Hazard function for cancer patients and cancer cell dynamics

Ivana Horová, Z. Zdeněk Pospíšil, J. Jiří Zelinka

www.elsevier.com/locate/yjtbi

Hazard function for cancer patients and cancer cell dynamics

Ivana Horová, Zdeněk Pospíšil, Jiří Zelinka

PII: S0022-5193(08)00308-1
DOI: doi:10.1016/j.jtbi.2008.06.014
Reference: YJTBI 5181

To appear in: *Journal of Theoretical Biology*

Received date: 6 February 2008
Revised date: 8 May 2008
Accepted date: 10 June 2008

Cite this article as: Ivana Horová, Zdeněk Pospíšil and Jiří Zelinka, Hazard function for cancer patients and cancer cell dynamics, *Journal of Theoretical Biology* (2008), doi:10.1016/j.jtbi.2008.06.014

HAZARD FUNCTION FOR CANCER PATIENTS AND CANCER CELL DYNAMICS

IVANA HOROVÁ[1], ZDENĚK POSPÍŠIL[2], JIŘÍ ZELINKA[1]

Abstract. The aim of the paper is to develop a procedure for an estimate of an analytical form of a hazard function for cancer patients. Although a deterministic approach based on cancer cell population dynamics yields the analytical expression, it depends on several parameters which should be estimated. On the other hand a kernel estimate is an effective nonparametric method for estimating of hazard functions. This method provides the pointwise estimate of the hazard function. Our procedure consists in two steps: in the first step we find the kernel estimate of the hazard function and in the second step the parameters in the deterministic model are obtained by the least squares method. A simulation study with different types of censorship is carried out and the developed procedure is applied to real data.

1. Introduction

Survival analysis belongs to classical parts of mathematical statistics and occupies an important place in the medical research. In the present paper we focus on estimating hazard functions under random censorship. We use the model where data are censored from the right[3]. This type of censoring is often met in many applications especially in medical research (see e.g Collett (2003), Hougaard (2001), Therneau and Grambsch (2000)). Nonparametric methods seem to be adequate for estimates of hazard functions because in contrast to the parametric modeling assumptions of unknown functions are much weaker, only smoothness and differentiability are required. Among these methods kernel estimates represent one of the most effective methods. Kernel estimates of hazard functions have been developed by many authors (see e.g. References in Horová and Zelinka (2007)).

Our approach is based on the method introduced by Tanner and Wong (1984), Müller and Wang (1990), Jiang and Marron (2003). These methods provide pointwise estimates of the hazard functions and are described in Section 3. Section 4 is devoted to a deterministic model which is defined as a solution of a dynamical problem and yields an analytical form of the hazard function for cancer patients. This dynamical problem is defined under assumption that the hazard is proportional to a rate of proliferation speed of cancer cells and uses the Gompertzian model of the tumor growth curve (Kozusko and Bajzer, 2003). But the deterministic model depends on some parameters which should be estimated. It could be done by a maximum likelihood method (see e.g. Hougaard (2001), Horová *et al.* (2008)), but in Section 5 we develop a procedure for finding the parameters in the deterministic model by means of the kernel estimate of the hazard function. Section 6 is devoted

[1]Research supported by MŠMT: LC06024
[2]Author supported by the Grant No. 201/01/0079 of the Grant Agency of the Czech Republic
[3]It is shortly reminded in Section 2.

to a simulation study. Here the quality of obtained hazard functions for different types of censoring and different sample sizes are evaluated in terms of L_2-measure. In Section 7 the developed procedure is applied to real data sets. Discussion and conclusion are included in Section 8.

2. Random censorship model

Survival data are frequently censored thus it makes sense to define a random censorship model. Let $T_1, T_2, \ldots, T_n$ be independent and identically distributed lifetimes with the cumulative distribution function F. Let $C_1, C_2, \ldots, C_n$ be independent and identically distributed censoring times with the cumulative distribution function G which are usually assumed to be independent of lifetimes. In the random censorship model we observe pairs (X_i, δ_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I\{X_i = T_i\}$ indicates whether the observation is censored or not. It follows that $\{X_i\}$ are independent and identically distributed with the cumulative distribution function L satisfying $\overline{L}(x) = \overline{F}(x)\,\overline{G}(x)$ where $\overline{H} = 1 - H$ is a survival function for any cumulative distribution function H.

The survival process can be also characterized by the hazard function $\lambda = \lambda(x)$, i.e. the probability that an individual dies at time x, conditional on he or she having survived to that time. If the lifetime distribution F has a density f, for $\overline{F}(x) > 0$ the hazard function is defined by

$$\lambda(x) = \frac{f(x)}{\overline{F}(x)}. \tag{1}$$

Since $\overline{F}(0) = 1$, the survival function can be expressed by the formula

$$\overline{F}(x) = \exp\left(-\int_0^x \lambda(t)dt\right). \tag{2}$$

Let cohort of the initial size N_0 die out with the time dependent death rate $\mu = \mu(x)$, i.e. the size of the cohort $N = N(x)$ at time x evolves according to the differential equation

$$N'(x) = -\mu(x)\,N(x), \qquad N(0) = N_0$$

whose the solution is given by

$$N(x) = N_0 \exp\left(-\int_0^x \mu(t)dt\right). \tag{3}$$

In this connection the survival function $\overline{F}$ is defined as

$$\overline{F}(x) = \frac{N(x)}{N_0}. \tag{4}$$

Hence, the death rate μ equals the hazard function λ. Consequently

$$\lambda(x) = -\frac{N'(x)}{N(x)}. \tag{5}$$

3. Kernel estimates of the hazard function

These estimates have been dealt with many authors, (see e.g Horová and Zelinka (2007) and references therein). Our approach is based on the model introduced by Tanner and Wong (1984), Müller and Wang (1990) and Jiang and Marron (2003).

Let $[0,T]$, $T>0$ be an interval for which $L(T)<1$, L is the cumulative distribution function of X_i's.

First let us make some assumptions:

1° $\lambda \in C^{k_0}[0,T]$, $k_0 \geq 2$, $C^{k_0}[0,T]$ denotes the class of functions having continuous derivatives up to the order k_0.

2° Let K be a real valued function on $\mathbb{R}$ satisfying conditions

$$(i) \qquad support(K) = [-1,1],\ K(-1) = K(1) = 0$$

$$(ii) \qquad K \in Lip[-1,1] : |K(x) - K(y)| \leq q|x-y|,\ 0 < q,\ \forall x, y \in [-1,1]$$

$$(iii) \qquad \int_{-1}^{1} x^j K(x)dx = \begin{cases} 1, & j = 0 \\ 0, & 0 < j < k \\ \beta_k \neq 0, & j = k,\ k \leq k_0 \end{cases}$$

Such a function is called a kernel of order k and a class of these kernels is denoted by S_k.

3° Let $\{h(n)\}$ be a non-random sequence of positive numbers satisfying $\lim\limits_{n\to\infty} h(n) = 0$, $\lim\limits_{n\to\infty} n\,h(n) = \infty$. These numbers are called bandwidths or smoothing parameters. For the sake of simplicity the dependence $h(n)$ on n will be omitted in following considerations.

4° Denote $K_h(x) = \frac{1}{h}K(\frac{x}{h})$.

The kernel estimate of the hazard function λ at the point $x \in [0,T]$ with the kernel K and bandwidth h is denoted by $\hat{\lambda}_{h,K}(x)$ and defined by

$$\hat{\lambda}_{h,K}(x) = \sum_{i=1}^{n} K_h\left(x - X_{(i)}\right) \frac{\delta_{(i)}}{n-i+1}, \tag{6}$$

where $X_{(i)}$ denotes the ith order statistics of $X_1,\ldots,X_n$ and $\delta_{(i)}$ the corresponding censoring status. The kernel K plays a role of a weight function and the bandwidth h controls the smoothness of the estimate. Under assumptions given above $\hat{\lambda}_{h,K}(x)$ yields a consistent estimate of $\lambda(x)$, i.e. $\hat{\lambda}_{h,K}(x) \xrightarrow{p} \lambda(x)$, $x \in [0,T]$ (see e.g. Müller and Wang (1990)).

The global quality of the estimate (6) can be described by the Mean Integrated Square Error (MISE):

$$\mathrm{MISE}(\hat{\lambda}_{h,K}) = \int_0^T E(\hat{\lambda}_{h,K}(x) - \lambda(x))^2 dx =$$

$$= \frac{V(K)\Lambda}{nh} + h^{2k}\beta_k^2 D_k + o\left(h^{2k} + \frac{1}{nh}\right),$$

where E denotes the expectation of a random variable and

$$\beta_k = \int_{-1}^{1} x^k K(x)dx, \quad V(K) = \int_{-1}^{1} K^2(x)dx,$$

$$\Lambda = \int_0^T \frac{\lambda(x)}{\overline{L}(x)} dx\,, \qquad D_k = \int_0^T \left(\frac{\lambda^{(k)}}{k!}\right)^2 dx\,.$$

Now we focus on the leading term $\overline{\mathrm{MISE}}(\lambda_{h,K})$ of $\mathrm{MISE}(\lambda_{h,K})$

$$\overline{\mathrm{MISE}}(\hat{\lambda}_{h,K}) = \frac{V(K)\Lambda}{nh} + h^{2k}\beta_k^2 D_k. \tag{7}$$

There is not any problem to choose a suitable kernel. There exists a class of optimal kernels minimizing $\overline{\mathrm{MISE}}(\hat{\lambda}_{h,K})$ with respect to K (see e.g. Müller (1988), Marron and Nolan (1989), Horová *et al.* (2002)). Here we recommend to use kernels of order two, namely the Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)\, I_{[-1,1]}$ or quartic kernel $K(x) = \frac{15}{16}(1-x^2)^2\, I_{[-1,1]}$, I is an indicator function.

The problem of choosing how much to smooth, i.e. how to choose a bandwidth, is of a crucial importance in kernel estimates.

It is easy to find that the asymptotically optimal bandwidth $h_{opt,k}$ minimizing $\overline{\mathrm{MISE}}(\hat{\lambda}_{h,K})$ with respect to h is given by

$$h_{opt,k}^{2k+1} = \frac{V(K)\Lambda}{2nk\beta_k^2 D_k}, \tag{8}$$

i.e. $h_{opt,k} = O(n^{-\frac{1}{2k+1}})$.

The formula (8) provides simple insight into "good" bandwidth. But an obvious problem of finding this optimal bandwidth is that $h_{opt,k}$ depends on the unknowns Λ and D_k. In the random censorship model modified cross-validation methods (see e.g. Uzunogullari and Wang (1992)) or modified likelihood methods (Tanner and Wong, 1984) could be applied. In our paper we use a special iterative method based on a suitable approximation of MISE (see Horová and Zelinka (2007), Horová *et al.* (2006)).

Let us denote with $\hat{h}_{opt,k}$ an estimate of $h_{opt,k}$. The influence of the bandwidth to the estimate is shown in Figure 1. Simulated data for $n = 100$ and censorship type II (see section 6) are used. A small bandwidth leads to the undersmoothed estimate of λ and a large bandwidth yields the oversmoothed estimate of λ.

At the end of this section we recall the formula for the asymptotic $(1-\alpha)$ confidence interval given by

$$\hat{\lambda}_{h,K}(x) \pm \left\{\frac{\hat{\lambda}_{h,K}(x)V(K)}{(1-L_n(x))hn}\right\}^{1/2} \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \tag{9}$$

where Φ is the normal cumulative distribution function and L_n is the modified empirical survival function of observation times

$$L_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} I_{\{X_i \le x\}}.$$

4. Deterministic model of hazard function

Let us assume that the hazard is proportional to the rate of proliferation of cancer cells. Let $y = y(x)$ denote a time dependent size of cancer cells population and thus

$$\lambda(x) = \rho y'(x) \tag{10}$$

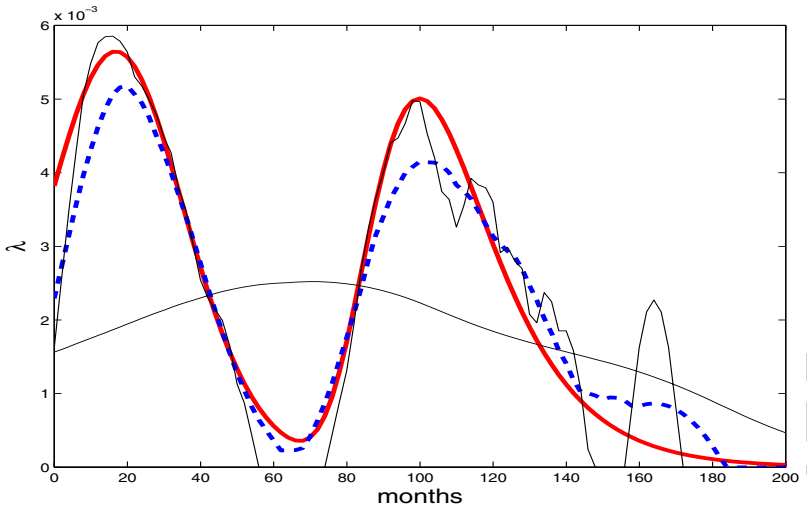

FIGURE 1. Influence of the bandwidth to the kernel estimate of λ
solid line – true hazard function
dashed line – estimate with the optimal bandwidth $h_{opt} = 19.643$
thin solid lines – undersmoothed ($h = 7.5$) and oversmoothed ($h = 100$) estimates

where ρ denotes the positive rate of proportionality. It can be shown (Kozusko and Bajzer, 2003) that under some not very restrictive conditions the classical Gompertzian model could be an appropriate model of the cancer cells growth. This model yields y as a solution of the differential equation

$$y' = -ay \log \frac{y}{b}, \quad y(0) = y_0. \tag{11}$$

The parameters y_0 and b denote the initial and the maximal possible size of cancer cells population, respectively. The parameter a can be interpreted as the maximal possible rate of increase of the tumor. Taking into account the assumption (10) and the solution of the initial problem (11) we arrive at the formula for the hazard function λ:

$$\lambda(x) = \lambda(x, a, t^*, \lambda^*) = \lambda^* \exp\left(1 - a(x - t^*) - e^{-a(x-t^*)}\right) \tag{12}$$

where

$$\lambda^* = \frac{\rho a b}{2}, \quad t^* = \frac{1}{a} \log\left(\log \frac{b}{y_0}\right)$$

and $\lambda(t^*) = \lambda^*$, λ^* and t^* denote the maximal hazard and the time of its achieving. The parameters a, λ^* are positive, t^* is non-negative.

Simple qualitative properties of this hazard function λ are in a good accordance with clinical observations, at least for some types of cancer. In particular, the fact that the force of mortality does not decrease immediately after surgery agrees with the property

$$\arg\max \lambda(x) = t^* > 0 \text{ for } b \gg y_0.$$

Further, the relation

$$\int_0^\infty \lambda(x)dx = \frac{e}{a}\left[1 - \exp\left(-e^{at^*}\right)\right] < \infty$$

expresses the fact that not all of patients die of cancer.

In this paper we want to show that the proposed hazard function can appropriately fit observed survival data. The parameters a, t^*, λ^* can be estimated by means of the kernel estimates. Such a procedure will be described in the next section.

The actual hazard function can be more complicated. In such a case, we suppose that the cohort of patients is split up into l subcohorts of the sizes $N_1,\ldots,N_l$ and that the each of subcohort size evolves with its special dynamics:

$$N_i(x) = \alpha_i N_0 \exp\left\{-\int_0^x \lambda_i(t)dt\right\},\ i = 1,\ldots,l$$

where $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$.

The splitting up of the patients into subcohorts (i.e. values of parameters α_i) may be carried out with respect to some clinical indications. The hazard function for the complete cohort is given by the formula

$$\lambda_c(x) = -\frac{N'(x)}{N(x)} = \frac{\sum_{i=1}^{l} \alpha_i \lambda_i(x) \exp\left\{-\int_0^x \lambda_i(t)dt\right\}}{\sum_{i=1}^{l} \alpha_i \exp\left\{e^{-\int_0^x \lambda_i(t)dt}\right\}} \tag{13}$$

and λ_c is called a *composed* hazard function.

In terms of parameters α_i, a_i, t_i^*, λ_i^*, $i = 1,\ldots,l$, the composed hazard function can be expressed as

$$\lambda_c(x) = \lambda_c(x, \alpha_1, a_1, t_1^*, \lambda_1^*, \ldots, \alpha_l, a_l, t_l^*, \lambda_l^*)$$

and l four-tuples of parameters should be estimated.

5. Estimates of deterministic model

The aim of this section is to propose the parameters estimate procedure of the composed hazard function (13) by means of the kernel estimates of this function.

For our purpose it is sufficient to use the kernel estimate $\hat{\lambda}_{h,K}$ with the kernel of order two and the bandwidth $\hat{h}_{opt,2} = O(n^{-1/5})$.

Let $\hat{\lambda}_j = \hat{\lambda}_{h,K}(x_j)$, $j = 1,\ldots,s$ denote the kernel estimate of the hazard function at the point x_j. The parameters of λ_c can be estimated by the least squares method, i.e.

$$\begin{aligned}(\hat{\alpha}_1,\ldots,\hat{\alpha}_{l-1},\hat{a}_1,\ldots,\hat{a}_l,\hat{\lambda}_1^*,\ldots,\hat{\lambda}_l^*,\hat{t}_1^*,\ldots,\hat{t}_l^*) = \\ = \arg\min\left\{\sum_{j=1}^{s}\left(\hat{\lambda}_j - \lambda_c(x_j,\alpha_1,a_1,t_1^*,\lambda_1^*,\ldots,\alpha_l,a_l,t_l^*,\lambda_l^*)\right)^2\right\}\end{aligned} \tag{14}$$

where $\alpha_i > 0$, $a_i > 0$, $t_i^* \geq 0$, $\lambda_i^* > 0$, $i = 1, \ldots, l$ and

$$\alpha_l = 1 - \sum_{i=1}^{l-1} \alpha_i.$$

This procedure yields the estimate of the analytical form of the hazard function and it is denoted by

$$\hat{\lambda}_c(x) = \lambda_c(x, \hat{\alpha}_1, \hat{a}_1, \hat{t}_1^*, \hat{\lambda}_1^*, \ldots, \hat{\alpha}_l, \hat{a}_l, \hat{t}_l^*, \hat{\lambda}_l^*). \tag{15}$$

The minimum (14) is located by the Newton method (procedure `nlm` from the R language) and the initial approximations are obtained by the maximal likelihood method from the observed (Kaplan and Meier, 1958) survival function (see Pospíšil (2005) for details).

6. Simulation study

The suggested method is tested on 12 sets of simulated data.

Survival data are simulated using the hazard function (13) with $l = 2$ and the following parameters:

parameter	$i = 1$	$i = 2$
α_i	0.2	0.8
a_i	0.02	0.06
t_i^*	70	100
λ_i^*	0.1	0.005

The data are generated for three patient groups sized – n =50, 100, 250, and four types of censoring time. Let T_{max} denotes the maximal time of the simulated death. Then censoring times are generated as a sample from the uniform distribution on the interval (t_m, t_M) where t_m, t_M depends on T_{max}. The applied types of censorship are

I. $t_m = T_{max}$, $t_M = 2T_{max}$
II. $t_m = \frac{1}{2}T_{max}$, $t_M = 2T_{max}$
III. $t_m = 0$, $t_M = 2T_{max}$
IV. $t_m = 0$, $t_M = T_{max}$

For each of the 12 simulated data sets the kernel estimation of the hazard function $\lambda_{h,K}$ is computed and then the parameters of the hazard function λ_c are estimated. Let the estimated parameters be $\hat{\alpha}_1, \hat{\alpha}_2, \hat{a}_1, \hat{a}_2, \hat{t}_1^*, \hat{t}_2^*, \hat{\lambda}_1^*, \hat{\lambda}_2^*$ and put $\hat{\lambda}_c(x) = \lambda_c(x, \hat{\alpha}_1, \hat{\alpha}_2, \hat{a}_1, \hat{a}_2, \hat{t}_1^*, \hat{t}_2^*, \hat{\lambda}_1^*, \hat{\lambda}_2^*)$. The quality of data fit is measured by the "average L_2-distance" defined by

$$ErrK = \frac{1}{T_{max}} \int_0^{T_{max}} \left(\lambda_c(x) - \hat{\lambda}_{h,K}(x)\right)^2 dx, \; ErrP = \frac{1}{T_{max}} \int_0^{T_{max}} \left(\lambda_c(x) - \hat{\lambda}_c(x)\right)^2 dx. \tag{16}$$

We use value $T_{max} = 200$ for all simulated data sets. The results are summarized in Table 1.

Let us consider the composed hazard function

$$\lambda_c(x) = \lambda_c(x, 0.2, 0.02, 70, 0.1, 0.8, 0.06, 100, 0.005).$$

n		Type of censorship			
		I	II	III	IV
50	$ErrK$	$3.41{\cdot}10^{-6}$	$3.41{\cdot}10^{-6}$	$3.65{\cdot}10^{-6}$	$4.56{\cdot}10^{-6}$
	$ErrP$	$3.29{\cdot}10^{-6}$	$3.31{\cdot}10^{-6}$	$3.58{\cdot}10^{-6}$	$6.7{\cdot}10^{-7}$
100	$ErrK$	$3.98{\cdot}10^{-7}$	$5.06{\cdot}10^{-7}$	$3.32{\cdot}10^{-6}$	$2.04{\cdot}10^{-6}$
	$ErrP$	$3.88{\cdot}10^{-7}$	$4.93{\cdot}10^{-7}$	$3.28{\cdot}10^{-6}$	$2.04{\cdot}10^{-6}$
250	$ErrK$	$1.86{\cdot}10^{-7}$	$1.85{\cdot}10^{-7}$	$2.73{\cdot}10^{-7}$	$2.48{\cdot}10^{-6}$
	$ErrP$	$1.8{\cdot}10^{-7}$	$1.81{\cdot}10^{-7}$	$1.81{\cdot}10^{-7}$	$1.66{\cdot}10^{-6}$

TABLE 1. L_2-distance (16) of actual and estimated hazard functions

$n = 50$	Type of censorship							
	I		II		III		IV	
p	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ
$\alpha_1 = 0.2$	0.05	0.774	0.05	0.767	0.04	0.806	0.12	0.380
$\alpha_2 = 0.8$	0.95	0.194	0.95	0.192	0.96	0.201	0.88	0.095
$a_1 = 0.02$	0.016	0.188	0.016	0.187	0.016	0.221	0.015	0.267
$a_2 = 0.06$	0.020	0.667	0.020	0.664	0.019	0.690	0.060	0.000
$t_1^* = 70$	77	0.096	77	0.103	77	0.105	128	0.823
$t_2^* = 100$	78	0.222	77	0.232	78	0.218	100	0.000
$\lambda_1^* = 0.1$	0.0822	0.178	0.0817	0.183	0.0791	0.209	0.8034	7.034
$\lambda_2^* = 0.005$	0.0023	0.545	0.0022	0.560	0.0019	0.614	0.0050	0.000

TABLE 2. Estimated parameters for distinct simulated data sets, p denotes any parameter, $\hat{p}$ its estimate, $\gamma = |\hat{p} - p|/p$.

$n = 100$	Type of censorship							
	I		II		III		IV	
p	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ
$\alpha_1 = 0.2$	0.17	0.128	0.17	0.134	0.29	0.467	0.24	0.203
$\alpha_2 = 0.8$	0.83	0.032	0.83	0.034	0.71	0.117	0.76	0.051
$a_1 = 0.02$	0.020	0.005	0.020	0.009	0.003	0.831	0.002	0.883
$a_2 = 0.06$	0.048	0.197	0.043	0.277	0.041	0.324	0.044	0.270
$t_1^* = 70$	73	0.050	74	0.060	490	6.000	836	10.942
$t_2^* = 100$	102	0.018	104	0.036	158	0.575	109	0.093
$\lambda_1^* = 0.1$	0.0946	0.054	0.0929	0.071	0.0725	0.275	0.5647	4.647
$\lambda_2^* = 0.005$	0.0038	0.231	0.0037	0.251	0.0005	0.893	0.0016	0.670

TABLE 3. Estimated parameters for distinct simulated data sets, p denotes any parameter, $\hat{p}$ its estimate, $\gamma = |\hat{p} - p|/p$.

Tables 2, 3 and 4 bring the estimates of the parameters obtained by the proposed method (14) including the relative errors: p denotes the parameter, $\hat{p}$ its estimate and $\gamma = |\hat{p} - p|/p$ is the relative error.

It can be seen that the estimation of parameters is of the correct order (relative error less than 0.5) provided that the patient cohort is large enough (at least 100

$n = 250$	Type of censorship							
	I		II		III		IV	
p	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ	$\hat{p}$	γ
$\alpha_1 = 0.2$	0.19	0.058	0.19	0.065	0.18	0.115	0.18	0.123
$\alpha_2 = 0.8$	0.81	0.015	0.81	0.016	0.82	0.029	0.82	0.031
$a_1 = 0.02$	0.028	0.407	0.027	0.351	0.022	0.080	0.023	0.147
$a_2 = 0.06$	0.053	0.109	0.052	0.141	0.060	0.003	0.056	0.067
$t_1^* = 70$	49	0.293	52	0.256	70	0.005	66	0.063
$t_2^* = 100$	101	0.013	102	0.016	98	0.023	99	0.008
$\lambda_1^* = 0.1$	0.0664	0.336	0.0706	0.294	0.1055	0.055	0.1141	0.141
$\lambda_2^* = 0.005$	0.0043	0.145	0.0043	0.131	0.0052	0.042	0.0079	0.583

TABLE 4. Estimated parameters for distinct simulated data sets, p denotes any parameter, $\hat{p}$ its estimate, $\gamma = |\hat{p} - p|/p$.

patients) and the censorship is not excessively severe (i.e. for the censorship types I and II).

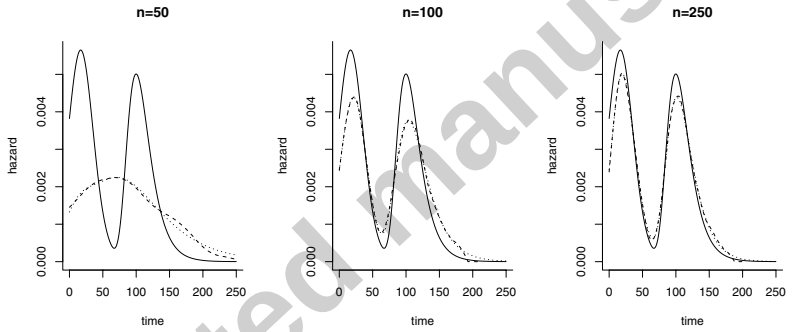

FIGURE 2. Results for simulated data and the censorship of the type II
solid line – $\lambda_c(x, 0.2, 0.02, 70, 0.1, 0.8, 0.06, 100, 0.005)$
dashed line – $\hat{\lambda}_{h,K}$, $h = 102.9599, 28.9217, 20.9116$, successively
dotted line – $\lambda_c(x, \cdot, \hat{\alpha}_1, \hat{\alpha}_2, \hat{a}_1, \hat{a}_2, \hat{t}_1^*, \hat{t}_2^*, \hat{\lambda}_1^*, \hat{\lambda}_2^*)$ (see Tables 2, 3 and 4)

Since visualization is an important component of data analysis the graphical representation of the estimated hazard functions $\hat{\lambda}_c$ together with the kernel estimates $\hat{\lambda}_{h,K}$ for simulated data of type II censorship is given on Figure 2.

7. APPLICATION

The first data set we are going to deal with have been kindly provided by the Masaryk Memorial Cancer Institute in Brno, Czech Republic (Soumarová *et al.*, 2002).

This data set (BRB) include 236 patients with breast carcinoma. The study has been based on the records of women who received both the breast conservative surgical treatment and radiotherapy as well at the Masaryk Memorial Cancer Institute in Brno in the period 1983–1994. The patients with breast carcinoma of the I and

II clinical stage are only included in this study. Of the complete set of 236 patients, 47 (19.9%) died of cancer.

The second data set (BRCB) comprises 152 patients with the same diagnosis and the treatment, but the patients were treated at the Hospital of České Budějovice in the period 1990 – 2005. Of the complete set of patients, 32 (21.1%) died of cancer (Dolečková *et al.*, 2006).

The characteristics of treated data sets are summarized in Table 5.

Set of patients		n	T	n_d	p_d
Breast carcinoma, Brno	BRB	236	220	47	19.9
Breast carcinoma,České Budějovice	BRCB	152	172	32	21.1

TABLE 5. Characteristics of treated patients data sets; n — number of patients, T — maximal follow up in months, n_d — number of deaths, p_d — percents of deaths.

Numbers of deaths in separate years for BRB and BRCB data are presented in Figures 3 and 4.

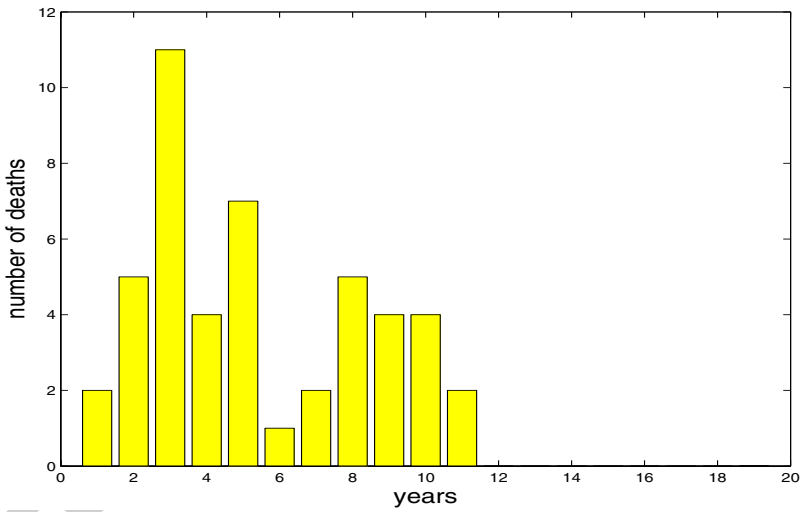

FIGURE 3. Deaths for BRB data

Our procedure is realized in two steps:

1° Find the kernel estimate $\hat{\lambda}_{h,K}$ with the kernel $K(x) = 3/4(1-x^2)\, I_{[-1,1]}$ and the corresponding optimal bandwidth $\hat{h}_{opt,2}$.
Compute values $\hat{\lambda}_{h,K}(x_j)$, $j = 1, \ldots, s$.

2° Use the least squares method (14) to obtain the parameters of the function $\hat{\lambda}_c$.

In order to compare the kernel estimate of the hazard function and the deterministic form of it obtained by (12) both estimates including the confidence intervals are displayed in the same figure. The 95%-confidence intervals are used, it means

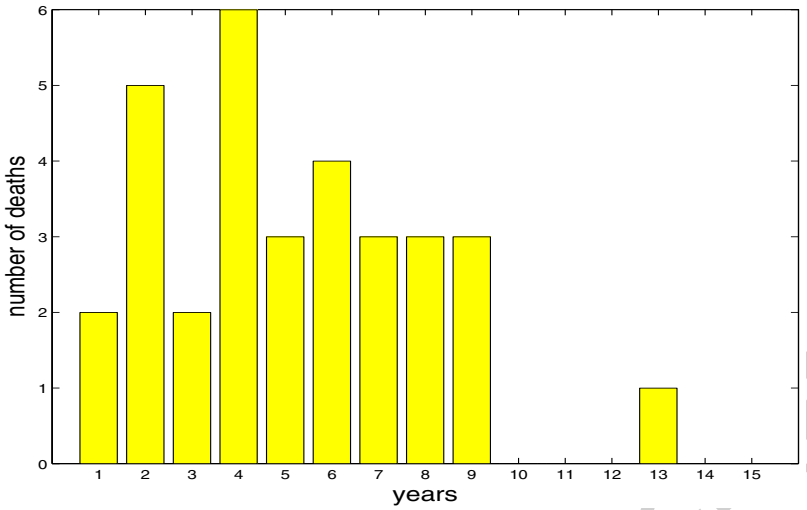

FIGURE 4. Deaths for BRCB data

that the probability that the value $\lambda_c(x)$ lies in the interval is 0.95 (Figures 5 and 6).

Since the estimated deterministic function $\hat{\lambda}_c$ lies within the confidence limits of the kernel estimate, we can conclude that the proposed method could provide a suitable tool for the analysis of survival data for cancer patients.

Table 6 brings the estimates of the parameters obtained by minimization process (14).

	BRB		BRCB	
$\hat{p}$	$i=1$	$i=2$	$i=1$	$i=2$
$\hat{\alpha}_i$	0.8983	0.1017	0.6174	0.3826
$\hat{a}_i$	0.04874	0.02125	0.03517	0.01985
$\hat{t}_i^*$	31.29	161.5	25.2	86.13
$\hat{\lambda}_i^*$	0.00319	0.1254	0.002236	0.004196

TABLE 6. Estimated parameters of composed hazard functions for BRB and BRCB data

The composed hazard function for the BRB data takes the form

$$\hat{\lambda}_c(x) = \hat{\lambda}_{BRB}(x) = \frac{0.8983\,\hat{\lambda}_1(x)\,e^{-\int_0^x \hat{\lambda}_1(t)dt} + 0.1017\,\hat{\lambda}_2(x)\,e^{-\int_0^x \hat{\lambda}_2(t)dt}}{0.8983\,e^{-\int_0^x \hat{\lambda}_1(t)dt} + 0.1017\,e^{-\int_0^x \hat{\lambda}_2(t)dt}}$$

for

$$\hat{\lambda}_1(x) = 0.00319\,\exp\{1 - 0.04874(x - 31.29) - e^{-0.04874(x-31.29)}\}$$

and

$$\hat{\lambda}_2(x) = 0.1254\,\exp\{1 - 0.02125(x - 161.5) - e^{-0.02125(x-161.5)}\}.$$

The composed hazard function for the BRCB data can be expressed as

$$\hat{\lambda}_c(x) = \hat{\lambda}_{BRCB}(x) = \frac{0.6174\,\hat{\lambda}_1(x)\,e^{-\int_0^x \hat{\lambda}_1(t)dt} + 0.3826\,\hat{\lambda}_2(x)\,e^{-\int_0^x \hat{\lambda}_2(t)dt}}{0.6174\,e^{-\int_0^x \hat{\lambda}_1(t)dt} + 0.3826\,e^{-\int_0^x \hat{\lambda}_2(t)dt}},$$

$$\hat{\lambda}_1(x) = 0.002236\,\exp\{1 - 0.03517(x - 25.2) - e^{-0.03517(x-25.2)}\},$$

$$\hat{\lambda}_2(x) = 0.004196\,\exp\{1 - 0.01985(x - 86.13) - e^{-0.01985(x-86.13)}\}.$$

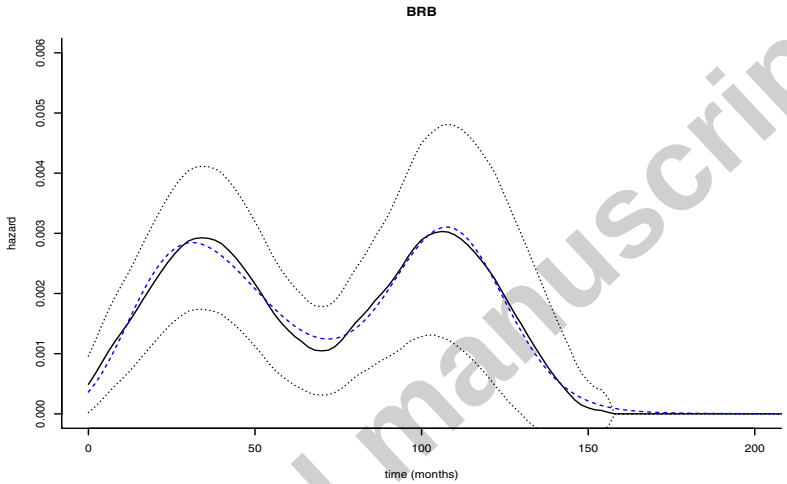

FIGURE 5. Deterministic and kernel estimates of hazard function for BRB data
solid line – kernel estimation $\hat{\lambda}_{h,K}$, $\hat{h}_{opt,2} = 25.2488$
dashed line – composed hazard function $= \hat{\lambda}_{BRB}$
dotted line – confidence intervals for the kernel estimate

8. Discussion and conclusion

The standard method for survival or hazard function parameters estimations from observed data is the maximum likelihood method (see e.g. Hougaard (2001)). The method was adopted and applied for the hazard function of the form (13) (Horová *et al.*, 2008). A disadvantage of the maximum likelihood method is that the numerical minimization need not converge for all simulated data. The proposed method of parameters identification — the kernel smoothing of the hazard function and the subsequent minimization of (14) — converged for all simulated data. Moreover, it identifies parameters more precisely. This fact can be demonstrated by simulations. We provided 100 simulations, in each of them we generated survival data for 250 patients using the hazard function (13) with $l = 2$, with parameters listed in Table 7, and with type III censorship. Subsequently, we estimated the parameters for each of the 100 simulated data sets by the maximum likelihood method

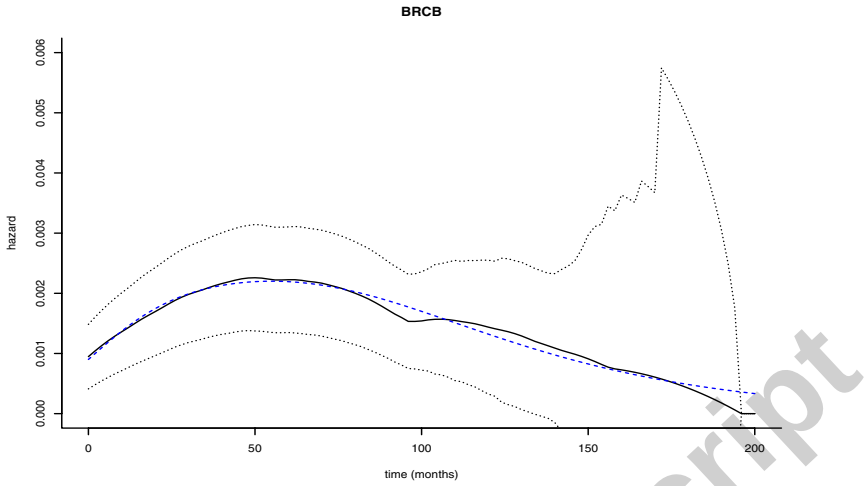

FIGURE 6. Deterministic and kernel estimates of hazard function for BRCB data
solid line – kernel estimation $\hat{\lambda}_{h,K}$, $\hat{h}_{opt,2} = 43.0545$
dashed line – composed hazard function $= \hat{\lambda}_{BRCB}$
dotted line – confidence intervals for the kernel estimate

and by the method proposed in this paper. Table 7 shows average and extremal values of parameters obtained by the both methods.

	Maximum likelihood method			Proposed method		
p	mean	minimum	maximum	mean	minimum	maximum
$\alpha_1 = 0.2$	0.3576	0.336	0.38	0.1948	0.1668	0.2363
$\alpha_2 = 0.8$	0.6424	0.62	0.664	0.8052	0.7637	0.8332
$a_1 = 0.02$	0.07066	0.06671	0.07568	0.02968	0.01913	0.04604
$a_2 = 0.06$	0.05376	0.04409	0.06986	0.05386	0.04433	0.06504
$t_1^* = 70$	16.65	14.31	17.88	51.55	25.02	74.38
$t_2^* = 100$	99.98	94.49	104.8	100.9	96.71	105.4
$\lambda_1^* = 0.1$	0.02099	0.01864	0.0229	0.07052	0.02831	0.1073
$\lambda_2^* = 0.005$	0.005961	0.004551	0.007433	0.004404	0.003482	0.005695

TABLE 7. Characteristics of parameters estimated for 100 simulated survival data sets by the maximum likelihood method and by the proposed method.

The parameters a_i in the hazard function λ_c (13) should characterize an intrinsic property of disease — rate of cancer cells proliferation; roughly saying, $|\log\log 2/a|$ is the doubling time for the cells. We can take notice of the fact, that a_i's estimated for both of the analyzed data sets are not very different, see Table 6. Since these data were collected on patients with the same type of cancer, this observation suggests that these parameters characterize the disease in real terms.

To verify this hypothesis, more patient sets with the same diagnosis and treatment (i.e. the breast carcinoma and the breast conservative surgical treatment with subsequent radiotherapy) should be examine.

Another hypothesis, namely that the parameters a_i characterize a type of cancer can be tested by examination of survival data for cancer patients with a different diagnosis.

The results of the application to the real data do not argue against our assumptions that the hazard is proportional to the rate of cancer cell proliferation. This fact can serve as a starting point for a future collaboration with oncologists or molecular biologists to test the hypothesis in more details.

References

Collett, D.: *Modelling Survival Data in Medical Research,* Chapman &Hall/CRC, Boca Raton-London-New York-Washington,D.C., 2003.

Dolečková M., Horová I., Budíková M. and Hon Z. (2006). Breast Carcinoma: Statistical Evaluation (*in Czech*). *Proceedings of XXX. BOD*, 41–45.

Hougaard, P.: *Analysis of Multivariate Data,* Springer-Verlag, New York-Berlin-Heidelberg, 2001.

Horová I., Zelinka J.: Kernel Estimates of Hazard Functions for Biomedical Data Sets, Chapter in *Statistical Methods for Biostatistics and Related Fields*, 63–86, 2007.

Horová I., Vieu P, Zelinka J.: Optimal Choice of Nonparametric Estimates of a Density and of its Derivatives. *Statistics & Decision*, **20**, 355–378, 2002.

Horová I., Zelinka J., Budíková M.: Estimates of Hazard Functions for Carcinoma Data Sets, *Environmetrics*, **17**, 239–255, 2006.

Horová I., Pospíšil, Z., Zelinka J.: Semiparametric Estimation of Hazard Function for Cancer Patients, *accepted to Sankhya*, 2008.

Jiang J., Marron J. S.: SiZer for Censored Density and Hazard Estimation *preprint*, 2003.

Kaplan E. L., Meier P.: Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481, 1958.

Kozusko, F., Bajzer, Ž.: Combining Gompertzian Growth and Cell Population Dynamics, *Mathematical Biosciences*, **185**, 153–167, 2003.

Marron J. S., Nolan D.: Canonical Kernels for Density Estimation *Stat. & Probab. Lett.*, **7**, 195–199, 1989.

Müller, H.G.: *Nonparametric Analysis of Longitudinal Data*. Lecture Notes in Statistics 46, Springer-Verlag Berlin, Heidelberg, 1988.

Müller H. G., Wang J. L.: Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data: An Alternative Change-Point Models, *Biometrika*, **77**, 2, 305-314, 1990.

Pospíšil, Z.: Gompertzian hazard function. 4^{th} *International Conference Aplimat*, Slovak University of Technology, 341–346, Bratislava, 2005.

Soumarová R., Horová H., Růžičková J., Čoupek P., Šlampa P., Šeneklová Z., Petráková K., Budíková M. and Horová I.: *Local and Distant Failure in Patients with Stage I and II Carcinoma of the Breast Treated with Breast-Conserving Surgery and Radiation Therapy* (in Czech, English summary). Radiační onkologie, **2**(1), 17–24, 2002.

Tanner M. A., Wong W. H.: Data-Based Nonparametric Estimation of the Hazard Function with Applications to Model Diagnostis and Exploratory Analysis. *Journal of the Am. Stat. Association*, **79**, 35, 174-182, 1984.

Therneau, T. M., Grambsch P. M.: *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York, 2000.

Uzunogullari U., Wang J. L.: A Comparision of Hazard Rate Estimators for Left Truncated and Right Censored Data, *Biometrika*, **79**, 2, 297-310, 1992.

Masaryk University, Faculty of Science, Department of Mathematics and Statistics, Kotlářská 2, CZ-611 37 Brno, Czech Republic