flexsurv: a platform for parametric survival modelling in R

Christopher H. Jackson
MRC Biostatistics Unit, Cambridge, UK
chris.jackson@mrc-bsu.cam.ac.uk

Abstract

flexsurv is an R package for fully-parametric modelling of survival data. Any parametric time-to-event distribution may be fitted if the user supplies a probability density or hazard function, and ideally also their cumulative versions. Standard survival distributions are built in, including the three and four-parameter generalized gamma and F distributions. Any parameter of any distribution can be modelled as a linear or log-linear function of covariates. The package also includes the spline model of Royston and Parmar (2002), in which both baseline survival and covariate effects can be arbitrarily flexible parametric functions of time. The main model-fitting function, `flexsurvreg`, uses the familiar syntax of `survreg` from the standard **survival** package. Censoring or left-truncation are specified in `Surv` objects. Estimates and confidence intervals for any function of the model parameters can be printed or plotted. **flexsurv** also enhances the **mstate** package (de˜Wreede, Fiocco, and Putter 2011) by providing cumulative hazards for fully-parametric multi-state models. This article explains the methods and design principles of the package, giving several worked examples of its use.

Keywords:˜survival.

1. Motivation and design

The Cox model for survival data is ubiquitous in medical research, since the effects of predictors can be estimated without needing to supply a baseline survival distribution that might be inaccurate. However, fully-parametric models have many advantages, and even the originator of the Cox model has expressed a preference for parametric modelling (see Reid 1994). Fully-specified models help to understand the pattern of the change in hazard through time, and help with prediction and extrapolation. For example, the mean survival $E(T) = \int_0^\infty S(t)dt$, used in health economic evaluations (Latimer 2013), needs the survivor function $S(t)$ to be fully-specified for all times t.

flexsurv allows parametric distributions of arbitrary complexity to be fitted to survival data, gaining the convenience of parametric modelling, while avoiding the risk of model misspecification. Built-in choices include spline-based models with any number of knots (Royston and Parmar 2002) and 3–4 parameter generalized gamma and F distribution families. Any user-defined model may be employed by supplying at minimum an R function to compute the probability density or hazard, and ideally also its cumulative form. Any parameters may be modelled in terms of covariates, and any function of the parameters may be printed or

plotted in model summaries.

flexsurv is intended as a general platform for survival modelling in R. The `survreg` function in the R package **survival** (Therneau 2014) only supports two-parameter (location/scale) distributions, though users can supply their own distributions if they can be parameterised in this form. Many other contributed R packages can fit survival models, e.g. **eha** (Broström 2014) and **VGAM** (Yee and Wild 1996), though these are either limited to specific distribution families, not specifically designed for survival analysis, or (**ActuDistns**, Nadarajah and Bakar 2013) contain only the definitions of distribution functions. **flexsurv** enables distribution functions provided by such packages to be used as survival models.

It is similar in spirit to the Stata packages **stpm2** (Lambert and Royston 2009) for spline-based survival modelling, and **stgenreg** (Crowther and Lambert 2013) for fitting survival models with user-defined hazard functions using numerical integration. Though in **flexsurv**, slow numerical integration can be avoided if the analytic cumulative distribution or hazard can be supplied, and optimisation can also be speeded by supplying analytic derivatives. **flexsurv** also has features for multi-state modelling and interval censoring, and general output reporting. It employs functional programming to work with user-defined or existing R functions.

§2 explains the general model that **flexsurv** is based on. §3 gives examples of its use for fitting built-in survival distributions with a fixed number of parameters, and §4 explains how users can define new distributions. §5 concentrates on classes of models where the number of parameters can be chosen by the user, such as splines. In §6 a simple use of **flexsurv** for parametric multi-state modelling is described. Finally §7 suggests some potential future extensions.

2. General parametric survival model

2.1. Definitions

The general model that **flexsurv** fits has probability density for death at time t:

$$f(t|\mu(\mathbf{z}), \boldsymbol{\alpha}(\mathbf{z})), \quad t \geq 0 \tag{1}$$

The cumulative distribution function $F(t)$, survivor function $S(t) = 1 - F(t)$, cumulative hazard $H(t) = -\log S(t)$ and hazard $h(t) = f(t)/S(t)$ are also defined (suppressing the conditioning for clarity). $\mu = \alpha_0$ is the parameter of primary interest, which usually governs the mean or *location* of the distribution. Other parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_R)$ are called "ancillary" and determine the shape, variance or higher moments.

Covariates All parameters may depend on a vector of covariates $\mathbf{z}$ through link-transformed linear models $g_0(\mu) = \gamma_0 + \boldsymbol{\beta}_0'\mathbf{z}$ and $g_r(\alpha_r) = \gamma_r + \boldsymbol{\beta}_r'\mathbf{z}$. $g()$ will typically be log() if the parameter is defined to be positive, or the identity function if the parameter is unrestricted.

Suppose that the location parameter, but not the ancillary parameters, depends on covariates. If the hazard function factorises as $h(t|\boldsymbol{\alpha}, \mu(\mathbf{z})) = \mu(\mathbf{z})h_0(t|\boldsymbol{\alpha})$, then this is a *proportional hazards* (PH) model, so that the hazard ratio between two groups (defined by two different values of $\mathbf{z}$) is constant over time t.

Alternatively, if $S(t|\mu(\mathbf{z}), \boldsymbol{\alpha}) = S(\mu(\mathbf{z})t|\boldsymbol{\alpha})$ then it is an *accelerated failure time* (AFT) model, so that the effect of covariates is to speed or slow the passage of time. For example, doubling the value of a covariate with coefficient $\beta = \log(2)$ would give half the expected survival time.

Data and likelihood Let $t_i : i = 1, \ldots, n$ be a sample of times from individuals i. Let $c_i = 1$ if t_i is an observed death time, or $c_i = 0$ if this is censored. Most commonly, t_i may be right-censored, thus the true death time is known only to be greater than t_i. More generally, the survival time may be interval-censored on (t_i, t_i^{max}). If there is right-censoring but no left-censoring then t_i^{max} is infinite, so that $S(t_i^{max}) = 0$, or if there is left-censoring but no right-censoring then $t_i = 0$.

Also let s_i be corresponding left-truncation (or delayed-entry) times, meaning that individual i is only observed conditionally on having survived up to s_i, thus $s_i = 0$ if there is no left-truncation. Time-dependent covariates (§3.1) and multi-state models (§6) can be represented through left-truncation.

The likelihood for the parameters $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ in model (1), given the corresponding data vectors, is

$$l(\{\boldsymbol{\theta}\}|\mathbf{t}, \mathbf{c}, \mathbf{s}, \mathbf{t}^{max}) = \left\{ \prod_{i:\ c_i=1} f_i(t_i) \prod_{i:\ c_i=0} (S_i(t_i) - S_i(t_i^{max})) \right\} / \prod_i S_i(s_i) \tag{2}$$

The individuals are independent, so that **flexsurv** does not currently support shared frailty, clustered or random effects models (see §7).

3. Fitting standard parametric survival models

An example dataset used throughout this paper is from 686 patients with primary node positive breast cancer, available in the package as `bc`. This was originally provided with **stpm** (Royston 2001), and analysed in much more detail by Sauerbrei and Royston (1999) and Royston and Parmar (2002). The first two records are:

```
> library(flexsurv)
> bc[1:2,]
```

```
  censrec rectime group   recyrs
1       0    1342  Good 3.676712
2       0    1578  Good 4.323288
```

The main model-fitting function is called `flexsurvreg`. Its first argument is an R *formula* object. The left hand side of the formula gives the response as a survival object, using the `Surv` function from the **survival** package.

```
> fs1 <- flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, dist="weibull")
```

Here, this indicates that the response variable is `recyrs`. This represents time of death or cancer recurrence when `censrec` is 1, or (right-)censoring when `censrec` is 0. The covariate

`group` is a factor representing a prognostic score, with three levels `"Good"` (the baseline), `"Medium"` and `"Poor"`. All of these variables are in the data frame `bc`. If the argument `dist` is a string, this denotes a built-in survival distribution. In this case we fit a Weibull survival model.

Printing the fitted model object gives estimates and confidence intervals for the model parameters and other useful information. Note that these are the *same parameters* as represented by the R distribution function `dweibull`: the `shape` α and the `scale` μ of the survivor function $S(t) = \exp(-(t/\mu)^\alpha)$, and `group` has a linear effect on $\log(\mu)$.

```
> fs1

Call:
flexsurvreg(formula = Surv(recyrs, censrec) ~ group, data = bc,     dist = "weibull")

Estimates:
             data mean  est      L95%     U95%     se       exp(est)  L95%
shape             NA     1.3797   1.2548   1.5170   0.0668       NA        NA
scale             NA    11.4229   9.1818  14.2110   1.2728       NA        NA
groupMedium   0.3338    -0.6136  -0.8623  -0.3649   0.1269   0.5414    0.4222
groupPoor     0.3324    -1.2122  -1.4583  -0.9661   0.1256   0.2975    0.2326
             U95%
shape             NA
scale             NA
groupMedium   0.6943
groupPoor     0.3806

N = 686,  Events: 299,  Censored: 387
Total time at risk: 2113.425
Log-likelihood = -811.9419, df = 4
AIC = 1631.884
```

The same model can be fitted using `survreg` in **survival**:

```
> survreg(Surv(recyrs, censrec) ~ group, data=bc, dist="weibull")

Call:
survreg(formula = Surv(recyrs, censrec) ~ group, data = bc, dist = "weibull")

Coefficients:
(Intercept) groupMedium   groupPoor
  2.4356168  -0.6135892  -1.2122137

Scale= 0.7248206

Loglik(model)= -811.9   Loglik(intercept only)= -873.2
        Chisq= 122.53 on 2 degrees of freedom, p= 0
n= 686
```

The maximised log-likelihoods are the same, however the parameterisation is different: the first coefficient `(Intercept)` reported by `survreg` is $\log(\mu)$, and `survreg`'s `"scale"` is `dweibull`'s (thus `flexsurvreg`)'s 1 / `shape`. The covariate effects $\boldsymbol{\beta}$, however, have the same "accelerated failure time" interpretation, as linear effects on $\log(\mu)$. The multiplicative effects $\exp(\boldsymbol{\beta})$ are printed in the output as `exp(est)`.

3.1. Additional modelling features

If we also had left-truncation times in a variable called `start`, the response would be `Surv(start,recyrs,censrec)`. Or if all responses were interval-censored between lower and upper bounds `tmin` and `tmax`, then we would write `Surv(tmin,tmax,type="interval2")`.

Note that just as in `survreg`, time-dependent covariates can be represented in "counting process" form — as a series of left-truncated survival times. For each individual there would be multiple records, each corresponding to an interval where the covariate is assumed to be constant. The response would be of the form `Surv(start,stop,censrec)`, where `start` and `stop` are the limits of each interval, and `censrec` indicates whether a death was observed at `stop`. The function `survSplit` in **survival** is a utility to construct data like this.

Relative survival models (Nelson, Lambert, Squire, and Jones 2007) can be implemented by supplying the variable in the data that represents the expected mortality rate in the `bhazard` argument to `flexsurvreg`. Case weights and data subsets can also be specified, as in standard R modelling functions, using `weights` or `subset` arguments.

3.2. Built-in models

`flexsurvreg`'s currently built-in distributions are listed in Table 1. In each case, the probability density $f()$ and parameters of the fitted model are taken from an existing R function of the same name but beginning with the letter `d`. For the Weibull, exponential (`dexp`), gamma (`dgamma`) and log-normal (`dlnorm`), the density functions are provided with standard installations of R. These density functions, and the corresponding cumulative distribution functions (with first letter `p` instead of `d`) are used internally in `flexsurvreg` to compute the likelihood.

flexsurv provides some additional survival distributions, including a Gompertz distribution with unrestricted shape parameter (`dist="gompertz"`), and the three- and four-parameter families described below. For all built-in distributions, **flexsurv** also defines functions beginning with `h` giving the hazard, and `H` for the cumulative hazard.

Generalized gamma This three-parameter distribution includes the Weibull, gamma and log-normal as special cases. The original parameterisation from Stacy (1962) is available as `dist="gengamma.orig"`, however the newer parameterisation (Prentice 1974) is preferred: `dist="gengamma"`. This has parameters (μ,σ,q), and survivor function

$$\begin{array}{ll} 1 - I(\gamma, u) & (q > 0) \\ 1 - \Phi(z) & (q = 0) \end{array}$$

where $I(\gamma, u) = \int_0^u x^{\gamma-1} \exp(-x)/\Gamma(\gamma)$ is the incomplete gamma function (the cumulative gamma distribution with shape a and scale 1), Φ is the standard normal cumulative distribution, $u = \gamma \exp(|q|z)$, $z = (\log(t) - \mu)/\sigma$, and $\gamma = q^{-2}$. The Prentice (1974) parameterisation extends the original one to include a further class of models with negative q, and survivor

function $I(\gamma, u)$, where z is replaced by $-z$. This stabilises estimation when the distribution is close to log-normal, since $q = 0$ is no longer near the boundary of the parameter space. In R notation, [1] the parameter values corresponding to the three special cases are

```
dgengamma(x, mu, sigma, Q=0)     ==  dlnorm(x, mu, sigma)
dgengamma(x, mu, sigma, Q=1)     ==  dweibull(x, shape=1/sigma, scale=exp(mu))
dgengamma(x, mu, sigma, Q=sigma) ==  dgamma(x, shape=1/sigma^2,
                                            rate=exp(-mu) / sigma^2)
```

Generalized F This four-parameter distribution includes the generalized gamma, and also the log-logistic, as special cases. The variety of hazard shapes that can be represented is discussed by Cox (2008). It is provided here in alternative "original" (`dist="genf.orig"`) and "stable" parameterisations (`dist="genf"`) as presented by Prentice (1975). See `help(GenF)` and `help(GenF.orig)` in the package documentation for the exact definitions.

3.3. Covariates on ancillary parameters

The generalized gamma model is fitted to the breast cancer survival data. `fs2` is an AFT model, where only the parameter μ depends on the prognostic covariate `group`. In a second model `fs3`, the first ancillary parameter `sigma` (α_1) also depends on this covariate, giving a model with a time-dependent effect that is neither PH nor AFT. The second ancillary parameter `Q` is still common between prognostic groups.

```
> fs2 <- flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, dist="gengamma")
> fs3 <- flexsurvreg(Surv(recyrs, censrec) ~ group + sigma(group),
+                    data=bc, dist="gengamma")
```

Ancillary covariates can alternatively be supplied using the `anc` argument to `flexsurvreg`. This syntax is required if any parameter names clash with the names of functions used in model formulae (e.g. `factor()` or `I()`).

```
fs3 <- flexsurvreg(Surv(recyrs, censrec) ~ group, anc=list(sigma= ~ group),
                   data=bc, dist="gengamma")
```

Table 3 compares all the models fitted to the breast cancer data, showing absolute fit to the data as measured by the maximised -2×log likelihood $-2LL$, number of parameters p, and Akaike's information criterion $-2LL + 2p$ (AIC). The model `fs2` has the lowest AIC, indicating the best estimated predictive ability.

3.4. Plotting outputs

The `plot()` method for `flexsurvreg` objects is used as a quick check of model fit. By default, this draws a Kaplan-Meier estimate of the survivor function $S(t)$, one for each combination of categorical covariates, or just a single "population average" curve if there are no categorical covariates (Figure 1). The corresponding estimates from the fitted model are overlaid. Fitted values from further models can be added with the `lines()` method.

[1] The parameter called q here and in previous literature is called Q in `dgengamma` and related functions, since the first argument of a cumulative distribution function is conventionally named `q`, for quantile, in R.

	Parameters (location in red)	Density R function	dist
Exponential	rate	dexp	"exp"
Weibull	shape, scale	dweibull	"weibull"
Gamma	shape, rate	dgamma	"gamma"
Log-normal	meanlog, sdlog	dlnorm	"lnorm"
Gompertz	shape, rate	dgompertz	"gompertz"
Generalized gamma (Prentice 1975)	mu, sigma, Q	dgengamma	"gengamma"
Generalized gamma (Stacy 1962)	shape, scale, k	dgengamma.orig	"gengamma.orig"
Generalized F (stable)	mu, sigma, Q, P	dgenf	"genf"
Generalized F (original)	mu, sigma, s1, s2	dgenf.orig	"genf.orig"

Table 1: Built-in parametric survival distributions in **flexsurv**.

```
> plot(fs1, col="gray", lwd.obs=2, xlab="Years", ylab="Recurrence-free survival")
> lines(fs2, col="red", lty=2)
> lines(fs3, col="red")
> legend("bottomleft", col=c("black","gray","red","red"),
+        lty=c(1,1,2,1), bty="n", lwd=rep(2,4),
+        c("Kaplan-Meier","Weibull","Generalized gamma (AFT)",
+          "Generalized gamma (time-varying)"))
```

`scale="hazard"` can be used to plot hazards from parametric models against kernel density estimates obtained from **muhaz** (Hess 2010; Mueller and Wang 1994). Figure 2 shows more clearly that the Weibull model is inadequate for the breast cancer data: the hazard must be increasing or decreasing — while the generalized gamma can represent the increase and subsequent decline in hazard seen in the data. Similarly, `scale="cumhaz"` plots cumulative hazards.

```
> plot(fs1, type="hazard", col="gray", lwd.obs=2, xlab="Years", ylab="Hazard")
> lines(fs2, type="hazard", col="red", lty=2)
> lines(fs3, type="hazard", col="red")
> legend("topright", col=c("black","gray","red","red"),
+        lty=c(1,1,2,1),  bty="n", lwd=rep(2,4),
+        c("Kernel density estimate","Weibull","Gen. gamma (AFT)",
+          "Gen. gamma (time-varying)"))
```

The numbers plotted are available from the `summary.flexsurvreg()` method. Confidence intervals are produced by simulating a large sample from the asymptotic normal distribution of the maximum likelihood estimates of $\{\boldsymbol{\beta}_r : r = 0, \ldots, R\}$ (Mandel 2013), via the function `normboot.flexsurvreg`.

In this example, there is only a single categorical covariate, and the `plot` and `summary` methods return one observed and fitted trajectory for each level of that covariate. For more complicated models, users should specify what covariate values they want summaries for, rather than relying on the default [2]. This is done by supplying the `newdata` argument, a data frame or list containing covariate values, just as in standard R functions like `predict.lm`.

[2] If there are only factor covariates, all combinations are plotted. If there are any continuous covariates,

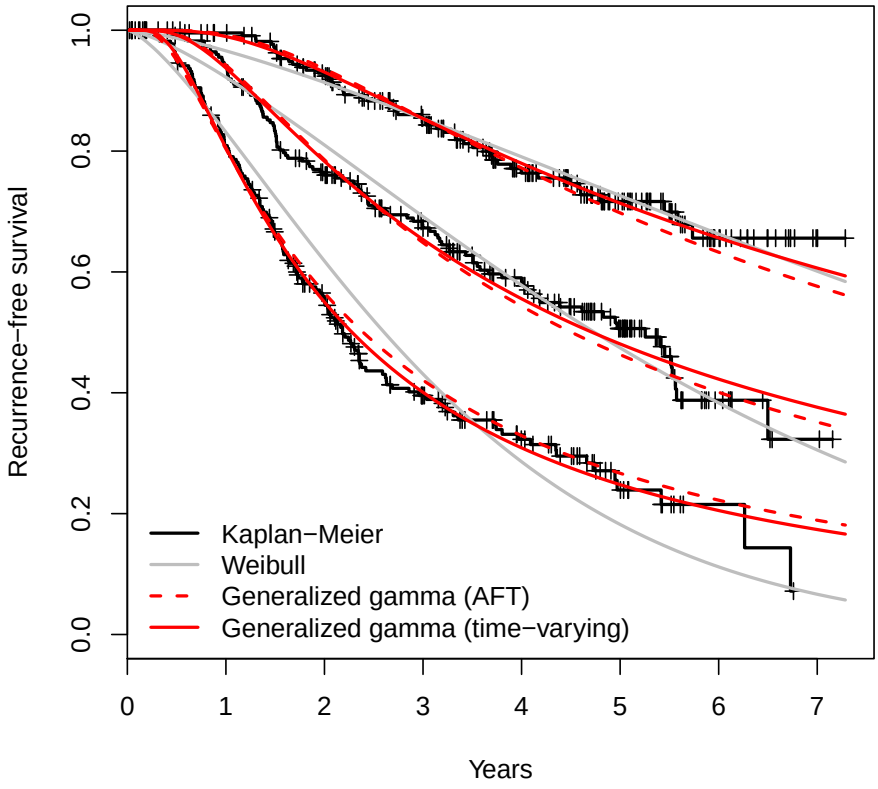

Figure 1: Estimated survival from parametric models and Kaplan-Meier estimates.

This `plot()` method is only for casual exploratory use. For publication-standard figures, it is preferable to set up the axes beforehand (`plot(...,type="n")`), and use the `lines()` methods, or construct plots by hand using the data available from `summary.flexsurvreg()`.

3.5. Custom model summaries

Any function of the parameters of a fitted model can be summarised or plotted by supplying the argument `fn` to `summary.flexsurvreg` or `plot.flexsurvreg`. This should be an R function, with optional first two arguments `t` representing time, and `start` representing a left-truncation point (if the result is conditional on survival up to that time). Any remaining arguments must be the parameters of the survival distribution. For example, median survival under the Weibull model `fs1` can be summarised as follows

```
> median.weibull <- function(shape, scale) {
+     qweibull(0.5, shape=shape, scale=scale)
```

these methods by default return a "population average" curve, with the linear model design matrix set to its average values, including the 0/1 contrasts defining factors, which doesn't represent any specific covariate combination.

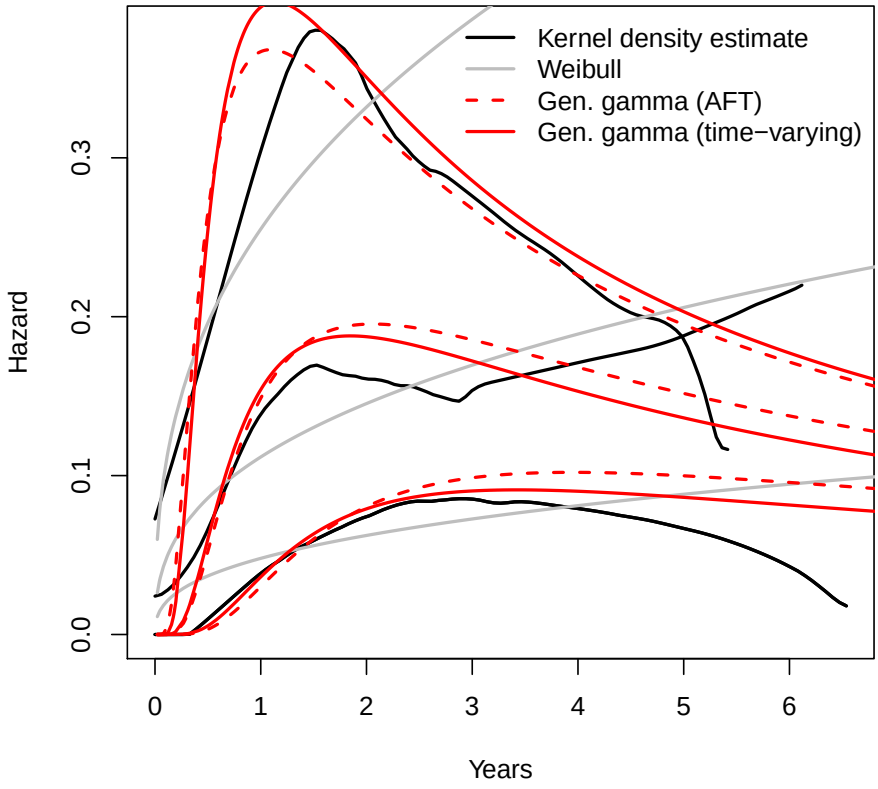

Figure 2: Estimated hazards from parametric models and kernel density estimates.

```
+ }
> summary(fs1, fn=median.weibull, t=1, B=10000)

group=Good
  time     est     lcl      ucl
1    1 8.75794 7.12019 10.86043

group=Medium
  time      est      lcl      ucl
1    1 4.741585 4.120987 5.469989

group=Poor
  time      est      lcl      ucl
1    1 2.605819 2.311429 2.935884
```

Although the median of the Weibull has an analytic form as $\mu \log(2)^{1/\alpha}$, the form of the code given here generalises to other distributions. The argument `t` (or `start`) can be omitted from

`median.weibull`, because the median is a time-constant function of the parameters, unlike the survival or hazard.

`10000` random samples are drawn to produce a slightly more precise confidence interval than the default — users should adjust this until the desired level of precision is obtained. A useful future extension of the package would be to allow users to supply derivatives of their custom summary function if possible, so that the delta method can be used to obtain approximate confidence intervals without simulation.

3.6. Computation

The likelihood is maximised in `flexsurvreg` using the optimisation methods available through the standard R `optim` function. By default, this is the `"BFGS"` method (Nash 1990) using the analytic derivatives of the likelihood with respect to the model parameters, if these are available, to improve the speed of convergence to the maximum. These derivatives are built-in for the exponential, Weibull, Gompertz, and hazard- and odds-based spline models (see §5.1). For custom distributions (see §4), the user can optionally supply functions with names beginning `"DLd"` and `"DLS"` respectively (e.g. `DLdweibull,DLSweibull`) to calculate the derivatives of the log density and log survivor functions with respect to the transformed baseline parameters γ.

4. Custom survival distributions

flexsurv is not limited to its built-in distributions. Any survival model of the form (1–2) can be fitted if we can provide either the density function $f()$ or the hazard $h()$. Many contributed R packages provide probability density and cumulative distribution functions for positive distributions. Though survival models may be more naturally characterised by their hazard function, representing the changing risk of death through time. For example, for survival following major surgery we may want a "U-shaped" hazard curve, representing a high risk soon after the operation, which then decreases, but increases naturally as survivors grow older.

To supply a custom distribution, the `dist` argument to `flexsurvreg` is defined to be an R list object, rather than a character string. The list has the following elements.

`name` Name of the distribution. In the first example below, we use a logistic distribution, and the name is `"llogis"`. Then there is assumed to be at least either

- a function to compute the probability density, which would be called called `dllogis` here, or
- a function to compute the hazard, called `hllogis`.

Ideally there will also be a function called `pllogis` for the cumulative distribution (if `d` is given), or `H` for the cumulative hazard (to complement `h`). If there is no analytic form for $F(t)$ or $H(t)$, then **flexsurv** can compute these internally by numerical integration, as in **stgenreg** (Crowther and Lambert 2013). The default options of the built-in R routine `integrate` for adaptive quadrature are used, though these may be changed using the `integ.opts` argument to `flexsurvreg`. Models specified this way will take an order of

magnitude more time to fit, and the fitting procedure may be unstable. An example is given in §5.2.

These functions must be *vectorised*, and the density function must also accept an argument `log`, which when `TRUE`, returns the log density. See the examples below.

In some cases, R's scoping rules may not find the functions in the working environment. They may then be supplied through the `dfns` argument to `flexsurvreg`.

`pars` Character vector naming the parameters of the distribution $\mu, \alpha_1, ..., \alpha_R$. These must match the arguments of the R distribution function or functions, in the same order.

`location` Character: quoted name of the location parameter μ. The location parameter will not necessarily be the first one, e.g. in `dweibull` the `scale` comes after the `shape`.

`transforms` A list of functions $g()$ which transform the parameters from their natural ranges to the real line, for example, `c(log,identity)` if the first is positive and the second unrestricted. [3]

`inv.transforms` List of corresponding inverse functions.

`inits` A function which provides plausible initial values of the parameters for maximum likelihood estimation. This is optional, but if not provided, then each call to `flexsurvreg` must have an `inits` argument containing a vector of initial values, which is inconvenient. Implausible initial values may produce a likelihood of zero, and a fatal error message (`initial value in 'vmmin' is not finite`) from the optimiser.

Each distribution will ideally have a heuristic for initialising parameters from summaries of the data. For example, since the median of the Weibull is $\mu \log(2)^{1/\alpha}$, a sensible estimate of μ will commonly be the median log uncensored survival time divided by $\log(2)$, with $\alpha = 1$, assuming that in practice the true value of α is not often far from 1. Then we define the function, of one argument `t` assumed to be the uncensored survival times, returning the initial values for the Weibull `shape` and `scale` respectively.

```
inits = function(t) c(1, median(t[t>0]) / log(2))
```

More complicated initial value functions may use other data such as the covariate values and censored observations: for an example, see the function `flexsurv.splineinits` in the package source that computes initial values for spline models (§5.1).

Example: Using functions from a contributed package The following custom model uses the log-logistic distribution functions (`dllogis` and `pllogis`) available in the package **eha**. The survivor function is $S(t|\mu, \alpha) = 1/(1 + (t/\mu)^\alpha)$, so that the odds $(1 - S(t))/S(t)$ of having died are a linear function of log time.

```
> library(eha)
> custom.llogis <- list(name="llogis",  pars=c("shape","scale"), location="scale",
+                       transforms=c(log, log), inv.transforms=c(exp, exp),
+                       inits=function(t){ c(1, median(t)) })
> fs4 <- flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, dist=custom.llogis)
```

[3] This is a *list*, not an *atomic vector* of functions, so if the distribution only has one parameter, we should write `transforms=c(log)` or `transforms=list(log)`, not `transforms=log`.

This fits the breast cancer data better than the Weibull, since it can represent a peaked hazard, but less well than the generalized gamma (Table 3).

Example: Wrapping functions from a contributed package Sometimes there may be probability density and similar functions in a contributed package, but in a different format. For example, **eha** also provides a three-parameter Gompertz-Makeham distribution with hazard $h(t|\mu, \alpha_1, \alpha_2) = \alpha_2 + \alpha_1 \exp(t/\mu)$. The shape parameters α_1, α_2 are provided to `dmakeham` as a vector argument of length two. However, `flexsurvreg` expects distribution functions to have one argument for each parameter. Therefore we write our own functions that wrap around the third-party functions.

```
> dmakeham3 <- function(x, shape1, shape2, scale, ...)  {
+     dmakeham(x, shape=c(shape1, shape2), scale=scale, ...)
+ }
> pmakeham3 <- function(q, shape1, shape2, scale, ...)  {
+     pmakeham(q, shape=c(shape1, shape2), scale=scale, ...)
+ }
```

`flexsurvreg` also requires these functions to be *vectorized*, as the standard distribution functions in R are. That is, we can supply a vector of alternative values for one or more arguments, and expect a vector of the same length to be returned. The R base function `Vectorize` can be used to do this here.

```
> dmakeham3 <- Vectorize(dmakeham3)
> pmakeham3 <- Vectorize(pmakeham3)
```

and this allows us to write, for example,

```
> pmakeham3(c(0, 1, 1, Inf), 1, c(1, 1, 2, 1), 1)

[1] 0.0000000 0.9340120 0.9757244 1.0000000
```

We could then use `dist=list(name="makeham3", pars=c("shape1","shape2","scale"),...)` in a `flexsurvreg` model, though in the breast cancer example, the second shape parameter is poorly identifiable.

Example: Changing the parameterisation of a distribution We may want to fit a Weibull model like `fs1`, but with the proportional hazards parameterisation $S(t) = \exp(-\mu t^{\alpha})$, so that the covariate effects reported in the printed `flexsurvreg` object can be interpreted as hazard ratios or log hazard ratios without any further transformation. Here instead of the density and cumulative distribution functions, we provide the hazard and cumulative hazard.[4]

```
> detach("package:eha")
> hweibullPH <- function(x, shape, scale = 1, log=FALSE){
```

[4]The **eha** package needs to be detached first so that **flexsurv**'s built-in `hweibull` is used, which returns `NaN` if the parameter values are zero, rather than failing as **eha**'s does.

```
+     hweibull(x, shape=shape, scale=scale^{-1/shape}, log=log)
+ }
> HweibullPH <- function(x, shape, scale=1, log=FALSE){
+     Hweibull(x, shape=shape, scale=scale^{-1/shape}, log=log)
+ }
> custom.weibullPH <- list(name="weibullPH",
+                          pars=c("shape","scale"), location="scale",
+                          transforms=c(log, log), inv.transforms=c(exp, exp),
+                          inits = function(t){
+                              c(1, median(t[t>0]) / log(2))
+                          })
> fs6 <- flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, dist=custom.weibullPH)
> fs6$res["scale","est"] ^ {-1/fs6$res["shape","est"]}

[1] 11.42286

> - fs6$res["groupMedium","est"] / fs6$res["shape","est"]

[1] -0.6135897
```

The fitted model is the same as `fs1`, therefore the maximised likelihood is the same, and the parameter estimates of `fs6` can be transformed to those of `fs1` as shown.

A slightly more complicated example is given in the package vignette `flexsurv-examples` of constructing a proportional hazards generalized gamma model.

5. Arbitrary-dimension models

flexsurv also supports models where the number of parameters is arbitrary. In the models discussed previously, the number of parameters in the model family is fixed (e.g. three for the generalized gamma). In this section, the model complexity can be chosen by the user, given the model family. We may want to represent more irregular hazard curves by more flexible functions, or use bigger models if a bigger sample size makes it feasible to estimate more parameters.

5.1. Royston and Parmar spline model

In the spline-based survival model of Royston and Parmar (2002), a transformation $g(S(t, z))$ of the survival function is modelled as a natural cubic spline function of log time: $g(S(t, z)) = s(x, \boldsymbol{\gamma})$ where $x = \log(t)$. This model can be fitted in **flexsurv** using the function `flexsurvspline`, and is also available in the Stata package **stpm2** (Lambert and Royston 2009) (historically **stpm**, Royston (2001, 2004)).

Typically we use $g(S(t, \mathbf{z})) = \log(-\log(S(t, \mathbf{z}))) = \log(H(t, \mathbf{z}))$, the log cumulative hazard, giving a proportional hazards model.

Spline parameterisation The complexity of the model, thus the dimension of $\boldsymbol{\gamma}$, is governed by the number of *knots* in the spline function $s()$. Natural cubic splines are piecewise

Model	$g(S(t,\mathbf{z}))$	In `flexsurvspline`	With $m=0$
Proportional hazards	$\log(-\log(S(t,\mathbf{z})))$ (log cumulative hazard)	`scale="hazard"`	Weibull `shape` γ_1, `scale` $\exp(-\gamma_0/\gamma_1)$
Proportional odds	$\log(S(t,\mathbf{z})^{-1}-1)$ (log cumulative odds)	`scale="odds"`	Log-logistic `shape` γ_1, `scale` $\exp(-\gamma_0/\gamma_1)$
Normal / probit	$\Phi^{-1}(S(t,\mathbf{z}))$ (inverse normal CDF, `qnorm`)	`scale="normal"`	Log-normal `meanlog` $-\gamma_0/\gamma_1$, `sdlog` $1/\gamma_1$

Table 2: Alternative modelling scales for `flexsurvspline`, and equivalent distributions for $m=0$ (with parameter definitions as in the R `d` functions referred to elsewhere in the paper)

cubic polynomials defined to be continuous, with continuous first and second derivatives at the knots, and also constrained to be linear beyond boundary knots k_{min}, k_{max}. As well as the boundary knots there may be up to $m \geq 0$ *internal* knots $k_1, \ldots, k_m$. Various spline parameterisations exist — the one used here is from Royston and Parmar (2002).

$$s(x,\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \ldots + \gamma_{m+1} v_m(x) \tag{3}$$

where $v_j(x)$ is the jth *basis* function

$$v_j(x) = (x-k_j)_+^3 - \lambda_j (x-k_{min})_+^3 - (1-\lambda_j)(x-k_{max})_+^3, \qquad \lambda_j = \frac{k_{max}-k_j}{k_{max}-k_{min}}$$

and $(x-a)_+ = max(0, x-a)$. If $m=0$ then there are only two parameters γ_0, γ_1. In fact if $g()$ is the log cumulative hazard, this is equivalent to a Weibull model. Table 2 explains two further choices of $g()$, and the parameter values and distributions they simplify to for $m=0$. The probability density and cumulative distribution functions for all these models are available as `dsurvspline` and `psurvspline`.

Covariates on spline parameters Covariates can be placed on any parameter γ through a linear model (with identity link function). Most straightforwardly, we can let the intercept γ_0 vary with covariates $\mathbf{z}$, giving a proportional hazards or odds model (depending on $g()$).

$$g(S(t,z)) = s(x,\boldsymbol{\gamma}) + \boldsymbol{\beta}^T \mathbf{z}$$

The spline coefficients $\gamma_j : j = 1, 2 \ldots$, the "ancillary" parameters, may also be modelled as linear functions of covariates $\mathbf{z}$, as

$$\gamma_j(\mathbf{z}) = \gamma_{j0} + \gamma_{j1} z_1 + \gamma_{j2} z_2 + \ldots$$

giving a model where the effects of covariates are arbitrarily flexible functions of time: a non-proportional hazards or odds model.

Spline models in flexsurv The argument `k` to `flexsurvspline` defines the number of internal knots m. Knot locations are chosen by default from quantiles of the log uncensored

death times, or users can supply their own locations in the `knots` argument. Initial values for numerical likelihood maximisation are chosen using the method described by Royston and Parmar (2002) of Cox regression combined with transforming an empirical survival estimate.

For example, the best-fitting model for the breast cancer dataset identified in Royston and Parmar (2002), a proportional odds model with one internal spline knot, is

```
> sp1 <- flexsurvspline(Surv(recyrs, censrec) ~ group, data=bc, k=1,
+                       scale="odds")
```

A further model where the first ancillary parameter also depends on the prognostic group, giving a time-varying odds ratio, is fitted as

```
> sp2 <- flexsurvspline(Surv(recyrs, censrec) ~ group + gamma1(group),
+                       data=bc, k=1, scale="odds")
```

These models give qualitatively similar results to the generalized gamma in this dataset (Figure 3), and have similar predictive ability as measured by AIC (Table 3). Though in general, an advantage of spline models is that extra flexibility is available where necessary.

```
> plot(sp1, type="hazard", ylim=c(0, 0.5), xlab="Years", ylab="Hazard")
> lines(sp2, type="hazard", col="red", lty=2)
> lines(fs2, type="hazard", col="blue")
> legend("topright", c("black","red","red","blue"), lty=c(1,1,2,1), lwd=rep(2,4),
+        c("Kernel density estimate","Spline (proportional odds)",
+          "Spline (time-varying)","Generalized gamma (time-varying)"))
```

Note that the log hazard ratios under the proportional hazards spline model with one internal knot are practically the same as under a standard Cox model.

```
> sp3 <- flexsurvspline(Surv(recyrs, censrec) ~ group, data=bc, k=1, scale="hazard")
> sp3$res[c("groupMedium","groupPoor"),c("est","se")]
```

```
                 est        se
groupMedium 0.8334517 0.1712042
groupPoor   1.6111788 0.1640933
```

```
> cox3 <- coxph(Surv(recyrs, censrec) ~ group, data=bc)
> coef(summary(cox3))[,c("coef","se(coef)")]
```

```
                coef  se(coef)
groupMedium 0.8401002 0.1713926
groupPoor   1.6180720 0.1645443
```

```
> res <- t(sapply(list(fs1, fs2, fs3, fs4, sp1, sp2),
+                 function(x)rbind(-2*x$loglik, x$npars, x$AIC)))
> rownames(res) <- c("Weibull (fs1)","Generalized gamma (fs2)",
+                    "Generalized gamma (fs3)","Log-logistic (fs4)",
+                    "Spline (sp1)", "Spline (sp2)")
> colnames(res) <- c("-2 log likelihood","Parameters","AIC")
```

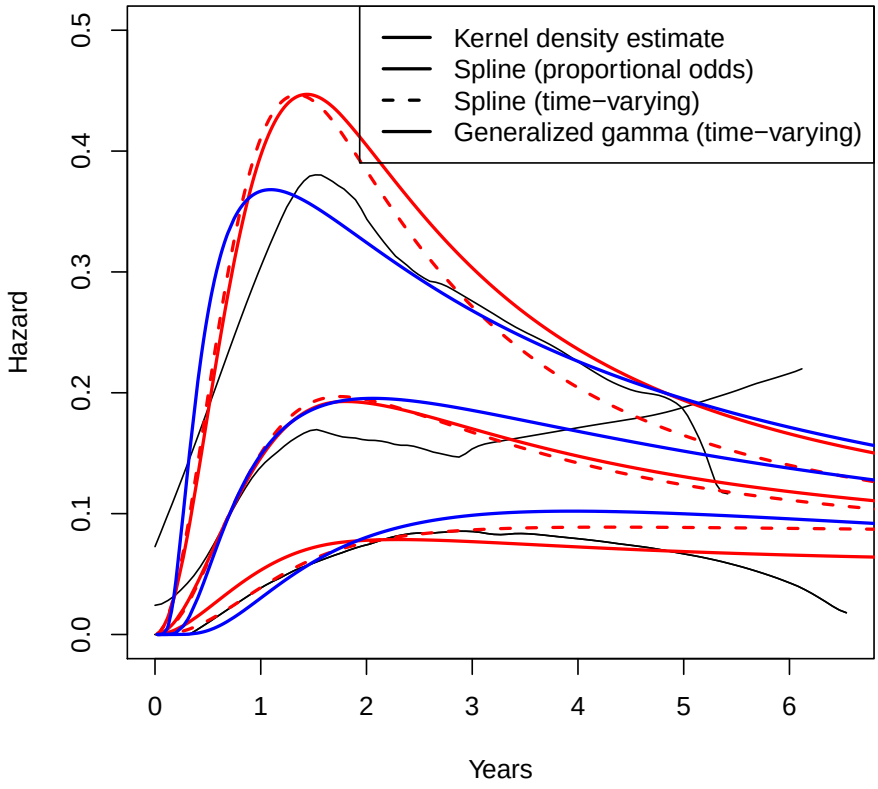

Figure 3: Comparison of spline and generalized gamma fitted hazards.

5.2. Implementing new general-dimension models

The spline model above is an example of the general parametric form (1), but the number of parameters, $R+1$ in (1), $m+2$ in (3), is arbitrary. **flexsurv** has the tools to deal with any model of this form. `flexsurvspline` works internally by building a custom distribution and then calling `flexsurvreg`. Similar models may in principle be built by users using the same method. This relies on a functional programming trick.

Creating distribution functions dynamically The R distribution functions supplied to custom models are expected to have a fixed number of arguments, including one for each scalar parameter. However, the distribution functions for the spline model (e.g. `dsurvspline`) have an argument `gamma` representing the *vector* of parameters γ, whose length is determined by choosing the number of knots. Just as the *scalar parameters* of conventional distribution functions can be supplied as *vector arguments* (as explained in §4), similarly, the vector parameters of spline-like distribution functions can be supplied as *matrix arguments*, representing alternative parameter values.

To convert a spline-like distribution function into the correct form, **flexsurv** provides the utility

```
> res
```

	-2 log likelihood	Parameters	AIC
Weibull (fs1)	1623.884	4	1631.884
Generalized gamma (fs2)	1575.137	5	1585.137
Generalized gamma (fs3)	1572.434	7	1586.434
Log-logistic (fs4)	1598.105	4	1606.105
Spline (sp1)	1577.964	5	1587.964
Spline (sp2)	1574.848	7	1588.848

Table 3: Comparison of parametric survival models fitted to the breast cancer data

`unroll.function`. This converts a function with one (or more) vector parameters (matrix arguments) to a function with an arbitrary number of scalar parameters (vector arguments). For example, the 5-year survival probability for the baseline group under the model `sp1` is

```
> gamma <- sp1$res[c("gamma0","gamma1","gamma2"),"est"]
> 1 - psurvspline(5, gamma=gamma, knots=sp1$knots)

[1] 0.6896969
```

An alternative function to compute this can be built by `unroll.function`. We tell it that the vector parameter `gamma` should be provided instead as three scalar parameters named `gamma0`, `gamma1`, `gamma2`. The resulting function `pfn` is in the correct form for a custom `flexsurvreg` distribution.

```
> pfn <- unroll.function(psurvspline, gamma=0:2)
> 1 - pfn(5, gamma0=gamma[1], gamma1=gamma[2], gamma2=gamma[3], knots=sp1$knots)

[1] 0.6896969
```

Users wishing to fit a new spline-like model with a known number of parameters could just as easily write distribution functions specific to that number of parameters, and use the methods in §4. However the `unroll.function` method is intended to simplify the process of extending the **flexsurv** package to implement new model families, through wrappers similar to `flexsurvspline`.

Example: splines on alternative scales An alternative to the Royston-Parmar spline model is to model the log *hazard* as a function of time instead of the log cumulative hazard. Crowther and Lambert (2013) demonstrate this model using the Stata **stgenreg** package. An advantage explained by Royston and Lambert (2011) is that when there are when there are multiple time-dependent effects, time-dependent hazard ratios can be interpreted independently of the values of other covariates.

This can also be implemented in `flexsurvreg` using `unroll.function`. A disadvantage of this model is that the cumulative hazard (hence the survivor function) has no analytic form, therefore to compute the likelihood, the hazard function needs to be integrated numerically.

This is done automatically in `flexsurvreg` (just as in **stgenreg**) if the cumulative hazard is not supplied.

Firstly, a function must be written to compute the hazard as a function of time `x`, the vector of parameters `gamma` (which can be supplied as a matrix argument so the function can give a vector of results), and a vector of knot locations. This uses **flexsurv**'s function `basis` to compute the natural cubic spline basis (3).

```
> hsurvspline.lh <- function(x, gamma, knots){
+     if(!is.matrix(gamma)) gamma <- matrix(gamma, nrow=1)
+     lg <- nrow(gamma) # return vector of length of longest argument
+     nret <- max(length(x), lg)
+     gamma <- apply(gamma, 2, function(x)rep(x,length=nret))
+     x <- rep(x, length=nret)
+     loghaz <- rowSums(basis(knots, log(x)) * gamma)
+     exp(loghaz)
+ }
```

The equivalent function is then created for a three-knot example of this model (one internal and two boundary knots) that has arguments `gamma0`, `gamma1` and `gamma2` corresponding to the three columns of `gamma`,

```
> hsurvspline.lh3 <- unroll.function(hsurvspline.lh, gamma=0:2)
```

```
> custom.hsurvspline.lh3 <- list(
+     name = "survspline.lh3",
+     pars = c(paste0("gamma",0:2)),
+     location = c("gamma0"),
+     transforms = rep(c(identity), 3), inv.transforms=rep(c(identity), 3)
+     )
```

To complete the model, the internal knot is placed at the median uncensored log survival time, and boundary knots are placed at the minimum and maximum. These are passed to `hsurvspline.lh` through the `aux` argument of `flexsurvreg`.

```
> dtime <- log(bc$recyrs)[bc$censrec==1]
> ak <- list(knots=quantile(dtime, c(0, 0.5, 1)))
```

Initial values must be provided in the call to `flexsurvreg`, since the custom distribution list did not include an `inits` component. For this example, "default" initial values of zero suffice, but the permitted values of γ_2 are fairly tightly constrained (from -0.5 to 0.5 here) using the `"L-BFGS-B"` bounded optimiser from R's `optim` (Nash 1990). Without the constraint, extreme values of γ_2, visited by the optimiser, cause the numerical integration of the hazard function to fail.

```
> sp4 <- flexsurvreg(Surv(recyrs, censrec) ~ group, data=bc, aux=ak,
+                    inits=c(0, 0, 0, 0, 0), dist=custom.hsurvspline.lh3,
+                    method="L-BFGS-B", lower=c(-Inf,-Inf,-0.5), upper=c(Inf,Inf,0.5),
+                    control=list(trace=1,REPORT=1))
```

This takes around ten minutes to converge, so is not presented here, though the fit is poorer than the equivalent spline model for the cumulative hazard. The 95% confidence interval for γ_2 of (0.16, 0.37) is firmly within the constraint.

Other arbitrary-dimension models Another potential application is to fractional polynomials (Royston and Altman 1994). These are of the form $\sum_{m=1}^{M} \alpha_m x^{p_m} log(x)^n$ where the power p_m is in the standard set $\{2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ (except that $log(x)$ is used instead of x^0), and n is a non-negative integer. They are similar to splines in that they can give arbitrarily close approximations to a nonlinear function, such as a hazard curve, and are particularly useful for expressing the effects of continuous predictors in regression models. See e.g. Sauerbrei, Royston, and Binder (2007), and several other publications by the same authors, for applications and discussion of their advantages over splines. The R package **gamlss** (Rigby and Stasinopoulos 2005) has a function to construct a fractional polynomial basis that might be employed in **flexsurv** models.

Polyhazard models (Louzada-Neto 1999) are another potential use of this technique. These express an overall hazard as a sum of latent cause-specific hazards, each one typically from the same class of distribution, e.g. a *poly-Weibull* model if they are all Weibull. For example, a U-shaped hazard curve following surgery may be the sum of early hazards from surgical mortality and later deaths from natural causes. However, such models may not always be identifiable without external information to fix or constrain the parameters of particular hazards (Demiris, Lunn, and Sharples 2011).

6. Multi-state models

A *multi-state model* represents how an individual moves between multiple states through time. Survival analysis is a special case with two states, "alive" and "dead". *Competing risks* are a further special case, where there are multiple causes of death, that is, one starting state and multiple possible destination states.

Given that an individual is in state $S(t)$ at time t, their next state, and the time of the change, are governed by a set of *transition intensities*

$$q_{rs}(t, \mathbf{z}(t), \mathcal{F}_t) = \lim_{\delta t \to 0} P(S(t + \delta t) = s | S(t) = r, t, \mathbf{z}(t), \mathcal{F}_t)/\delta t$$

for states $r, s = 1, \ldots, R$, which for a survival model are equivalent to the hazard $h(t)$. The intensity represents the instantaneous risk of moving from state r to state s, and is zero if the transition is impossible. It may depend on the time t, patient characteristics $\mathbf{z}(t)$, and possibly also the "history" of the process up to that time, $\mathcal{F}_t$: the states previously visited or the length of time spent in them.

Data Instead of a single event time, there may now be a series of event times $t_1, \ldots, t_n$ for an individual. The last of these may be an observed or right-censored event time. Note *panel data* are not considered here — that is, observations of the state of the process at a series of times (Kalbfleisch and Lawless 1985). In panel data, we do not necessarily know the time of each transition, or even whether transitions of a certain type have occurred at all between a pair of observations. Multi-state models for that type of data (and also exact event times)

can be fitted with the **msm** package for R (Jackson 2011), but are restricted to (piecewise) exponential event time distributions.

Alternative time scales In semi-Markov (clock-reset) models, $q_{rs}(t)$ depends on the time t since entry into the current state, but otherwise, the time since the beginning of the process is forgotten. Any software to fit survival models can also fit this kind of multi-state model.

In an inhomogeneous Markov (clock-forward) model, t represents the time since the beginning of the process, but the intensity $q_{rs}(t)$ does not depend further on $\mathcal{F}_t$. Again any survival modelling software can be used, with the additional requirement that it can deal with left-truncation or *counting process* data.

These approaches are equivalent for competing risks models, since there is at most one transition for each individual, so that the time since the beginning of the process equals the time spent in the current state. Therefore no left-truncation is necessary.

Example For illustration, consider a simple three-state example, previously studied by Heng, Sharples, McNeil, Stewart, Wreghitt, and Wallwork (1998). Recipients of lung transplants are are risk of bronchiolitis obliterans syndrome (BOS). This was defined as a decrease in lung function to below 80% of a baseline value defined in the six months following transplant. A three-state "illness-death" model represents the risk of developing BOS, the risk of dying before developing BOS, and the risk of death after BOS. BOS is assumed to be irreversible, so there are only three allowed transitions (Figure 4), each with an intensity function $q_{rs}(t)$.

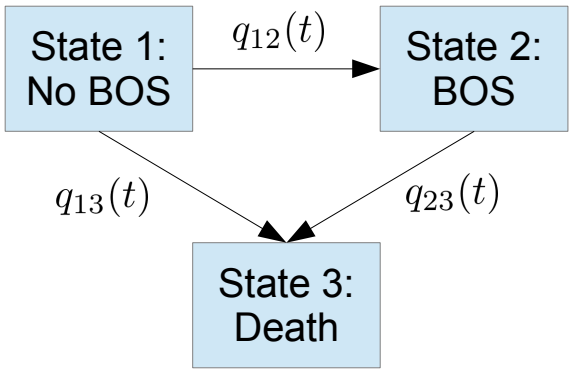

Figure 4: Three-state multi-state model for bronchiolitis obliterans syndrome (BOS).

6.1. Representing multi-state data as survival data

Putter, Fiocco, and Geskus (2007) discuss how to implement multi-state models by manipulating the data into the a suitable form for survival modelling software — an overview is given here. For each permitted $r \to s$ transition in the multi-state model, there is a corresponding "survival" (time-to-event) model, with hazard rates defined by $q_{rs}(t)$. For a patient who enters state r at time t_j, their next event at t_{j+1} is defined by the model structure to be one of a set of competing events $s_1, \ldots, s_{n_r}$. This implies there are n_r corresponding survival models for this state r, and $\sum_r n_r$ models over all states r. In the BOS example, there are $n_1 = 2, n_2 = 1$ and $n_3 = 0$ possible transitions from states 1, 2 and 3 respectively.

The data to inform the n_r models from state r consists firstly of an indicator for whether the transition to the corresponding state $s_1, \ldots, s_{n_r}$ is observed or censored at t_{j+1}. If the individual moves to state s_k, the transitions to all other states in this set are censored at this time. This indicator is coupled with:

- (for a semi-Markov model) the time elapsed $dt_j = t_{j+1} - t_j$ from state r entry to state s entry. The "survival" model for the $r \to s$ transition is fitted to this time.
- (for an inhomogeneous Markov model) the start and stop time (t_j, t_{j+1}), as in §3.1. The $r \to s$ model is fitted to the right-censored time t_{j+1} from the *start of the process*, but is conditional on not experiencing the $r \to s$ transition until after the state r entry time. In other words, the $r \to s$ transition model is *left-truncated* at the state r entry time.

The **mstate** R package (de˜Wreede, Fiocco, and Putter 2010; de˜Wreede *et˜al.* 2011) has a utility `msprep` to produce data of this form from "wide-format" datasets where rows represent individuals, and times of different events appear in different columns. **msm** (Jackson 2011) has a similar utility `msm2Surv` for producing the required form given longitudinal data where rows represent state observations.

The outcomes of two patients in the BOS data are given by

```
> bosms3[18:22,]

An object of class 'msdata'

Data:
   id from to    Tstart     Tstop     years status trans
18  7    1  2 0.0000000 0.1697467 0.1697467      1     1
19  7    1  3 0.0000000 0.1697467 0.1697467      0     2
20  7    2  3 0.1697467 0.6297057 0.4599589      1     3
21  8    1  2 0.0000000 8.1615332 8.1615332      0     1
22  8    1  3 0.0000000 8.1615332 8.1615332      1     2
```

Each row represents an observed (`status=1`) or censored (`status=0`) transition time for one of three time-to-event models indicated by the categorical variable `trans` (defined as a factor). Times are expressed in years, with the baseline time 0 representing six months after transplant. Values of `trans` of 1, 2, 3 correspond to no BOS→BOS, no BOS→death and BOS→death respectively. The first row indicates that the patient (`id` 7) moved from state 1 (no BOS) to state 2 (BOS) at 0.17 years, but (second row) this is also interpreted as a censored time from state 1 to state 3, potential death before BOS onset. This patient then died, given by the third row with `status` 1 for `trans` 3. Patient 8 died before BOS onset, therefore at 8.2 years their potential BOS onset is censored (fourth row), but their death before BOS is observed (fifth row).

6.2. Fitting parametric multi-state models

Three multi-state models are fitted to the BOS data using `flexsurvreg`. The first two use the "clock-reset" time scale. The first is a simple time-homogeneous Markov model where

all transition intensities are constant through time, so that the clock-forward and clock-reset scales are identical. The time to the next event is exponentially-distributed, but with a different rate q_{rs} for each transition type `trans`. The second is a semi-Markov model where the times to BOS onset, death without BOS and the time from BOS onset to death all have Weibull distributions, with a different shape and scale for each transition type. The third is a clock-forward, inhomogeneous Markov version of the Weibull model: the 1→2 and 1→3 transition models are the same, but the third has a different interpretation, now the time *from baseline* to death with BOS has a Weibull distribution.

```
> crexp <- flexsurvreg(Surv(years, status) ~ trans, data=bosms3,
+                      dist="exp")
> crwei <- flexsurvreg(Surv(years, status) ~ trans + shape(trans),
+                      data=bosms3, dist="weibull")
> cfwei <- flexsurvreg(Surv(Tstart, Tstop, status) ~ trans + shape(trans),
+                      data=bosms3, dist="weibull")
```

The equivalent Cox models are also fitted using `coxph` from the **survival** package. These specify a different baseline hazard for each transition type through a function `strata` in the formula, so since there are no other covariates, they are essentially non-parametric. Note that the `strata` function is not currently understood by `flexsurvreg` — the user must say explicitly what parameters, if any, vary with the transition type, as in `crwei`.

```
> crcox <- coxph(Surv(years, status) ~ strata(trans), data=bosms3)
> cfcox <- coxph(Surv(Tstart, Tstop, status) ~ strata(trans), data=bosms3)
```

In all cases, if there were other covariates, they could simply be included in the model formula. Typically, covariate effects will vary with the transition type, so that an interaction term with `trans` would be included. Some post-processing might then be needed to combine the main covariate effects and interaction terms into an easily-interpretable quantity (such as the hazard ratio for the r, s transition). Alternatively, **mstate** has a utility `expand.covs` to expand a single covariate in the data into a set of transition-specific covariates, to aid interpretation (see de˜Wreede *et˜al.* 2011).

Any parametric distribution can be fitted, just as for standard survival models with `flexsurvreg`. Spline models may also be fitted with `flexsurvspline`, and if hazards are assumed proportional, they are expected to give similar results to the Cox model. A restriction is that all transition-specific models must be from the same parametric family. Though to enable a mixture of simpler and more complex models, we could choose a very flexible family, such as the generalized gamma or a spline, and use the `fixedpars` argument to `flexsurvreg` to fix parameters for certain transitions at values for which the flexible family collapses to a simpler one (e.g. §3.2, Table 2).

6.3. Obtaining cumulative transition-specific hazards

The `mstate` package enables *semi-parametric* multi-state modelling. Models must be fitted with `coxph`, which also provides a semi-parametric estimate of each transition-specific baseline hazard (de˜Wreede *et˜al.* 2010, 2011). These imply piecewise-constant estimates of each cumulative $r \to s$ transition-specific hazard function $H_{rs}(t) = \int_0^t q_{rs}(u)du$. These estimates, and

their covariances, are provided by **mstate**'s function `msfit`, and form the basis of prediction from the model. For the Cox models for the BOS data,

```
> library(mstate)
> tmat <- rbind(c(NA,1,2),c(NA,NA,3),c(NA,NA,NA))
> mrcox <- msfit(crcox, trans=tmat)
> mfcox <- msfit(cfcox, trans=tmat)
```

`tmat` describes the transition structure, as a matrix of integers whose r, s entry is i if the ith transition type is r, s, and has `NA`s on the diagonal and where the r, s transition is disallowed.

flexsurv provides an analogous function `msfit.flexsurvreg` to produce cumulative hazards from *fully-parametric* multi-state models in the same format. This is a short wrapper around `summary.flexsurvreg(...,type="cumhaz")`, previously mentioned in §3.4. The difference from **mstate**'s method is that hazard estimates can be produced for any grid of times `t`, at any level of detail and even beyond the range of the data, since the model is fully parametric. The argument `newdata` can be used in the same way to specify a desired covariate category, though in this example there are no covariates in addition to the transition type. The name of the (factor) covariate indicating the transition type can also be supplied through the `tvar` argument, in this case it is the default, `"trans"`.

```
> tgrid <- seq(0,14,by=0.1)
> mrwei <- msfit.flexsurvreg(crwei, t=tgrid, trans=tmat)
> mrexp <- msfit.flexsurvreg(crexp, t=tgrid, trans=tmat)
> mfwei <- msfit.flexsurvreg(cfwei, t=tgrid, trans=tmat)
```

These can be plotted (Figure 5) to show the fit of the parametric models compared to the non-parametric estimates. Both models appear to fit adequately, though give diverging extrapolations after around 6 years when the data become sparse. The Weibull clock-reset model has an improved AIC of 1091, compared to 1099 for the exponential model. For the 2–3 transition, the clock-forward and clock-reset models give slightly different hazard trajectories.

```
> cols <- c("black","red","blue")
> plot(mrcox, xlab="Years after baseline", lwd=3, xlim=c(0,14), cols=cols)
> for (i in 1:3){
+     lines(tgrid, mrexp$Haz$Haz[mrexp$Haz$trans==i], col=cols[i], lty=2, lwd=2)
+     lines(tgrid, mrwei$Haz$Haz[mrwei$Haz$trans==i], col=cols[i], lty=3, lwd=2)
+ }
> lines(mfcox$Haz$time[mfcox$Haz$trans==3], mfcox$Haz$Haz[mfcox$Haz$trans==3],
+       type="s", col="darkgreen", lty=1, lwd=2)
> lines(tgrid, mfwei$Haz$Haz[mfwei$Haz$trans==3], col="darkgreen", lty=3, lwd=2)
> legend("topleft", inset=c(0,0.2), lwd=2, col=c("darkgreen"),
+        c("2 -> 3 (clock-forward)"), bty="n")
> legend("topleft", inset=c(0,0.3), c("Non-parametric","Exponential","Weibull"),
+        lty=c(1,2,3), lwd=c(3,2,2), bty="n")
```

6.4. Prediction from multi-state models

Since `msfit.flexsurvreg` returns transition hazards in the same format as **mstate**'s `msfit`

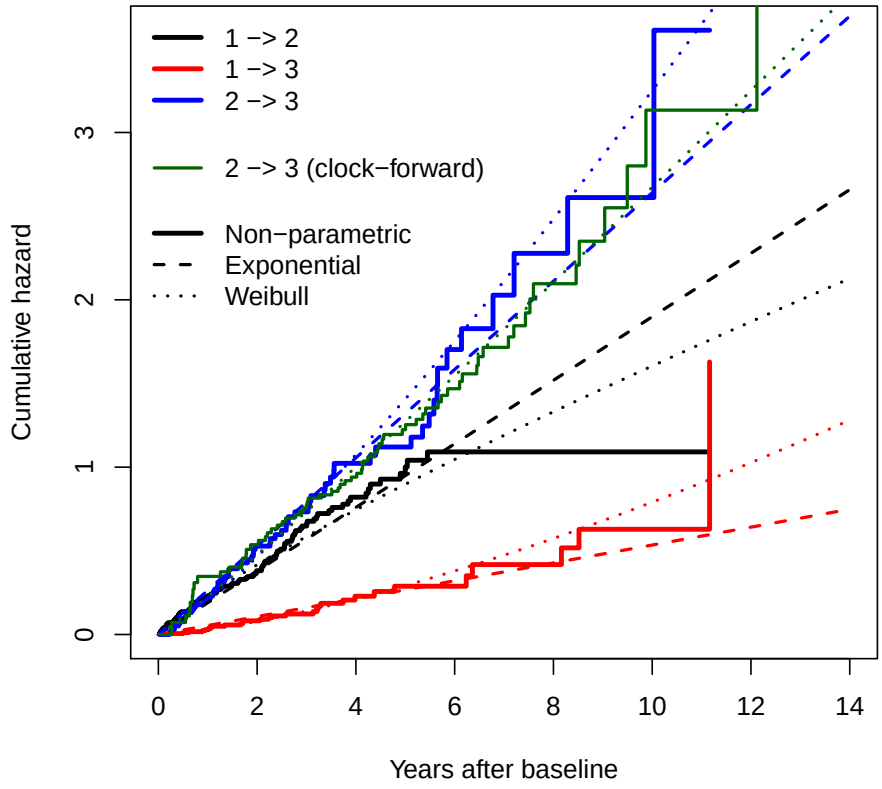

Figure 5: Cumulative hazards for three transitions in the BOS multi-state model (clock-reset), under non-parametric, exponential and Weibull models. For the 2–3 transition, an alternative clock-forward scale is shown for the non-parametric and Weibull models.

method, the functions of **mstate** for prediction can then be used. For example, the *transition probabilities* of the multi-state model are the probabilities of occupying each state s at time $t > t_0$, given that the individual is in state r at time t_0.

$$P(t_0, t) = P(S(t) = s | S(t_0) = r)$$

For an inhomogeneous Markov model, these are related to the transition intensities via the Kolmogorov forward equation

$$\frac{dP(t_0, t)}{dt} = P(t_0, t)Q(t)$$

with initial condition $P() = I$ (Cox and Miller 1965). An approximate solution (e.g. Aalen, Borgan, and Gjessing 2008) is given by a product integral

$$P(t_0, t) = \prod_{i=0}^{m-1} \{I + Q(t_i)dt\}$$

where a fine grid of times $t_0, t_1, \ldots, t_m = t$ is chosen to span the prediction interval, and $Q(t_i)dt$ is the increment in the cumulative hazard matrix between times t_i and t_{i+1}. Q may also depend on other covariates, as long as these are known in advance.

In **mstate** these can be calculated with the `probtrans` function, applied to the cumulative hazards returned by `msfit`. For Cox models, the time grid is naturally defined by the observed survival times, giving the Aalen-Johansen estimator (Andersen, Borgan, Gill, and Keiding 1993). Here, the probability of remaining alive and free of BOS is estimated at 0.27 at 5 years and 0.17 at 10 years.

```
> ptc <- probtrans(mfcox, predt=0, direction="forward")[[1]]
> ptc[c(165, 193),]
```

```
        time   pstate1    pstate2   pstate3        se1        se2        se3
165 4.999316 0.2727122 0.29427877 0.4330090 0.03740559 0.03882367 0.04036366
193 9.872690 0.1740995 0.03975934 0.7861412 0.04031056 0.02224179 0.04462605
```

For parametric models, using a similar discrete-time approximation was suggested by Cook and Lawless (2014). This is achieved by passing the object returned by `msfit.flexsurvreg` to `probtrans` in **mstate**. It can be made arbitrarily accurate by choosing a finer resolution for the grid of times when calling `msfit.flexsurvreg`. Under the Weibull model, the probability of remaining alive and free of BOS is estimated at 0.3 at 5 years and 0.09 at 10 years: the discrepancy from the Cox model is more marked at 10 years when the data are more sparse (Figure˜5).

```
> ptw <- probtrans(mfwei, predt=0, direction="forward")[[1]]
> ptw[ptw$time %in% c(5,10),]
```

```
    time    pstate1   pstate2   pstate3        se1        se2        se3
51     5 0.29984307 0.2543251 0.4458318 0.03222771 0.03467057 0.03655434
101   10 0.08853812 0.1194529 0.7920090 0.02688788 0.03212213 0.04035576
```

An alternative approach is to solve the Kolmogorov forward equation numerically, as in Titman (2011). This is implemented in the function `pmatrix.fs`, using the **deSolve** package (Soetaert, Petzoldt, and Setzer 2010). This returns the full transition probability matrix $P(t_0, t)$ from time $t_0 = 0$ to a time or set of times t specified in the call. This is more accurate than the time grid approximation unless the grid is very fine. The first row of each matrix below is close to `pstate1`–`pstate3` in the corresponding row of `ptw`.

```
> pmatrix.fs(cfwei, t=c(5,10), trans=tmat)
```

```
$`5`
          [,1]      [,2]      [,3]
[1,] 0.3042166 0.2521711 0.4436123
[2,] 0.0000000 0.2804165 0.7195835
[3,] 0.0000000 0.0000000 1.0000000

$`10`
```

```
           [,1]       [,2]      [,3]
[1,] 0.09116559 0.12048039 0.7883540
[2,] 0.00000000 0.06903914 0.9309609
[3,] 0.00000000 0.00000000 1.0000000
```

On the other hand, there is no analytic formula for confidence intervals, unlike in `probtrans`. Therefore these must be obtained by simulation from the asymptotic distribution of the maximum likelihood estimates, and by default this is not done. See `help(pmatrix.fs)` for full details.

For semi-Markov models, the Kolmogorov equation does not apply, since the transition intensity matrix $Q(t)$ is not known in advance at time t, but depends on when the transitions occur between time t_0 and t. $P(t_0, t)$ can then be estimated by simulation, given the estimated cumulative hazards, using **mstate**'s function `mssample` (Fiocco, Putter, and van Houwelingen 2008). 5000 samples are sufficient in this case to give estimates of transition probabilities that are accurate to within around 0.01, and similar to the clock-forward estimates. Bootstrapping would be required for standard errors, which would be computationally expensive given that it currently takes about 20 seconds to generate 5000 samples.

```
> mssample(mrcox$Haz, trans=tmat, clock="reset", M=5000, tvec=c(5, 10))
> mssample(mrwei$Haz, trans=tmat, clock="reset", M=5000, tvec=c(5, 10))
```

7. Potential extensions

Models where multiple survival times are assumed to be correlated within groups, sometimes called (shared) frailty models (Hougaard 1995), would be a useful extension to **flexsurv**. See, e.g. Crowther, Look, and Riley (2014) for a recent application based on parametric models. These might be implemented by exploiting tractability for specific distributions, such as gamma frailties, or by adjusting standard errors to account for clustering, as implemented in `survreg`. More complex random effects models would require numerical integration, and Crowther *et˜al.* (2014) provide Stata software based on Gauss-Hermite quadrature. Alternatively, a probabilistic modelling language such as Stan (Stan Development Team 2014) or BUGS (Lunn, Jackson, Best, Thomas, and Spiegelhalter 2012) would be naturally suited to complex extensions such as random effects on multiple parameters or multiple hierarchical levels.

flexsurv is intended as a platform for parametric survival modelling. Extensions of the software to deal with different models may be written by users themselves, through the facilities described in §4 and §5.2. These might then be included in the package as built-in distributions, or at least demonstrated in the package's other vignette `flexsurv-examples`. Each new class of models would ideally come with

- guidance on what situations the model is useful for, e.g. what shape of hazards it can represent
- some intuitive interpretation of the model parameters, their plausible values in typical situations, and potential identifiability problems. This would also help with choosing initial values for numerical maximum likelihood estimation, ideally through an `inits` function in the custom distribution list (§4).

The examples in this paper were run using version 0.4 of **flexsurv**, which will eventually be available from http://CRAN.R-project.org/package=flexsurv. Development versions are available on https://github.com/chjackson/flexsurv-dev, and contributions are welcome.

A. Acknowledgements

Thanks to Milan Bouchet-Valat for help with implementing covariates on ancillary parameters, Andrea Manca for motivating the development of the package, and all users who have reported bugs and given suggestions.

References

Aalen O, Borgan O, Gjessing H (2008). *Survival and event history analysis: a process point of view*. Springer.

Andersen PK, Borgan O, Gill RD, Keiding N (1993). *Statistical models based on counting processes*. Springer, New York.

Broström G (2014). *eha: Event History Analysis*. R package version 2.4-1, URL http://CRAN.R-project.org/package=eha.

Cook RJ, Lawless JF (2014). "Statistical issues in modeling chronic disease in cohort studies." *Statistics in Biosciences*, **6**(1), 127–161.

Cox C (2008). "The generalized F distribution: An umbrella for parametric survival analysis." *Statistics in Medicine*, **27**, 4301–4312.

Cox DR, Miller HD (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Crowther MJ, Lambert PC (2013). "stgenreg: A Stata package for general parametric survival analysis." *Journal of Statistical Software*, **53**, 1–17.

Crowther MJ, Look MP, Riley RD (2014). "Multilevel mixed effects parametric survival models using adaptive Gauss–Hermite quadrature with application to recurrent events and individual participant data meta-analysis." *Statistics In Medicine*.

de˜Wreede L, Fiocco M, Putter H (2010). "The `mstate` package for estimation and prediction in non-and semi-parametric multi-state and competing risks models." *Computer Methods and Programs in Biomedicine*, **99**(3), 261–274.

de˜Wreede LC, Fiocco M, Putter H (2011). "`mstate`: an R package for the analysis of competing risks and multi-state models." *J Stat Softw*, **38**, 1–30.

Demiris N, Lunn D, Sharples L (2011). "Survival extrapolation using the poly-Weibull model." *Statistical Methods in Medical Research*.

Fiocco M, Putter H, van Houwelingen HC (2008). "Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models." *Statistics in Medicine*, **27**(21), 4340–4358.

Heng D, Sharples L, McNeil K, Stewart S, Wreghitt T, Wallwork J (1998). "Bronchiolitis Obliterans Syndrome: Incidence, Natural History, Prognosis, and Risk Factors." *The Journal of Heart and Lung Transplantation*, **17**(12), 1255–1263.

Hess K (2010). *muhaz: Hazard Function Estimation in Survival Analysis*. R package version 1.2.5, S original by K. Hess and R port by R. Gentleman, URL http://CRAN.R-project.org/package=muhaz.

Hougaard P (1995). "Frailty models for survival data." *Lifetime Data Analysis*, **1**(3), 255–273.

Jackson CH (2011). "Multi-State Models for Panel Data: The msm Package for R." *Journal of Statistical Software*, **38**(8).

Kalbfleisch J, Lawless J (1985). "The Analysis of Panel Data under a Markov Assumption." *Journal of the American Statistical Association*, **80**(392), 863–871.

Lambert PC, Royston P (2009). "Further development of flexible parametric models for survival analysis." *Stata Journal*, **9**(2), 265.

Latimer NR (2013). "Survival analysis for economic evaluations alongside clinical trials — extrapolation with patient-level data inconsistencies, limitations, and a practical guide." *Medical Decision Making*, **33**(6), 743–754.

Louzada-Neto F (1999). "Polyhazard models for lifetime data." *Biometrics*, **55**, 1281–1285.

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D (2012). *The BUGS Book: a practical introduction to Bayesian analysis*. CRC Press.

Mandel M (2013). "Simulation-based confidence intervals for functions with complicated derivatives." *The American Statistician*, **67**(2), 76–81.

Mueller HG, Wang JL (1994). "Hazard rates estimation under random censoring with varying kernels and bandwidths." *Biometrics*, **50**, 61–76.

Nadarajah S, Bakar SAA (2013). "A new R package for actuarial survival models." *Computational Statistics*, **28**(5), 2139–2160.

Nash JC (1990). *Compact numerical methods for computers: linear algebra and function minimisation*. CRC Press.

Nelson CP, Lambert PC, Squire IB, Jones DR (2007). "Flexible parametric models for relative survival, with application in coronary heart disease." *Statistics in medicine*, **26**(30), 5486–5498.

Prentice RL (1974). "A log gamma model and its maximum likelihood estimation." *Biometrika*, **61**(3), 539–544.

Prentice RL (1975). "Discrimination among some parametric models." *Biometrika*, **62**(3), 607–614.

Putter H, Fiocco M, Geskus RB (2007). "Tutorial in biostatistics: competing risks and multi-state models." *Statistics in Medicine*, **26**, 2389–2430.

Reid N (1994). "A conversation with Sir David Cox." *Statistical Science*, pp. 439–455.

Rigby RA, Stasinopoulos DM (2005). "Generalized additive models for location, scale and shape,(with discussion)." *Applied Statistics*, **54**, 507–554.

Royston P (2001). "Flexible parametric alternatives to the Cox model, and more." *Stata Journal*, **1**(1), 1–28.

Royston P (2004). "Flexible parametric alternatives to the Cox model: update." *The Stata Journal*, **4**(1), 98–101.

Royston P, Altman DG (1994). "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling." *Applied Statistics*, pp. 429–467.

Royston P, Lambert PC (2011). "Flexible parametric survival analysis using Stata: beyond the Cox model." *Stata Press books*.

Royston P, Parmar M (2002). "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**(1), 2175–2197.

Sauerbrei W, Royston P (1999). "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**(1), 71–94.

Sauerbrei W, Royston P, Binder H (2007). "Selection of important variables and determination of functional form for continuous predictors in multivariable model building." *Statistics in medicine*, **26**(30), 5512–5528.

Soetaert K, Petzoldt T, Setzer RW (2010). "Solving Differential Equations in R: Package deSolve." *Journal of Statistical Software*, **33**(9), 1–25. ISSN 1548-7660. URL http://www.jstatsoft.org/v33/i09.

Stacy EW (1962). "A generalization of the gamma distribution." *Annals of Mathematical Statistics*, (33), 1187–92.

Stan Development Team (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.4*. URL http://mc-stan.org/.

Therneau T (2014). "A Package for Survival Analysis in S." R package version 2.37-7. http://CRAN.R-project.org/package=survival.

Titman AC (2011). "Flexible nonhomogeneous Markov models for panel observed data." *Biometrics*, **67**(3), 780–787.

Yee TW, Wild CJ (1996). "Vector generalized additive models." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 481–493.