Grund, Birgit; Polzehl, Jörg

Working Paper

Semiparametric lack-of-fit tests in an additive hazard regression model

SFB 373 Discussion Paper, No. 1999,98

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Grund, Birgit; Polzehl, Jörg (1999) : Semiparametric lack-of-fit tests in an additive hazard regression model, SFB 373 Discussion Paper, No. 1999,98, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10046856

This Version is available at:
https://hdl.handle.net/10419/61767

Semiparametric lack-of-fit tests in an additive hazard regression model

Birgit Grund* and Jörg Polzehl†

October 7, 1999

Abstract

In the semiparametric additive hazard regression model of McKeague and Sasieni (1994), the hazard contributions of some covariates are allowed to change over time, without parametric restrictions (Aalen model), while the contributions of other covariates are assumed to be constant. In this paper, we develop tests that help to decide which of the covariate contributions indeed change over time. The remaining covariates may be modelled with constant hazard coefficients, thus reducing the number of curves that have to be estimated nonparametrically. Several bootstrap tests are proposed. The behavior of the tests is investigated in a simulation study. In a practical example, the tests consistently identify covariates with constant and with changing hazard contributions.

Keywords: Aalen model, hazard regression, lack-of-fit, confidence bands, parametric bootstrap, semiparametric, survival analysis.
AMS-Classification: Primary 62G05, Secondary 62G25

*School of Statistics, University of Minnesota, St. Paul, MN 55108, U.S.A.
†Weierstraß-Institut für Angewandte Analysis und Stochastik, D-10117 Berlin, Germany.
[0]The research on this paper was supported in part by the Sonderforschungsbereich 373 at the Humboldt University Berlin. The Technical Report was printed using funds made available by the Deutsche Forschungsgemeinschaft. The research of the first author was supported in part by NSF grant number DMS-9501893.

1 Introduction

In exploring time-to-event data with covariates, the nonparametric additive hazard model of Aalen (1980) has received increasing attention in recent years. The major attraction is that the regression coefficients are modelled as curves over time, without parametric restrictions. They are estimated nonparametrically. However, this flexibility comes at a high price: the nonparametric curve estimates for the regression coefficients tend to be highly variable, and even with reasonably large sample sizes only a few coefficients can be fitted in any one model.

In the current paper, we introduce, for the first time, quantifiable lack-of-fit tests that help to decide whether hazard contributions of (baseline) covariates are indeed changing over time. If not, the corresponding regression coefficients could be fitted as constants, resulting in a substantial reduction in the variability of the estimates.

The traditional model choice problem, whether or not a covariate should be included in an Aalen model at all, has been widely investigated. Tests have been developed, among others, by Aalen (1980, 1989), and McKeague and Sasieni (1994); see Andersen et al. (1993) for an overview. However, it is important to distinguish covariates with hazard contributions that change over time from those with significant, but constant hazard contributions. The traditional model selection tests are not suited for this task and can not be easily adapted. They rely on asymptotic distributions to calculate P-values, and the corresponding necessary asymptotic distributions for our test problem (regression coefficient is constant versus a nonparametric alternative) are not available.

In our tests, P-values are obtained through bootstrap resampling. We show, in a simulation study, that the bootstrap tests hold the nominal significance level and are sensitive in detecting alternatives for the considered models. In a practical example, the influence of four covariates on the length of stay in nursing homes is modelled. All tests consistently indicate that the hazard contribution of two of the covariates is approximately constant, while the other two have a short-term effect only.

We now specify the data structure, model, and test problem in detail.

Data. We consider right-censored time-to-event data. The observations are triplets (t_i, δ_i, z_i), $i = 1, \ldots, n$, where t_i denotes the observed survival time of the i^{th} individual, z_i is the vector of covariate values, and δ_i denotes the censoring indicator.

The observed survival time of the i^{th} individual is given by $t_i = \min(y_i, c_i)$, where y_i denotes the uncensored survival time, generated from the conditional distribution $y|z_i$, and c_i denotes the random censoring time. The censoring indicator takes the value $\delta_i = 1$ when $t_i = y_i$, and $\delta_i = 0$, otherwise. Throughout the paper, we assume that the censoring time c is independent of the covariates, and of the conditional distribution of

event times, $y|z$.

Model. We assume that hazard for an individual with covariate values $z = (u', x')'$, at time t, is determined by the semiparametric additive model of McKeague and Sasieni (1994),

$$\lambda(t|z) = \alpha(t)'u + \beta' x. \tag{1}$$

The unknown hazard regression coefficient vectors $\alpha(t) = \big(\alpha_0(t), \ldots, \alpha_p(t)\big)'$ and $\beta = (\beta_{p+1}, \ldots, \beta_{p+q})'$ express the contributions of the covariate subvectors $u \in I\!R^{p+1}$ and $x \in I\!R^q$, respectively. The subvector u includes all covariates with potentially non-constant hazard contributions. We define $u = (1, u_1, \ldots, u_p)'$, so that the "baseline hazard" component $\alpha_0(t)$ is estimated nonparametrically.

The hazard model includes both the additive hazard model by Aalen (1989), $\lambda(t|u) = \alpha(t)'u$, and the parametric additive hazard model $\lambda(t|x) = \alpha_0(t) + \beta' x$, investigated by Lin and Ying (1994).

McKeague and Sasieni (1994) derived semiparametric likelihood estimates for β, and for the vector of cumulative hazard coefficient functions, $A(t) = \int_0^t \alpha(s)ds$, the integral defined componentwise. These estimates are described in more detail in our Section 2. They will be used throughout the paper, whenever estimates for the cumulative hazard coefficients $A(t)$ and βt, for the cumulative hazard function, $\Lambda(t|z) = \int_0^t \lambda(s)ds = A(t)'u + t\beta' x$, or for the survival function, $S(t|z) = \exp\{-\Lambda(t|z)\}$, are needed.

Test problem. We are testing the hypothesis that certain hazard coefficients in the model are constant, $H_0 : \alpha_p(t) = \beta_p$, *for some constant* β_p *and for all* $t > 0$, versus a nonparametric alternative that $\alpha_p(t)$ is not constant. More generally, we suggest methods for testing composite hypotheses,

$$\begin{aligned} H_0 :&\ \ \alpha_k(t) = \beta_k \text{ (constant), for } k = p-r, \ldots, p, \text{ for all } t > 0, \\ H_a :&\ \ \text{at least one of the coefficients } \alpha_k(t),\ k = p-r, \ldots, p, \text{ is not constant,} \end{aligned}$$

within the hazard regression model $\lambda(t|z) = \alpha_0(t) + \sum_{k=1}^{p} \alpha_k(t) z_k + \sum_{k=p+1}^{q} \beta_k z_k$.

Discussion. McKeague and Sasieni (1994) suggest an intuitive test for the simple hypothesis $H_0 : \alpha_p(t) = \beta_p$ based on pointwise confidence intervals. They fit a full Aalen model, and plot the nonparametric estimates $\hat{A}_p(t)$ for the covariate of interest versus the corresponding parametric estimates $\hat{\beta}_p t$. Pointwise asymptotic confidence intervals for $A_p(t)$ are computed using the asymptotic distribution of $\hat{A}_p(t)$. If the straight line $\hat{\beta}_p t$ leaves the series of pointwise confidence intervals around $\hat{A}_p$ at any time t, the parametric model is rejected.

The major advantage of this visual test is that we can see immediately *where* the straight line estimate leaves the confidence intervals around $\hat{A}_p(t)$. However, the proposed procedure has serious weaknesses: (i) the coverage probability of the confidence "bands" formed by the pointwise confidence intervals is unknown, and may be considerably lower than the stated coverage of the confidence intervals; (ii) the asymptotic distribution of $\hat{A}_k(t)$ may not adequately reflect the variability of the finite-sample distribution; and (iii) the sampling error due to estimating β_k is correlated with the estimation error in $\hat{A}_k(t)$, and should be taken into account for the testing procedure. Finally, centering the confidence interval around $\hat{A}_k(t)$ may neglect the estimation bias.

Our bootstrap tests overcome limitations (i)–(iii) of the McKeague and Sasieni (1994) confidence intervals. The tests are based on different measures of distance between curves, in particular the supremum norm, L_2 norm, and a weighted L_2 norm. Visual decision aids that are similar to the pointwise confidence intervals proposed by McKeague and Sasieni (1994) can be constructed easily from our supremum test statistics, with the advantage that we provide true confidence *bands*, with simultaneous coverage.

Additionally, we propose model checking procedures based on the survival function $S(t)$, and the cumulative hazard $\Lambda(t)$. These tests can be used to check several components simultaneously, without incurring multiple comparisons problems.

P-values are calculated through bootstrap resampling. While bootstrap in classical regression is well-investigated, properties of bootstrap procedures for censored data are mostly unknown. However, the conditional bootstrap that we used to calculate the P-values provides acceptable results in our simulation study.

Related problems in classical regression have been investigated by Härdle and Mammen (1993), and Horowitz and Härdle (1994). They propose tests for the appropriateness of a parametric fit versus a nonparametric regression estimate. Härdle and Mammen (1993) show that the convergence of the distribution of their test statistic to its asymptotic limit is slow, and bootstrap provides a better adjustment of P-values.

In the following section, we shortly describe estimation in the semiparametric hazard model (1). Our test statistics are introduced in Section 3. In Section 4, the bootstrap procedure that we use for calculating P-values is described in detail. Finite sample properties of the tests are illustrated in a simulation study in Section 5, and in Section 6, the tests are applied to a real-life data set on the length of stay in nursing homes.

2 Estimation in the semiparametric hazard model

Semiparametric likelihood estimates for the cumulative hazard coefficient $A(t)$ and for β were derived by McKeague and Sasieni (1994). Since these estimates are central to our tests, we include a short summary here.

The likelihood function for $\alpha(t)$ and β is given by

$$l(\alpha, \beta) = \sum_{i=1}^{n} \left\{ \delta_i \log\left(\lambda_i(t_i)\right) - \int I_{[t_i \geq t]} \lambda_i(t) dt \right\}, \tag{2}$$

where $\lambda_i(t) = \lambda_i(t|z_i) = \alpha(t)'u_i + \beta' x_i$, the vectors $u \in I\!R^{p+1}$ and $x \in I\!R^q$ consist of the covariates of the i^{th} individual with changing and with constant hazard contributions, respectively, and $I_{[t_i \geq t]}$ is the indicator function of the event $t_i \geq t$.

The likelihood function (2) is maximized by the values $A(t)$ and β that solve the equation system

$$A(t) = \int_0^t (U'WU)^{-1}(U'WdN - U'WX\beta ds), \tag{3}$$

$$\beta = \left(\int X'HX dt \right)^{-1} \int X'HdN, \tag{4}$$

where $U = U(t)$ denotes the $n \times (p+1)$ matrix with the i^{th} row equal to $I_{[t_i \geq t]} u_i'$; this means, at any time point t, the row is *zero* for individuals that are *not* in the risk set at time t. Similarly, $X = X(t)$ is the $n \times q$ matrix for the covariate components corresponding to β, with rows $I_{[t_i \geq t]}\ x_i'$; $W = W(t) = \mathrm{diag}\left\{\lambda_i(t)\right\}^{-1}$ is an $n \times n$ diagonal matrix; $N(t) = (N_1(t), \ldots, N_n(t))'$ denotes the n-dimensional counting process corresponding to our sample, with the component $N_i(t) = I_{[t_i \leq t, \delta_i = 1]}$ jumping from zero to one at event time t_i; and matrix $H = H(t) = W - WU(U'WU)^{-1}U'W$.

In the current paper, we estimate $A(t)$ and β in four steps:

Step 1: Fit a full Aalen model with all covariates, $\lambda(t|z) = \alpha_F(t)'z$. The resulting estimates, $\tilde{A}_F(t)$, of the cumulative hazard coefficients $A_F(t) = \int_0^t \alpha_F(s)ds$ are step functions, with jumps in the event times.

Step 2: Use a local polynomial smoother on the pairs $\left(t_i, \tilde{A}_F(t_i)\right)$, and estimate the hazard coefficient vector $\alpha_F(t)$ as derivative.

Step 3: Calculate estimates $\hat{\lambda}(t|z)$ and $\hat{W}(t) = \mathrm{diag}\left\{\hat{\lambda}_i(t)\right\}^{-1}$ using the estimates of $\alpha_F(t)$ obtained in Step 2. For simplicity, we calculate $\hat{W}(t)$ only at event points, and define the components of $\hat{W}(t)$ as left-continuous step functions between even points.

Step 4: In (3) and (4), replace W by $\hat{W}$, and solve for β and $A(t)$. These are the final estimates $\hat{\beta}$ and $\hat{A}(t)$ for β and $A(t)$, respectively.

Note that the components of the resulting $\hat{A}(t)$ are step functions in t, with jumps at exactly the event times t_i. In Step 4, the integrals in formulae (3) and (4) resolve into sums, since U, X and $\hat{W}$ are step functions with jumps at the event points.

Remark 2.1 *The estimation algorithm is a slight modification of Method 1 in McKeague and Sasieni (1994, Section 2.3), in that we use a local linear fit instead of kernel smoothing. Note that under the null hypothesis, one or several components of $A_F(t)$ are straight lines, and we expect their estimates to be close to straight lines as well. The design points, in our case the event times t_i, are far from equidistant. By choosing a local linear fit, we ensure that whatever "linearity" appears in the estimates $\tilde{A}_F(t)$ is not disturbed by smoothing. In terms of asymptotic properties, local linear smoothers tend to have a lower integrated mean squared error than kernel estimates under non-equidistant design.*

Remark 2.2 *McKeague and Sasieni (1994) have shown that the above estimates for β and $A(t)$ are efficient, the distribution of $n^{-1/2}(\hat{\beta} - \beta)$ converges to a q-variate normal distribution with mean zero, and that $n^{-1/2}(\hat{A}(t) - A(t))$ converges in distribution to a p-variate Gaussian process with mean zero. They provided consistent estimates for the covariance matrix and the covariance process, respectively, of the limiting distributions.*

3 Test statistics

For the sake of brevity, let us consider the simple null hypothesis $H_0 : \alpha_p(t) = \beta_p$. This means, we are interested in one given covariate, z_p, and whether the influence of this covariate upon the hazard is additive with a constant coefficient. The remaining coefficients are either estimated nonparametrically, or considered constant. We introduce three classes of tests which are based on the cumulative hazard coefficient $A_p(t)$; on the cumulative hazard function $\Lambda(t)$; and on the survival function $S(t)$, respectively. P-values are calculated through bootstrap resampling, described in Section 4.

Tests $\mathbf{T}^A$: The first class of tests focusses on the cumulative hazard coefficient $A_p(t)$, over a certain time interval of interest, $[\underline{t},\ \bar{t}]$.

$$\mathbf{T}^A_{max} = \max_{\underline{t} \le t_i \le \bar{t}} \frac{\left| \hat{A}_p(t_i) - \hat{\beta}_p t_i \right|}{\sqrt{\widehat{\mathrm{var}}\ \hat{A}_p(t_i)}}, \tag{5}$$

$$\mathbf{T}_q^A = \sum_{\underline{t} \le t_i \le \bar{t}} \left(\hat{A}_p(t_i) - \hat{\beta}_p t_i \right)^2,$$

$$\mathbf{T}_s^A = \sum_{\underline{t} \le t_i \le \bar{t}} \frac{\left(\hat{A}_p(t_i) - \hat{\beta}_p t_i \right)^2}{\widehat{\text{var}}\, \hat{A}_p(t_i)}. \tag{6}$$

All three test statistics measure the difference between $\hat{A}_p(t)$ and $\hat{\beta}_p t$, the likelihood estimates of the cumulative hazard coefficient of the covariate z_p under the alternative and under the null hypothesis, respectively, at the event times t_i. The calculation of likelihood estimates is described in Section 2. The nonparametric estimate $\hat{A}_p(t)$ is obtained by fitting the model $\lambda(t|z) = \alpha_0(t) + \sum_{k=1}^{p} \alpha_k(t) z_k + \sum_{k=p+1}^{q} \beta_k z_k$, while β_k is estimated in $\lambda(t|z) = \alpha_0(t) + \sum_{k=1}^{p-1} \alpha_k(t) z_k + \sum_{k=p}^{q} \beta_k z_k$. In the notation of (1), we are splitting z into the subvectors $u = (1, z_1, \ldots, z_p)'$ and $x = (z_{p+1}, \ldots, z_q)'$ when estimating $A_p(t)$, and into subvectors $u = (1, z_1, \ldots, z_{p-1})'$ and $x = (z_p, \ldots, z_q)'$ in order to estimate β_p. Although the hazard models under H_0 and H_a differ only in the one covariate z_p, the estimates for all the hazard coefficients are affected. Nevertheless, the tests in class $\mathbf{T}^A$ include only the coefficients of z_p.

The test statistics (5) and (6) are standardized with $\widehat{\text{var}}\, \hat{A}_p(t_i)$, an estimate of the variance of $\hat{A}_p(t_i)$. We obtain the variance estimates through bootstrap, described in detail in Section 4.

The interval $[\underline{t},\ \bar{t}]$ is to be defined by the user. Typically, $\bar{t}$ should be chosen small enough to avoid excessive variation of the nonparametric estimate $\hat{A}(t)$ for large values of t, due to a small risk set. The natural choice for the lower boundary seems to be $\underline{t} = 0$. However, depending on the computational implementation, it may be advisable to drop the first few event points and choose some small $\underline{t} > 0$ in order to avoid boundary effects both in the nonparametric estimate of $A(t)$ and for the bootstrap variance estimates close to $t = 0$.

Test statistic (5) uses a (weighted) supremum norm to measure the distance between the curves $\hat{A}_p(t)$ and $\hat{\beta}_p t$. Therefore, this test should be sensitive to large deviations from linearity that occur for a short period of time. The other two tests build on a weighted L_2 norm, and measure the distance between $\hat{A}_p(t)$ and $\hat{\beta}_p t$ cumulatively. The summation over the event points in $[\underline{t},\ \bar{t}]$ corresponds to a weighting with the density of event times. Standardization in (5) and (6) lowers the influence of late events, and of any other events with large variability in the estimate of $A_p(t)$.

Remark 3.1 *The test* $\mathbf{T}^A_{max}$*, when used with a rejection region of* $|\mathbf{T}^A_{max}| \geq 1.96$*, is related to the visual test procedure suggested by McKeague and Sasieni (1994), which is based on asymptotic confidence intervals for* $A_i(t)$*. While the confidence intervals are constructed with estimates of the asymptotic variance of* $\hat{A}_i(t)$*, the* $\mathbf{T}^A_{max}$ *test uses bootstrap estimates of the finite-sample variances. If a confidence band around the line* $\hat{\beta}t$ *is desired as a visual decision aid, we suggest to plot the intervals* $\beta t_i \pm \hat{q}_{0.05}\sqrt{\widehat{\text{var}}\hat{A}(t_i)}$*, at all event times* t_i*. Here,* $\hat{q}_{0.05}$ *denotes the bootstrap estimate of the* 95^{th} *percentile of the* $\mathbf{T}^A_{max}$ *distribution.*

Tests $\mathbf{T}^\Lambda$: Test statistics based on the cumulative hazard function $\Lambda(t|z)$ are defined as follows:

$$\mathbf{T}^\Lambda_{max} = \max_{\underline{t} \leq t_i \leq \bar{t}} \frac{\left|\hat{\Lambda}(t_i|z_i; H_a) - \hat{\Lambda}(t_i|z_i; H_0)\right|}{\sqrt{\widehat{\text{var}}\ \hat{\Lambda}(t_i|z_i; H_a)}},$$

$$\mathbf{T}^\Lambda_q = \sum_{\underline{t} \leq t_i \leq \bar{t}} \left(\hat{\Lambda}(t_i|z_i; H_a) - \hat{\Lambda}(t_i|z_i; H_0)\right)^2,$$

$$\mathbf{T}^\Lambda_s = \sum_{\underline{t} \leq t_i \leq \bar{t}} \frac{\left(\hat{\Lambda}(t_i|z_i; H_a) - \hat{\Lambda}(t_i|z_i; H_0)\right)^2}{\widehat{\text{var}}\ \hat{\Lambda}(t_i|z_i; H_a)}.$$

Here, $\hat{\Lambda}(t\,|z; H_a)$ and $\hat{\Lambda}(t\,|z; H_0)$ are the estimates of the cumulative hazard function at time t, given covariates z, under the alternative H_a and under the null hypothesis H_0, respectively. In the simple test problem, subvectors u and x are of dimension p and q+1 under the null hypothesis, and of dimension p+1 and q under the alternative. Again, $\widehat{\text{var}}\ \hat{\Lambda}(t|\ z; H_a)$ denotes bootstrap estimate of the variance of $\hat{\Lambda}(t|\ z; H_a)$.

Tests $\mathbf{T}^S$: The test statistics in class $\mathbf{T}^S$ are based on the survival function $S(t|z)$,

$$\mathbf{T}^S_{max} = \max_{\underline{t} \leq t_i \leq \bar{t}} \frac{\left|\hat{S}(t_i|z_i; H_a) - \hat{S}(t_i|z_i; H_0)\right|}{\sqrt{\widehat{\text{var}}\ \hat{S}(t_i|z_i; H_a)}},$$

$$\mathbf{T}^S_q = \sum_{\underline{t} \leq t_i \leq \bar{t}} \left(\hat{S}(t_i|z_i; H_a) - \hat{S}(t_i|z_i; H_0)\right)^2,$$

$$\mathbf{T}_s^S = \sum_{\underline{t} \le t_i \le \bar{t}} \frac{\left(\hat{S}(t_i|z_i; H_a) - \hat{S}(t_i|z_i; H_0)\right)^2}{\widehat{\text{var}}\ \hat{S}(t_i|z_i; H_a)}.$$

Note that all test statistics in the classes $\mathbf{T}^\Lambda$ and $\mathbf{T}^S$ can easily be extended to test hypothesis on several covariates simultaneously, by estimating Λ and S under the appropriate hypotheses. For the $\mathbf{T}^\Lambda$ and $\mathbf{T}^S$ tests, summation over event times now corresponds to a weighting with respect to the joint density of event times and covariates.

In all three classes, the test statistics $\mathbf{T}^\bullet_{max}$ can be expected to suffer from the high variability of the nonparametric estimates $\hat{A}(t)$, more so than the test statistics based on the L_2-norm, since the supremum distance is particularly sensitive to spurious bumps in the nonparametric curve estimates. In our simulation study, we include modifications of the $\mathbf{T}^\bullet_{max}$ statistics where the coefficient estimates $\hat{A}(t)$ are replaced by smoothed versions, $\tilde{A}(t; h)$, designed to decrease the variability. The corresponding test statistics will be denoted by $\mathbf{T}^\bullet_{max,h}$, indicating the use of a global bandwidth h.

4 Bootstrap sampling

We obtain P-values for the proposed test statistics through bootstrap resampling. There are several ways to draw bootstrap samples from censored data. For an overview we refer to Burr (1994) and Davison and Hinkley (1997). In the current paper, we use "parametric bootstrap", where bootstrap samples are drawn from estimates of the conditional distributions of y and of c, given the observed pairs $(z_i, \delta_i),\ i = 1, \ldots n$.

Let $G(t)$ denote the distribution function of the censoring times, c_i. We generate a bootstrap sample, $(t_i^*, \delta_i^*, z_i),\ i = 1, \ldots n$, in three steps:

Step 1: For each $z_i,\ i = 1, \ldots n$, draw one y_i^* from an estimate of the conditional distribution function of the event times, $\hat{F}(t|z_i) = 1 - \exp\left\{-\hat{\Lambda}(t|z_i)\right\}$. In our implementation, the estimate $\hat{F}(t|z_i)$ is defined as $\hat{\Lambda}(t|z_i) = \hat{\Lambda}(t|z_i, H_0) = \hat{A}(t)'u_i + t\hat{\beta}'x_i$ at the event times, and as a linear interpolation for time points t between events. The parameters $A(t)$ and β are estimated under the null hypothesis model. Sometimes, the resulting $\hat{F}(t|z_i)$ is not a valid distribution function. In this case, we project the estimate onto the space of distribution functions. Usually, this means that, for any given z_i, the final estimate $\hat{F}(t|z_i)$ is forced to be monotonously non-decreasing.

Step 2: If $\delta_i = 0$, set $c_i^* = c_i$. If $\delta_i = 1$, generate c_i^* from the conditional distribution

$$\hat{G}(t|c > t_i) = \frac{\hat{G}(t) - \hat{G}(t_i)}{1 - \hat{G}(t_i)},$$

where $\hat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution $G(t)$.

Step 3: Set $t_i^* = \min(y_i^*, c_i^*)$. Set $\delta_i^* = 1$ if $t_i^* = y_i^*$, and $\delta_i^* = 0$, else.

This approach is related to one of the resampling methods suggested in Davison and Hinkley (1997, page 351) for Cox's proportional hazards model.

P-values for a test statistic $\mathbf{T}$ are calculated as follows. Let B denote the number of bootstrap samples. Let $\mathbf{T}^{*b} = \mathbf{T}\left((t_1^{*b}, \delta_1^{*b}, z_1), \ldots, (t_n^{*b}, \delta_n^{*b}, z_n)\right)$ denote the value of the test statistic $\mathbf{T}$ computed with the b^{th} bootstrap sample, $\left\{(t_i^{*b}, \delta_i^{*b}, z_i),\ i = 1, \ldots, n\right\}$, and $s = \mathbf{T}\big((t_1, \delta_1, z_1), \ldots, (t_n, \delta_n, z_n)\big)$ the value of $\mathbf{T}$ applied to the original sample. Then, bootstrap P-values are obtained as empirical quantiles of the sample $\left\{\mathbf{T}^{*b},\ b = 1, \ldots, B\right\}$, i. e.,

$$\hat{P}(\mathbf{T} > s) = \frac{1}{B} \sum_{b=1}^{B} I_{[\mathbf{T}^{*b} > s]}.$$

For a general discussion of bootstrap P-values, we refer to Shao and Tu (1995, Chapter 3).

In defining the test statistics $\mathbf{T}_{max}^{\bullet}$ and $\mathbf{T}_{s}^{\bullet}$, we used bootstrap estimates of variances. The variance of $\hat{A}_p(t)$ is estimated by

$$\widehat{\text{var}}\ \hat{A}_p(t) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{A}_p^{*b}(t) - \frac{1}{B} \sum_{b=1}^{B} \hat{A}_p^{*b}(t) \right)^2.$$

This same variance estimate is used as well in the bootstrap statistics $\mathbf{T}^{*b}$, in order to avoid the computational burden of double bootstrap. The variance estimates for $\hat{\Lambda}(t|z)$ and for $\hat{S}(t|z)$ are defined accordingly.

5 Simulation study

The aim of the simulation study is: (a) to determine how well the proposed test procedures hold the nominal significance level; (b) to compare the power of the tests under various alternatives; and (c) to investigate the effect of smoothing on the tests $\mathbf{T}_{max}^{\bullet}$.

All simulations have the following parameters in common:

- For each model, the simulation is based on $N = 500$ samples of size $n = 250$. For each of the N "data" samples, $B = 500$ bootstrap samples are drawn to calculate P-values and variance estimates.

- The covariate values $z_i,\ i = 1 \ldots, n$ are generated, independently, from a uniform $U(-0.5, 0.5)$ distribution. In models with two covariates, the components are independent, $U(-0.5, 0.5)$.

- The censoring distribution is exponential with $\lambda = 0.05$, with end point censoring at $t = 5$. The censoring times are independent of covariate values.

- The event times are generated, independently, from the conditional distribution $F(t|z_i) = 1 - \exp\{-\Lambda(t|z_i)\}$, where Λ is given by the underlying hazard model.

- The "interval of interest" for the test statistics is $0.25 = \underline{t} \leq t \leq \bar{t} = 3$.

We investigate three basic hazard models, in which covariate z_1 has a *constant effect*, a *short-term effect*, and a *delayed impact effect* on the hazard. Each of the three base models is run with one and with two covariates, resulting in a total of six models. In each case, we are testing the hypotheses $H_0 :\ \alpha_1(t) = \beta_1$ versus $H_a :\ \alpha_1(t)$ *is not constant.* In models with two covariates, the second covariate, z_2, is introduced as a nuisance parameter; the hazard function does not depend on z_2. A description of the hazard models follows.

Constant effect model, one covariate. With one covariate, the hazard function of the constant effect model is given by $\lambda(t|z) = \alpha_0(t) + \alpha_1(t)z$, with

$$\alpha_0(t) = 0.4, \quad \text{and} \quad \alpha_1(t) = 0.4, \quad \text{for all } t \geq 0. \tag{7}$$

Note that for the constant effect model, the null hypothesis, $H_0 :\ \alpha_1(t) = \beta_1$, is true, since the hazard coefficient $\alpha_1(t) = 0.4$ is constant.

Figure 1 illustrates the constant effects model. In panel (a), both the cumulative baseline hazard, $A_0(t)$, and the cumulative hazard coefficient, $A_1(t)$, are plotted. The two lines coincide. Panel (b) shows one dataset of size $n = 250$, generated from the constant effects model. The points (t_i, z_i) are marked "+" when $\delta_i = 1$ (event), and "$\times$" when $\delta_i = 0$ (censored). Note that the hazard is three times higher for observations with $z = 0.5$ than for observations with $z = -0.5$, although this is difficult to detect from the visual impression of the sample.

Panels (c) and (d) describe how the cumulative hazard, $\Lambda(t|z)$, and the conditional cumulative survival function, $S(t|z)$, respectively, behave for different covariate values.

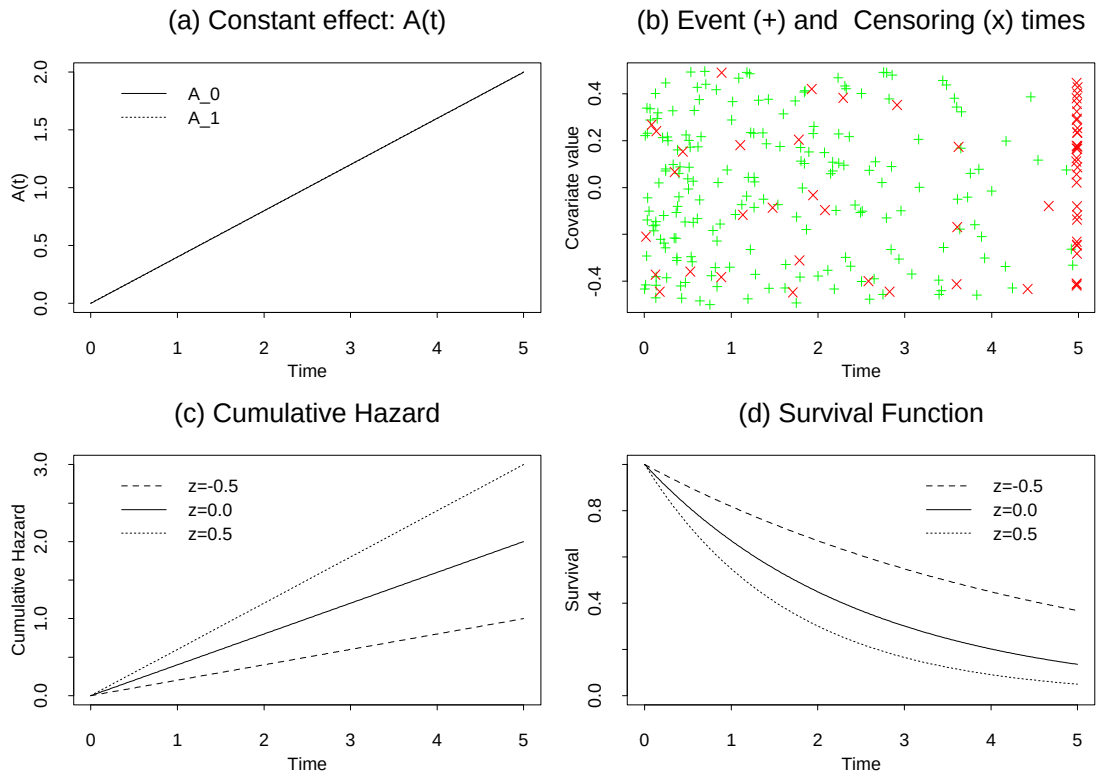

Figure 1: *Constant effect model.* (a) Cumulative baseline hazard $A_0(t)$ and cumulative hazard coefficient $A_1(t)$ (identical); (b) event (+) and censoring times (×) for one $n = 250$ dataset; (c) cumulative hazard, and (d) survival function for the minimum ($z = -0.5$), mean ($z = 0$), and maximum ($z = 0.5$) of covariate values.

The solid line corresponds to $z = 0$, the mean covariate value, and the dashed and dotted lines display Λ and S for the minimal ($z = -0.5$) and maximal ($z = 0.5$) covariate values. The tests $\mathbf{T}^{\Lambda}$ and $\mathbf{T}^{S}$ are based on estimates of these functions.

Figure 2 displays estimates obtained under both the null hypothesis, $\lambda(t|z) = \alpha_0(t) + \beta_1 z$, and the alternative, $\lambda(t|z) = \alpha_0(t) + \alpha_1(t)z$. In the upper row, the estimates of the cumulative hazard coefficients $A_0(t)$ and $A_1(t)$ under H_0 (dotted lines) and under H_a (dashed lines) are plotted along with the true functions (solid lines). The bottom row shows estimates of the cumulative hazard $\Lambda(t\,|z)$ and of the survival function $S(t\,|z)$, for covariate values $z = -0.5$, $z = 0$ and $z = 0.5$. All the estimates are computed with the sample displayed in Figure 1 (b). Note that the estimate of $A_1(t)$ exhibits much higher variability than the estimate of the baseline cumulative hazard, $A_0(t)$. Clearly, $A_1(t) = 0.4t$ is estimated better under H_0 than under H_a. Our tests measure the difference between the estimated curves under H_0 and under H_a over the interval $0.25 \leq t \leq 3$.

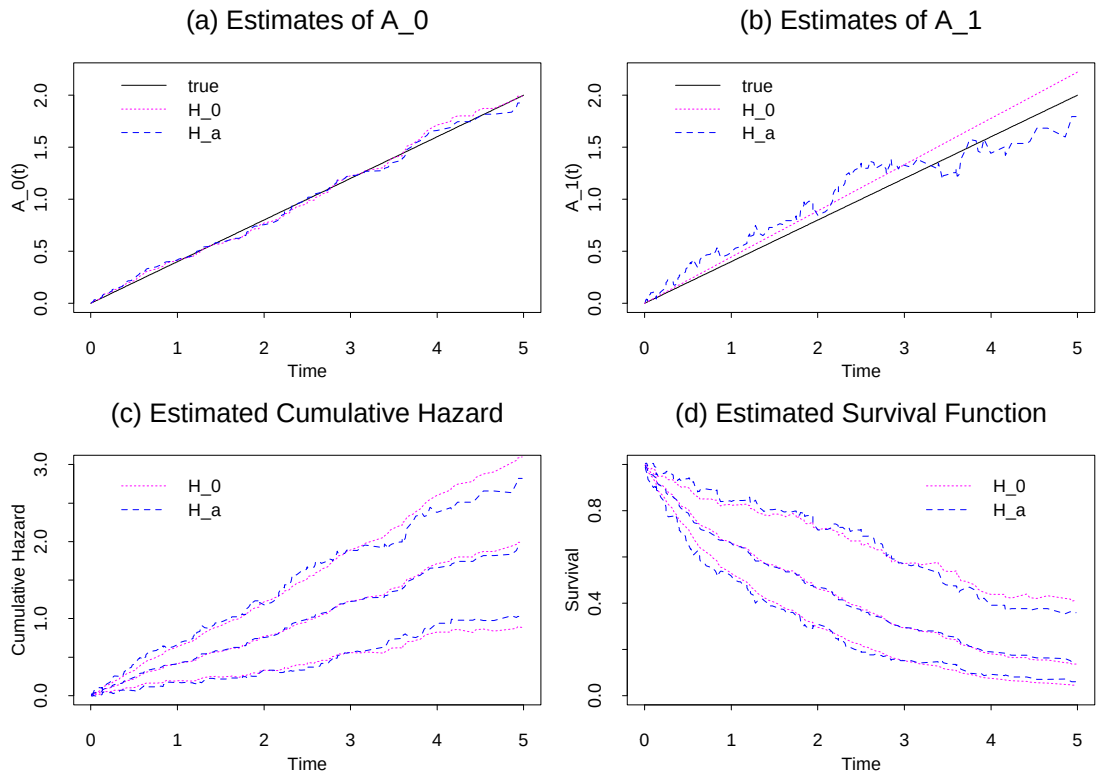

Figure 2: Estimates in the *constant effect* model: (a) and (b) show A_0 and A_1 (solid lines) and their estimates under H_0 and under H_a; (c) and (d) show estimates of $\Lambda(t|z)$ and $S(t|z)$ under H_0 and under H_a for the minimum, mean and maximum of covariate values. Estimates are based on the sample displayed in Figure 1 (b)

Figure 3 displays the empirical distribution function of the bootstrap P-values, calculated from $N = 500$ samples. In order to save space, we restrict the plot to the interval $[0, 0.5]$. A 45° line would indicate that our bootstrap estimates of the P-values are close to the true P-values. To be exact, a 45° line for a test $\mathbf{T}$ means that the bootstrap estimates for the P-values, each based on $B = 500$ bootstrap resamples per data sample (horizontal axis), would take values j/N, $j = 1, \ldots, N$, and thus ideally reflect the distribution of $\mathbf{T}$ under H_0. Nominal significance levels of 0.05 and 0.1 are emphasized by vertical dotted lines.

Note that all of our tests approximate 45° lines reasonably well, indicating that the bootstrap P-values are reliable estimates. The tests based on $A_1(t)$ perform best. In the classes $\mathbf{T}^{\Lambda}$ and $\mathbf{T}^{S}$, the tests $\mathbf{T}^{\bullet}_{q}$ and $\mathbf{T}^{\bullet}_{s}$ get overall closer to the targeted significance level than the $\mathbf{T}^{\bullet}_{max}$ tests. When there is a deviation from the nominal significance level, P-value are overestimated. This means, the proposed bootstrap tests are conservative, under the given model.

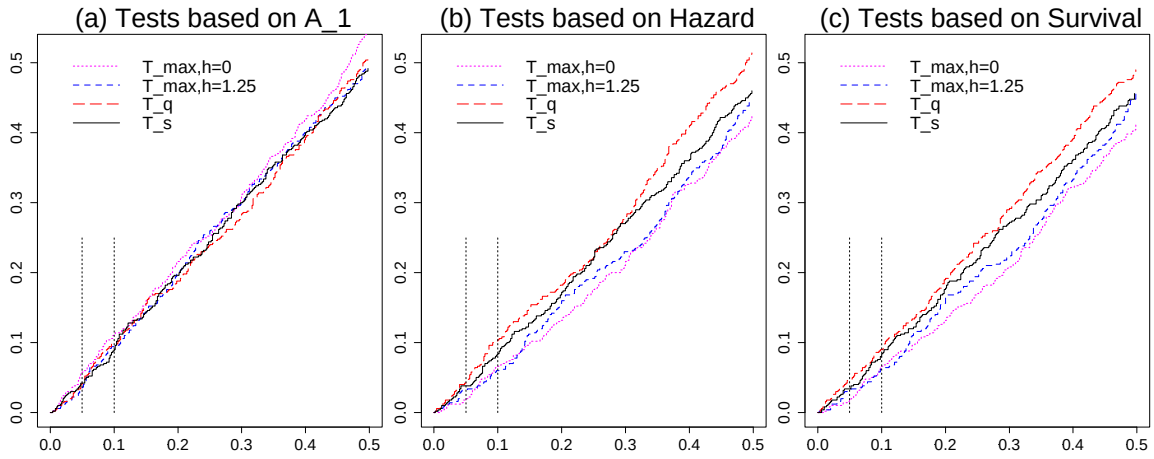

Figure 3: *Constant effect model.* Empirical distribution of bootstrap P-values based on $N = 500$ samples for the three classes of tests. (a) Tests $\mathbf{T}^A_{max,h=0}$, $\mathbf{T}^A_{max,h=1.25}$, $\mathbf{T}^A_q$, and $\mathbf{T}^A_s$; (b) tests $\mathbf{T}^\Lambda_{max,h=0}$, $\mathbf{T}^\Lambda_{max,h=1.25}$, $\mathbf{T}^\Lambda_q$, and $\mathbf{T}^\Lambda_s$; and (c) tests $\mathbf{T}^S_{max,h=0}$, $\mathbf{T}^S_{max,h=1.25}$, $\mathbf{T}^S_q$, and $\mathbf{T}^S_s$. Vertical lines mark nominal significance levels of 0.05 and 0.1.

	$\mathbf{T}^A_{max}$	$\mathbf{T}^A_q$	$\mathbf{T}^A_s$	$\mathbf{T}^\Lambda_{max}$	$\mathbf{T}^\Lambda_q$	$\mathbf{T}^\Lambda_s$	$\mathbf{T}^S_{max}$	$\mathbf{T}^S_q$	$\mathbf{T}^S_s$
one covariate	0.058	0.044	0.046	0.020	0.044	0.038	0.018	0.046	0.036
two covariates	0.042	0.062	0.046	0.014	0.058	0.038	0.014	0.042	0.036

Table 1: *Constant effect model.* Observed significance levels under a nominal level of $\alpha = 0.05$. The standard error for each of the values is less than 0.01.

Constant effect model, two covariates. In this model, a second covariate, $z_2 \sim U[-0.5, 0.5]$, is added as a nuisance parameter. The hazard function that generates the (t_i, δ_i) data is still given by (7), with $z = z_1$; it does not depend on z_2. However, for the purpose of testing, the hazard coefficient of z_2 is included nonparametrically. The null hypothesis is $H_0 : \alpha_1(t) = \beta_1$, under the hazard model $\lambda(t|z) = \alpha_0(t) + \alpha_1(t)z_1 + \alpha_2(t)z_2$ with unknown $\alpha_2(t)$.

In both constant effect models, the data is generated under the null hypothesis. Therefore, these models allow us to study the actual significance level of our bootstrap tests. Table 1 summarizes the results for $N = 500$ samples. The numbers in the body of the table are the proportion of samples for which the P-value was less or equal to 0.05. The standard error for each of the entries is less than 0.01.

According to Table 1, the targeted significance level of $\alpha = 0.05$ is approximated best by the tests in class $\mathbf{T}^A$, and tests $\mathbf{T}^\Lambda_q$ and $\mathbf{T}^S_q$. Tests $\mathbf{T}^\Lambda_{max}$ and $\mathbf{T}^S_{max}$ tend to be overly conservative. For most of the tests, adding a second covariate as nuisance

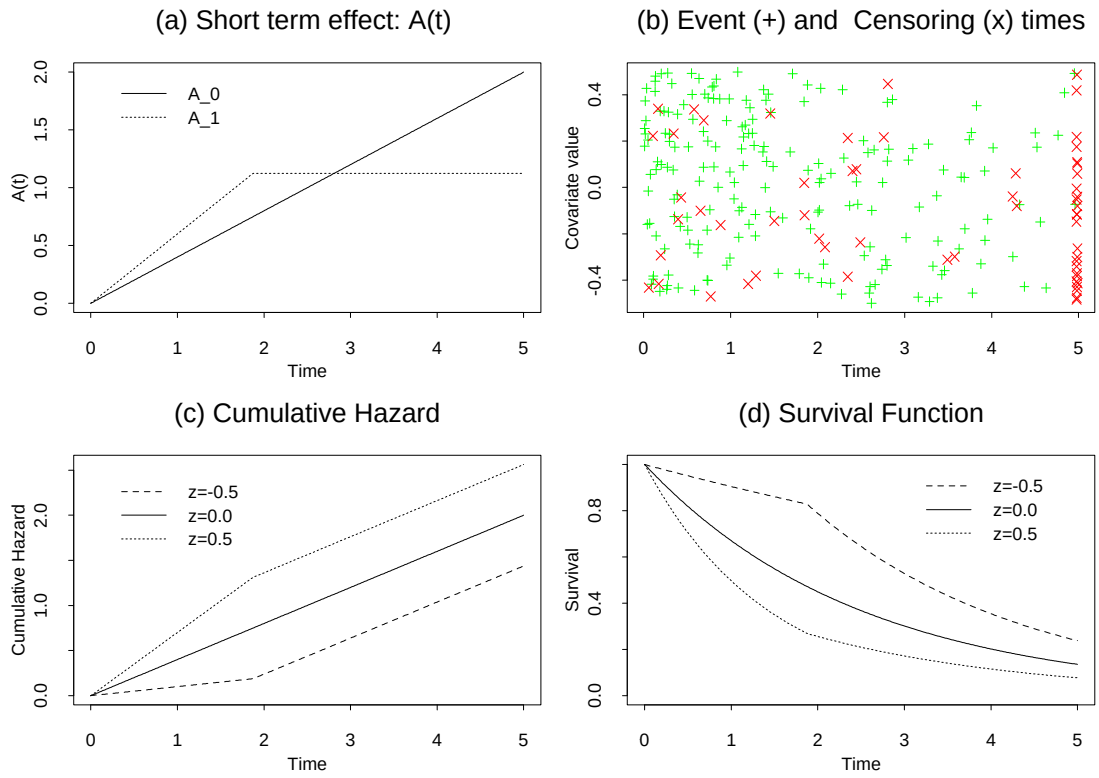

Figure 4: *Short-term effect model.* (a) Cumulative baseline hazard $A_0(t)$, and cumulative hazard coefficient $A_1(t)$; (b) event (+) and censoring times (×) for one $n = 250$ dataset; (c) cumulative hazard, and (d) survival function for the minimum ($z = -0.5$), mean ($z = 0$), and maximum ($z = 0.5$) of covariate values.

parameter does not change the actual significance level much. While the significance level seems to decrease for the test $\mathbf{T}^A_{max}$, and increase for $\mathbf{T}^A_q$, the differences are less than two standard deviations.

Short-term effect model. In the short-term effect model, the hazard function $\lambda(t|z) = \alpha_0(t) + \alpha_1(t)z$ is defined by

$$\alpha_0(t) = 0.4, \quad \text{and} \quad \alpha_1(t) = \begin{cases} 0.6, & \text{for } t \leq 1.875, \\ 0, & \text{for } t > 1.875. \end{cases} \tag{8}$$

Here, covariate effects on the hazard are constant over approximately the first half of the time range of interest, and then vanish.

Figure 4 describes the short-term effect model. Panel (a) shows the cumulative hazard coefficients $A_0(t)$ and $A_1(t)$. In Panel (b), one sample of size $n = 250$ is displayed, similar to Figure 1 (b). The data reflect that for large covariate values, the events tend to occur earlier; according to the model, the hazard for observations with $z = 0.5$ is 7 times higher than for observations with $z = -0.5$, for $t \leq 1.875$.

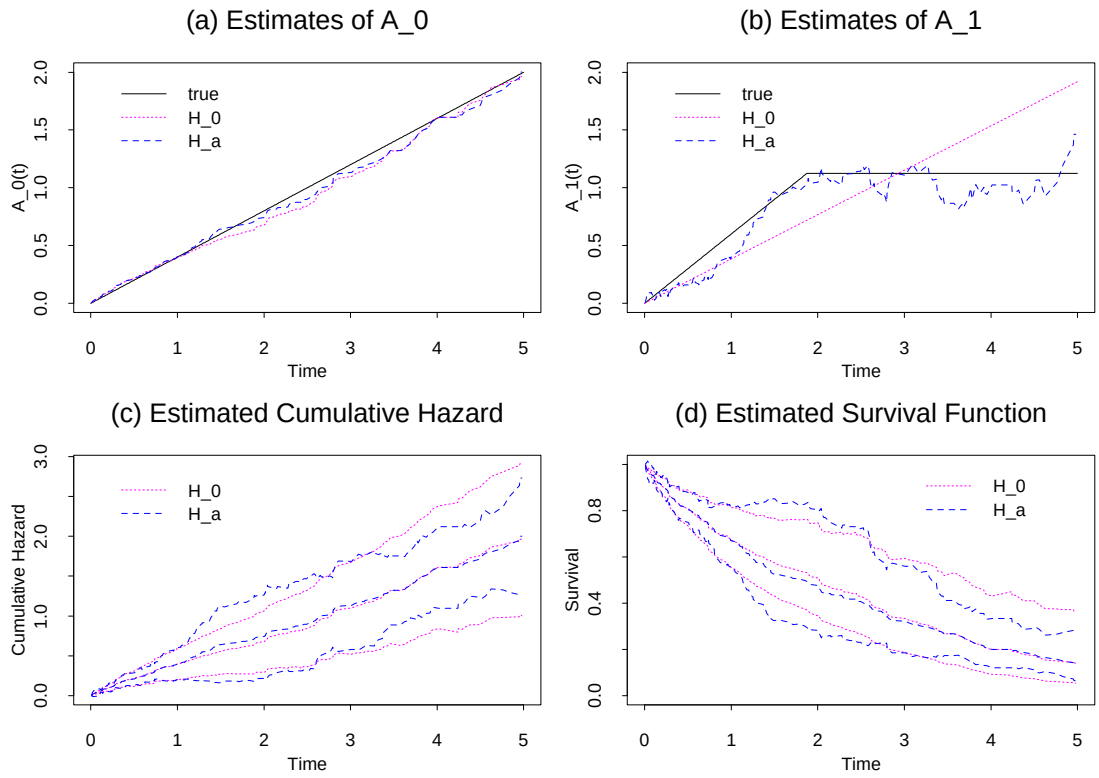

Figure 5: Estimates in the *short-term effect model*: (a) and (b) show A_0 and A_1 (solid lines) and their estimates under H_0 and under H_a; (c) and (d) show estimates of $\Lambda(t|z)$ and $S(t|z)$ under H_0 and under H_a for the minimum, mean and maximum of covariate values. Estimates are computed from the sample displayed in Figure 4(b)

The bottom row of Figure 4 gives the cumulative hazard function $\Lambda(t|z)$ and the survival function $S(t|z)$ for covariate values $z = -0.5$, $z = 0$ and $z = 0.5$.

Estimates for the short-term effect model are presented in Figure 5. In the upper row, the true cumulative hazard coefficients $A_0(t)$ and $A_1(t)$ are plotted as solid lines, and their estimates under H_0 and H_a as dotted and dashed lines, respectively. The estimates are computed from the data sample shown in Figure 4(b). Panel (b) of Figure 5 illustrates why the test statistics in class $\mathbf{T}^A$ are sensitive in identifying the short-term effect model versus the null hypothesis. The nonparametric estimate $\hat{A}_1(t)$ tracks the true function $A_1(t)$, although with substantial variation. In contrast, the estimate $\hat{\beta}_1 t$, under H_0, has to be a straight line. The $\mathbf{T}^A$ tests summarize the difference between these two curve estimates over the range $0.25 \leq t \leq 3$.

Panels (c) and (d) of Figure 5 again display estimates of the cumulative hazard $\Lambda(t|z)$ and of the survival function $S(t|z)$ under H_0 (dotted line) and H_a (dashed lines), for covariate values $z = -0.5$, $z = 0$ and $z = 0.5$. The $\mathbf{T}^\Lambda$ and $\mathbf{T}^S$ tests measure weighted

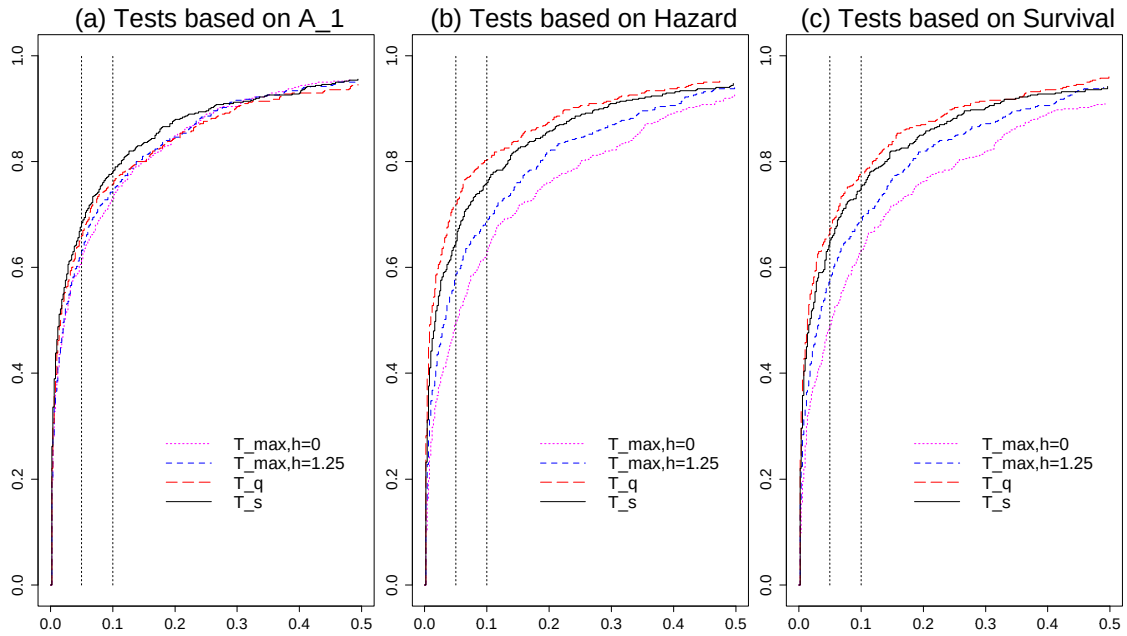

Figure 6: *Short-term effect model.* Empirical distribution function of bootstrap P-values based on $N = 500$ samples. Vertical lines mark nominal significance levels of 0.05 and 0.1. (a) Tests $T^A_{max,h=0}$, $T^A_{max,h=1.25}$, T^A_q, and T^A_s; (b) tests $T^\Lambda_{max,h=0}$, $T^\Lambda_{max,h=1.25}$, T^Λ_q, and T^Λ_s; (c) tests $T^S_{max,h=0}$, $T^S_{max,h=1.25}$, T^S_q, and T^S_s.

L_2 differences between these curves.

Figure 6 shows the empirical distribution of bootstrap P-values for the short-term effect model. These curves show the power of the corresponding tests, since the short-term effect model is included in the alternative H_a. The vertical dotted lines indicate nominal significance levels of 0.05 and 0.1. Numerical values for the power of the tests for a significance level of 0.05 are provided in Table 2 below.

The *short-term effect model with two covariates* is defined by the same hazard function (8). The second covariate, $z_2 \sim U[-0.5, 0.5]$, is introduced as nuisance parameter, with regression coefficient $\alpha_2(t) = 0$.

Delayed impact model. The hazard coefficients are defined as:

$$\alpha_0(t) = \begin{cases} 0.2, & \text{for } t \le 1.5, \\ 0.6, & \text{for } t > 1.5, \end{cases} \quad \text{and} \quad \alpha_1(t) = \begin{cases} 0, & \text{for } t \le 1.5, \\ 0.6, & \text{for } t > 1.5. \end{cases}$$

Here, the effect of the covariate on the hazard is delayed; the covariate value contributes to the hazard only starting at $t = 1.5$.

Figure 7 illustrates the model parameters, similar to Figures 1 and 4 above. Other

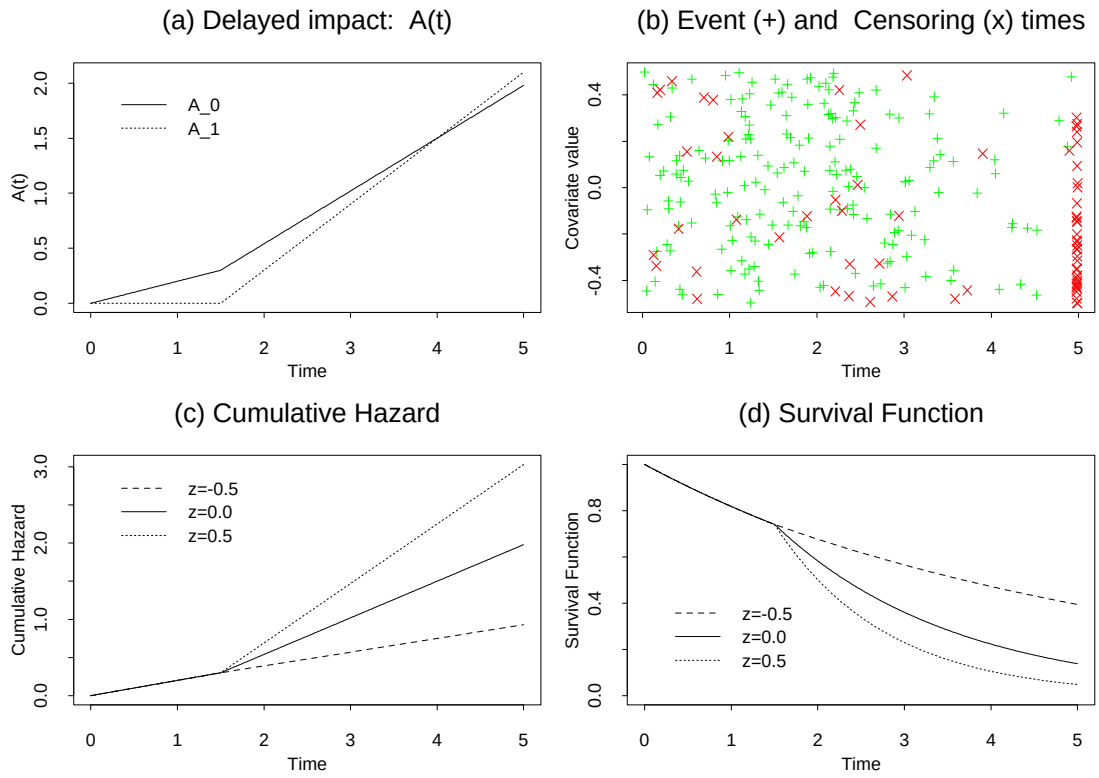

Figure 7: *Delayed impact model.* (a) Cumulative baseline hazard $A_0(t)$, and cumulative hazard coefficient $A_1(t)$; (b) event (+) and censoring times (×) for one $n = 250$ dataset; (c) cumulative hazard, and (d) survival function for the minimum ($z = -0.5$), mean ($z = 0$), and maximum ($z = 0.5$) of covariate values.

than in the previous models, the baseline hazard is not constant. We choose this particular form for our simulations to ensure that the risk set is not overly depleted before even the covariate effect sets in at $t = 1.5$.

Figure 8 displays the estimates obtained in the delayed impact model under H_0 and H_a, similar to Figures 2 and 5 for the previous models.

The empirical distributions of bootstrap P-values for the various tests are plotted in Figure 9. Again, curves represent the power of the corresponding tests, since the delayed impact model is part of the alternative, H_a. Vertical dotted lines denote the nominal significance levels 0.05 and 0.1. Numerical values for the power of the tests at a significance level of 0.05 are provided in Table 2.

Table 2 allows to compare the power of our tests under four alternatives, at a significance level of 0.05. The numbers in the body of the table are the proportion of samples for which the bootstrap P-values are less or equal to 0.05, based on $N = 500$ samples of size $n = 250$. The standard error of the entries is less than 0.023.

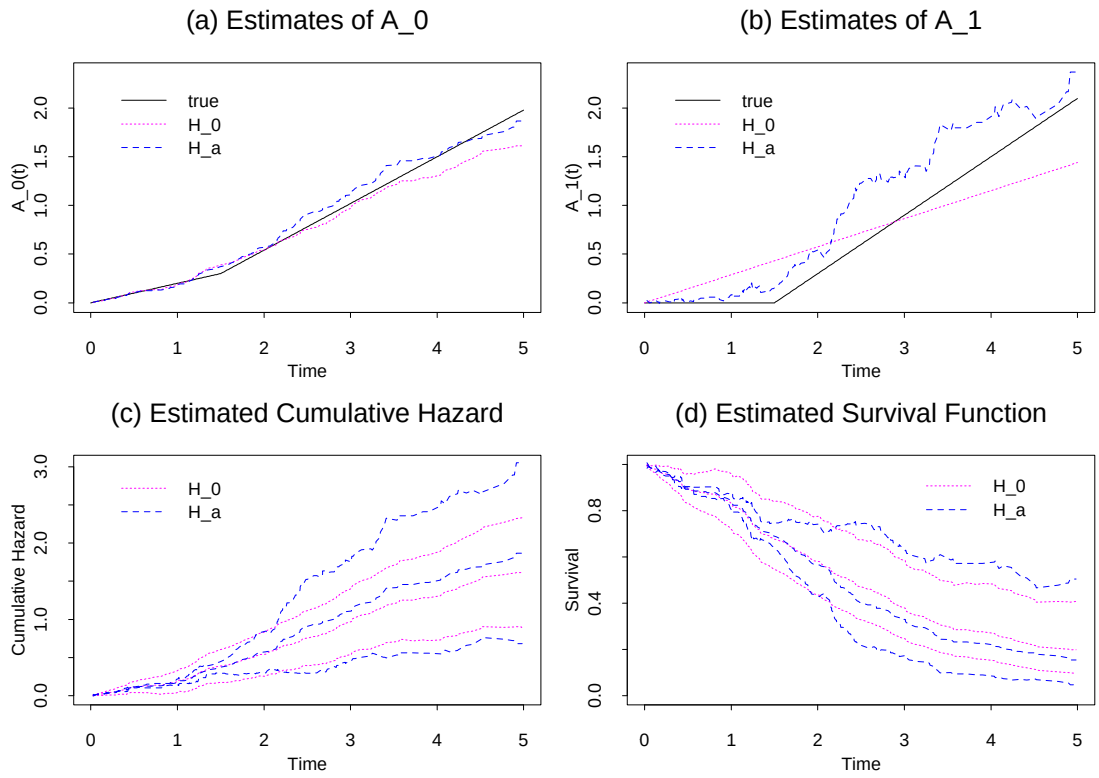

Figure 8: Estimates in the *delayed impact model*: (a) and (b) show A_0 and A_1 (solid lines) and their estimates under H_0 (dotted) and under H_a (dashed); (c) and (d) show estimates of $\Lambda(t|z)$ and $S(t|z)$ under H_0 and under H_a for the minimum, mean and maximum of covariate values. Estimates are based on the sample displayed in Figure 7(b).

	$\mathbf{T}^A_{max}$	$\mathbf{T}^A_q$	$\mathbf{T}^A_s$	$\mathbf{T}^\Lambda_{max}$	$\mathbf{T}^\Lambda_q$	$\mathbf{T}^\Lambda_s$	$\mathbf{T}^S_{max}$	$\mathbf{T}^S_q$	$\mathbf{T}^S_s$
				short-term effect model					
one covariate	0.626	0.668	0.688	0.502	0.726	0.666	0.498	0.676	0.654
two covariates	0.516	0.646	0.696	0.432	0.664	0.656	0.428	0.708	0.644
				delayed impact model					
one covariate	0.798	0.698	0.846	0.694	0.714	0.838	0.680	0.846	0.842
two covariates	0.756	0.774	0.854	0.590	0.802	0.826	0.576	0.868	0.828

Table 2: Power of the tests under four alternatives, for a significance level of 0.05, estimated from $N = 500$ samples. The standard error of the entries is less than 0.023.

Overall, the test statistics $\mathbf{T}^\bullet_s$ and $\mathbf{T}^\bullet_q$ seem to be somewhat more sensitive in identifying the alternatives. The better power corresponds to their being less conservative under H_0, as seen in better approximation of the 0.05 significance level under the constant ef-

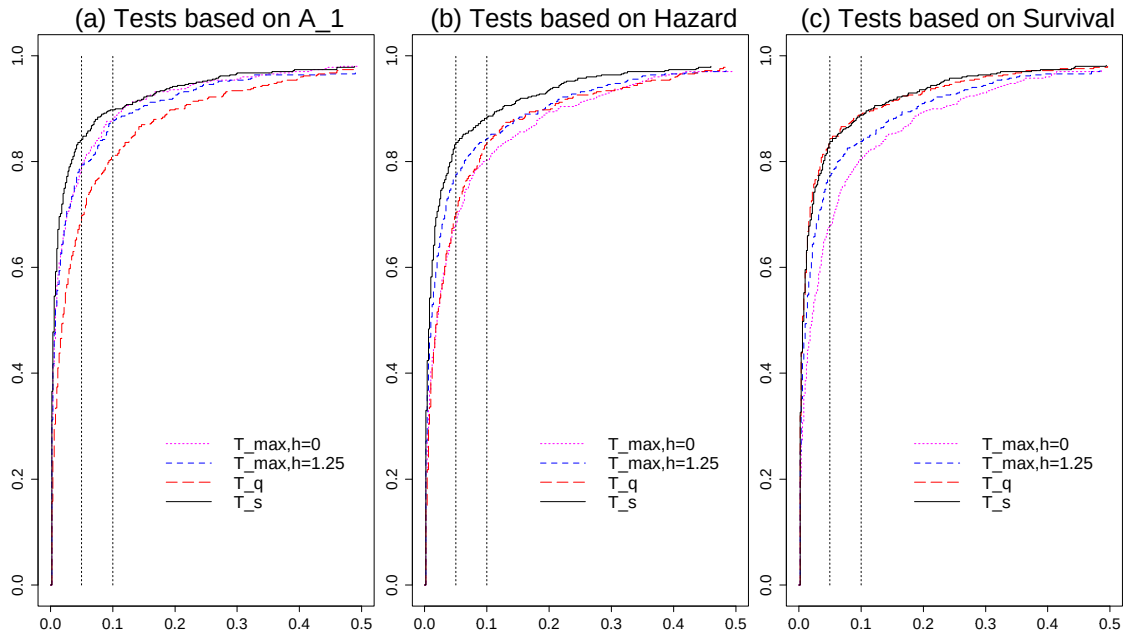

Figure 9: *Delayed impact model.* Empirical distribution functions of bootstrap P-values based on $N = 500$ samples. Vertical lines mark nominal significance levels of 0.05 and 0.1. (a) Tests $T^A_{max,h=0}$, $T^A_{max,h=1.25}$, T^A_q, and T^A_s; (b) tests $T^\Lambda_{max,h=0}$, $T^\Lambda_{max,h=1.25}$, T^Λ_q, and T^Λ_s; (c) tests $T^S_{max,h=0}$, $T^S_{max,h=1.25}$, T^S_q, and T^S_s.

fects model. Among the two classes, there is no clear winner between the $\mathbf{T}^A$, $\mathbf{T}^S$, and $\mathbf{T}^\Lambda$ statistics. Surprisingly, adding a second covariate does not reduce the power of these tests significantly. However, the second covariate reduces the power of the $\mathbf{T}^\bullet_{max}$ tests. Note that this table provides the power of our tests just for the four specific alternatives, and for only one significance level. The power of the tests for different significance levels is illustrated in Figures 6 and 9.

Finally, we investigate the effect of local linear smoothing on the power of the $\mathbf{T}^\bullet_{max}$ tests more closely. We consider tests $\mathbf{T}^\bullet_{max,h}$, where the semiparametric likelihood estimates $\hat{A}(t)$ are replaced by a smoothed version, $\tilde{A}(t;h)$. We use a local linear smoother with Epanechnikov kernel and bandwidth h, applied to the pairs $(t_i,\ \hat{A}_1(t_i))$, where the t_i are the event times only.

The results are summarized in Table 3. As in Table 2, the numbers in the body of the table are the proportion of samples for which the bootstrap P-values were less or equal 0.05, out of $N = 500$ samples of size $n = 250$. They estimate the actual significance level of the tests under the constant effect model (first three columns), and the power of the tests for the short-term effect and the delayed impact models (middle three and last three columns, respectively). The standard errors of the power

	constant effect			*short-term effect*			*delayed impact*		
h	$\mathbf{T}^A_{max,h}$	$\mathbf{T}^\Lambda_{max,h}$	$\mathbf{T}^S_{max,h}$	$\mathbf{T}^A_{max,h}$	$\mathbf{T}^\Lambda_{max,h}$	$\mathbf{T}^S_{max,h}$	$\mathbf{T}^A_{max,h}$	$\mathbf{T}^\Lambda_{max,h}$	$\mathbf{T}^S_{max,h}$
h=0	0.058	0.020	0.018	0.626	0.502	0.498	0.798	0.694	0.680
h=0.625	0.038	0.034	0.032	0.634	0.550	0.548	0.798	0.758	0.764
h=1.25	0.038	0.032	0.032	0.636	0.596	0.584	0.796	0.780	0.776
h=2.5	0.034	0.028	0.028	0.602	0.584	0.584	0.758	0.796	0.794

Table 3: *Effects of smoothing* on the performance of the tests $\mathbf{T}^\bullet_{max,h}$. Left three columns: actual significance level, for the constant effect model; middle columns: power of the tests under the short-term effect model, for a significance level of 0.05; right columns: power of the tests under the delayed impact model. The standard errors of the table entries are less than 0.023, and less than 0.01 for the first three columns.

estimates are less than 0.023, and for the first three columns, less than 0.01. The rows with $h = 0$ correspond to the original, unsmoothed tests. The hazard models each have one covariate.

It appears that moderate smoothing ($h = 0.625$ and $h = 1.25$) is beneficial, although the improvements are modest. There is virtually no difference in performance between these two bandwidths, which suggests that a sophisticated (global) bandwidth selection procedure might not add much practical value. Caution is advised against choosing the bandwidth too large; in this case, smoothing will obscure the features that distinguish the underlying hazard model from the null hypothesis. The effect of smoothing on the $\mathbf{T}^\bullet_{\max}$ tests can also be observed in Figures 3, 6 and 9.

Further simulations have shown that moderate smoothing has little or no effect on the tests $\mathbf{T}^\bullet_s$ and $\mathbf{T}^\bullet_q$, while oversmoothing may prove detrimental. The different effects of smoothing on the tests can be explained by the fact that smoothing the $\hat{A}(t)$ diminishes spurious bumps in the curve estimates. The supremum distance is particularly sensitive to these bumps, more so than the integrated quadratic distance used by the tests $\mathbf{T}^\bullet_s$ and $\mathbf{T}^\bullet_q$.

6 Example

In this section, we analyze a subset of the *Nursing home data* presented in Morris et. al. (1994). The original dataset consists of a treatment group (nursing homes received financial incentives for admitting certain kinds of patients), and a control group (no incentives). For each patient, the length of stay in the nursing home was recorded,

along with their age, gender, marital status, and a health status index. Our analysis is restricted to the control group, which includes $n = 889$ patients. Study duration in the control group was one year; censoring occurred only when patients remained in the nursing home at the end of the study.

Our goal is to determine how the length of stay for a patient is influenced by the four covariates in a fitted semiparametric additive hazard model, and which of the covariates, if any, should *not* be modeled by a *constant* hazard coefficient.

We are testing $H_0 : \alpha_k(t) = \beta_k$ for each of the covariates separately, while estimating the coefficients of the other three covariates nonparametrically.

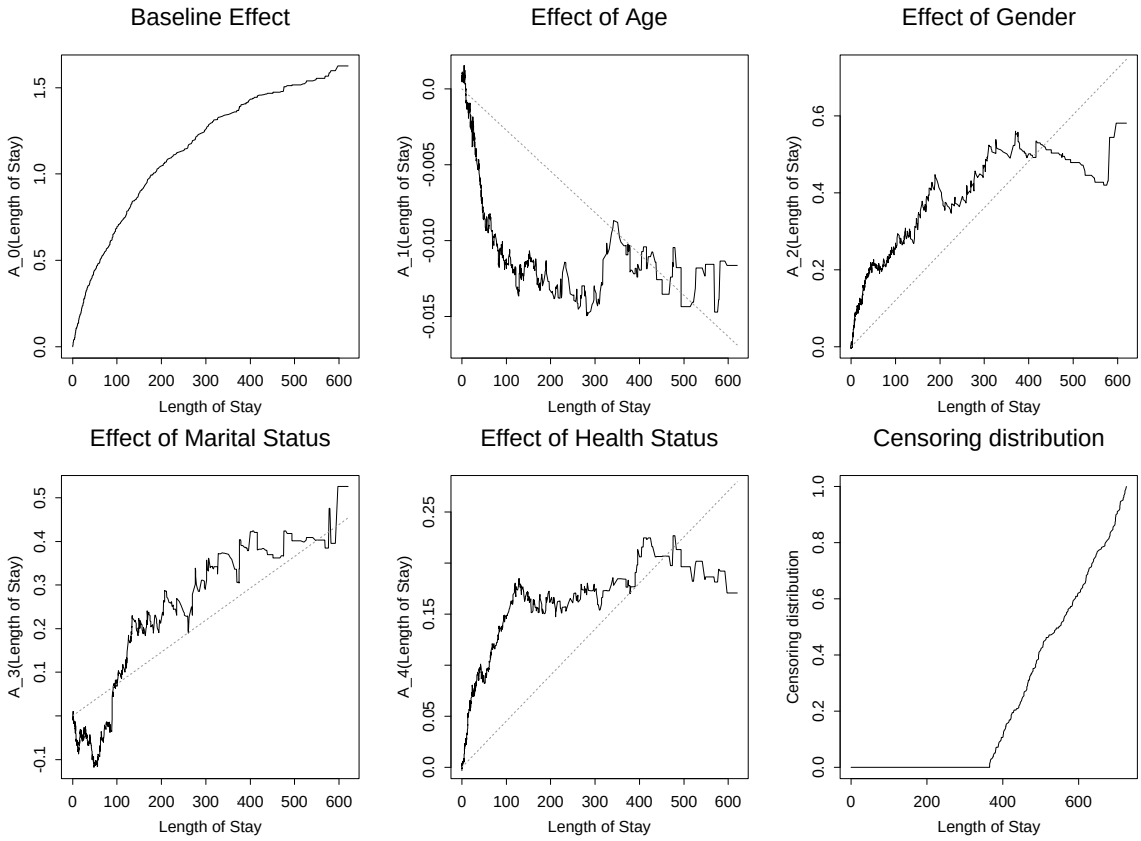

Figure 10: *Nursing home data.* Estimates of the $A_i(t)$ when all hazard coefficients are estimated nonparametrically (solid lines), and of $\beta_i t$ when a semiparametric model with one constant coefficient is fitted (dashed lines). Estimates of the cumulative baseline hazard, and of the censoring distribution.

Figure 10 shows the estimates $\hat{A}_k(t)$ of the cumulative hazard coefficients under the fully nonparametric additive model (Aalen model; solid lines) and under the semiparametric model (dashed lines). Additionally, estimates of the cumulative baseline hazard, $\hat{A}_0$, and of the censoring distribution are displayed. Note that our test statistics re-

	$\mathbf{T}^A_{max}$	$\mathbf{T}^A_q$	$\mathbf{T}^A_s$	$\mathbf{T}^\Lambda_{max}$	$\mathbf{T}^\Lambda_q$	$\mathbf{T}^\Lambda_s$	$\mathbf{T}^S_{max}$	$\mathbf{T}^S_q$	$\mathbf{T}^S_s$
Age	0.090	0.038*	0.042*	0.062	0.038*	0.042*	0.062	0.034*	0.048*
Gender	0.042*	0.018*	0.006*	0.106	0.016*	0.008*	0.100	0.006*	0.008*
Married	0.196	0.326	0.144	0.156	0.312	0.218	0.164	0.164	0.210
Health	0.758	0.768	0.680	0.724	0.780	0.656	0.730	0.688	0.660

Table 4: P-values obtained in testing for constant versus nonparametric hazard contributions of the covariates.

flect the distance between the solid and dashed lines, over the "interval of interest" $5 \leq t \leq 600$ (measured in days).

P-values for our bootstrap tests are summarized in Table 4. Each P-value is based on $B = 500$ bootstrap samples. Significant results (P-value ≤ 0.05) are marked by an asterisk.

The tests $\mathbf{T}^\bullet_q$ and $\mathbf{T}^\bullet_s$ consistently identify the hazard coefficients of *Age* and *Gender* as significantly nonconstant (P-value ≤ 0.05). The $\mathbf{T}^\bullet_{max}$ tests mostly are not significant. This results is consistent with the lower power of $\mathbf{T}^\bullet_{max}$ in the simulation study. The hazard contribution of *Married* and *Health* do not change over time significantly, and might be modelled as constant.

The plots of the cumulative hazard coefficients in Figure 10 help to understand how the non-constant hazard coefficients change with time. The panels for *Age* and *Gender* imply that both covariates have a strong differential effect early on, which vanishes later. In particular, the coefficient for *Age* suggests that older patients are less likely to have short-term stays. For more extended stays, age is not a factor in predicting the length of the stay.

References

[1] Aalen, O.(1980), A model for nonparametric regression analysis of counting processes, *in:* Klonecki, W., Kozek, A., and Rosinski, J. (Eds.), *Mathematical Statistics and Probability Theory*, Lecture Notes in Statistics **2**, 1-25, Springer-Verlag, New York.

[2] Aalen, O. (1989), A Linear Regression Model for the Analysis of Life Times, *Statist. Med.* **8**, 907-925.

[3] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993), *Statistical models based on counting processes*, Springer, New York.

[4] Burr, D. (1994), A comparison of certain bootstrap confidence intervals in the Cox model, *J. Amer. Statist. Assoc.* **89**, 1290-1302.

[5] Davison, A. and Hinkley, D. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

[6] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications,* Chapman & Hall, London.

[7] Horowitz, J. and Härdle, W. (1994), Testing a parametric model against a semiparametric alternative, *Econometric Theory* **10**, 821-848.

[8] Härdle, W. and Mammen, E. (1993), Comparing nonparametric versus parametric regression fits, *Ann. Statist.* **21**, 1926-1947.

[9] Lin, D.Y. and Ying, Z. (1994), Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61-71.

[10] McKeague and Sasieni (1994), A partly parametric additive risk model, *Biometrika* **81**, 501-14.

[11] Morris, C.N., Norton, E.C. and Zhou, X.H. (1994), Parametric duration analysis of nursing home usage, *in:* Lange, Ryan, Billard, Brillinger, Conquest and Greenhouse (Eds.), *Case Studies in Biometry,* Wiley, New York.

[12] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.