Evaluation of Risk Factors for Parametrization of Cancer Models

Andreas Bauer[1], Günther Zauner[2], Christoph Urach[2], Felix Breitenecker[1]
[1]Vienna University of Technology, Institute for Analysis and Scientific Computing, Austria
[2]dwh Simulation Services, Vienna, Austria
andreas.e101.bauer@tuwien.ac.at

Cancer is the second most cause of death in Austria and around 38000 people are diagnosed with cancer each year [1]. The goal of this paper is to analyze methods for evaluation of risk factors in order to parametrize a micro simulation model for cancer prevalence. The focus of this paper is on modeling the survival time. This is done by the methods of survival analysis and model selection. Firstly, the survival function is estimated by the Kaplan-Meier estimate. Afterwards, a Cox proportional hazards regression is performed with all possible sets of parameters. These models are tested by twos with the likelihood ratio test in order to compare them. Another approach is the so-called Lasso method. This method puts a constraint on the sum of the absolute values of the regression coefficients and in most cases forces some of the coefficients to go to zero. The Akaike Information Criterion is also applied. All three methods are compared and the parameters which are supported, at least to a certain extent, by all of them are included in the estimation of the survival time of the prevalence model.

1 Introduction

Cancer is the second most cause of death in Austria and around 38000 people are diagnosed with cancer each year [1]. So, it is of great importance to find out the risk factors on one side, but also to model the incidence and prevalence to be able to evaluate health policy measures on the other side. An important step for doing a simulation is to find out the potential influences on the course of the disease and to quantify them in order to parametrize the model. The goal of this paper is to test methods for identification of possible influence factors on the course of cancer and to do a survival analysis for finding out the factors on which the course of the disease depends. Also, methods of model selection are used. These analyses will be used for the parametrization of a micro-simulation model for cancer prevalence later on.

2 Data

The following six categories are examined to find out, if they are possible influences on the development or the course of cancer: sex, age at the diagnosis date, chronic diseases X, Y and Z and the stage of cancer at the date of the diagnosis. In Table 1, an overview of these categories with according types and ranges is presented.

Number	Category	Type	Range
1	Age	ratio	23-83
2	Sex	nominal	0,1
3	Chronic disease Y	ordinal	0,1
4	Chronic disease Z	ordinal	0,1
5	Chronic disease X	ordinal	0,1
6	Stage of disease	ordinal	2-4

Table 1. Overview of categories with according types and ranges

3 Methods

3.1 Survival Analysis

In order to examine the survival time of the individuals depending on the possible influence factors, methods of survival analysis are applied. These methods allow the estimation and the analysis of the survival function and the hazard function. The survival function $S(t)$is defined as the probability that an individual will survive up to time t and the hazard function $h(t)$ is defined as the instanteneous rate of death at time t.

Another important aspect regarding survival analysis is censoring. In the field of survival analysis often the data collection ends before the event of interest has

occurred for all individuals. For those individuals, the survival time cannot be determined. The only thing that is known is that the survival time exceeds the time of the observation of the particular patient.

The Kaplan-Meier estimate is an estimate for the survival function $S(t)$. It makes use of the information of the exact date of the occurrence of death. The estimated survival probability s_t at time t is:

$$s_t = \frac{n_t - d_t}{n_t} \quad (1)$$

n_t is the number of people alive at time t and d_t is the number of people that died at time t. So, s_t is simply the ratio of the people alive who survive time t. Thus, the probability of surviving up to a certain point of time t_j is calculated with the so-called product-limit formula [2]:

$$S(t_j) = \prod_{i=1}^{j} S(t_i) \quad (2)$$

A common approach to do regression analysis on survival data is the so-called Cox regression, also known as proportional hazards regression. It assumes that the ratio of the hazards comparing different exposure groups remains constant over time. This is called the proportional hazards assumption. The mathematical form of the proportional hazards model is:

$$h(t) = h_0(t) * \exp\left(\sum_{i=1}^{n} b_i * x_i\right) \quad (3)$$

$h_0(t)$ denotes the baseline hazard which refers to a particular group of individuals (for example, the individuals with value zero in all binary categories, with mean age and with stage of illness two), n is the number of covariates, x_i is the value of the ith covariate and b_i is the corresponding regression coefficients [3,4].

3.2 Model Selection

The methods of model selection can be used to find the significant covariates for our model depending on given data. The goal of model selection is to eliminate some of the covariates from the full model with six covariates to get a simpler model which still explains most of the effects correctly. In order to find an appropriate model, three approaches are considered: Likelihood ratio tests, Lasso – Method and Akaike Information Criterion (AIC).

Firstly, the Cox regression is performed with all possible sets of parameters. That means the parameter sets of the models are all possible subsets of the full set with six parameters.

For each two nested models the likelihood ratio test is applied. With this test we examine, if the bigger model of the two significantly provides additional information in comparison to the smaller nested model. The significance level is set to 0.05.

Another approach to select a model is to use the Lasso-method. The regression coefficients of the Cox regression are calculated as usual by minimizing the partial log-likelihood, but additionally the sum of the absolute values of the regression coefficients is bounded by a constant in order to force some of the coefficient to shrink to zero. This results in a sequence of models depending on the size of the constraint. There are various ways to determine the "best" size of the constraint. It can be either chosen arbitrarily or automatically based on the data. For instance, the use of an approximate generalized cross-validation (GCV) statistic is a common tool for automatic constraint selection [5].

Another criterion to select a set of variables is the so-called "Akaike Information Criterion" (AIC). The AIC value is calculated as follows:

$$AIC = -2\log\mathcal{L}(t_0|y) + 2K \quad (4)$$

The first summand is the negative of twice the numerical value of the log-likelihood at its maximum point t_0 given data y and the second summand is twice the number of parameters of the model. The smaller the AIC value of a model is, the better it is, because the AIC value can be interpreted as a kind of information loss [6]. The AIC can also be used for automatic choice of a constraint for the abovementioned Lasso-method.

4 Results

The Kaplan-Meier estimate was calculated for various groups of the population. Figure 1 shows the Kaplan-Meier estimates for male and female individuals in comparison. We can see that the estimate for the males is lower than the estimate for the females until about 800 days after the diagnosis, when only 10 percent of the individuals are left alive.

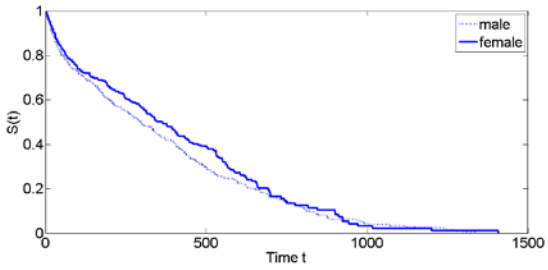

Figure 1. Comparison of Kaplan-Meier estimates for male and female individuals

The coefficients for the Cox regression with all six covariates included are shown in Table 2.

b_1	b_2	b_3	b_4	b_5	b_6
-0.11	0.02	-0.05	0.09	-0.25	0.02

Table 2. Coefficients of Cox regression in order: sex, age, chronic disease Y, chronic disease Z, chronic disease X, stage of illness

The p-values of the Cox regression show that the p-value of the term age is the lowest, so we start with the model with only term age. The likelihood ratio test shows that the model with added terms chronic disease X and sex and the model with added term chronic disease X, both given the term age, are statistically significant, while any other extension given the term age is not significant.

Figure 2 shows the values of the six regression coefficients of the Cox model plotted over the l_1- norm of the coefficient vector. On the x-axis above the plot also the number of non-zero coefficients is displayed.

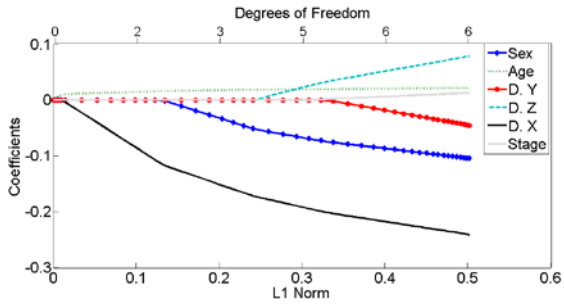

Figure 2. Cox regression coefficients over the norm of coefficients vector.

In Figure 2, we see that the smaller the norm of the vector gets, the smaller is the number of non-zero coefficients. The coefficient that vanishes at last, when the norm of the coefficient vector goes to zero, is the coefficient of the parameter age, right after the coefficients of the parameters chronic disease X and sex. The other three coefficients are eliminated earlier.

The AIC value is calculated for the models with all 64 possible sets of variables. In order to compare different models, AIC differences are computed, because the relative values of the AIC are more meaningful than the absolute values. The AIC differences are computed by subtracting the AIC value of the model with the least AIC value from the AIC values of each model.

In Table 3, the five models with the least AIC values and the AIC differences are listed.

Parameters	**AIC**	**AIC difference**
1; 2; 5	6788.75	0
2; 5	6789.15	0.40
1; 2; 3; 5	6790.27	1.52
1; 2; 4; 5	6790.41	1.66
1; 2; 5; 6	6790.59	1.84

Table 3. Parameters sets with lowest AIC values and AIC differences, numbers referred to numbering of categories in Table 1

Table 3 shows that the model with parameters sex, age and chronic disease X has the lowest AIC–value followed by the model with parameters age and chronic disease X. The other three listed models are also substantially supported by the AIC.

5 Conclusion and Outlook

The categories age, sex and chronic disease X are found to be significant by likelihood ratio tests using the Cox regression. The AIC ranks this model as first too. The Lasso-Method shows that these three parameters are the last three parameters that are left, when the norm of the coefficient vector declines. Since the set of parameters is not very big, in uncertain situations, where it is not obvious, if a certain parameter should be included or not, the parameter will be included in the model to avoid the situation that a substantial effect is eliminated from the model by accident. So all used methods suggest that these three parameters definitely should be included in the estimation of the survival time for the prevalence model. For the other categories, further analysis will be done to determine, if they also will be included in the future model.

6 References

[1] *Jahrbuch der Gesundheitsstatistik*. Statistik Austria, Austria, 2012.

[2] B. R. Kirkwood and J. A. C. Sterne. *Essential Medical Statistics*, 2nd edition. Wiley-Blackwell, United Kingdom, 2003.

[3] G. Rodriguez. *Lecture Notes on Generalized Linear Models*. Princeton University, United Kingdom, 2007.

[4] D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman & Hall, United Kingdom, 1984.

[5] R. Tibshirani. *The Lasso Method for Variable Selection in the Cox Model*. Statistics in Medicine Volume 16, p. 385-395, 1997.

[6] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edition*. Springer, United States of America, 2002.