separately for tamoxifen alone and tamoxifen plus radiation for two event types: relapse and malignancy. The other variables were set up for theirs median values: pathsize=1.5; hist=1; hrlevel=1; hgb=135; nodediss=1; age=67. As we could observe the probability of relapse is greater in the group of patients treated with tamoxifen alone. Probability of malignancy is less for the group of patients treated with tamoxifen during the first 6 years of observations, later the probability in this group is greater then for patients treated with tamoxifen and radiation.

Figure 4 shows the influence of histology for the probability of relapse and malignancy. Other variable were set up for their median values (see description of figure 3). Two histological types were examined: $hist = 1$ (ductal) and $hist = 4$ (mixed). In figure 4a) we can observe significant differences between two CIF functions calculated for two types of histology. The patients with ductal histology treated only with tamoxifen have greater probability of relapse than patients with mixed histology. Such differences are not visible on figures representing the probability of malignancy (both for $tx = 1$ and $tx = 2$) and for the probability of relapse in group of patients treated with tamoxifen plus radiation.

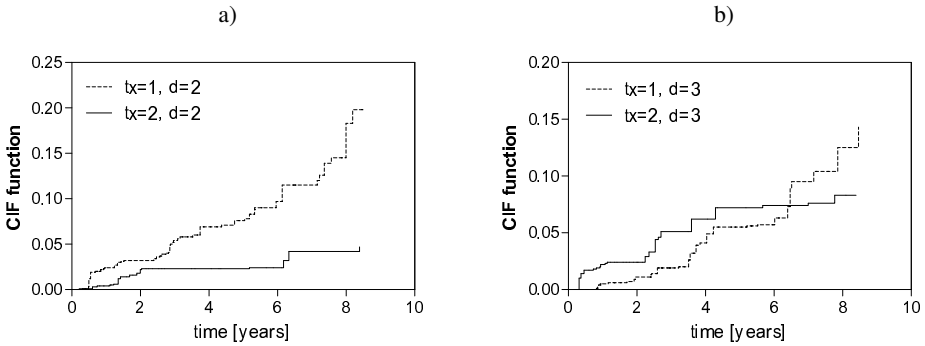

Fig. 3. CIF functions calculated for a) relapse ($d = 2$); b) malignancy ($d = 3$).

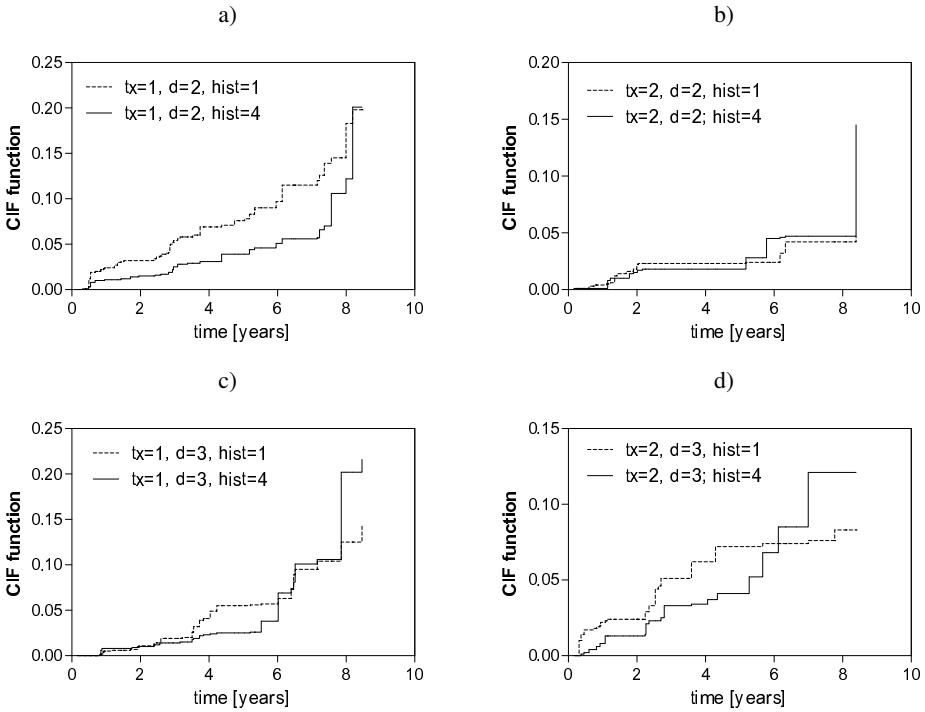

Fig. 4. CIF functions calculated for two types of histology: ductual and mixed for a) relapse ($d = 2$) and $tx = 1$; b) relapse and $tx = 2$; c) malignancy ($d = 3$) and $tx = 1$; d) malignancy and $tx = 3$.

4.2 Follicular cell lymphoma data

Lymphoma patient dataset was created at Princess Margaret Hospital, Toronto [10]. In the experiments we use the subset of 541 patients having follicular type lymphoma, registered for treatment at the hospital between 1967 and 1996, with early stage disease (I or II) and treated with radiation alone or with radiation and chemotherapy. Each patient is described by four variables, described in table 2.

Table 2. Description of variables in follicular type lymphoma data

Variable name	Description
Variables assessed at the time of diagnosis	
age	Age (years)
hgb	Haemoglobin (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

The event of interest is failure from the disease: no response to treatment or relapse. Competing risk type of event is death without failure. There are 272 event of interest and 76 observations with death without relapse.

On the base of lymphoma data the possibility of examination of the impact of continuous variables for probability of event occurrence is shown. For this purpose, described above algorithm of ensemble of dipolar trees generation should return the median value or the value of any other centile of the CIF function calculated for a new observation.

In figures 5 - 8 we can observe the quartiles of CIF function calculated for the first event (no response to treatment or relapse) for patients treated with radiation alone. In figures 5 and 6 the influence of age and hemoglobin for lower quartiles calculated for CIF functions for patients with clinical stage I and II is presented. The probability of relapse at a given value of the lower quartile is equal to 0.25. So the higher values of this statistics are connected with better prognosis for the patient.

For people with clinical stage I (Fig. 5) the lowest values of the first quartile are for haemoglobin at range 100-120. Here the influence of age is not visible. The failure prediction is better for young patients with haemoglobin in range 145-160 and for older patients (age: 50-65) and haemoglobin equal to 130-140. Here we can see the interaction of age and haemoglobin. For patients with clinical stage II (Fig. 6) we

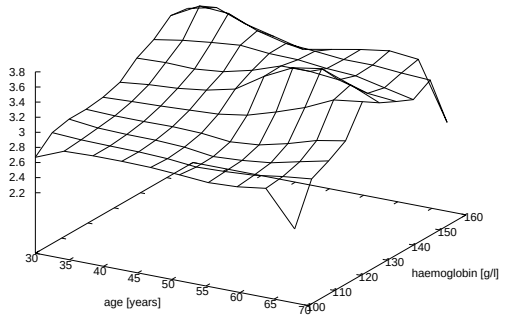

Fig. 5. The influence of age and hemoglobin for lower quartiles calculated for CIF functions ($d = 1$) for patients with clinical stage I

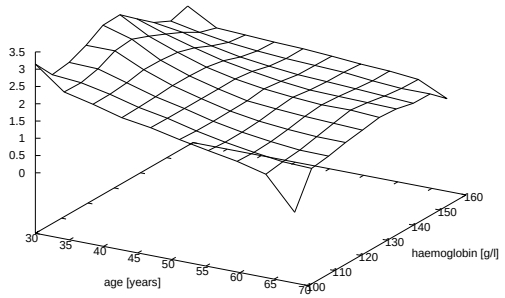

Fig. 6. The influence of age and hemoglobin for lower quartiles calculated for CIF functions ($d = 1$) for patients with clinical stage II

can observe the influence of age (younger people have better prognosis) and there is no influence of haemoglobin.

In figures 7 and 8 the influence of age and haemoglobin for the median values of CIF functions are presented. We can see that on average the medians are greater for people with clinical stage I (Fig. 7) than for people with clinical stage II (Fig. 8). In case of clinical stage I the best prediction is for younger people with greater value

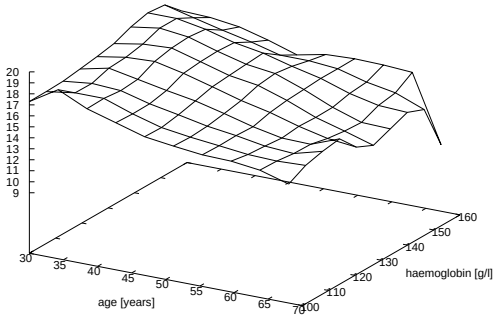

Fig. 7. The influence of age and hemoglobin for median values calculated for CIF functions for patients with clinical stage I

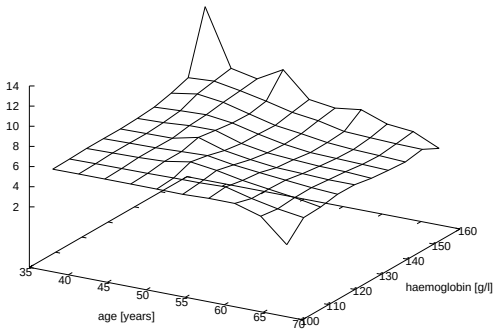

Fig. 8. The influence of age and hemoglobin for median values calculated for CIF functions for patients with clinical stage II

of haemoglobin. The impact of two examined continuous variables is not significant for patient with clinical stage II.

5. Conclusions

In the paper the possibilities of assessing the variables influence for the event occurrence is presented. The methodology based on the ensemble of dipolar survival trees is applied for this purpose. The experiments were performed on two real datasets. The

first one - breast cancer data - is served as an example of discrete variables assessing. For this purpose the cumulative incidence functions were drawn for different values of discrete variables. In this case, two types of treatment and histology were used. For the other dataset, follicular type lymphoma data, the influence of two continuous variables for relapse occurrence were assessed. The surfaces of the lower quartile and median values calculated for the CIF functions for different values of age and haemoglobin were analyzed.

As one could see, presented graphs may suggest the influence of a given variable for failure occurrence exists and also may help to establish if the assumptions of statistical methods are fulfilled for examined data.

References

[1] L. Bobrowski, M. Krętowska, M. Krętowski, Design of neural classifying networks by using dipolar criterions, Proc. of the Third Conference on Neural Networks and Their Applications, Kule, Poland, 1997, pp. 689-694.

[2] F.M. Callaghan, Classification trees for survival data with competing risks, Univeristy of Pittsburgh, PhD. thesis, 2008.

[3] D.R. Cox, Regression models and life tables (with discussion), Journal of the Royal Statistical Society B **34**, 1972, pp. 187-220.

[4] A. W. Fyles, D. R. McCready, L. A Manchul., M. E. Trudeau, P. Merante, M. Pintilie, L. M. Weir, and I. A. Olivotto, Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer, New England Journal of Medicine 351, 2004, pp. 963-970.

[5] N.A. Ibrahim, A. Kudus, I. Daud, M.R. Abu Bakar, Decision tree for competing risks survival probability in breast cancer study, World Academy of Science, Engineering and Technology, 38, 2008, pp. 15-19.

[6] N.A. Ibrahim, A. Kudus, Decision tree for prognostic classification of multivariate survival data and competing risks, in: Strangio M. A. (Eds.), Recent Advances in Technologies, 2009.

[7] H. Ishwaran, U.B. Kogalur, R.D. Moore, S.J. Gange, B.M. Lau, Random survival forests for competing risks, 2010.

[8] M. Krętowska, Random forest of dipolar trees for survival prediction, L.Rutkowski et al. (Eds.), ICAISC 2006, LNAI 4029, 2006, pp. 909-918.

[9] M. Krętowska, Competing risks and survival tree ensemble, (submitted).

[10] M. Pintilie, Competing Risks: A Practical Perspective, John Willey & Sons, 2006.

[11] H. Putter, M. Fiocco, R.B. Geskus, Tutorial in biostatistics: Competing risks and multi-stage models, Statistics in Medicine **26**, 2007, pp. 2389-2430.

Małgorzata Krętowska

DANE Z KONKURENCYJNYM RYZYKIEM - GRAFICZNA REPREZENTACJA WPŁYWU CZYNNIKÓW RYZYKA

Streszczenie W pracy przedstawione zostały możliwości graficznej weryfikacji hipotez dotyczących wpływu poszczególnych cech na czas wystąpienia porażki. Jako narzędzie prognostyczne zostały wykorzystane predyktory złożone, w których dipolowe drzewa przeżycia służą jako pojedyncze predyktory. Algorytm tworzenia predyktorów złożonych wykorzystuje informację pochodzącą z obserwacji cenzorowanych, jak również jest przystosowany do danych z konkurencyjnym ryzykiem.
Eksperymenty zostały wykonane przy użyciu dwóch zbiorów danych: zbiór opisujący pacjentki z rakiem piersi i drugi - opisujący pacjentów z chłoniakiem grudkowym. Pierwszy z analizowanych zbiorów posłużył jako przykład do badania wpływu zmiennych dyskretnych. W tym celu wyznaczone zostały dystrybuanty (ang. cumulative incidence function) dla wyróżnionych dwóch zdarzeń konkurencyjnych i dwóch cech: rodzaju leczenia oraz typu histologicznego raka. W przypadku zbioru z chłoniakiem grudkowym badane były cechy ciągłe: wiek oraz wartość hemoglobiny. Analiza tych danych opierała się na wyznaczeniu wartości kwartyla pierwszego oraz mediany z funkcji dystrybuanty, wyznaczonej dla czasu nawrotu choroby.

Słowa kluczowe: dane z konkurencyjnym ryzykiem, analiza przeżyć, predyktory złożone, kryterium dipolowe

Artykuł zrealizowano w ramach pracy badawczej S/WI/2/08.