bshazard: A Flexible Tool for Nonparametric Smoothing of the Hazard Function

by Paola Rebora, Agus Salim and Marie Reilly

Abstract The hazard function is a key component in the inferential process in survival analysis and relevant for describing the pattern of failures. However, it is rarely shown in research papers due to the difficulties in nonparametric estimation. We developed the **bshazard** package to facilitate the computation of a nonparametric estimate of the hazard function, with data-driven smoothing. The method accounts for left truncation, right censoring and possible covariates. B-splines are used to estimate the shape of the hazard within the generalized linear mixed models framework. Smoothness is controlled by imposing an autoregressive structure on the baseline hazard coefficients. This perspective allows an 'automatic' smoothing by avoiding the need to choose the smoothing parameter, which is estimated from the data as a dispersion parameter. A simulation study demonstrates the capability of our software and an application to estimate the hazard of Non-Hodgkin's lymphoma in Swedish population data shows its potential.

Introduction

The hazard function is the basis of the inferential process in survival analysis, and although relevant for describing the pattern of failures, is often neglected in favor of survival curves in clinical papers. The most widely applied model in survival analysis (the Cox model) allows valid comparisons in terms of hazard ratios without distributional assumptions concerning the baseline hazard function, whose nonparametric estimate is rarely shown. Thus the reference against which the relative hazard is estimated is usually ignored and a crude measure of absolute risk is sometimes provided by the Kaplan-Meier estimator that is indirectly related to the shape of the hazard function.

In the methodological literature, some methods have been developed to obtain a nonparametric hazard estimate, including kernel-based (Müller and Wang, 1994; Hess et al., 1999) and spline-based estimators (O'Sullivan, 1988; Cai et al., 2002). However, specific statistical software commands accounting for the characteristics of survival data are lacking. An exception is the R package **muhaz** (Hess and Gentleman, 2010) that estimates the hazard function from right-censored data using kernel-based methods, but this package does not accommodate left-truncated data nor does it allow for adjustment for possible covariates. Flexible parametric survival models can also be used to describe and explore the hazard function (Royston and Parmar, 2002) and in R these have been implemented in the package **flexsurv** (Jackson, 2014). In these models a transformation of the survival function is modeled as a natural cubic spline function of the logarithm of time (plus linear effects of covariates). However this approach relies on an appropriate choice of the number of knots to be used in the spline.

We have developed the **bshazard** package to obtain a nonparametric smoothed estimate of the hazard function based on B-splines and generalized linear mixed models (GLMM). This perspective enables 'automatic' smoothing, as the smoothing parameter can be estimated from the data as a dispersion parameter (Lee et al., 2006; Pawitan, 2001; Eilers and Marx, 1996). Our implementation accommodates the typical characteristics of survival data (left truncation, right censoring) and also possible covariates. In the following sections we briefly review the methodology and demonstrate the performance of the package in numerical examples using simulated data. We illustrate the use of the package in an application to Swedish population data, where we estimate the incidence of Non-Hodgkin's lymphoma (NHL) in sisters of patients.

Methodological background

In this section we briefly review the methodology for smoothing the hazard rate inspired by Eilers and Marx (1996) and described in Chapter 9 of Lee et al. (2006).

Let T denote the possibly right censored failure time random variable; we are interested in the estimate of the hazard function defined as

$$h(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \cdot P(t < T \leq t + \Delta t | T \geq t)$$

so that $h(t)\Delta t$ is the probability of an event in the infinitesimal interval $(t, t + \Delta t)$, given survival to

time t.

Consider a sample of subjects that are observed from a common origin (e.g. start of exposure, time of diagnosis) to the occurrence of the event of interest. Observation can be right censored if the last follow-up time is before the occurrence of the event and/or left truncated if the observation starts after the origin (time 0). By partitioning the time axis into very small intervals, the number of events in each interval is approximately Poisson with mean $\mu(t) = h(t)P(t)$, where $P(t)$ is the total person-time in the interval (in the simplest case without censoring $P(t)$ will be the product of the number of individuals at risk at the beginning of the interval and the length of the interval). The time at risk of each subject can thus be split into n bins or time intervals (common to all subjects), similar to the life-table approach to survival analysis. For a data set with one record per individual including entry time, exit time and censoring indicator, this splitting of time can be implemented easily by the `splitLexis` function in the **Epi** package. Using (common) break points supplied by the user, the function divides each data record into disjoint follow-up intervals each with an entry time, exit time and censoring indicator, stacking these as separate 'observations' in the new data set. The bins should be small enough so that $h(t)$ can be considered approximately constant within each interval.

We will use t to denote the vector of the midpoints of all bins, with t_i representing the midpoint for the i^{th} bin ($i = 1, \ldots, n$). The vectors $y(t)$ and $P(t)$ represent the total number of events observed and the total person-time in each interval. Using the Poisson likelihood to approximate the general likelihood for survival data (Lee et al., 2006; Lambert and Eilers, 2005; Whitehead, 1980), the hazard can be estimated by modeling the expected number of events, $\mu(t)$, in each interval as a Poisson variable by using $P(t)$ as an offset term:

$$\log[\mu(t)] = f(t) + \log[P(t)],$$

where $f(t)$ denotes the logarithm of the hazard. In this context it is straightforward to account for possible covariates in a proportional hazards scheme. After the splitting, data can be aggregated according to the bin (time) and also to the covariate values for each subject, so that the final data is organised with one record for each bin and covariate combination. Denoting by X the design matrix which contains the covariate values (fixed in time) and by β the corresponding vector of coefficients, the model becomes:

$$\log[\mu(t)] = X\beta + f(t) + \log[P(t)].$$

Note that the coefficients β do not vary with time, so that subjects with different values of the covariates X have proportional hazard rates, i.e. $\log[h(t, X = x_1)] = \log[h(t, X = x_0)] + (x_1 - x_0)\beta$.

Smoothers, such as regression splines, can be used to estimate the function $f(t) = log[h(t)]$, and in particular B-splines provide a numerically efficient choice of basis functions (De Boor, 1978). B-splines consist of polynomial pieces of degree m, joined at a number of positions, called knots, along the time axis. The total number of knots (k) and their positions are arbitrarily chosen and are quite crucial for the final estimate, since the function can have an inflection at these locations. By using B-splines to estimate $f(t)$, the expected number of events above can be rewritten as:

$$\log[\mu(t)] = X\beta + Zv + \log[P(t)], \tag{1}$$

where Z is the matrix whose q columns are the B-splines, i.e. the values of the basis functions at the midpoints of the time bins (that will be repeated for each covariate combination) and v is a vector of length q of coefficients whose magnitude determines the amount of inflection at the knot locations. The number of basis functions $q = k + m - 1$, where k is the total number of knots, including minimum and maximum, and m is the degree of the polynomial splines. Thus the problem of estimating the hazard function reduces to the estimation of coefficients in a Generalised Linear Model framework.

B-splines are advantageous because the smoothed estimate is determined only by values at neighboring points and has no boundary effect. Nevertheless the major criticism of this approach is the arbitrary choice of the knots which determine the locations at which the smoothed function can bend and this has been subject to various investigations and discussions (Kooperberg and Stone, 1992). With a small number of knots, a B-spline may not be appealing because it is not smooth at the knots (this problem may be avoided by higher-degree B-splines) and can lead to underfitting of the data. The number of knots can be increased to provide a smoother estimate, but a large number of knots may lead to serious overfitting. O'Sullivan proposed using a relatively large number of knots and preventing overfitting by a penalty term to restrict the flexibility of the fitted curve (O'Sullivan, 1986, 1988), which is analogous to likelihood-based mixed effects modeling (Eilers and Marx, 1996; Lee et al., 2006; Pawitan, 2001). In fact, the penalized least squares equation for a quadratic penalty corresponds to a mixed model log-likelihood, where the set of second-order differences of the coefficients of the B-splines (denoted $\Delta^2 v$) have autoregressive structure and are normally distributed with mean 0 and variance $\sigma_v^2 I_{q-2}$ where I_{q-2} is the identity matrix with dimension $q - 2$ and q is the number of basis functions (Lee et al., 2006).

Intuitively, since the coefficients v of the B-splines determine the change at knot locations (if the B-splines are of degree 1, they determine the change in slope) they also determine the amount of smoothing. Assuming the coefficients are normally distributed with mean 0 helps to control the amount of smoothing and has the advantage of allowing an automatic smoothing in the sense that the smoothing parameter can be estimated directly from the data as a dispersion parameter (Eilers and Marx, 1996). Thus the main criticism of the use of B-splines is overcome: the choice of knots is no longer crucial for the final estimate and in fact a large number of equally spaced knots can be chosen (more than 40 knots are rarely needed) and overfitting is prevented by the penalty (Lee et al., 2006).

More formally, for model (1) the element z_{ij} represents the value of the j^{th} basis function ($j = 1,\ldots,q$) at the midpoint of the i^{th} bin ($i = 1,\ldots,n$) and the $q-2$ second-order differences of the coefficients are

$$\Delta^2 v = \begin{pmatrix} v_3 - 2v_2 + v_1 \\ v_4 - 2v_3 + v_2 \\ \cdots \\ v_q - 2v_{q-1} + v_{q-2} \end{pmatrix}$$

Assuming these to be normally distributed with mean 0 and variance $\sigma_v^2 I_{q-2}$ and conditioning on these random effects, the number of observed events y is assumed to follow a Poisson distribution with overdispersion: $E(y_i|v) = \mu_i$ and $V(y_i|v) = \mu_i\phi$, where ϕ represents the dispersion parameter.

The Extended Quasi-Likelihood approximation (Lee et al., 2006) facilitates an explicit likelihood formulation (also for overdispersed Poisson data):

$$\log L(\phi, \sigma_v^2, v) = \sum \left\{ -\frac{1}{2}\log\left[2\pi\phi V(y(t_i))\right] - \frac{1}{2\phi} d[y(t_i), \mu(t_i)] \right\} - \frac{q-2}{2}\log(2\pi\sigma_v^2) - \frac{1}{2\sigma_v^2} v^T R^{-1} v,$$

where $R^{-1} = (\Delta^2(I))^T \Delta^2(I)$, A^T denotes the transpose matrix of A and $d[y(t_i), \mu(t_i)]$ is the deviance function defined by:

$$d[y(t_i), \mu(t_i)] = 2\int_{\mu(t_i)}^{y(t_i)} \frac{y(t_i) - u}{V(u)} du.$$

This log-likelihood can be seen as a penalized likelihood where the term $v^T R^{-1} v$ is the roughness penalty; the smoothing parameter is determined by $\lambda = \frac{\phi}{\sigma_v^2}$ with a large λ implying more smoothing and a small λ denoting rough v. The extreme situation of $\lambda = 0$ corresponds to no smoothing and is analogous to consider v as a vector of fixed parameters.

The parameter ϕ, representing the dispersion parameter in the Poisson model, is usually assumed to be 1. However, if we use $\phi = 1$ when in fact $\phi > 1$ (overdispersion), we are likely to undersmooth the data, while the estimate is not influenced by underdispersion (Lee et al., 2006).

The advantage of the mixed model approach is that we have an established procedure for estimating λ, equivalent to estimating variance components in mixed models. In this setting the Iterative Weighted Least Squares (IWLS) numerical algorithm works reliably and is usually used. This algorithm uses a Taylor approximation to the extended likelihood and the application to mixed models is described in detail in Chapters 6 and 17 in Pawitan (2001).

In our application, the following iterative algorithm is used:

1. Given initial/last estimated values of λ, ϕ and $\mu(t_i)$, estimate v and β.
2. Given $\widehat{v}$ and $\widehat{\beta}$, update the estimates of λ and ϕ.
3. Iterate between 1 and 2 until convergence.

We begin by implementing step 1 of the IWLS algorithm as follows. Given a fixed value of λ (starting with $\lambda = 10$ is a good starting value), we compute the working vector Y

$$Y_i = z_i^T v^0 + x_i^T \beta^0 + \log(P(t_i)) + \frac{y(t_i) - \mu(t_i)^0}{\mu(t_i)^0},$$

where z_i is the i^{th} row of Z, $y(t_i)$ and $P(t_i)$ are the number of observed events and the total person-time in the i^{th} interval and x_i denotes the i^{th} row of the matrix X_c of covariate values centered at their mean values. For the starting values $\mu(t_i)^0$ we take the average over all time intervals of the raw hazard, computed as the number of events divided by the person-time at risk in the interval, $\frac{y(t_i)}{P(t_i)}$. As for the coefficients, we take $v^0 = \log[\mu(t)^0]$ and $\beta^0 = 0$. Defining W as the variance of the working vector (Y) with elements $w_i = \mu(t_i)^0$, the updating formula for the random effects is the solution to

$$\left(Z^T W Z + \lambda R^{-1}\right) v = Z^T W \left(Y - \log(P(t)) - X_c\beta\right)$$

where $R^{-1} = \left(\Delta^2(I)\right)^T \Delta^2(I)$ and β is the solution to

$$\left(X_c^T W X_c\right) \beta = X_c^T W \left(Y - \log(P(t)) - Zv\right).$$

Note that if λ is set to 0, v is estimated as a vector of fixed parameters as mentioned above.

At this point we can introduce a quantity to describe the complexity of the smoothing, that is the effective number of parameters (or degrees of freedom) associated with v, denoted by df and given by:

$$df = \text{trace}\left\{\left(Z^T W Z + \lambda R^{-1}\right)^{-1} Z^T W Z\right\}.$$

When λ is set to 0, df is equal to the number of basis functions q, while it decreases with increasing penalisation.

For step 2, given $\widehat{v}$ and $\widehat{\beta}$, the dispersion parameter is updated by the method of moments (Wedderburn, 1974) as follows:

$$\widehat{\phi} = \text{var}\left[\frac{y(t_i) - \widehat{\mu(t_i)}}{\sqrt{\widehat{\mu(t_i)}}}\right].$$

An estimated variance greater than 1 would suggest overdispersion in the data. When we believe overdispersion is not present (ϕ close to 1), we suggest fixing ϕ at 1, especially when adjusting for covariates where there is a greater risk of overfitting.

The quantity σ_v^2 can be updated (Lee et al., 2006) by:

$$\widehat{\sigma}_v^2 = \frac{\widehat{v}^T R^{-1} \widehat{v}}{df - 2}.$$

Once convergence is reached, a point-wise confidence band for the smooth estimate is computed for the linear predictor $\log[\widehat{\mu}(t)]$ with variance matrix (assuming the fixed parameters are known at the estimated values) $H = Z(Z^T W Z + \lambda R^{-1})^{-1} Z^T$. This is then transformed to the hazard scale by:

$$\frac{\exp\left\{\log[\widehat{\mu}(t)] \pm z_{\alpha/2}\sqrt{H}\right\}}{P(t)},$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of a standard normal distribution.

Getting started

This section provides explanations of the input data and output data of the package **bshazard**, which estimates the hazard function nonparametrically. In order to use package **bshazard**, the following R packages need to be installed: **survival**, **Epi** and **splines** (Therneau, 2014; Carstensen et al., 2011). After installation, the main function can be called by:

```
library(bshazard)
bshazard(formula, data, nbin, nk, degree, l0, lambda, phi, alpha, err, verbose)
```

The only mandatory argument is `formula`, which should contain a survival object (interpreted by the **survival** package) with the time of event (possibly in a counting process format), the censoring indicator and (optionally) the covariates. For example, the function

```
output <- bshazard(formula = Surv(time_of_entry, time_of_exit,
  censoring_indicator ~ covariates))
```

will produce a smoothed estimate of the hazard accounting for `covariates` with the following default settings:

- the time axis split at each distinct observed time in the data (censoring or event);
- B-splines with 31 knots;
- degree of B-splines 1;
- smoothing parameter estimated from the data (with starting value 10);
- overdispersion parameter estimated from the data;
- 95% confidence intervals.

By providing various arguments in the function call, the user can override the default settings for the data format (e.g. specify a data frame), the number of bins (`nbin`), the number of knots (`nk`), degree of splines (`degree`), smoothing and overdispersion parameters (`lambda` and `phi`), confidence level (`alpha`) and convergence criterion (`err`). Detailed explanations are provided in the function's help file.

The `output` of the `bshazard` function includes a data frame with the original data split in time intervals (`raw.data`), vectors containing the estimated hazard and confidence limits and the parameter estimates (coefficients of the covariates, $\hat{\phi}$, $\hat{\sigma}_v^2$, df).

The package also includes the following functions:

- `summary(output)` prints the values of the estimated hazard, with confidence limits, for each observed time and the parameter estimates;
- `plot(output)` and `lines(output)` plot the hazard, with confidence limits.

Numerical examples

In order to test the flexibility of the proposed algorithm we simulated data from a non monotone function that could represent, for example, the seasonality of flu incidence:

$$h(t) = b \cdot \left[h_1^{p_1} p_1 t^{p_1-1} e^{-(h_1 \cdot t)^{p_1}}\right] + (1-b) \cdot \left[h_2^{p_2} p_2 t^{p_2-1} e^{-(h_2 \cdot t)^{p_2}}\right], \quad (2)$$

where $h(t)$ is the hazard function, b is a Bernoulli random variable with parameter 0.6, $h_1 = 1.2$, $p_1 = 2$, $h_2 = 0.3$, $p_2 = 5$. The choice of the parameter values was inspired by Cai et al. (2002). We considered samples of 500 subjects and for each subject we also simulated a censoring time as $U(0,5)$. Under this model we simulated 500 random samples and, for each sample, we estimated the hazard function by:

```
fit <- bshazard(Surv(exit_time, cens) ~ 1, data = dati, nbin = 100)
```

The choice to pre-bin the data in 100 time intervals was for simplicity of comparison of different estimates of hazard from different simulations at the same time point. The hazard function estimate did not change when using different numbers of bins or different numbers of knots (data not shown). The results of this simulation are summarised in Figure 1. The mean estimate of the hazard function is very close to the true hazard. For comparison, we also estimated the hazard using `muhaz(exit_time, cens, max.time = 3, bw.method = "g", n.est.grid = 100)` and plotted its mean estimate in Figure 1, where it can be seen to be very close to the true hazard. Under the same distribution we also simulated a left truncation time l as $U(-1,1)$, with $l < 0$ considered as 0 (late entry for half of the subjects). In this simulation, only subjects with event/censoring times greater than the left truncation time were valid for analysis. This setting provided results very similar to the previous setting (data not shown).

In a second simulation, we included a covariate X generated as a standard normal random variable that influenced the hazard rate according to the model:

$$h(t) = 0.5t \cdot \exp(X).$$

Under this model we again simulated 500 random samples and, for each sample, estimated the hazard function by:

```
fit <- bshazard(Surv(entry_time, exit_time, cens) ~ x, data = dati, nbin = 30)
```

Results are shown in Figure 2. The mean estimate of the hazard function is again very close to the true hazard. Note that when adjusting for continuous covariates, the hazard is estimated under the assumption of a linear effect of the covariates centered at their mean values. In this case, a comparison with the `muhaz` function was not possible given that it does not accommodate covariates or late entry.

Results from non-Hodgkin's lymphoma data

In this section we use the **bshazard** package to estimate the hazard of cancer diagnosis using data from Swedish population registers. For all individuals born in Sweden since 1932, the Multi-Generation Register maintained by Statistics Sweden (Statistics Sweden, 2009) identifies the biological parents, thus enabling the construction of kinships within families. Using the individually unique national registration number, individuals in this register can be linked to the National Cancer Register, which records all malignancies diagnosed since 1958. In our application we use data from Lee et al. (2013) who analyzed 3,015 sisters of 1,902 non-Hodgkin's lymphoma (NHL) female patients and 15,697 sisters of 3,836 matched controls who were cancer free at the time of diagnosis of the corresponding case and who matched the case with respect to age, year of birth, sex and county of residence.

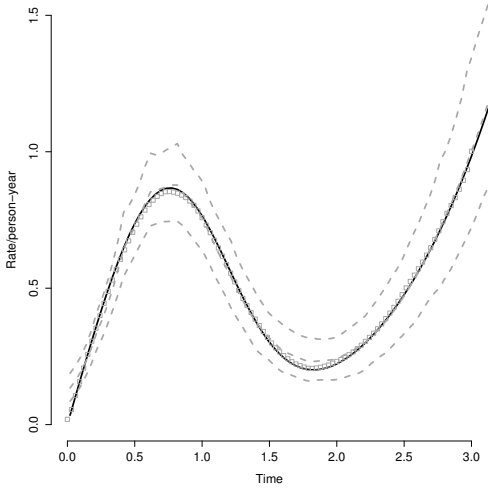

Figure 1: Numerical example; estimated and true hazard function ($h(t) = b[1.2^2 2te^{-(1.2t)^2}] + (1-b)[0.3^5 5t^4 e^{-(0.3t)^5}]$) with right censoring for $n = 500$. The solid black line is the true hazard function, the dashed gray lines are the mean estimate from `bshazard` and the empirical pointwise 95% confidence interval. Squares represent the mean estimate from the `muhaz` function.

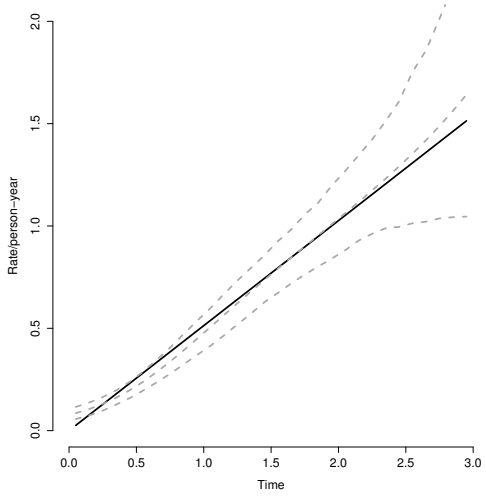

Figure 2: Numerical example; estimated and true hazard function ($h(t) = 0.5t\exp(X)$) with left truncation and right censoring for $n = 500$. The solid black line represents the true hazard function, and the dashed gray lines are the mean estimate and the empirical pointwise 95% confidence interval.

Sisters are at risk from birth to the age at NHL diagnosis or censoring due to death, emigration or end of study (2007). Individuals born before the start of the cancer register (1958) were considered at risk from their age at the start of the cancer register, resulting in delayed entry. In the study period, 32 NHL diagnoses were observed in the exposed group (sisters of subjects with a diagnosis of NHL) and 39 in the unexposed group (sisters of cancer-free subjects).

In order to illustrate the automatic smoothing, we first concentrated on sisters of cancer-free subjects and computed the smoothed hazard using the usual B-splines (i.e. setting the smoothing parameter $\lambda = 0$):

```
fit_notexp_l0 <- bshazard(Surv(entry_age, exit_age, cens) ~ 1, lambda = 0,
  data = sis[nexpo])
```

where `[nexpo]` selects only the sisters of control subjects.

The resulting hazard function, plotted in Figure 3, has several bumps, so we proceeded to estimate the smoothing parameter from the data, again using the same number of knots (31 as default):

```
fit_notexp <- bshazard(Surv(entry_age, exit_age, cens) ~ 1, data = sis[nexpo])
```

The resulting estimate of λ was 11,311.84 with 2.40 degrees of freedom and $\widehat{\phi} = 0.82$ and the dotted line in Figure 3 presents the corresponding hazard estimate. The estimate of the hazard function was unchanged when using different numbers of knots or by setting the overdispersion parameter to 1 (in fact no overdispersion was found $\widehat{\phi} = 0.82$).

Lee et al. found that sisters of female NHL patients have hazard ratio of NHL of 4.36 (95% confidence interval [2.74; 6.94]) compared to sisters of controls (Lee et al., 2013). For comparison with their results we estimated the risk of NHL adjusting for 'exposure' (i.e. being a sister of a case rather than a control):

```
fit_adj_exp <- bshazard(Surv(entry_age, exit_age, cens) ~ case, data = sis)
```

where the variable `case` indicates whether the subject is a sister of a case. We obtained a very similar hazard ratio, $\exp(\widehat{\beta}) = 4.35$, as expected. Note that the code provides the hazard for a subject with covariate value equal to the mean of all subjects:

$$\widehat{h}(t;\overline{x}) = \frac{\widehat{\mu}(t,\overline{x})}{P(t)} = \frac{\exp\left[\overline{x}\widehat{\beta} + z\widehat{v} + \log(P(t))\right]}{P(t)}.$$

Since this estimate is not meaningful for categorical variables, we obtained separate hazard

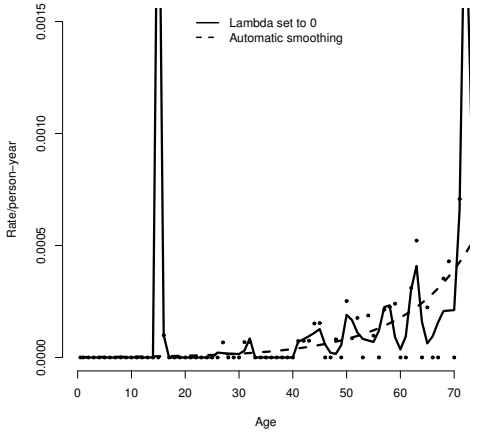

Figure 3: Hazard estimate of NHL in sisters of controls with smoothing parameter set to 0 (continuous line) and smoothing estimated from the data (dashed line). Dots represent the raw hazard.

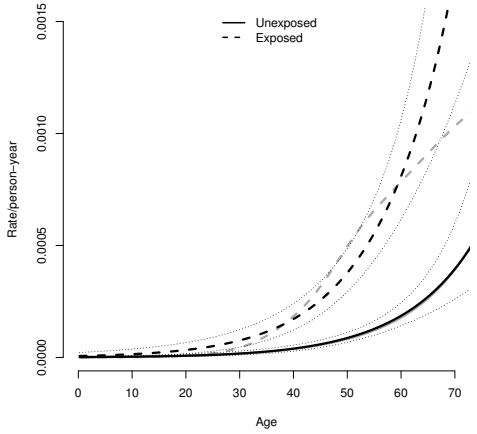

Figure 4: Smoothed hazard of NHL in unexposed (continuous line) and exposed (dashed line) sisters obtained from the model with exposure as a covariate. Dotted lines represent confidence intervals. For reference, the stratified estimates obtained from separate models for exposed and unexposed sisters are presented in grey.

estimates for unexposed and exposed subjects from:

$$\widehat{h}(t;x) = \frac{\exp\left[x\widehat{\beta} + z\widehat{v} + \log(P(t))\right]}{P(t)} = \widehat{h}(t;\overline{x}) \cdot \exp\left[(x-\overline{x})\widehat{\beta}\right].$$

The function `plot` in the package **bshazard** calls an object of class 'bahazard' and allows one to easily plot separate curves for each set of covariate values. The estimates plotted in Figure 4 were obtained using `plot(fit_adj_exp, overall = FALSE)` and assume proportionality of the hazard in exposed and unexposed sisters. As a reference, we also performed stratified analyses obtaining estimates separately for exposed and unexposed sisters and these are plotted in grey in Figure 4. As expected, the hazard estimates are similar to the separate estimates, but are constrained by the assumption of proportionality. This is especially evident in the exposed group: the hazard increase seems to level off after age 55, but this is not detected by the hazard estimate obtained under the joint adjusted model.

The adjustment to the hazard estimate is particularly advantageous for continuous variables. The proposed method allows inclusion of more than one covariate, so in the NHL application the hazard of exposed and unexposed subjects could be further adjusted for calendar time by:

```
fit_adj_exp_caly <- bshazard(Surv(entry_age, exit_age, cens) ~ case + yob, data = sis)
```

This yielded hazard estimates that were essentially unchanged and are not reported here.

Discussion

We have implemented a software package in R to obtain a nonparametric smoothed estimate of the hazard function based on B-splines from the perspective of generalized linear mixed models. The Poisson model leads to an elegant form for the log of the hazard (Lambert and Eilers, 2005). We adopted the discrete approach to survival analysis allowing staggered entry, intermittent risk periods and large data sets to be handled with ease. The code is easy to use and accommodates the typical characteristics of survival data (left truncation, right censoring) by using the **survival** package. It also accounts for possible covariates, but the user should be aware that covariates are assumed to have a constant effect on the hazard (proportional hazards). It is of note that the model is also valid for estimating the rate function over time where an individual can have repeated events (Cook and Lawless, 2002). For such applications, the code can be used without change and data should be included in a counting process format. The package **bshazard** is available from http://CRAN.R-project.org/package=bshazard.

The main advantage of our function is that the user can obtain the estimated hazard by a simple line of code and that the extent of smoothing is data-driven (i.e. the user does not need to specify any smoothing parameter). The number of degrees of freedom gives an intuitive interpretation of the amount of smoothing and is also useful for making comparisons between different smoothers.

To prepare the data for analysis the package uses the `splitLexis` function (**Epi** package). The choice of time intervals does not affect the smoothed estimate as long as the bins are small enough for the assumption of constant hazard within the bin to be reasonably satisfied. For large numbers of observations the splitting can be time consuming, especially when accounting for many covariates. Nevertheless, in our relatively large data-set of NHL sisters, the most complex hazard estimate, adjusting for two covariates (`fit_adj_exp_caly`) was obtained in less than one minute. Interval censored data are not included in the code at this time, but the package can still be used if the censored intervals are relatively short. In this situation we could choose the bins in such a way that every censored interval is completely included in one bin, avoiding the problem of the specification of the exact event time, but some assumptions on person-time at risk will be needed. With small data sets and in the presence of covariates, estimation of both smoothing and overdispersion parameters can cause some convergence problem; in this case if there is not strong evidence of overdispersion we suggest fixing ϕ at 1.

The possibility to estimate the hazard function with a simple command provides a useful tool for a deeper understanding of the process being studied, both in terms of the magnitude of the risk and the shape of the curve (Rebora et al., 2008). For example, in a previous paper, we found that sisters of subjects diagnosed with NHL have a hazard ratio of 4.36 (95% confidence interval [2.74; 6.94]) for NHL compared to sisters of controls (Lee et al., 2013), but did not show at which age the risk was higher. Reanalyzing the data using **bshazard** revealed how the magnitude of the risk varied with age. This important information is often neglected in epidemiological studies, in large part due to the lack of simple and accessible software tools. An important area of application is in the presence of time-dependent variables, when an absolute measure of risk cannot be obtained by the Kaplan-Meier estimator. For example, in a comparison between the effect on disease-free survival of chemotherapy and transplantation, which occur at different time points, the Kaplan-Meier method will tend to overestimate the survival of the transplanted group, since these patients have to survive until transplant (immortal time bias). In such situations, a hazard estimate is particularly useful for presenting the instantaneous risk of an event over time given that it conditions on the subjects at risk at each time.

In summary, the **bshazard** package can enhance the analysis of survival data in a wide range of applications. The advantage of automatic smoothing and the close relationship with the `survfit` function make the package very simple to use.

Acknowledgments

The authors would like to thank Yudi Pawitan who provided the initial code for smoothing.

Bibliography

T. Cai, R. J. Hyndman, and M. P. Wand. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, 11(4):784–798, 2002. [p114, 118]

B. Carstensen, M. Plummer, E. Laara, and M. H. e. al. *Epi: A Package for Statistical Analysis in Epidemiology*, 2011. URL `http://CRAN.R-project.org/package=Epi`. R package version 1.1.20. [p117]

R. J. Cook and J. F. Lawless. Analysis of repeated events. *Statistical Methods in Medical Research*, 11(2): 141–166, 2002. [p120]

C. De Boor. *A Practical Guide to Splines*. New York: Springer Verlag, 1978. [p115]

P. H. C. Eilers and B. D. Marx. Flexible Smoothing with *B*-splines and Penalties. *Statistical Science*, 11 (2):89–102, 1996. [p114, 115, 116]

K. Hess and R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*, 2010. URL `http://CRAN.R-project.org/package=muhaz`. R package version 1.2.5. [p114]

K. R. Hess, D. M. Serachitopol, and B. W. Brown. Hazard function estimators: A simulation study. *Statistics in Medicine*, 18(22):3075–3088, 1999. [p114]

C. Jackson. *flexsurv: Flexible Parametric Survival and Multi-State Models*, 2014. URL http://CRAN.R-project.org/package=flexsurv. R package version 0.5. [p114]

C. Kooperberg and C. J. Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992. [p115]

P. Lambert and P. H. C. Eilers. Bayesian proportional hazards model with time-varying regression coefficients: A penalized poisson regression approach. *Statistics in Medicine*, 24(24):3977–3989, 2005. [p115, 120]

M. Lee, P. Rebora, M. G. Valsecchi, K. Czene, and M. Reilly. A unified model for estimating and testing familial aggregation. *Statistics in Medicine*, 32(30):5353–5365, 2013. [p118, 119, 121]

Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*, volume 106. Chapman & Hall/CRC, 2006. [p114, 115, 116, 117]

H.-G. Müller and J.-L. Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61–76, 1994. [p114]

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986. [p115]

F. O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2):363–379, 1988. [p114, 115]

Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001. [p114, 115, 116]

P. Rebora, K. Czene, and M. Reilly. Timing of familial breast cancer in sisters. *Journal of the National Cancer Institute*, 100(10):721–727, 2008. [p121]

P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002. [p114]

Statistics Sweden. The Multi-Generation Register 2008. A Description of Contents and Quality, 2009. [p118]

T. Therneau. *survival: A Package for Survival Analysis in S*, 2014. URL http://CRAN.R-project.org/package=survival. R package version 2.37-7. [p117]

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974. [p117]

J. Whitehead. Fitting Cox's regression model to survival data using GLIM. *Applied Statistics*, 29(3): 268–275, 1980. [p115]

Paola Rebora
Department of Health Sciences
University of Milano-Bicocca
Via Cadore 48 20900 Monza, Italy
paola.rebora@unimib.it

Agus Salim
Department of Mathematics & Statistics
La Trobe University
Bundoora, Australia
a.salim@latrobe.edu.au

Marie Reilly
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
BOX 281, 171 77 Stockholm, Sweden
marie.reilly@ki.se