Cohort effect in survival analysis: a change-point perspective

Olivier Bouaziz, Grégory Nuel

Cohort effect in survival analysis: a change-point perspective.

Olivier Bouaziz
MAP5, Université Paris Descartes
45 rue des Saints Pères, 75270 Paris Cedex 06
Grégory Nuel
LPMA, CNRS 7599
4 place Jussieu, F-75005 Paris

March 11, 2016

Abstract

In a survival analysis context we suggest a new method to treat the so-called cohort effet as a breakpoint problem where each stratum has its own baseline hazard function and proportional hazard parameter. Our approach combines various baseline hazard estimators (exponential, Weibull, piecewise constant, and non parametric) with the proportional hazard model into a EM-based segmentation approach using constrained HMM. The method provides an automatic procedure to estimate the hazard function in each stratum as well as the full posterior distribution of breakpoint locations. The new method is applied both to simulated data and to a diabetes cohort where it proves its consistence and usefulness.

Keywords: Cohort effect; Constrained HMM; Cox model; EM; Survival analysis

1 Introduction

In epidemiology, it is well known that survival data are subject to the so-called cohort effect (see Yang and Land, 2013, for the definition of cohort effect for instance). In a cohort study, individuals recruited or born at different dates might have heterogeneous hazard rates. While this phenomenon can be continuous in nature with a slow change of hazard rates over time, it is often due to a radical change in the treatment or prevention strategy. For instance, if one is interested in the time until first bacterial infection, the discover of penicillin might represent a breakpoint in the survival study. Patients born at an early date such that they could not benefit from the penicillin treatment would have a different survival distribution compared to other patients who might have had access to penicillin treatment. Other examples include tritherapy in HIV patients, national screening policy for patients with cancer such as breast cancer for instance. This cohort

effect naturally lead us to consider breakpoint models that take into account the survival heterogeneity between patients.

In the literature, one classical method to avoid bias caused by a possible heterogeneity in the survival distribution consists in adjusting with respect to the variable year of birth. For example, in Andersen et al (1993), the authors studied a dataset on nephropathy for diabetics (introduced in Example I.3.11 of their book) using a multi-state model, where each transition intensity models was adjusted with respect to the calendar time of disease onset (see Table VII.2.1 page 520 of their book). The authors concluded that "it is seen that all intensities decrease with t_0 (the calendar year of onset of diabetes), indicating a general medical improvement over time". More recently, adjusting with respect to the year of birth is also one of the recommendation of Kratz (2011) where the authors mention that "Survival analysis tracks length of time without reference to calendar time. This is the reason that many analysis adjust for year of diagnosis or birth cohort (i.e., year or period of years of birth)".

A more sophisticated method to take into account a cohort effect is the Age-Period-Cohort approach which consists in adjusting a proportional hazard model on various cohort-orientated covariates (age, date of birth or recruitment, location, etc.). We refer the reader to Yang and Land (2013) for a thorough review of all recent works for Age-Period-Cohort analysis.

An other alternative consists in dividing the ordered dataset in arbitrary segments (typically, every decade) and estimate some survival quantities on each segment. This is what is done in Bergh et al (1999) for example, where the authors stratified their study on each year of birth and computed odds ratio on each stratum. A refinement of this stratification method could be to merge any adjacent segments whose survival distributions are similar. To our knowledge, there exists no methods that perform automatic location in the change of survival distribution and at the same time allows estimation on each stratum.

From a statistical point of view we consider this situation as a change-point model where abrupt changes occur either in terms of baseline hazard rates or in terms of proportional factors. In such a model, we aim at two objectives: first we want to estimate the hazard rates and the proportional factors in each homogenous region through a Cox model (see Cox et al, 1972). Secondly, we want to accurately provide the number and location of the breakpoints. Recently a constrained Hidden Markov Model (HMM) method was suggested in the context of breakpoint analysis (see Luong et al, 2013). This method allows to perform a full change-point analysis in a segment-based model (one parameter by segment) providing linear EM estimates of the parameter and a full specification of the posterior distribution of change points. In this paper we adapt this method to the context of survival analysis with hazard rate estimates, where the estimation is performed through the EM algorithm (see Dempster et al, 1977) to provide update of the estimates and the posterior distribution at each iteration step.

In the classical Cox model, the baseline is usually left unspecified. This allows great flexibility in the model while the Cox's partial likelihood provides efficient estimation of the regression parameters. Estimation of the cumulative baseline is performed through

the Breslow estimator (see Breslow, 1972). However, in our context classical estimation methods will not lead to consistent estimators due to numerical instabilities. In order to provide estimation in a Cox model with nonparametric baseline, a smooth estimator of the baseline is required. Therefore, different strategies are proposed throughout this article to model the baseline in the Cox model. Three parametric baseline models are studied, the exponential baseline, the Weibull baseline and the piecewise constant hazard baseline, and the nonparametric baseline model is implemented with a kernel type estimator for the baseline.

In Section 2.1, the stratified Cox model is presented along with some discussions of basic assumptions on the model. In Section 2.2, the likelihood of the model is presented. It can be seen as a weighted likelihood where the weights correspond to the posterior probability of each individual to be in each segment given the data and the previous update of the model parameter. The EM algorithm is then introduced as an iterated method to perform estimation in this context. In Section 3, computation of the weights is derived. In Section 4, maximisation of the log-likelihood is developed for a fixed weight. All three parametric baseline hazards and the nonparametric baseline are considered in this section. Section 5 gives a summary of the implementation of the proposed algorithm along with some discussions on the calibration of the algorithm parameters. A simulation study is presented in Section 6 and a real data analysis on survival of diabetic patients is studied in Section 7. Finally, Section 8 concludes this article with some general comments on the proposed methods.

2 Model and estimation procedure

2.1 The stratified hazard rate model

Let T^* represent the survival time of interest associated with its counting process $N^*(t) = I(T^* \leq t)$ and its at risk process $Y^*(t) = I(T^* \geq t)$ for $t \geq 0$. Let $\boldsymbol{X}$ represent a p-dimensional covariate row vector. In practice, T^* might be censored by a random variable C so that we observe $(T = T^* \wedge C, \Delta = I(T^* \leq C), \boldsymbol{X})$. Introduce the observed counting and at risk processes denoted respectively by $N(t) = I(T \leq t, \Delta = 1)$ and $Y(t) = I(T \geq t)$ and let τ be the endpoint of the study. The data consist of n independent replications $(T_i, \Delta_i, \boldsymbol{X}_i)_{i=1,\dots,n}$ associated with their counting process $N_i(t)$ and at risk process $Y_i(t)$, for $t \in [0, \tau]$.

The cohort effect is modelized through the random variables R and its n i.i.d. replications $R_1, R_2, \dots, R_n$ which represent a segment index associated to each individual. We suppose that the population is composed of K segments such that for $i = 1, \dots, n$, $R_i \in \{1, 2, \dots, K\}$. Without loss of generality, we also assume that the $R_i s$ are ordered. For example, if the population is a mixture of three subpopulations such that we have $n = 10$ and two breakpoints occurring after positions 3 and 7 then $R_{1:10} = 1112222333$.

The goal of this paper is to study a hazard Cox model stratified with respect to the

segment index. This model is defined in the following way:

$$\mathbb{E}[dN^*(t)|Y^*(t), \boldsymbol{X}, R] = Y^*(t) \sum_{k=1}^{K} \lambda_k(t) \exp(\boldsymbol{X}\boldsymbol{\beta}_k) I(R = k) dt, \tag{1}$$

where the λ_k represent unknown baseline hazard functions and the $\boldsymbol{\beta}_k$ unknown regression parameters associated to each segment index. Let $\Lambda_k(t) = \int_0^t \lambda_k(s)ds$ represents the cumulative baseline hazard function of the kth segment index. We denote by $\boldsymbol{\theta} = (\Lambda_1, \ldots, \Lambda_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ the model parameter we aim to estimate. Note that if the $R_i s$ were observed and if $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$, this model would reduce to the classical stratified Cox model (see for instance Martinussen and Scheike, 2006, page 190).

In order to make inference on the model parameter we will assume that the endpoint τ is defined such that, for all t in $[0, \tau]$, $\mathbb{P}[T > t] > 0$. This assumption is common in survival analysis settings to prevent us from classical estimation problems that occur in the right tail of the distribution of T, see for instance Andersen et al (1993). We will also suppose the following independent censoring assumption. For all t in $[0, \tau]$, assume that

$$\mathbb{E}[dN^*(t)|Y^*(t), \boldsymbol{X}, R] = \mathbb{E}[dN^*(t)|Y(t), \boldsymbol{X}, R]. \tag{2}$$

This is the classical independent right censoring assumption (see for instance Definition III.2.1. of Andersen et al, 1993 or page 53 of Martinussen and Scheike, 2006) adapted to our stratified model. This assumption means that the censoring variable does not carry any extra information about the probability of observing a new event given the covariate vector and the segment index. It will be trivially fulfilled if one would suppose the censoring variable to be independent of the event time conditionally on $\boldsymbol{X}$ and R. Note however that Equation (2) is slightly more general. In the absence of covariates, examples of data where C and T^* are dependent but Equation (2) is still valid can be found for instance in Exercise 1.8 in Fleming and Harrington (1991).

The main interest in Equation (2) lies in the fact that under independent censoring our model defined by Equation (1) is still verified if we replace the processes $N^*(t)$ and $Y^*(t)$ by their observed counterpart, namely $N(t)$ and $Y(t)$. This is a standard result in survival analysis. In the absence of covariates and of the segment index, the proof of this result can be found for example in Fleming and Harrington (1991), pages 27-29.

Therefore, an estimation procedure on the model parameter can be carried out using the observed data. Let

$$e_i(k; \boldsymbol{\theta}) = \mathbb{P}(T_i, \Delta_i, \boldsymbol{X}_i | R_i = k; \boldsymbol{\theta})$$

represents the contribution of the ith individual to the likelihood conditionally to its segment index being equal to k. From standard arguments on likelihood constructions in the context of survival analysis (see for instance Andersen et al, 1993), we have under

independent and non informative censoring:

$$\log e_i(k;\boldsymbol{\theta}) = \int_0^\tau \left\{\log\left(\lambda_k(t)\right) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right\} dN_i(t) - \int_0^\tau Y_i(t)\lambda_k(t)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)dt, \quad (3)$$

where the equality holds true up to a constant that does not depend on the model parameter $\boldsymbol{\theta}$. Since the segment indexes are not observed, the conditional likelihood of our model with respect to the segment indexes cannot be directly computed. To overcome this problem, an Expectation-Maximization (EM) algorithm procedure is developed in the next section.

2.2 The EM algorithm

By considering the segmentation $R_{1:n} = R_1, R_2, \ldots, R_n$ as a latent variable, the EM-algorithm (see Dempster et al, 1977) consists in performing alternatively until convergence the following two-steps.

Expectation Step: compute the conditional expected log-likelihood

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \int_{R_{1:n}} \mathbb{P}(R_{1:n}|\text{data};\boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(R_{1:n}, \text{data};\boldsymbol{\theta}) dR_{1:n}$$

where $\boldsymbol{\theta}_{\text{old}}$ denote the previous value of the parameter and data $= (T_{1:n}, \Delta_{1:n}, \boldsymbol{X}_{1:n})$.

Maximization Step: update parameter with

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}). \quad (4)$$

Assuming that the prior segmentation distribution $\mathbb{P}(R_{1:n};\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, we easily get:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_i(k;\boldsymbol{\theta}_{\text{old}}) \log e_i(k;\boldsymbol{\theta}) \quad (5)$$

where for any $i \in \{1,\ldots,n\}$, $k \in \{1,\ldots,K\}$ and $\boldsymbol{\theta}$ we define:

$$w_i(k;\boldsymbol{\theta}) = \mathbb{P}(R_i = k|\text{data};\boldsymbol{\theta}).$$

Our EM algorithm hence alternates two steps. First, the E-Step which consists in computing the weights $w_i(k;\boldsymbol{\theta}_{\text{old}})$. This is done in Section 3 using a constrained Hidden Markov Model (HMM). Then for the M-Step, Equation (4) needs to be solved. This is done in Section 4 using the weighted log-likelihood expression given by Equation (5).

3 Computation of the posterior segment distribution

As suggested in Luong et al (2013), the posterior segmentation distribution can be obtained using the constrained HMM. For completeness, we give all the necessary information to implement this constrained HMM. The basic idea consists in modeling the segmentation variable $R_{1:n}$ using a Markov chain over $\{1, \dots, K, K+1\}$ where $K+1$ is an absorbing (technical junk) state. The segmentation always start with $R_1 = 1$ and its transition matrix $\mathbb{P}(R_i|R_{i-1})$ is given by the following matrix (in the particular case where $K = 4$):

$$\left(\begin{array}{ccccc} 1-\eta_i(1) & \eta_i(1) & 0 & 0 & 0 \\ 0 & 1-\eta_i(2) & \eta_i(2) & 0 & 0 \\ 0 & 0 & 1-\eta_i(3) & \eta_i(3) & 0 \\ 0 & 0 & 0 & 1-\eta_i(4) & \eta_i(4) \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

where $\eta_i(k) = \mathbb{P}(R_i = k+1|R_{i-1} = k)$ is a prior distribution. In order to obtain a valid segmentation of n points into K segments, one must add the constraint that $\{R_n = K\}$, this is why the model can be seen as a constrained HMM. A very natural choice for the prior distribution is to use $\eta_i(k) = \text{constant} \in [0, 1]$ which leads to a uniform prior distribution over the space of segmentations. But more sophisticated prior might be use: priors forbidding change-points at certain locations (this might for example be useful for dealing with ties in data ordering), priors incorporating knowledge on most likely breakpoint locations, or even using posterior segmentation distribution from a previous study as a prior.

For any given parameter $\boldsymbol{\theta}$, we then introduce the following forward and backward quantities: $F_i(k;\boldsymbol{\theta}) = \mathbb{P}(\text{data}_{1:i}, R_i = k;\boldsymbol{\theta})$ and $B_i(k;\boldsymbol{\theta}) = \mathbb{P}(\text{data}_{i+1:n}, R_n = K|R_i = k;\boldsymbol{\theta})$ for all $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$. These quantities can be computed recursively using the following recursions:

$$F_i(k;\boldsymbol{\theta}) = F_{i-1}(k-1;\boldsymbol{\theta})\eta_i(k-1)e_i(k;\boldsymbol{\theta}) + F_{i-1}(k;\boldsymbol{\theta})(1-\eta_i(k))e_i(k;\boldsymbol{\theta}) \tag{6}$$

$$B_{i-1}(k;\boldsymbol{\theta}) = (1-\eta_i(k))e_i(k;\boldsymbol{\theta})B_i(k;\boldsymbol{\theta}) + \eta_i(k)e_{i+1}(k+1;\boldsymbol{\theta})B_i(k+1;\boldsymbol{\theta}) \tag{7}$$

and we can derive from them posterior distributions of interest:

$$\mathbb{P}(R_i = k|\text{data};\boldsymbol{\theta}) = w_i(k;\boldsymbol{\theta}) \propto F_i(k;\boldsymbol{\theta})B_i(k;\boldsymbol{\theta}) \tag{8}$$

$$\mathbb{P}(\text{BP}_k = i|\text{data};\boldsymbol{\theta}) \propto F_i(k;\boldsymbol{\theta})\eta_{i+1}(k)e_{i+1}(k+1;\boldsymbol{\theta})B_{i+1}(k+1;\boldsymbol{\theta}) \tag{9}$$

where $\{\text{BP}_k = i\} = \{R_i = k, R_{i+1} = k+1\}$. It is hence clear that Equation (8) allows to compute the marginal weights used in the EM algorithm (Section 2.2) while Equation (9) gives the marginal distribution of the k^{th} breakpoint. Note that the full posterior segmentation distribution can be proved to be an heterogeneous Markov chain which transition can be derived immediatelty from Equations (8) and (9) (see Luong et al, 2013, for more details).

Let us finally point out that the likelihood also can be derived from the forward-backward quantities and for any $i \in \{1, \ldots, n\}$ as:

$$\mathbb{P}(\text{data}|\boldsymbol{\theta}) = \frac{\sum_{R_{1:n}} \mathbb{P}(\text{data}, R_{1:n}, R_n = K|\boldsymbol{\theta})}{\sum_{R_{1:n}} \mathbb{P}(R_{1:n}, R_n = K|\boldsymbol{\theta})} = \frac{\sum_{k=1}^{K} F_i(k;\boldsymbol{\theta}) B_i(k;\boldsymbol{\theta})}{\sum_{k=1}^{K} F_i^0(k) B_i^0(k)} \tag{10}$$

where F^0 and B^0 are obtained through recursions (6) and (7) by replacing all $e_i(k;\boldsymbol{\theta})$ by 1:

$$F_i^0(k) = F_{i-1}^0(k-1)\eta_i(k-1) + F_{i-1}^0(k)(1-\eta_i(k))$$
$$B_{i-1}^0(k) = (1-\eta_i(k))B_i^0(k) + \eta_i(k)B_i^0(k+1).$$

These quantities depend only on η, n and K, thus they do not need to be updated during the EM algorithm.

4 Log-likelihood maximization with known weights

Suppose you have at hand some preliminary estimator $\boldsymbol{\theta}_{\text{old}}$. In Section 3, we showed how to use this quantity to estimate the marginal posterior probability $w_i(k;\boldsymbol{\theta}_{\text{old}})$ of position i to be in the k^{th} segment given the data and under $\boldsymbol{\theta}_{\text{old}}$. From the expression of the $e_i(k,\boldsymbol{\theta})$ derived in (3), Equation (4) can be solved by maximizing a simple weighted log-likelihood. When the weights are all equal to 1, statistical inference has already been studied, either in a fully parametric case if one assumes a parametric form for the baseline hazard rate (see for instance Kalbfleisch and Prentice, 2002) or in a semiparametric way if the baseline hazard rate is left unspecified which corresponds to the well known Cox model. In the latter case, a weighted log-likelihood has also been briefly studied in Therneau and Grambsch (2000), pages 161-168. But in both parametric and semiparametric cases, our weighted log-likelihood estimation procedure is very similar to the standard estimation techniques used in the absence of weights.

From Equation (1) applied to the observed counting and at-risk processes, one gets the following decomposition:

$$N_i(t)I(R_i = k) - \int_0^t Y_i(s)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)I(R_i = k)d\Lambda_k(s) = I(R_i = k)M_i(t),$$

where $M_i(t)$ is a martingale with respect to the filtration $\sigma(N_i(s), Y_i(s), \boldsymbol{X}_i, I(R_i = k) : 0 \leq s \leq t, k = 1, \ldots, K)$. Taking the expectation conditionally on $\{N_{1:n}(t), Y_{1:n}(t), \boldsymbol{X}_{1:n} : 0 \leq t \leq \tau\}$, summing over the n individuals and taking the differential of this expression gives:

$$\sum_{i=1}^{n} \{dN_i(t)w_i(k;\boldsymbol{\theta}_{\text{old}}) - Y_i(t)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)w_i(k;\boldsymbol{\theta}_{\text{old}})d\Lambda_k(t)\} = \sum_{i=1}^{n} w_i(k;\boldsymbol{\theta}_{\text{old}})dM_{ik}(t). \tag{11}$$

This is the weighted version of the standard martingale decomposition. In Section 4.3 a weighted Nelson-Aalen estimator will be derived from this equality.

In the next sections, we introduce different estimators obtained from different choices for the baseline hazard rate in a Cox model. We propose three parametric families for the baseline hazard: the exponential, the Weibull and the piecewise constant hazard cases. We also study the nonparametric case that arises when the baseline hazard is left unspecified. This case is the most flexible since it does not require any particular form for the baseline hazard but, as shown in Section 5.2, this model involves a smoothing parameter in the estimation procedure in order to consistently estimate the model parameter and the posterior segment distribution.

4.1 The exponential and Weibull baseline hazards

In this model, we assume that the baseline hazard in the k^{th} segment index belongs to the Weibull family with shape parameter λ_k and scale parameter p_k. That is, $\lambda_k(t) = p_k(t/\lambda_k)^{p_k-1}/\lambda_k$, $\Lambda_k(t) = (t/\lambda_k)^{p_k}$ and $S_k(t) = \exp(-(t/\lambda_k)^{p_k})$.

Equation (3) can then be written in the following way:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i\left(\log(p_k) - p_k\log(\lambda_k) + (p_k-1)\log(T_i) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \left(\frac{T_i}{\lambda_k}\right)^{p_k}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

The exponential family is derived as a special case of the Weibull case by setting $p_k = 1$ for all $k = 1, \ldots, K$. In that case, Equation (3) reduces to:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i\left(-\log(\lambda_k) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \left(\frac{T_i}{\lambda_k}\right)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

Computation of the estimates through Equation (4) is done via the **survreg** function in the `survival` R package. The gradient vector and Hessian matrix can directly be derived from the expression of the log-likelihood and the estimates can then be computed using the Newton-Raphson algorithm. A weight option is also available in the **survreg** function which allows to compute estimates that precisely maximize the log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ presented in Equation (4).

The models obtained under these families of baseline hazard functions have the nice property that they both belong to the class of parametric Cox models and of parametric Accelerated Failure Time models (Kalbfleisch and Prentice, 2002). Moreover, the two parameters of the Weibull family make the baseline hazard quite flexible. As a matter of fact, the Weibull model will provide a fairly good fit to any true baseline hazard that is monotone with time. However, these families of model will not properly fit a model with true baseline hazard having a bathtub shape (i.e a $\cup$ shape) or an upside down bathtub shape (i.e. a $\cap$ shape) which are common types of baseline that can occur in practice.

The model introduced in the next section does not assume any specific shape for the baseline hazard and consequently will be able to fit any class of baseline hazard functions. However, this model requires to specify in advance a number of cutpoints and makes the approximation that the hazard is constant between each cutpoint.

4.2 The piecewise constant baseline hazard

In this model, the baseline hazard on each segment index is assumed to be piecewise constant on L cuts represented by $c_0, c_1, \ldots, c_L$, with the convention that $c_0 = 0$ and $c_L = +\infty$. Let $I_l(t) = I(c_{l-1} < t \leq c_l)$. We suppose that

$$\lambda_k(t) = \sum_{l=1}^{L} I_l(t)\alpha_l^k,$$

$$\Lambda_k(t) = \alpha_1^k t I_1(t) + \sum_{l=2}^{L} (\alpha_1^k c_1 + \cdots + \alpha_{l-1}^k (c_{l-1} - c_{l-2}) + \alpha_l^k (t - c_{l-1}))I_l(t),$$

$$S_k(t) = \exp(\alpha_1^k t) I_1(t) + \sum_{l=2}^{L} \exp(\alpha_1^k c_1 + \cdots + \alpha_{l-1}^k (c_{l-1} - c_{l-2}) + \alpha_l^k (t - c_{l-1}))I_l(t).$$

Equation (3) can then be written in the following form:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i \left(\log(\lambda_k(T_i)) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \int_0^\tau Y_i(t)\lambda_k(t)dt \exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

For computational purpose, it is interesting to note that the log-likelihood can be written in a Poisson regression form. Introduce $R_{i,l} = \int_0^\tau Y_i(t)I_l(t)dt = c_l \wedge T_i - c_{l-1}$, the total time individual i is at risk in the lth interval and $O_{i,l} = \int_0^\tau I_l(t)dN_i(t) = I_l(T_i)\Delta_i$, the number of events for individual i in the lth subinterval. Then, we have $\Delta_i \log(\lambda_k) = \sum_l O_{i,l}\log(\alpha_l^k)$, $\int_0^{+\infty} Y_i(t)\lambda_k(t)dt = \sum_l \alpha_l^k R_{i,l}$ and the log-likelihood can be written again as:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L} w_i(k;\boldsymbol{\theta}_{\text{old}}) \left\{O_{i,l}(\log(\alpha_l^k) + \boldsymbol{X}_i\boldsymbol{\beta}_k) - \alpha_l^k R_{i,l}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)\right\}.$$

This log-likelihood is proportional to the log-likelihood one would obtain in a Poisson regression, where the $O_{i,l}$ are the response variables and are assumed to follow, conditionally on the $\boldsymbol{X}_i$, a Poisson distribution with parameter equal to $\alpha_l^k R_{i,l}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)$. Therefore, the estimates can easily be computed using the **glm** function in the R software and specifying $\log(R_{i,l})$ as "offsets" in the model. See for instance Aalen et al (2008) p.223-225 for more details on the connection between piecewise-constant hazard model and Poisson regression. A weight option is also available in the **glm** function.

Finally, note that the exponential case could also be derived as a special case of the piecewise constant hazard family with $L = 1$.

As mentioned earlier, the piecewise constant hazard model is very useful when one does not know the shape of the baseline hazard a priori. However one must specify in advance the value of L in the model. A low value of L will make the estimator imprecise between two cut points, since the model makes the assumption that the hazard is constant on these intervals. On the other hand, as the number of cut points increases the specification of the baseline hazard becomes more flexible. But increasing the value of

L will also automatically increase the variance of the baseline hazard estimator. Therefore, a too large value of L should be avoided since it might lead to overfitting. So a balance should be kept between bias and variance by choosing an adequate number of cut points. The choice of L is discussed in more details in Section 5.3. In the next section, the nonparametric setting where the baseline hazard function is left completely unspecified is presented.

4.3 The nonparametric baseline hazard

In the absence of weights, this model has been widely used because of its great flexibility, the baseline hazard being estimated without making any assumption on its shape, and because it can easily be implemented in a straightforward manner. First, the regression parameter is estimated by maximizing the Cox partial likelihood which contains terms involving only the regression parameter (and not the baseline hazard). Secondly, the baseline hazard estimator is deduced by the martingale decomposition of the observed counting process. More details on the standard estimation procedure in the Cox model can be found for instance in Andersen et al (1993) or Fleming and Harrington (1991).

The martingale decomposition in Equation (11) suggests, for a fixed $\boldsymbol{\beta}_k$, the following weighted Nelson-Aalen type estimator for Λ_k:

$$\tilde{\Lambda}_k(t, \boldsymbol{\beta}_k) = \sum_{i=1}^{n} \int_0^t \frac{w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(s)}{\sum_j Y_j(s) \exp(\boldsymbol{X}_j \boldsymbol{\beta}_k) w_j(k; \boldsymbol{\theta}_{\text{old}})}.$$

Now, plugging-in this quantity into $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ gives the following weighted Cox partial likelihood:

$$Q^{\text{PL}}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^\tau$$
$$\left\{ \boldsymbol{X}_i \boldsymbol{\beta}_k + \log(w_i(k; \boldsymbol{\theta}_{\text{old}})) - \log \left(\sum_{j=1}^{n} Y_j(t) \exp(\boldsymbol{X}_j \boldsymbol{\beta}_k) w_j(k; \boldsymbol{\theta}_{\text{old}}) \right) \right\} w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t).$$

Introduce for $k = 1, \ldots, K$, $l = 0, 1, 2$, $S_k^{(l)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = \sum_j Y_j(t) \boldsymbol{X}_j^{\otimes l} \exp(\boldsymbol{X}_j \boldsymbol{\beta}) w_j(k; \boldsymbol{\theta}_{\text{old}})$ and $E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = S_k^{(1)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) / S_k^{(0)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})$. Then, on each stratum k, define the score function

$$U_k(\boldsymbol{\beta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n} \int_0^\tau \{\boldsymbol{X}_i - E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})\} w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t),$$

such that $\widehat{\boldsymbol{\beta}}_k$ verifies the equality $U_k(\widehat{\boldsymbol{\beta}}_k | \boldsymbol{\theta}_{\text{old}}) = 0$.

Introduce $V_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = S_k^{(2)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) / S_k^{(0)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) - E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})^{\otimes 2}$ and let

$$I_k(t, \boldsymbol{\beta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n} \int_0^\tau V_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t),$$

represents minus the derivative of the score function with respect to $\boldsymbol{\beta}$. Then, computation of the estimator $\widehat{\boldsymbol{\theta}}$ can be performed using the iterative Newton-Raphson algorithm. The m^{th} iteration step writes as follows:

$$\widehat{\boldsymbol{\beta}}_k^{(m)} = \widehat{\boldsymbol{\beta}}_k^{(m-1)} + I_k(t, \widehat{\boldsymbol{\beta}}_k^{(m-1)}|\boldsymbol{\theta}_{\text{old}})^{-1} U_k(\widehat{\boldsymbol{\beta}}_k^{(m-1)}|\boldsymbol{\theta}_{\text{old}}).$$

At convergence, we get the estimator $\widetilde{\boldsymbol{\theta}} = (\tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_K, \widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ where $\tilde{\Lambda}_k(t) = \tilde{\Lambda}_k(t, \widehat{\boldsymbol{\beta}}_k)$ are plug-in Nelson-Aalen estimators of the cumulative hazard functions. Note that the $\widetilde{\boldsymbol{\theta}}$ estimator can be computed with the **coxph** function in the R `survival` library. The weights option can be directly specified in this function.

Finally, as for the parametric models, computation of the new weights is done through the EM algorithm (see Section 3). Then, a simple idea could be to use plug-in estimators again, i.e. to replace $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}}$ in the expression of the $e_i(k;\boldsymbol{\theta})$. However, although this is a relevant strategy for the parametric models it will not lead to a consistent estimator for the Cox model. Because of the shape of the Nelson-Aalen estimators, which are stepwise functions, the information in the estimated partial likelihood (or equivalently in $e_i(k;\widetilde{\boldsymbol{\theta}})$), at a given time point is limited. To stabilize the solution, smoothing is needed. In Section 5.2, new kernel type estimators of the Λ_ks and λ_ks are derived and are used as plug-in estimates in order to compute the weights.

5 Practical implementation

5.1 Parametric baseline hazards

The parametric case is straightforward: the final estimators are obtained by alternating computation of the estimates through Equation (4) and computation of the weights through the posterior segment distribution calculated in Section 3.

The algorithm of our estimation procedure is as follows. First suppose you have at your disposal an initial weight function $w_i(k;\boldsymbol{\theta}_{\text{old}})$.

Step 1. Compute $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ from Equation (4). In the exponential or Weibull models, this can be done via the **survreg** function in R (see Section 4.1) and in the piecewise constant hazard model, this can be done via the **glm** function in R (see Section 4.2)

Step 2. Compute the new weights $w_i(k;\widehat{\boldsymbol{\theta}})$ using Equation (8) in Section 3.

Step 3. Let $\boldsymbol{\theta}_{\text{old}} = \widehat{\boldsymbol{\theta}}$ and return to Step 1.

5.2 Nonparametric baseline hazard

The nonparametric case requires one supplementary step. After the first step, smoothed versions of the baseline hazard and cumulative baseline hazard estimators need to be derived. The weighted log-likelihood and the weights are then computed using these smoothed estimators. We propose in this work to use kernel type estimators but our

method could be extended to any type of smoothing estimators such as wavelets, splines, k-nearest neighbor estimators, projection estimators etc.

Let K be a kernel such that $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2K(u)du < \infty$ and $\int K^2(u)du < \infty$. Let h be a bandwidth satisfying $h \to 0$ and $nh \to \infty$ as n tends to infinity. We now introduce the smoothed versions of $\tilde{\lambda}_k$ and $\tilde{\Lambda}_k$:

$$\hat{\lambda}_k(t) = \frac{1}{h}\sum_{i=1}^{n}\int K\left(\frac{u-t}{h}\right)d\tilde{\Lambda}_k(u) \text{ and } \hat{\Lambda}_k(t) = \int_0^t \hat{\lambda}_k(s)ds, \tag{12}$$

and we note $\widehat{\boldsymbol{\theta}} = (\hat{\Lambda}_1, \ldots, \hat{\Lambda}_K, \widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$. This new estimator is now used to estimate $e_i(k;\boldsymbol{\theta})$ and then to obtain estimators of the weights. From Equation (3) we have:

$$\log\left(e_i(k;\widehat{\boldsymbol{\theta}})\right) = \Delta_i\left(\log\left(\hat{\lambda}_k(T_i)\right) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_k\right) - \hat{\Lambda}_k(T_i)\exp(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_k). \tag{13}$$

Note that the weighted likelihood $Q(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}_{\text{old}})$ obtained from these $e_i(k;\widehat{\boldsymbol{\theta}})$ does not reduce to a partial likelihood like in Section 4.3 due to the use of smoothed hazard and cumulative hazard estimators. However this is not an important matter since our algorithm does not require the maximization of this likelihood: Equation (13) is only needed for the computation of the new weights from Equation (8) in Section 3 while the optimization step only involves the Cox partial likelihood and is easily performed through the Newton-Raphson algorithm.

The final algorithm of our estimation procedure is as follows. First suppose you have at your disposal an initial weight function $w_i(k;\boldsymbol{\theta}_{\text{old}})$.

Step 1. Compute $\widetilde{\boldsymbol{\theta}}$ using the Newton-Raphson algorithm described in Section 4.3. This can be done via the **coxph** function in R.

Step 2. Smooth the $\tilde{\lambda}_k$ and $\tilde{\Lambda}_k$ using Equation (12). This gives $\widehat{\boldsymbol{\theta}}$.

Step 3. Compute $\log\left(e_i(k;\widehat{\boldsymbol{\theta}})\right)$ as in Equation (13) and get the new weights $w_i(k;\widehat{\boldsymbol{\theta}})$ from Equation (8) in Section 3.

Step 4. Let $\boldsymbol{\theta}_{\text{old}} = \widehat{\boldsymbol{\theta}}$ and return to Step 1.

5.3 Choice of the parameters and stopping rule to find the correct model

These algorithms need to be initialized by either choosing initial model parameters or by directly choosing initial weight functions. We propose the following *ad-hoc* method to initialize the weights for a sample of size n and K segments. First divide the sample in K segments and for any individual i in segment k, choose $w_i(k,\boldsymbol{\theta}_{\text{old}}) = p$ with p a high number between 0 and 1 (for instance, take $p = 0.7$). For any individual j that is not in segment k, choose $w_j(k,\boldsymbol{\theta}_{\text{old}}) = 1 - p$.

In all models, the Newton-Raphson algorithm is initialized by taking the null vector for $\widehat{\boldsymbol{\beta}}_k^{(0)}$. Step 2 in the parametric models and step 3 in the Cox model are performed using the R package `postCP` developed by Luong et al (2013).

The exponential and Weibull baseline hazard models only require the initialization of either the model parameters or the weights. On the opposite, the piecewise constant baseline hazard model and the nonparametric baseline model require an extra parameter to be chosen. In both models, the estimation procedure is not very sensitive to the choice of this parameter, especially in terms of breakpoints detection. In particular, the number of cut points in the piecewise constant hazard is set by default to 3 and as shown in the simulation section, this leads to very performant breakpoints selection. Increasing the number of cut points does usually not make the breakpoints detection more accurate. These 3 breakpoints can be chosen for instance from the data as the quantiles of the event times of order 0.25, 0.5 and 0.75 respectively. The same phenomena happens for the choice of the bandwidth in the nonparametric model: detecting the correct number of breakpoints is not much affected by the choice of the bandwidth. However, it might still be of interest to find an optimal bandwidth if one wants to give a precise estimation of the baseline hazard. This problem is classical for density estimation and has been studied for nonparametric estimation of baseline hazards by Andersen et al (1993). Equations (4.2.25) and (4.2.26) of their book suggest that a bandwidth of order $n^{-1/5}$ would give the best compromise between bias and variance trade-off in the estimation of the baseline hazard. In particular asymptotic normality of order $(nh)^{1/2}$ would be achieved with such a bandwidth as expressed by their theorem IV.2.4. More discussions about how to choose the bandwidth from the data can be found in Andersen et al (1993), see in particular their Examples IV.2.3, IV.2.4 and IV.2.5. Since the interest in the choice of the bandwidth is limited in our context we will not pursue this discussion here but as a rule of thumb we recommend the user to choose $h = n^{-1/5}$ in real data situations.

One other important issue is to find the correct number of breakpoints in the dataset. A simple solution consists to start with a model with one breakpoint and increment the number of breakpoints one by one. As presented in the real data analysis for example (see Section 7) a visual inspection of the plots of the maximum a posteriori of the breakpoints can help to find the right model. However, the conclusion from theses plots can be subjective and it is therefore important to propose a numerical indicator that helps discriminating between different models. We propose the following BIC criterion designed to make a tradeoff between information provided by the data on a model and the complexity of the model:

$$\mathrm{BIC}(d) = -2\log \mathbb{P}(\mathrm{data}|\hat{\boldsymbol{\theta}}) + d\log(n)$$

where the likelihood $\mathbb{P}(\mathrm{data}|\hat{\boldsymbol{\theta}})$ can be computed using Equation (10), and d corresponds to the dimension of the model. The value of d is different for every model, it corresponds to the total number of parameters that need to be estimated. For the exponential baseline, $d = (p+1)K$, for the Weibull baseline, $d = (p+2)K$ and for the piecewise constant hazard baseline, $d = (p+L)K$. No such indicator can be derived for a nonparametric baseline hazard since in that case the number of parameters to be estimated equals infin-

ity. This BIC criterion is used in Section 7 for the exponential baseline to discriminate between different models and find the correct number of breakpoints.

6 Simulated data

In this section we evaluate the performance of our estimation technique through numerical experiments. We consider a Cox model as defined by Equation (1), with $K = 3$ segments and a binary covariate $\boldsymbol{X}$ distributed as a Bernoulli variable with parameter equal to 0.5. We consider different scenarios corresponding to different baseline hazards and different regression parameters:

Scenario 1. Exponential baselines, with $\lambda_1(t) = 1$, $\lambda_2(t) = 0.5$, $\lambda_3(t) = 0.7$ and $\beta_1 = 1.5$, $\beta_2 = -0.5$, $\beta_3 = -0.5$.

Scenario 2. Weibull baselines, $\lambda_1(t) = 5t^4$, $\lambda_2(t) = 2t$, $\lambda_3(t) = 2t$ and $\beta_1 = 1.5$, $\beta_2 = -1$, $\beta_3 = -5$.

Scenario 3. Piecewise constant baselines,

$$\begin{aligned}\lambda_1(t) &= 0.8\, I(0 < t \leq 1) + 1.2\, I(1 < t \leq 3) + 1.6\, I(3 < t),\\ \lambda_2(t) &= 1.2\, I(0 < t \leq 4) + 1.6\, I(4 < t \leq 6) + 2\, I(6 < t),\\ \lambda_3(t) &= 1.6\, I(0 < t \leq 5) + 2\, I(5 < t \leq 7) + 2.4\, I(7 < t),\end{aligned}$$

and $\beta_1 = 1.5$, $\beta_2 = -0.5$, $\beta_3 = -1.5$.

Scenario 4. Gompertz baselines, $\lambda_1(t) = e^{5t}$, $\lambda_2(t) = e^{2t}$, $\lambda_3(t) = e^{2t}$ and $\beta_1 = 1.5$, $\beta_2 = -0.5$, $\beta_3 = -1.5$.

In all three scenarios, the sample size n equals 3 000, and the data were simulated such that $R_1 = \cdots = R_{1000} = 1$, $R_{1001} = \cdots = R_{2000} = 2$ and $R_{2001} = \cdots = R_{3000} = 3$. Each scenario was calibrated such that the change in the hazard distribution between segments 1 and 2 was more important than the difference in the hazard distribution between segments 2 and 3. This is illustrated by Figure 1 which provides the plots of the conditional hazard rates in each scenario. The censoring variable was chosen as a uniform distribution such that approximately 50% of the observations were censored in each scenario. In Scenario 1, the censoring was distributed as a uniform distribution with parameters 0 and 2.4, such that 24%, 65% and 60% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 2, the censoring was distributed as a uniform distribution with parameters 0 and 1.8, such that 33%, 47% and 67% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 3, the censoring was distributed as a uniform distribution with parameters 0 and 1.5, such that 38%, 54% and 58% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 4, the censoring was distributed as a uniform distribution with parameters 0 and 0.9, such that 23%, 58% and 67% of individuals were respectively censored in segments 1, 2 and 3. For the piecewise constant hazard model estimator, as recommended in Section 5.3,

the cuts positions were chosen from the quartile of the data. This lead us to take the value $0.2, 0.5$ and 1.1 for Scenario 1, $0.4, 0.7, 1$ for Scenario 2, 0.15, 0.35 and 0.5 for Scenario 3 and 0.1, 0.2 and 0.4 for Scenario 4. For the nonparametric baseline hazard model estimator, as recommended in Section 5.3, the bandwidth was chosen equals to $3000^{-1/5} \approx 0.2$ in all scenarios. Finally we ran 1 000 replications of each of these scenarios and the results were reported in Tables 1.

In all scenarios, detection of the first breakpoint is usually very accurate where in many cases the average breakpoint location is exactly equal to the true breakpoint location, 1 000. The second breakpoint is more difficult to detect as shown by wider confidence intervals even though the average breakpoint location is usually close to the true breakpoint location, 2 000. The average value of the marginal probability of breakpoint detections also illustrate the uncertainty about the second breakpoint location: the probability for the first breakpoint location is in all cases much higher than for the second breakpoint location.

The most problematic breakpoint to find corresponds to the breakpoint from segment 2 to 3 under Scenario 1 and as a matter of fact none of the proposed methods manage to provide an accurate 95% confidence interval. In this scenario, for every estimation methods there was a probability of approximately 1 over 1 000 that the algorithm fails to find the second breakpoints leading to an error in the program.

It is interesting to notice that on the overall the true hazard distribution of the data does not seem to play any role in the detection power of our estimation methods as long as the change in the hazard distribution in two segments is big enough. For instance, in Scenario 4, which involves a simulation setup that does not correspond to any of the parametric baseline distributions proposed in the different estimation methods, all estimators find very accurate breakpoint locations with very narrow confidence intervals. The estimation performance of the regression parameter does not seem to be much affected by the data simulation setup neither, since the Weibull, piecewise constant and nonparametric baseline estimators show little difference in their estimation performance from one scenario to another. One exception is the exponential baseline estimator which seems to behave poorly in Scenarios 2 and 4 when looking at the regression parameter estimates and the confidence intervals for the second breakpoint compared to the other estimators.

Globally, all estimators are performant both in breakpoint detections and parameters estimation as long as the change in the hazard distribution is big enough from one segment to another. In that case, the nonparametric baseline estimator seems to give the biggest value of the probability of the breakpoint distribution. When only a slight change occurs between the hazard distribution of two segments, all the proposed methods are less precise and the exponential baseline estimator seems to be the less performant of all baseline estimators.

More simulation studies which are not reported here have been carried out. When the change in distribution between two segments increases, the probability of the marginal breakpoint distribution increases accordingly and can be almost equal to 1 in some situations. For instance the marginal probabilities of the breakpoints found in Section 7

seem to indicate a much drastic change in the survival distributions than the simulation setting presented here. Finally scenarios which mix different parametric survival distributions in each segment have also been investigated. These simulations lead to similar behaviour of our estimators and are therefore omitted.

7 Survival analysis of diabetic patients at the Steno memorial hospital

In this section we illustrate our method on a dataset on survival of diabetics patients at the Steno memorial hospital. The data are described in great details in Example I.3.11 in Andersen et al (1993) and were originally studied through a illness-death model where the illness state corresponded to the diabetic nephropathy status of the patients. Here, we will only focus our interest on the survival of the patients, that is the variable of interest is the time from diagnostic of diabetes of a patient until death. The data were collected between 1933 and 1981 and patients were included in the study if the diagnosis of diabetes mellitus was established before age 31 years and between 1933 and 1972. A total of 2 709 patients were followed from the first contact with the hospital until death, emigration or the 31st of December 1984. On these 2 709 patients 707 (26%) deaths were observed and the other 2 002 (74%) patients were considered right censored. Since most of the patients did not contact the hospital directly after the diagnostic of diabetes, patients in this dataset are also left truncated. This needs to be taken into account because it means that individuals have a delayed entry into the study and will be observed only if they did not die before attending the Steno hospital. Without appropriate methods to deal with left truncation our estimation techniques will tend to overestimate the survival of diabetics patients. Gender (coded as 0 for women and 1 for men) and the year of birth was recorded for every patients. The dataset is composed of approximately 56% of male and 44% of female. The years of birth range from 1903 to 1971. Our aim was to determine if there was any change in the hazard distribution according to the date of birth when adjusting by gender. The marginal survival curves and parameter estimates in a Cox model with exponential baseline hazard were also computed.

To accommodate our method for left truncation the individual at risk process $Y_i(t)$ needs to be replaced by $Y_i(t) = I(L_i \leq t \leq T_i)$ where L_i represents the left truncation variable for individual i. This will affect the value of the emission probability $e_i(k;\boldsymbol{\theta})$ (see Equation (3)) which in turn will affect the value of the a posteriori segment distribution $w_i(k;\boldsymbol{\theta})$ and the value of the weighted log likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$. The parameters are estimated by maximizing the log likelihood in Equation (4) as before. For example, in the exponential model, the logarithm of the emission probability is equal to:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i\left(-\log(\lambda_k) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \left(\frac{T_i - L_i}{\lambda_k}\right)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

Since only the year of birth is known (and not the exact date of birth) it means that a breakpoint can only occur when changing from one year of birth to another. To take

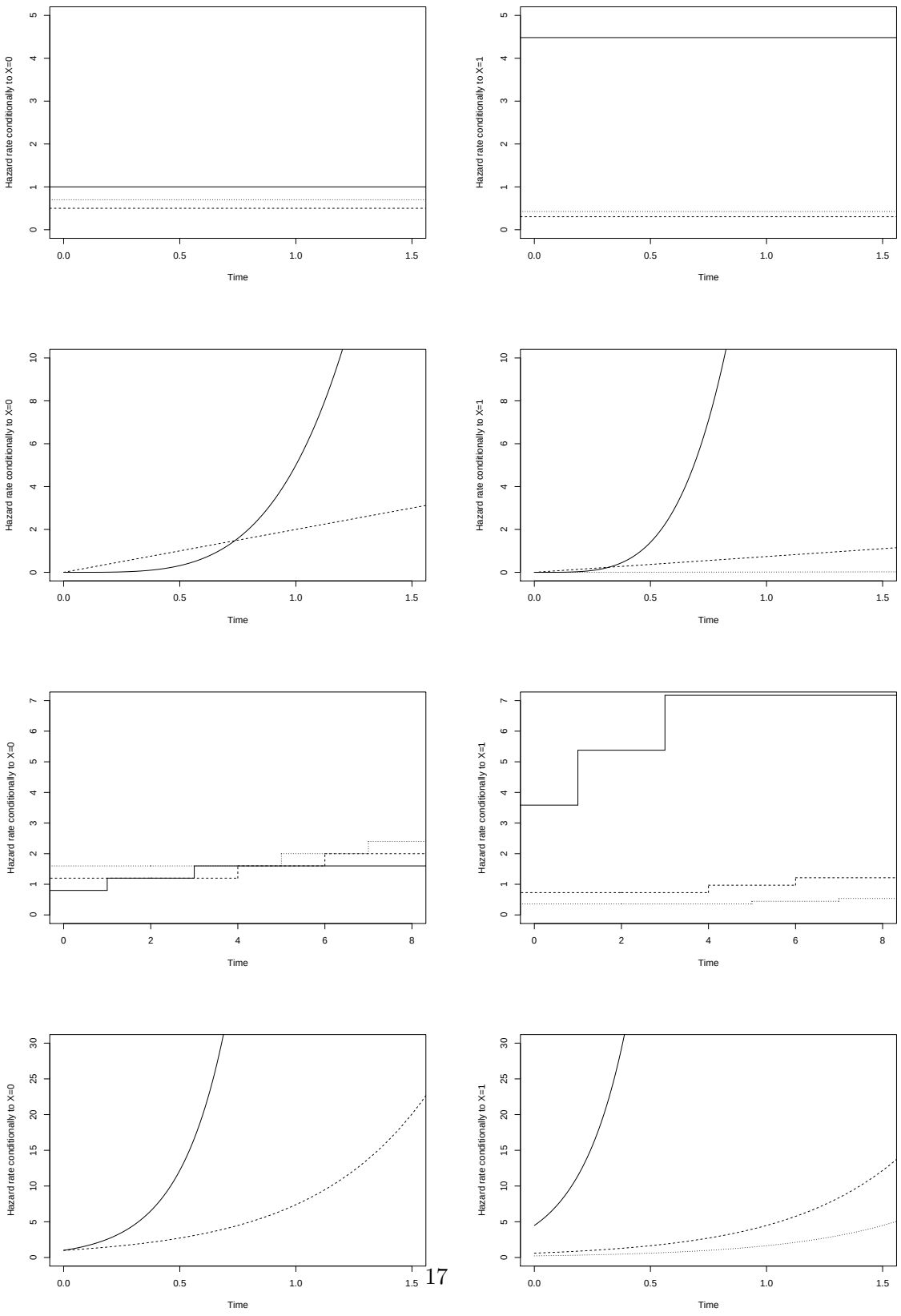

Figure 1: Conditional hazard rates in simulated data for Scenarios 1 to 4 from top to bottom. Solid line: hazard in segment 1. Dash line: hazard rate in segment 2. Dot line: hazard rate in segment 3.

Table 1: Bias, variance, MSE of $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$ and estimations of the maximum probability of breakpoints, average breakpoint locations along with their 95% confidence intervals from the exponential baseline model, the Weibull baseline model, the piecewise constant baseline model and the nonparametric baseline model.

Scenario 1

	Bias of $\hat{\boldsymbol{\beta}}$	Variance of $\hat{\boldsymbol{\beta}}$	MSE of $\hat{\boldsymbol{\beta}}$	Max proba bp12	Mean bp12	95% CI bp12	Max proba bp23	Mean bp23	95% CI bp23
Exponential	0.002	0.006	0.006	0.411	1000	994-1006	0.032	2120	1662-2974
	-0.002	0.015	0.015						
	-0.052	0.706	0.709						
Weibull	0.002	0.007	0.007	0.408	1000	994-1006	0.043	2216	1740-2981
	-0.002	0.011	0.011						
	-0.007	0.407	0.407						
Piecewise	0.003	0.007	0.007	0.402	1000	994-1006	0.069	2479	1800-2987
	0.000	0.009	0.009						
	-0.066	0.574	0.578						
Nonparametric	0.002	0.007	0.007	0.429	1001	996-1007	0.054	1954	1013-2995
	-0.069	0.820	0.825						
	-0.017	2.597	2.598						

Scenario 2

	Bias of $\hat{\boldsymbol{\beta}}$	Variance of $\hat{\boldsymbol{\beta}}$	MSE of $\hat{\boldsymbol{\beta}}$	Max proba bp12	Mean bp12	95% CI bp12	Max proba bp23	Mean bp23	95% CI bp23
Exponential	-1.207	0.000	1.458	0.054	998	973-1016	0.092	1943	1407-2002
	0.512	0.003	0.266						
	2.737	0.168	7.661						
Weibull	-0.010	0.008	0.008	0.309	1002	996-1020	0.154	1997	1978-2009
	-0.009	0.008	0.008						
	-0.043	0.255	0.257						
Piecewise	-0.187	0.007	0.042	0.323	1001	995-1008	0.192	1998	1983-2011
	0.031	0.007	0.008						
	0.007	0.304	0.304						
Nonparametric	0.000	0.010	0.010	0.332	1000	992-1008	0.195	1998	1983-2012
	-0.006	0.009	0.009						
	-0.122	0.708	0.723						

Scenario 3

	Bias of $\hat{\boldsymbol{\beta}}$	Variance of $\hat{\boldsymbol{\beta}}$	MSE of $\hat{\boldsymbol{\beta}}$	Max proba bp12	Mean bp12	95% CI bp12	Max proba bp23	Mean bp23	95% CI bp23
Exponential	-0.033	0.008	0.009	0.214	1001	986-1014	0.043	1997	1854-2119
	0.002	0.010	0.010						
	-0.007	0.016	0.016						
Weibull	-0.013	0.007	0.008	0.216	1001	986-1014	0.044	1994	1847-2111
	0.003	0.010	0.010						
	-0.007	0.015	0.015						
Piecewise	-0.007	0.008	0.008	0.217	1001	986-1014	0.046	1990	1844-2116
	0.006	0.011	0.011						
	-0.005	0.016	0.016						
Nonparametric	0.002	0.008	0.008	0.220	1002	991-1021	0.042	1997	1847-2131
	-0.001	0.010	0.010						
	-0.006	0.015	0.015						

Scenario 4

	Bias of $\hat{\boldsymbol{\beta}}$	Variance of $\hat{\boldsymbol{\beta}}$	MSE of $\hat{\boldsymbol{\beta}}$	Max proba bp12	Mean bp12	95% CI bp12	Max proba bp23	Mean bp23	95% CI bp23
Exponential	-0.639	0.002	0.410	0.238	1000	992-1006	0.027	1641	1015-2016
	0.196	0.020	0.058						
	0.575	0.035	0.366						
Weibull	-0.212	0.005	0.050	0.352	1000	994-1006	0.049	1994	1899-2079
	0.022	0.010	0.010						
	0.044	0.017	0.019						
Piecewise	-0.076	0.007	0.013	0.378	1000	994-1006	0.051	1989	1862-2080
	0.013	0.010	0.011						
	0.028	0.019	0.020						
Nonparametric	0.006	0.008	0.008	0.420	1000	991-1006	0.049	2009	1928-2137
	-0.004	0.011	0.011						
	-0.023	0.165	0.165						

this into account we first ordered all individuals with respect to their date of birth and the computation of the posterior distribution was constrained through the priors $\eta_i(k)$, defined in Section 3, such that $\eta_i(k) = 0$ for any k if individuals i and $i+1$ were born the same year. Other priors were set to 0.5. Since 0 is an absorbing state this ensured us to have change-points only for a new birth cohort.

The exponential model was first computed for one breakpoint, then we increased the numbers of breakpoints. The BIC criterion introduced in Section 5.3 was used as a stopping rule to find the correct number of breakpoints.

The maximum a posteriori of the breakpoints have been computed in Figure 2. For example, from the model with only one breakpoint it seems that the survival of diabetics patients was different for individuals born (strictly) before 1936 than for individuals born after 1936 (including the year 1936) with a probability of having a breakpoint equal to 88%. In each figure the peaks of the distribution of the breakpoint is high except for the model with four breakpoints where the probability of having a breakpoint in 1921 is equal to 27% while, in the same model, the probability of having a breakpoint in 1930, 1936 and 1947 is equal respectively to 63%, 61% and 67%. The same last three breakpoints were found in the model with three breakpoints with respective probabilities equal this time to 79%, 65% and 67%. All these results seem to indicate that the correct model is the three breakpoints model. This is confirmed by the value of the BIC criterion in each model in Table 2 where the lowest BIC value is obtained for the three breakpoints model.

In Table 2, parameter estimates for the Cox model with exponential value have been computed with gender as a covariate. The baseline values are slightly decreasing with respect to the birth cohort in the sense that men and women born at a latter birth period have a smaller hazard of death than individuals born at a earlier birth period. For the three breakpoints model, the effect of gender is positively associated to the hazard before 1930 (the regression parameter is equal to 0.345 and the hazard rate equals 1.41) and between 1936 and 1947 (the regression parameter is equal to 0.356 and the hazard rate equals 1.43) while the effect is nearly null for other periods (the hazard rate is equal to 0.95 for both other periods).

Non parametric survival estimates have been computed using a weighted Kaplan-Meier estimator in Figure 3. The curves show a clear increase in the survival of patients according to the year of birth. Patients born at a latter year have a greater survival than patients born at a earlier period at any age. For example, in the three breakpoints model, the survival 30 years after diagnoses of diabetes is equal to 53.2%, 62.4%, 72.5% and 85.1% for the respective birth cohorts $1903-1929$, $1930-1935$, $1936-1946$ and $1947-1971$.

The dataset has also been studied for the exponential model without adjusting by gender. The same breakpoints were found using the BIC criterion and the hazard and survival estimates were nearly identical.

Table 2: λ's and β's estimates in the Cox model adjusted by gender with exponential baseline for the models with one, two, three and four breakpoints along with BIC criterion for each model.

	One bp 1936	Two bp 1934, 47	Three bp 1930, 36, 47	Four bp 1921, 30, 36, 47
$\hat{\lambda}_1$	0.020	0.020	0.021	0.022
$\hat{\lambda}_2$	0.006	0.009	0.017	0.020
$\hat{\lambda}_3$		0.004	0.008	0.017
$\hat{\lambda}_4$			0.004	0.008
$\hat{\lambda}_5$				0.004
$\hat{\beta}_1$	0.213	0.260	0.345	0.426
$\hat{\beta}_2$	0.289	0.286	-0.053	0.240
$\hat{\beta}_3$		-0.034	0.356	-0.050
$\hat{\beta}_4$			-0.056	0.359
$\hat{\beta}_5$				-0.057
BIC	7342.539	7307.425	7306.889	7314.102

8 Discussion

In this article we proposed a new way to deal with cohort studies through a change-point model. In this model we suppose that abrupt changes can occur in the survival distribution of the event time. More specifically after specifying the number of segments, either the baseline hazard rates or the regression parameters are allowed to change in the different segments. Estimation in such a model is performed by an EM algorithm with use of constrained Hidden Markov Model (HMM) method as recently suggested by Luong et al (2013). The method proposes different specifications of the baseline and as shown by the simulation study, all different models provide both accurate estimates and accurate breakpoint locations. The method was also shown to adapt to more realistic problematics such as left truncation on the Steno memorial hospital dataset. Clearly, the methods developed here could be extended to a more complex setting. In particular, handling time dependent covariates or applying the method to recurrent events could be straightforward extensions. Also, the methodology should be directly applicable to other survival models such as the Accelerated Failure Time Model (see Kalbfleisch and Prentice, 2002; Wei, 1992) or the Aalen model (see Aalen, 1980; Scheike, 2002).

Parametric baseline models suffer from less flexibility compared to the nonparametric but as shown in the simulation study, they still detect the location breakpoints very accurately and provide very efficient estimates of the regression parameters even when the true dataset does not belong to the model estimator. These estimators also share the advantage to be parametric and consequently a BIC criterion could be derived. This is a very useful tool in order to find the correct number of breakpoints. This feature of parametric models was illustrated on the Steno memorial hospital dataset where it was

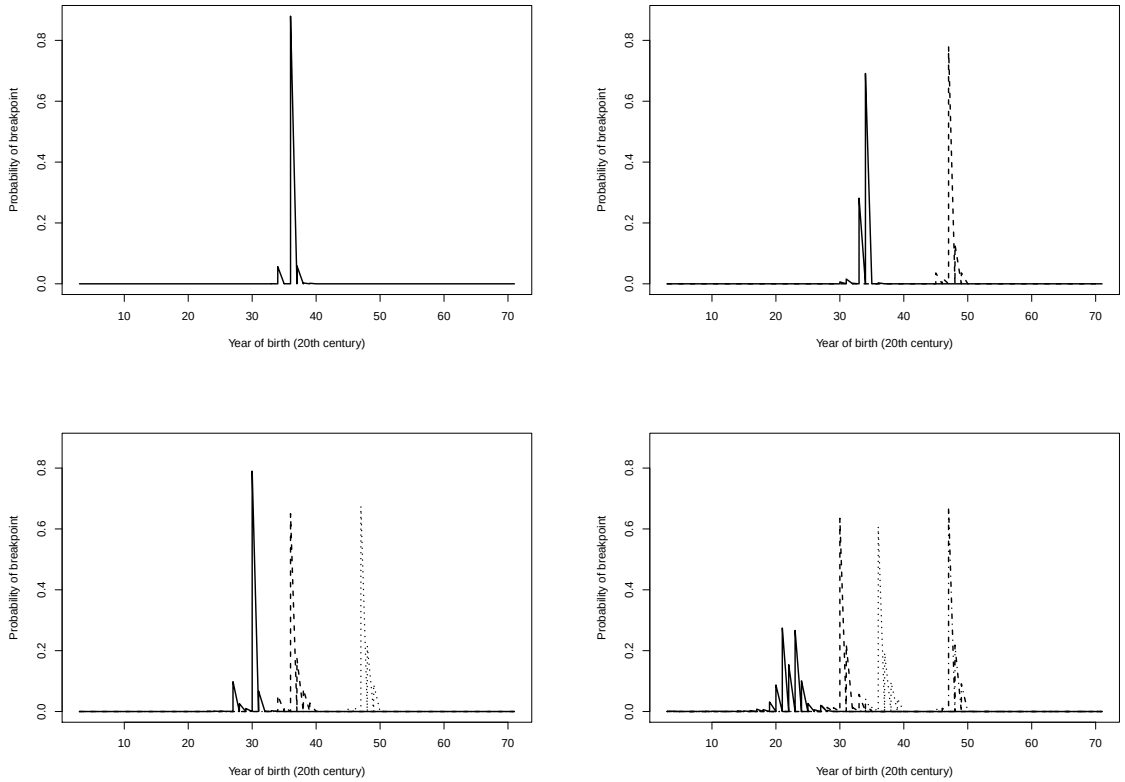

Figure 2: Marginal distributions of the breakpoints in the models with one, two, three and four breakpoints. The maximum a posteriori for the breakpoints are respectively: top left 1936, top right 1934 and 1947, bottom left 1930, 1936 and 1947, bottom right 1921, 1930, 1936 and 1947.

found that the three breakpoints model seems to be the correct model. On the other hand, no BIC criterion could be derived for the nonparametric model but as mentioned in the simulation study, this model gives the biggest value of the probability of the breakpoint distribution and consequently seems to be the most performant model to detect a change-point in the hazard distribution.

Acknowledgements The authors are very grateful to Professor Per Kragh Andersen for his valuable comments and for sharing with us the Steno memorial hospital dataset. This work is part of the DECURION project which was funded both by the IRESP and the french “Ligue nationale contre le Cancer.

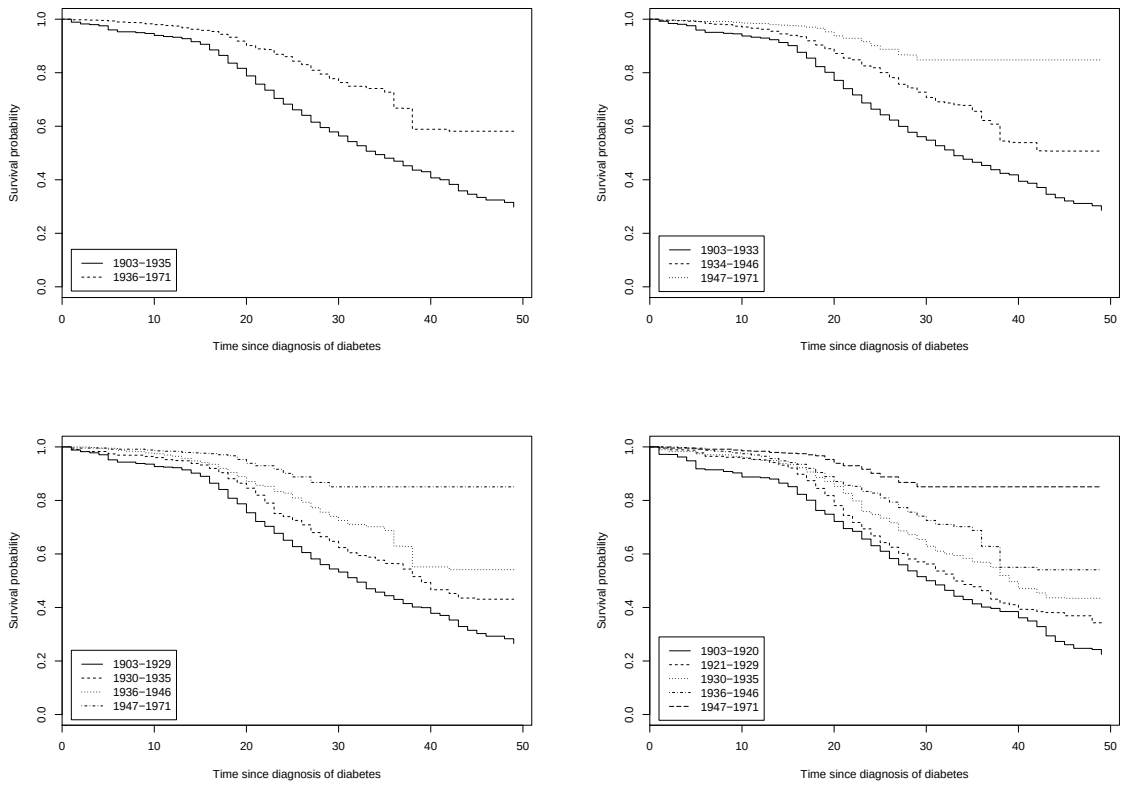

Figure 3: Weighted Kaplan-Meier estimators in the models with one, two and three breakpoints.

References

Aalen O (1980) A model for nonparametric regression analysis of counting processes. In: Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory, Springer-Verlag, New York, pp 1–25

Aalen OO, Borgan Ø, Gjessing HK (2008) Survival and Event History Analysis. Statistics for Biology and Health, Springer Science

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer Series in Statistics, Springer-Verlag, New York

Bergh T, Ericson A, Hillensjö T, Nygren KG, Wennerholm UB (1999) Deliveries and children born after in-vitro fertilisation in sweden 1982-95: a retrospective cohort study. Lancet 354(9190):1579–85

Breslow NE (1972) Discussion of the paper by D. R. Cox. Journal of the Royal Statistical Society Series B 34(2):187–220

Cox DR, Society S, Methodological SB (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B 34(2):187–220

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39(1):1–38

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley Series in Probability and Statistics, Wiley-Interscience (John Wiley & Sons), Hoboken, NJ

Kratz MH (2011) Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, 3rd edn. Cambridge University Press

Luong TM, Rozenholc Y, Nuel G (2013) Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. Computational Statistics and Data Analysis 68:129–140

Martinussen T, Scheike TH (2006) Dynamic regression models for survival data. Statistics for Biology and Health, Springer, New York

Scheike TH (2002) The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. Lifetime Data Analysis 8(3):247–262

Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Statistics for Biology and Health, Springer-Verlag, New York

Wei LJ (1992) The accelerated failure time model : a useful alternative to the cox regression. Statistics in medicine 11:1871–1879

Yang Y, Land KC (2013) Age-Period-Cohort Analysis. Interdisciplinary Statistics, Chapman et Hall