Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Simulating biologically plausible complex survival data

Michael J. Crowther[1*†] and Paul C. Lambert[1,2]

Simulation studies are conducted to assess the performance of current and novel statistical models in pre-defined scenarios. It is often desirable that chosen simulation scenarios accurately reflect a biologically plausible underlying distribution. This is particularly important in the framework of survival analysis, where simulated distributions are chosen for both the event time and the censoring time. This paper develops methods for using complex distributions when generating survival times to assess methods in practice. We describe a general algorithm involving numerical integration and root finding techniques to generate survival times from a variety of complex parametric distributions, incorporating any combination of time-dependent effects, time-varying covariates, delayed entry, random effects and covariates measured with error. User-friendly Stata software is provided.

Keywords: simulation, survival, time-varying covariates, time-dependent effects, delayed entry, measurement error

1. Introduction

Simulation studies are conducted to assess the performance of current and novel statistical models in pre-defined scenarios. The quality and reporting of simulation studies varies considerably which has led to the establishment of general guidelines for the development and reporting of simulation studies in medical research [1]. In order to establish certain properties, such as bias and coverage or robustness to deviations from underlying assumptions, it is often desirable that chosen simulation scenarios accurately reflect a biologically plausible distribution. This is particularly important in the framework of survival analysis, where distributions are chosen for both the event time and the censoring time.

Previous studies have introduced a highly flexible framework to simulate survival data for Cox proportional hazards models [2, 3]. Bender *et al.* noted that many simulation studies that generated survival data assumed an exponential distribution for the distribution of event times. Although many recent studies have gone beyond the standard exponential choice to a slightly more complex Weibull distribution [4, 5], these choices are often not flexible enough to fully reflect the underlying distributions observed in clinical data.

Often in clinical trials [6] and population based studies [7, 8], at least one turning point is observed in the underlying hazard function. Although hazard ratios can be insensitive to a poorly specified baseline hazard function, when interest lies in measures of absolute risk it is vital to accurately capture the baseline [9]. Through fully flexible parametric models we can both accurately capture complex hazard functions, but also simulate biologically plausible survival data. Such methods are becoming more commonplace as the benefits of a parametric approach, such as the

[1] *Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH.*
[2] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, S-171 77 Stockholm, Sweden.*
[*] *Correspondence to: Michael J. Crowther. Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH.*
[†] *E-mail: michael.crowther@le.ac.uk*

Contract/grant sponsor: Contract/grant sponsor name

reporting of measures of absolute risk, become recognised in applied research [10]. In order to assess such parametric approaches we require methods to simulate survival data from a variety of complex distributions, beyond standard distributions such as the exponential, Weibull and Gompertz.

Furthermore, a variety of extensions to the standard survival analysis framework, such as incorporation of time-dependent effects (non-proportional hazards), the occurrence of time-varying covariates, random covariate effects and covariates measured with error, all require suitable simulation techniques to assess statistical models developed for each setting. A further often observed phenomenon in survival analysis is the presence of informative censoring. Standard survival models make the assumption of no dependence between the survival and censoring mechanisms. Assessing the robustness of methods to deviations from this assumption is a key question in survival analysis [11].

In this paper we describe a general algorithm for the simulation of survival times. In Section 2 we introduce a motivating dataset which exhibits turning points in the underlying hazard function, which cannot be captured by standard parametric distributions. In Section 3 we briefly describe the methods of Bender *et al.* to simulate data from standard parametric distributions with an analytically tractable and invertible cumulative hazard function, which forms the basis for our simulation framework. In Sections 4 and 5 we describe a range of simulation scenarios, culminating in our general simulation algorithm to simulate survival data from complex distributions using root finding techniques with nested numerical integration. In Section 6 we describe how to incorporate time-dependent effects, both with standard and complex parametric distributions. In Section 7 we describe how to incorporate both binary and continuous time-varying covariates. In Section 8 we describe how to incorporate delayed entry, whilst in Section 9 we describe how the techniques can be applied to incorporate dependent censoring. The methods are illustrated using the publicly available `survsim` package in Stata [12, 13]. Finally, in Section 10, we conclude the paper with a discussion.

2. Motivating dataset

To illustrate various aspects of the simulation techniques, we use a data set of 686 women diagnosed with breast cancer in Germany [14], with 246 patients randomized to receive hormonal therapy and 440 to receive a placebo. Outcome is recurrence-free survival, with 299 patients experiencing the event of interest. We apply a Weibull proportional hazards model and a proportional hazards flexible parametric survival model with 5 degrees of freedom, adjusting for treatment in each. The flexible parametric model uses restricted cubic splines on the cumulative hazard scale as a way of getting a smooth but highly flexible parametric form incorporating turning points [9]. Figure 1 shows the fitted survival curves from each model overlaid on the Kaplan-Meier curves, by treatment group, showing the much improved fit from the more flexible model.

Furthermore, in Figure 2 we have the fitted hazard functions, across treatment group, for the Weibull and flexible parametric survival models, illustrating a marked difference in the estimated underlying shapes, clearly showing that the Weibull model is inappropriate.

3. Simulating survival times

3.1. Simulating from standard parametric distributions

A pivotal paper by Bender et al. [3] described a highly efficient and easy to implement technique to generate survival times from a variety of parametric distributions. We briefly describe the methods of Bender et al., as it forms the basis for the extensions below. The hazard function of a proportional hazards model can be expressed as

$$h(t) = h_0(t)\exp(X\beta) \tag{1}$$

where $h_0(t)$ is the baseline hazard function specified by some parametric distribution, X is a vector of time-independent covariates with corresponding regression coefficients, β. The corresponding cumulative hazard, $H(t)$, survival, $S(t)$ and cumulative distribution, $F(t)$, functions are obtained as follows

$$H(t|X) = H_0(t)\exp(X\beta), \quad \text{where} \quad H_0(t) = \int_0^t h_0(u)\mathrm{d}u \tag{2}$$

$$S(t|X) = \exp[-H(t)] \quad \text{and} \quad F(t|X) = 1 - \exp[-H(t)] \tag{3}$$

If we let T be the simulated survival time, Bender et al. [3] showed that by letting

$$F(T|X) = 1 - \exp[-H(T|X)] = u, \quad \text{where} \quad u \sim U(0,1) \tag{4}$$

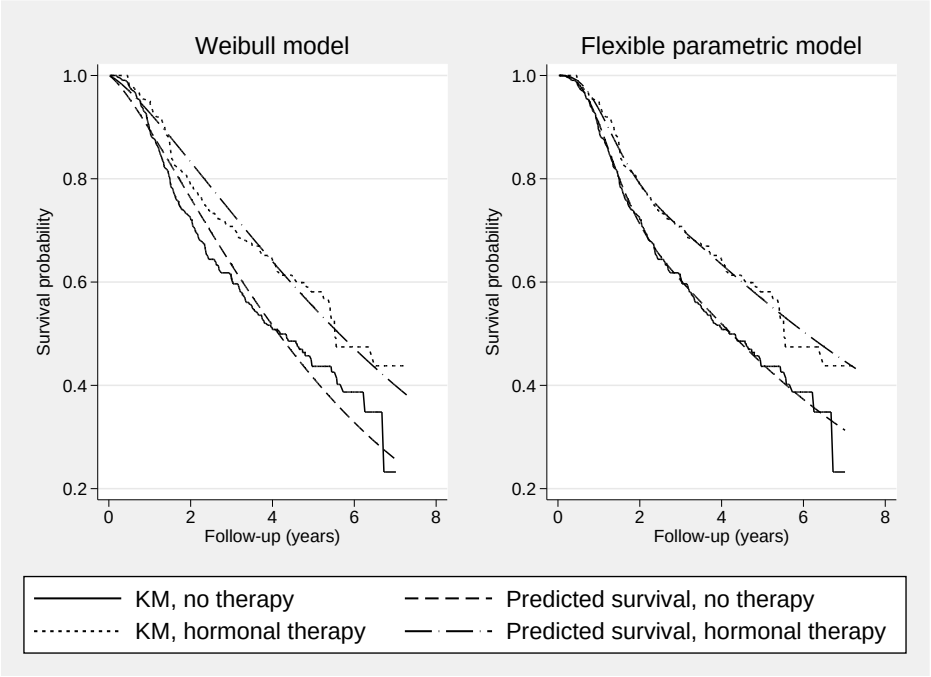

Figure 1. Predicted survival from Weibull and flexible parametric survival models overlaid on the Kaplan-Meier (KM) curves

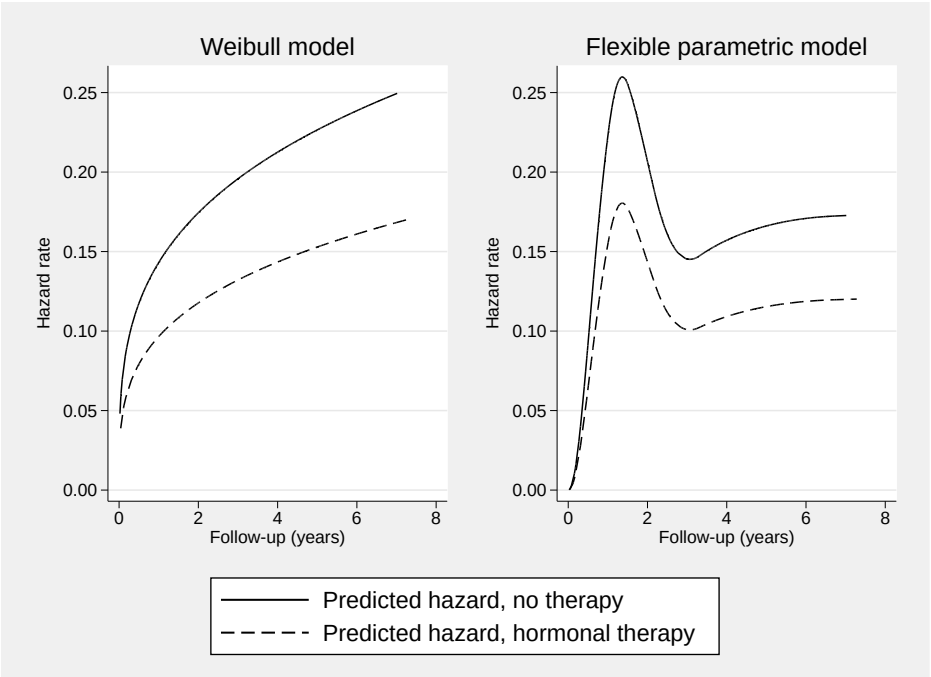

Figure 2. Predicted hazard functions from Weibull and flexible parametric survival models

or equivalently we can write

$$S(T|X) = u \tag{5}$$

Thus, if $h_0(T) > 0$, then Equation (5) can be re-arranged and directly solved for T, as long as $H_0(t)$ can be directly inverted.

$$T = H_0^{-1}[-\log(U)\exp(-X\beta)] \tag{6}$$

The data generating process then only requires draws from a uniform distribution, followed by application of Equation (5). The three standard choices for $h_0(T)$ are the exponential, Weibull and Gompertz distributions. These three choices can be considered restrictive in terms of the shapes of the baseline hazard function that can

be generated. However, these distributions remain appealing to researchers conducting simulation studies, perhaps because they all have invertible cumulative hazard functions, allowing the direct application of Equation (6).

Example 1 in the appendix illustrates Stata code for the simulation of proportional hazards data under an exponential, Gompertz or Weibull distribution.

4. A general framework for simulation of survival data

We now extend the method of Bender et al. to allow for more complex simulation studies where either the integral to obtain the cumulative hazard in Equation (2) is intractable or $H_0(t)$ is non-invertible. Figure 3 shows a schematic flow diagram illustrating the general framework for simulating survival data from a defined hazard or cumulative hazard function.

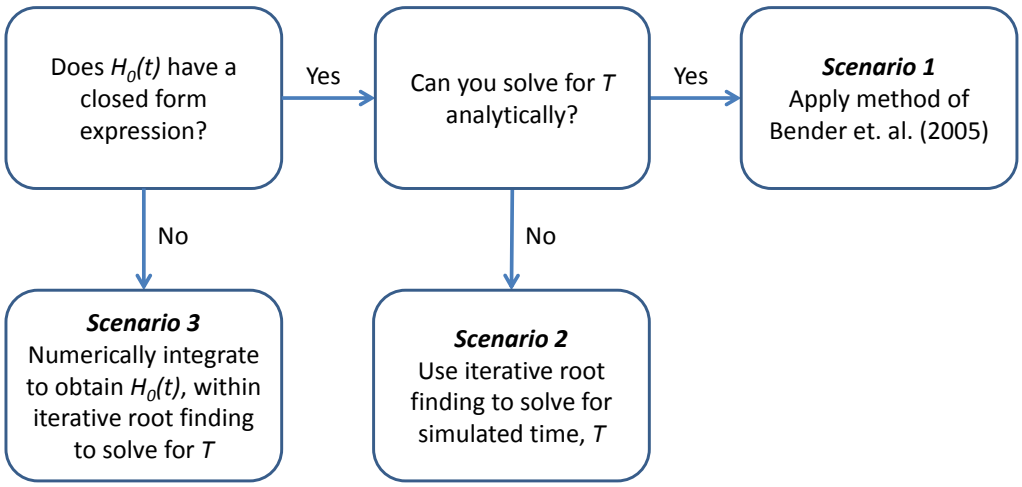

Figure 3. Schematic flow diagram of simulation techniques

4.1. Scenario 1

Scenario 1 involves the setting of Bender et al. (2005) described in Section 3.1, where the cumulative hazard function has a closed form expression and can be directly inverted to solve for T the simulated survival time.

4.2. Scenario 2

Scenario 2 arises when we wish to use a more complex baseline hazard function to simulate data under a proportional hazards model. In this case we assume the cumulative hazard function has a closed form expression. However, if we choose a more complex hazard and consequently cumulative hazard function, here we assume that we can no longer directly invert the cumulative hazard function, and therefore cannot directly solve for T, the simulated survival time. In this situation we proceed by applying root finding techniques. We describe this in more detail in Section 5.1.

4.3. Scenario 3

Finally, Scenario 3 arises when we wish to define a complex hazard function which cannot be integrated analytically to obtain the cumulative hazard function. To accommodate this setting we can use numerical integration techniques such as Gauss-Legendre quadrature. Following this, we once again have a cumulative hazard function which cannot be directly inverted to solve for the simulated survival time, T, therefore requiring root finding techniques as in Scenario 2. This results in a general 2-stage algorithm involving numerical integration nested within an iterative root finding procedure. We describe this in more detail in Section 5.3.

5. Simulating from complex baseline hazard functions

5.1. Root finding

The first extension we describe involves the situation where we wish to use a more complex baseline hazard function, to simulate data under a proportional hazards model. In this case we still assume that the cumulative hazard can be evaluated analytically for a given hazard function.

The step between Equation (5) and Equation (6) is reliant on being able to directly re-arrange Equation (5) to solve for T, the simulated survival time. When this condition fails we require iterative techniques to find the root of Equation (5). We illustrate this situation through example.

5.1.1. Example: 2-component mixture Weibull distribution We now begin to introduce some complexity in the parametric distribution used to generate survival times. Motivation for going beyond the standard parametric approaches described in Section 3.1 originates from the often observed situation of a turning point in a dataset's baseline hazard function, illustrated in Section 2. One such approach is to use a mixture distribution.

Here we define the overall baseline survival function of a 2-component parametric mixture model. Finite mixture survival models of this form have been used in standard survival analysis [15], and mixture and non-mixture cure models to obtain improved estimates of statistical cure [16]. We define parametric components additive on the survival scale

$$S_0(t) = pS_{01}(t) + (1-p)S_{02}(t) \tag{7}$$

where $S_{01}(t)$ and $S_{02}(t)$ are the survival function of any standard parametric distribution, and p represents the mixing parameter where $0 \leq p \leq 1$. For illustrative purposes we proceed by assuming a 2-component mixture Weibull distribution, with

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2}) \tag{8}$$

where $\{\lambda_1, \lambda_2\}$, and $\{\gamma_1, \gamma_2\}$ are scale and shape parameters, respectively. Transforming to the cumulative hazard scale

$$H_0(t) = -\log\left(p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\right) \tag{9}$$

and differentiating with respect to t, we obtain the baseline hazard function

$$h_0(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1}p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})} \tag{10}$$

Proportional hazards can then be induced

$$h(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1}p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})}\exp(\boldsymbol{X}\boldsymbol{\beta}) \tag{11}$$

where $\boldsymbol{X}$ is a vector of time-independent covariates with associated regression coefficients, $\boldsymbol{\beta}$. This model can be used to simulate survival data from a variety of functions with turning points, to better reflect observed clinical datasets. We illustrate some examples in Figure 4, based on those seen in real datasets [6, 14, 17].

Equation (11) can be directly integrated with respect to t to obtain the cumulative hazard function, and subsequently the survival function. However, we are still left with a survival function that when substituted into Equation (5), produces an equation which cannot be directly solved for t. We now describe two root finding techniques to accommodate this situation. We do note, however, that in this example we could simulate from the 2-component mixture distribution by generating a categorical latent variable and then drawing from the appropriate mixture component, thus avoiding iterative techniques.

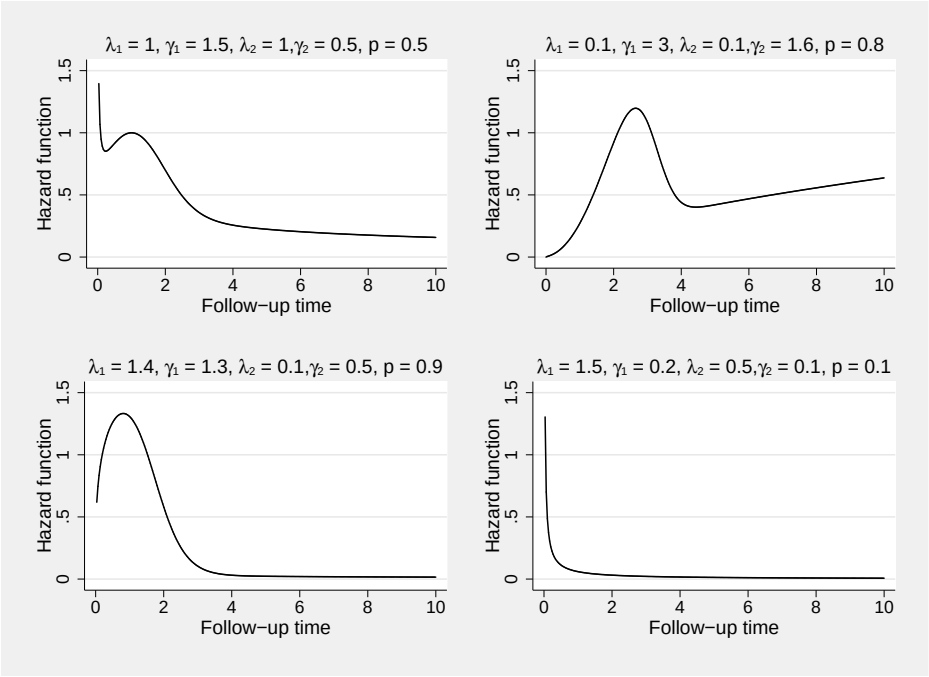

Figure 4. Example mixture Weibull hazard functions

5.1.2. Brent's univariate root-finding method To generate our survival times we need to solve for t the following

$$g(t) = S(t) - U = 0 \tag{12}$$

An efficient method to calculate the simulated survival times is to use Brent's univariate root finder. This algorithm combines the bisection method with linear or quadratic interpolation [18]. The algorithm is executed until a desired tolerance (we use a default of 1E-08) is met.

5.1.3. Newton-Raphson root finder An alternative method to Brent's root finder is to use Newton-Raphson iterations, which uses the first two terms of the Taylor series expansion of $g(t)$, our objective function. In our experience we have found Brent's method to be far superior in terms of reliability and accuracy compared to Newton-Raphson iterations, which can have convergence problems.

5.2. Example simulation study

We illustrate the root finding technique described in Section 5.1, applied to the 2-component mixture Weibull distribution. In each of 1000 repetitions, we generate 1000 survival times from a 2-component mixture Weibull baseline hazard function, with parameters $\lambda_1 = 0.3$, $\gamma_1 = 2.5$, $\lambda_2 = 0.025$, $\gamma_2 = 1.9$ and $p = 0.3$. These parameter values are chosen to closely approximate the observed hazard function seen in the example dataset above. We also include a binary treatment variable, drawn from $X_i \sim \text{Bin}(1, 0.5)$, with associated hazard ratio $\exp(\beta) = 0.7$, and apply administrative censoring at 5 years. Computation time to generate the 1000 datasets was 34 seconds on an Intel Core i5 2.5GHz CPU. To each simulated dataset we apply a Weibull survival model and the 2-component mixture Weibull model [19], monitoring estimates of the log hazard ratio. Furthermore, we monitor estimates of the survival probability and hazard rate in the reference group ($X_i = 0$), at time points $t = \{1, 2, 3, 4, 5\}$.

Results are presented in Table 1. The mixture Weibull model produces unbiased estimates and good coverage probabilities in the log hazard ratio, and the estimates of survival and hazard, indicating its ability to capture the complex underlying hazard function. In comparison, estimates from the Weibull model indicate minor bias in the log hazard ratio, with large bias observed in the estimates of survival and hazard, across the 5 time points, indicating its inability to effectively capture the underlying shape.

Example 2 in the appendix illustrates Stata code used to generate proportional hazards data from the described 2-component mixture Weibull model.

Table 1. Bias and coverage of the log hazard ratio and estimates of baseline survival and hazard at specific time points from both Weibull and mixture Weibull models.

		Weibull		Mixture Weibull	
	Truth	Bias	95% Coverage	Bias	95% Coverage
β	-0.357	-0.013	92.3	-0.002	93.7
Survival					
1 year	0.905	-0.019	43.9	0.001	94.7
2 year	0.693	0.055	5.4	-0.001	94.3
3 year	0.575	0.041	39.2	-0.000	95.2
4 year	0.494	0.004	93.7	0.001	94.3
5 year	0.411	-0.015	88.1	-0.000	94.1
Hazard					
1 year	0.220	-0.067	0.0	-0.002	93.8
2 year	0.250	-0.066	0.1	0.001	95.4
3 year	0.146	0.059	0.0	-0.001	94.2
4 year	0.166	0.055	1.5	-0.001	95.1
5 year	0.202	0.033	51.0	0.009	95.1

5.3. Numerical integration

We now describe the scenario where we define a more general functional form for the hazard function, which will require numerical integration techniques to evaluate the cumulative hazard function. This is then followed by the root finding technique described above. We once again illustrate through example.

5.3.1. Example: Fractional polynomials Fractional polynomials select powers from a pre-defined set, usually $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, which can be used to model continuous covariates which exhibit non-linearity [20]. In this case, the continuous 'covariate' of interest is survival time. Here we use fractional polynomials to define the log hazard function.

$$h(t) = h_0(t)\exp(X\beta) \tag{13}$$

where $h_0(t)$ is any general function which satisfies $h_0 > 0$ for $t > 0$. Here we expand $\log(h_0(t))$ into a fractional polynomial function with 2 turning points, in this case an FP3 model with powers $\{1, 0.5, 0.5\}$.

$$\log(h_0(t)) = -18 + 7.3t - 11.5t^{0.5}\log(t) + 9.5t^{0.5} \tag{14}$$

The assumed hazard function is shown in Figure 5. This provides a reasonable fit to the example dataset described in Section 2.

The next step to simulate survival times from this underlying hazard function is to calculate the cumulative hazard function; however, when we substitute Equation (14) into Equation (13) and attempt to integrate, we obtain an analytically intractable integral, therefore requiring numerical techniques.

5.3.2. Gaussian quadrature Numerical integration techniques, such as Gaussian quadrature [21], provide an approximation to an analytically intractable integral. Gaussian quadrature turns an integral into a weighted summation of a function evaluated at a set of pre-defined points called nodes. For example, integrating a hazard function

$$\int_0^t h(u)\mathrm{d}u \tag{15}$$

we first need to undertake a change of interval using

$$\int_0^t h(u)\mathrm{d}u = \frac{t}{2}\int_{-1}^{1} h\left(\frac{t}{2}z + \frac{t}{2}\right)\mathrm{d}z \tag{16}$$

We can now numerically integrate, using for example Gauss-Legendre quadrature, resulting in

$$\int_0^t h(u)\mathrm{d}u \approx \frac{t}{2}\sum_{i=1}^{m} w_i h\left(\frac{t}{2}z_i + \frac{t}{2}\right) \tag{17}$$

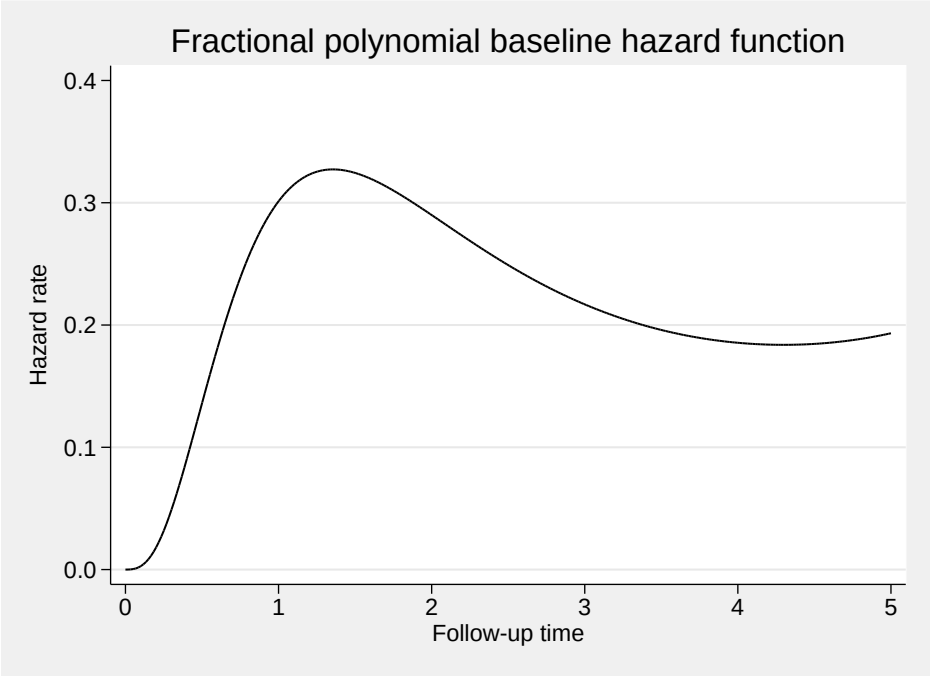

Figure 5. Fractional polynomial baseline hazard function

where w_i and z_i are a set weights and node locations. Under Gauss-Legendre quadrature the weights $w_i = 1$. The numerical accuracy of the approximation depends on the number of nodes, m. In our experience we have found that often 30 nodes are sufficient. The accuracy can be assessed by simulating survival times with an increasing number of nodes.

Now that we can calculate the cumulative hazard, we then apply one of the root finding procedures described in Section 5.1. The iterative algorithm, however, in this case now has multiple steps, including numerical integration nested within either Newton-Raphson steps or Brent's method.

5.4. Example simulation study

We now illustrate the algorithm by simulating survival data from the baseline hazard function defined in Equation (14). In this case, we can use the flexible parametric survival model, which models the baseline log cumulative hazard function using restricted cubic splines with 5 degrees of freedom, to assess how well it captures complex hazard functions. For comparison we also apply a Weibull proportional hazards model. For each of 1000 repetitions we simulate 1000 survival times, incorporating a binary and continuous covariate, representing gender, $X_{1i} \sim \text{Bin}(1, 0.5)$ and age, $X_{2i} \sim N(65, 12)$, with associated log hazard ratios of $\beta_1 = -0.5$ and $\beta_2 = 0.02$, respectively. We assume administrative censoring at 5 years. Computation time to generate the 1000 datasets was 144 seconds on an Intel Core i5 2.5GHz CPU. In each repetition we monitor estimates of the log hazard ratios for the effects of gender and age. Furthermore, we assess estimates for survival probabilities and hazard rates at $t = \{1, 2, 3, 4, 5\}$, estimated at $X_{1i} = 0$ and $X_{2i} = 65$.

Results are presented in Table 2. Under the Weibull model, we observe substantial bias of -0.082 for β_1, the treatment effect, compared to unbiased estimates under the flexible parametric survival model. Estimates of the hazard and survival functions are generally heavily biased under the Weibull model which poor coverage probabilities, compared to minimal bias and good coverage under the flexible parametric approach. Note, there is a small amount of bias for the hazard at 5 years under the flexible parametric survival model; however, given that this is not the true model, it generally performs very well.

Example 3 in the appendix shows Stata code used to generate data from the described fractional polynomial baseline hazard function.

Table 2. Bias and coverage of the log hazard ratio and estimates of baseline survival and hazard at specific time points from Weibull and flexible parametric models.

		Weibull		FPM	
	Truth	Bias	95% Coverage	Bias	95% Coverage
β_1	-0.500	-0.082	74.2	-0.001	93.1
β_2	0.020	0.003	75.4	0.000	94.2
Survival					
1 year	0.602	0.011	91.4	0.000	94.7
2 year	0.189	0.059	4.9	0.001	94.3
3 year	0.076	0.001	87.1	-0.000	94.2
4 year	0.037	-0.018	10.3	-0.001	94.4
5 year	0.018	-0.014	0.1	0.001	95.3
Hazard					
1 year	1.105	-0.365	0.0	-0.001	95.6
2 year	1.064	-0.009	84.9	-0.005	94.7
3 year	0.796	0.503	0.0	0.022	94.4
4 year	0.681	0.824	0.0	0.013	96.1
5 year	0.709	0.979	0.0	-0.064	91.2

PH - proportional hazards, FPM - flexible parametric model

6. Simulating time-dependent effects

The presence of non-proportional hazards, i.e. time-dependent effects, is commonplace in the analysis of time to event data [22]. This is often observed in the analysis of registry based data sources where follow-up time can be over many years [23]. Furthermore, evidence is often found of time-dependent treatment effects [24].

6.1. Standard parametric distributions

Under standard parametric distributions, the inclusion of time-dependent effects can be conducted so as to ensure an analytically tractable and invertible cumulative hazard function. For example, under an exponential distribution or a Gompertz distribution, covariates can be interacted with time, to result in a baseline hazard function which can still be directly integrated, and subsequently directly solved for the simulated survival time, t. Similarly, under a Weibull distribution, an interaction can be formed between covariates and log time.

For example, consider a binary covariate, X_1, which takes values 0 or 1. Under a Gompertz baseline hazard function, we can invoke non-proportional hazards by interacting X_1 with linear time, t:

$$h(t) = \lambda \exp(\gamma t + \beta_1 X_1 + \beta_2 X_1 t) \tag{18}$$

Equation (18) can be re-arranged,

$$h(t) = \lambda \exp\left[(\gamma + \beta_2 X_1)t + \beta_1 X_1\right] \tag{19}$$

integrated to obtain the cumulative hazard function

$$\begin{aligned} H(t) &= \int_0^t \lambda \exp\left[(\gamma + \beta_2 X_1)u + \beta_1 X_1\right] \mathrm{d}u \\ &= \frac{\lambda \exp(\beta_1 X_1)}{\gamma + \beta_2 X_1}\{\exp\left[(\gamma + \beta_2 X_1)t\right] - 1\} \end{aligned} \tag{20}$$

We therefore have, from Equation (5)

$$U = \exp\left(-\frac{\lambda \exp(\beta_1 X_1)}{\gamma + \beta_2 X_1}\{\exp\left[(\gamma + \beta_2 X_1)t\right] - 1\}\right) \tag{21}$$

which can be inverted and solved for t, the simulated survival time

$$t = \frac{1}{\gamma + \beta_2 X_1} \log\left(-\frac{\gamma + \beta_2 X_1}{\lambda \exp(\beta_1 X_1)} \log(U) + 1\right) \tag{22}$$

This of course can be extended to multiple time-dependent effects; however, if we wish to use a more complex distribution we once again have analytically intractable and non-invertible cumulative hazard functions.

6.2. Complex hazard functions

Incorporating time-dependent effects when simulating more complex hazard functions returns us to the scenario where we require both numerical integration and iterative root-finding procedures. For example, this arises when including a time-dependent effect into the 2-component mixture Weibull model.

$$h(t) = h_0(t)\exp(\beta_1(t)X) \tag{23}$$

where $\beta_1(t)$ is a function of time, t, such as a simple linear term, or something more complex such as a fractional polynomial or spline function.

6.3. Example simulation study

We now conduct a simulation study assessing the performance of the Weibull and flexible parametric models under proportional hazards, when simulating a time-dependent diminishing treatment effect. We further apply a flexible parametric model allowing for a time-dependent treatment effect. In all models we assess estimates of the hazard and survival functions in each treatment group. Survival times are simulated from the baseline hazard function shown in Equation (14).

For each of 1000 repetitions we simulate 1000 survival times, incorporating a binary and continuous covariate, representing treatment, $X_{1i} \sim \text{Bin}(1, 0.5)$ and age, $X_{2i} \sim N(65, 12)$, and assume administrative censoring at 5 years. We simulate a time-dependent treatment effect under the following

$$\beta_1(t) = -0.7 + 0.01t + 0.4\log(t) \tag{24}$$

and proportional age effect of $\beta_2 = 0.02$. The true hazard ratio of the treatment effect at 1, 2 and 5 years is 0.502, 0.668 and 0.994, respectively. Computation time to generate the 1000 datasets was 173 seconds on an Intel Core i5 2.5GHz CPU. In each repetition we also monitor estimates of the log hazard ratios for the effect of age.

Results are presented in Table 3. As in Section 5.4, we observe very poor performance when using a Weibull model, with large bias in estimates of the hazard and survival functions in both treatment groups. We observe improved performance under the proportional hazards flexible parametric model; however, some bias and poor coverage is seen in estimates of the hazard and survival functions, particularly in the treatment group, but as this model assumes proportional hazards and the true model has non-proportional hazards, this is to be expected. Under the flexible parametric model allowing for a time-dependent treatment effect we observe much reduced bias and improved coverage probabilities, indicating that the model has captured the complex time-dependent effect, even though we do not fit the true underlying model.

We include Stata code in Example 4 of the appendix, showing how to simulate survival times from this complex scenario.

7. Simulating time-varying covariates

Time-varying covariates occur frequently in medical research. In cancer clinical trials the occurrence of treatment switching or non-compliance, occurs when a patient switches from, for example, the standard therapy to the new treatment, often around the time of progression. An area of increasing interest in the biostatistical literature is the joint modelling of longitudinal and survival data, where we observe a repeatedly measured biomarker, and wish to investigate the association of this time-varying biomarker and survival.

Recently Austin [25] extended the methods of Bender et al. to simulate time-varying covariates of three types: first, a dichotomous time-varying covariate that can change at most once; second, a continuous time-varying covariate; third, a dichotomous time-varying covariate where subjects can switch groups multiple times. Austin derived closed form expressions, including time-independent covariates, under the exponential, Weibull and Gompertz distributions.

Under our simulation framework, we generalise the approach of Austin to incorporate any combination of time-varying covariates, with a user-defined baseline hazard function to incorporate more biologically realistic hazard functions.

Table 3. Bias and coverage of the log hazard ratio and estimates of baseline survival and hazard at specific time points from proportional hazards Weibull and flexible parametric models, and non-proportional hazards flexible parametric model.

		Weibull PH		FPM PH		FPM NPH	
	Truth	Bias	95% Coverage	Bias	95% Coverage	Bias	95% Coverage
β_2	0.020	0.003	74.5	0.000	94.0	0.000	94.6
Survival ($X_1 = 0$, $X_2 = 65$)							
1 year	0.602	0.035	39.0	0.030	61.7	0.000	94.9
2 year	0.189	0.062	5.0	0.006	91.3	0.001	94.0
3 year	0.076	-0.006	79.6	-0.010	78.9	-0.000	94.3
4 year	0.037	-0.022	0.7	-0.012	42.7	-0.001	95.5
5 year	0.018	-0.016	0.0	-0.008	38.4	0.001	95.7
Hazard ($X_1 = 0$, $X_2 = 65$)							
1 year	1.105	-0.374	0.0	-0.058	90.1	-0.005	95.4
2 year	1.064	0.062	74.5	0.093	81.7	-0.012	94.8
3 year	0.796	0.654	0.0	0.224	10.6	0.025	95.0
4 year	0.681	1.054	0.0	0.268	17.9	0.025	95.6
5 year	0.709	1.286	0.0	0.223	54.0	-0.046	95.5
Survival ($X_1 = 1$, $X_2 = 65$)							
1 year	0.803	0.180	0.0	0.168	61.7	-0.000	94.8
2 year	0.406	0.281	0.0	0.206	91.3	0.001	94.6
3 year	0.208	0.157	0.0	0.138	78.9	-0.001	94.8
4 year	0.112	0.061	0.0	0.086	42.7	-0.002	95.3
5 year	0.059	0.018	7.2	0.055	38.4	0.001	95.5
Hazard ($X_1 = 1$, $X_2 = 65$)							
1 year	0.554	-0.707	0.0	-0.511	0.0	0.000	94.1
2 year	0.711	-0.451	0.0	-0.408	0.0	-0.008	94.2
3 year	0.632	-0.007	90.2	-0.218	0.1	0.020	93.6
4 year	0.612	0.263	0.0	-0.144	37.5	0.011	95.5
5 year	0.705	0.377	0.0	-0.181	38.1	-0.079	88.6

PH - proportional hazards, FPM - flexible parametric model
NPH - non-proportional hazards

7.1. Simulating treatment switching

In this scenario we wish to simulate a time-varying binary covariate. We define X_1 to represent initial treatment a patient is randomised to, with treatment A ($X_1 = 0$) and treatment B ($X_1 = 1$). We assume patients were randomised to treatment arms at $t = 0$. For simplicity we allow patients to switch arm at most once. We also include a binary covariate which represents disease severity, X_2, with each patient having a 40% chance of having a bad prognosis ($X_2 = 1$), which increases a patient's event rate. Under a general baseline hazard function, $h_0(t)$, we have

$$h(t) = h_0(t) \exp\{\beta_1 [I(t \leq t_s)X_1 + I(t > t_s)(1 - X_1)] + \beta_2 X_2\} \tag{25}$$

where β_1 is the log hazard ratio for the effect of treatment, which in this case we assume is the same regardless of whether patients switched or not. In this example we assume that a patient initially randomised to treatment ($X = 1$) has a treatment effect of $\exp(\beta_1)$ until their switching time, t_S, after which their hazard ratio is 1. Thus, the time-dependence is introduced through the indicator functions $I(t \leq t_s)$ and $I(t > t_s)$. The switching times need to be generated: in the example code included in the appendix, we generate the potential switching times from a uniform distribution, which is dependent on disease severity (X_2). Endogenous/non-ignorable treatment switching can be created if the variable X_2 is deleted and not available for analysis. This scenario can be easily extended to allow for any number of switches. Alternatively, we first could generate a vector of survival times, $\mathbf{t_s}$, to represent time to progression.

7.2. Simulating a continuous time-varying biomarker and survival incorporating random effects and covariates measured with error

Here we wish to simulate a continuous biomarker exhibiting a linear trend, under the following model

$$W(t) = \beta_{0i} + \beta_{1i}t + \boldsymbol{\delta}\boldsymbol{u}_i \tag{26}$$

where

$$\boldsymbol{\beta_i} \sim N(\boldsymbol{\beta}, \boldsymbol{V}) \tag{27}$$

and $\boldsymbol{u}_i$ is a vector of baseline covariates with associated coefficients $\boldsymbol{\delta}$. By including our trajectory function, $W(t)$, in the linear predictor of our survival model, multiplied by an association parameter α, we use the simulation algorithm described in Section 5.3 to directly simulate survival times under a joint model framework [26].

$$h(t) = h_0(t)\exp\{\alpha W(t) + \boldsymbol{\phi}\boldsymbol{v_i}\} \tag{28}$$

where $h_0(t)$ is our user defined baseline hazard function, $\boldsymbol{v_i}$ is a vector of baseline covariates with associated coefficients $\boldsymbol{\phi}$.

Following the simulation of survival times, we can then construct any measurement schedule we desire for the longitudinal outcome, using Equation (26), and subsequently calculate the observed longitudinal measurements. To complete the joint model framework, measurement error in the longitudinal outcome can be incorporated simply by drawing the observed longitudinal values from $N(W(t), \sigma_e^2)$, where σ_e^2 is our measurement error variance. This example further illustrates the ease at which incorporating random covariate effects can be conducted. We include example Stata code in the appendix.

8. Delayed entry

It is becoming increasing popular in epidemiological contexts to use age as the timescale, as it is often an improved way of controlling for age than including it as a baseline covariate in a standard survival analysis [27]. Equation (17) can be extended to incorporate delayed entry (left truncation) using the following transformation

$$\int_{t_0}^{t} h(u)\mathrm{d}u = \frac{t-t_0}{2}\int_{-1}^{1} h\left(\frac{t-t_0}{2}z + \frac{t+t_0}{2}\right)\mathrm{d}z \approx \frac{t-t_0}{2}\sum_{i=1}^{n} w_i h\left(\frac{t-t_0}{2}z_i + \frac{t+t_0}{2}\right) \tag{29}$$

where t_0 denotes the time of entry. Equation (29) can be used as before to simulate survival times, accounting for a user specified entry time. We provide example code in the appendix (Example 7), where the user generates entry times from a normal distribution, N(30, 3), to represent age at entry. We can of course include any combination of time-dependent effects and time-varying covariates within our defined hazard function, $h(t)$.

9. Simulating a censoring distribution

In the previous examples we have assumed an administrative censoring time, i.e. maximum follow-up time that each patient can be observed. In practice, we often observe intermittent censoring, which we may wish to simulate. All of the scenarios and techniques described above can be used to generate censoring times. By simulating a set of event times and a second set of censoring times, for each patient, we can simply take the minimum to obtain the observed survival time, and consequently the event indicator. Furthermore, by making our censoring distribution dependent on covariates (be they baseline covariates, with time-dependent effects, or time-varying), we can incorporate informative censoring [11]. Alternatively, we could simulate our survival times, and then use draws from a uniform distribution between the minimum and maximum follow-up times to define a censoring fraction.

10. Discussion

We have described a general framework for the generation of survival data, incorporating any combination of complex baseline hazard functions, time-dependent effects, time-varying covariates, delayed entry, random effects and covariates measured with error. This centres on scenarios where we can't define the simulated survival time in a closed form expression.

Previous work in the simulation of survival includes using standard baseline distributions, such as the exponential, Weibull and Gompertz, with time-invariant covariates [3]. Mackenzie and Abrahamowicz described techniques to allow for time-dependent effects, and allowed specification of the marginal distribution of event times and covariate distributions [28]. A recent paper by Austin provided closed form expressions to incorporate 3 types of time-varying covariates, which built on work by Leemis et al [2], who described techniques to invert the cumulative hazard function with a single time-varying covariate. Furthermore, Sylvestre and Abrahamowicz describe two algorithms (permutation based and binomial model based) to generate survival times with time-varying covariates [29]. Finally, Royston [30] provided a method to simulate from parametric models that use restricted cubic splines on the log cumulative hazard scale.

Our approach relies on numerical integration to evaluate analytically intractable hazard functions. In our experience, 15 to 20 Gauss-Legendre quadrature points is often sufficient to provide accurate generation of survival times; however, we use 30 nodes as computation time is minimal. As with any estimation method which utilises numerical techniques, the accuracy of the generation process can be assessed by defining a seed and changing the tolerance of the root finder, and/or the number of quadrature nodes, and establishing that the generated survival times do not change.

We have illustrated our simulation approach through a variety of simulation studies and examples. In particular, by simulating from a complex underlying distribution, we have shown that substantial bias can be observed in estimates of the log hazard ratio for a treatment effect, when fitting a standard Weibull proportional hazards model.

Although in this article we have extolled the benefits of simulating from distributions beyond the standard choices, it must be stated that in many settings a simpler distribution may be adequate. For example, if fitting Cox models under proportional hazards and only the hazard ratio is of interest, then the baseline distribution used is inconsequential and therefore a simpler distribution should take preference. However, as described above, if evaluating parametric methods or incorporating time-dependent effects, then using a more complex distribution can provide much more realistic scenarios in order to fully assess the methods being evaluated.

To facilitate the use of the methods described in this article, user friendly Stata software has been developed [12, 13]. For each of the examples described in this article, we include example Stata code to simulate the survival times in the appendix. Further extensions not shown in this article include incorporating a cure proportion. This can be easily achieved by defining a mixture or non-mixture cure hazard function. This framework can also be applied in the generation of competing risks data, be it through cause-specific hazards or the approach of Beyersmann et al [31].

Given the inherent requirement of simulation studies to assess the statistical properties and performance of current and novel methods, we believe this framework can play an important role in allowing the generation of more biologically realistic survival data, incorporating much more complex scenarios. For example, the 2-component mixture distribution described in Section 5.1.1 has recently been used to simulate joint model data from a baseline (cumulative) hazard function, to assess the use of splines to capture complex baseline hazard functions [26]. Although we have concentrated on parametric survival models in this article, the framework is entirely applicable to examining the performance of the Cox model in any of the scenarios described [32].

Appendix

In this section we provide example code to generate survival times under each of the scenarios described in the article. In terms of notation used in the Stata code, `#t` is used to denote time in the user-defined (log) baseline hazard function, and colon operators are used to allow the use of Stata's matrix programming language, Mata. Further details can be found in the help file.

Example 1: standard parametric distributions with proportional hazards

We simulate 1000 survival times from either an exponential, Weibull or Gompertz baseline distribution, incorporating proportional effects of treatment and age, with log hazard ratios of -0.5 and 0.02, respectively.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a continuous age covariate
. gen trt = rbinomial(1,0.5)
. gen age = rnormal(65,12)
. //Simulate times from an exponential distribution
```

```
. survsim stime1, dist(exp) lambda(0.1) covariates(trt -0.5 age 0.02)
. //Simulate times from a Gompertz distribution
. survsim stime2, dist(gompertz) lambda(0.1) gamma(1.2) covariates(trt -0.5 age 0.02)
. //Simulate times from a Weibull distribution
. survsim stime3, dist(weibull) lambda(0.1) gamma(1.2) covariates(trt -0.5 age 0.02)
```

Example 2: 2-component mixture Weibull with proportional hazards

We simulate 1000 survival times from a 2-component mixture Weibull distribution, incorporating proportional effects of treatment and age, with log hazard ratios of -0.357 and 0.02, respectively. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a continuous age covariate
. gen treatment = rbinomial(1,0.5)
. gen age = rnormal(65,12)
. //Simulate times from a 2-component mixture Weibull with proportional hazards
. survsim time1, mixture lambdas(0.1 0.05) gammas(1 1.5) pmix(0.5)
> covariates(treatment -0.5 age 0.01) maxtime(5)
Warning: 444 survival times were above the upper limit of 5
         They have been set to 5 and can be considered censored
         You can identify them by _survsim_rc = 3
```

Example 3: User defined log hazard function using fractional polynomials

We simulate 1000 survival times from a user-defined log hazard function using fractional polynomials of time, incorporating proportional effects of treatment and age, with log hazard ratios of -0.5 and 0.02, respectively. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a continuous age covariate
. gen trt = rbinomial(1,0.5)
. gen age = rnormal(65,12)
. //Simulate times from a user-defined log hazard function
. survsim stime event, loghazard(-18 :+ 7.3:*#t:-11.5:*#t:^(0.5):*log(#t) :+ 9.5:*#t:^0.5)
> maxt(5) nodes(30) covariates(trt -0.5 age 0.02)
Warning: 68 survival times were above the upper limit of 5
         They have been set to 5 and can be considered censored
         You can identify them by _survsim_rc = 3
```

Example 4: User defined log hazard function using fractional polynomials, with a complex time-dependent treatment effect

We simulate 1000 survival times from a user-defined log hazard function using fractional polynomials of time, incorporating a proportional effect of age = 0.02, with a diminishing treatment effect. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a continuous age covariate
. gen trt = rbinomial(1,0.5)
. gen age = rnormal(65,12)
. //Simulate times from a user-defined log hazard function
. survsim stime event, loghazard(-18 :+ 7.3:*#t:-11.5:*#t:^(0.5):*log(#t) :+ 9.5:*#t:^0.5)
> maxt(5) nodes(30) cov(trt -0.7 age 0.02) tde(trt 1) tdefunc(0.01:*#t :+ 0.4:*log(#t))
Warning: 43 survival times were above the upper limit of 5
         They have been set to 5 and can be considered censored
         You can identify them by _survsim_rc = 3
```

Example 5: User defined log hazard function using fractional polynomials, with a binary time-varying covariate

We simulate 1000 survival times from a user-defined log hazard function using fractional polynomials of time, incorporating a proportional effect of age = 0.02, a binary time-varying treatment group and a binary disease

severity indicator. Treatment switching is dependent on disease severity. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a bad prognosis indicator
> variable
. gen trt = runiform()>0.5
. gen badprog = runiform()<0.4
. gen age = rnormal(65,12)
. //Generate an indicator variable for patients who swap treatment group, dependent
> on bad prognosis
. gen sp = cond(badprog==1,runiform()<0.4,runiform()<0.2)
. //Generate the time of swapping treatment group
. gen tswap = cond(sp==1,4.5*runiform() + 0.5,5)
. //Define the treatment effect
. local trteffect = -0.5
. survsim test1 event1, loghazard(-2.3:+2:*#t:-#t:^(2):+0.12:*#t:^3
> :`trteffect´:*((#t:<=tswap):*trt :+ (#t:>tswap):*(1:-trt))) maxt(5) nodes(30)
> cov(badprog `=log(1.5)´ age 0.02)
Warning: 73 survival times were above the upper limit of 5
         They have been set to 5 and can be considered censored
         You can identify them by _survsim_rc = 3
```

Example 6: User defined log hazard function using fractional polynomials, with a continuous time-varying covariate

We simulate 1000 survival times from a user-defined log hazard function using fractional polynomials of time, incorporating a continuous time-varying biomarker. We further construct the observed biomarker values incorporating measurement error. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and a continuous age covariate
. gen trt = runiform()>0.5
. gen age = rnormal(65,12)
. //Define the association between the biomarker and survival
. local alpha = 0.25
. //Generate the random intercept and random slopes for the longitudinal submodel
. gen b0 = rnormal(0,1)
. gen b1 = rnormal(1,0.5)
. survsim stime event, loghazard(-2.3:+2:*#t:-#t:^(2):+0.12:*#t:^3 :+ `alpha´ :* (b0 :+ b1 :* #t))
> maxt(5) nodes(30) mint(0.03) cov(trt -0.5 age 0.02)
Warning: 45 survival times were above the upper limit of 5
         They have been set to 5 and can be considered censored
         You can identify them by _survsim_rc = 3
. //Generate observed biomarker values at times 0, 1, 2, 3 , 4 years
. gen id = _n
. expand 5
. bys id: gen meastime = _n-1
. //Remove observations after event or censoring time
. bys id: drop if meastime>=stime
. //Generate observed biomarker values incorporating measurement error
. gen response = b0 + b1*meastime + rnormal(0,0.5)
```

Example 7: User-defined log hazard function with delayed entry

We simulate 1000 survival times from a user-defined log hazard function using fractional polynomials of time, conditional on a vector of entry times, incorporating a proportional effect of treatment = -0.5. Administrative censoring is assumed at 5 years through use of the `maxt()` option.

```
. //Simulate 1000 survival times
. set obs 1000
obs was 0, now 1000
. //Generate a binary treatment group indicator and age at entry
. gen treatment = rbinomial(1,0.5)
```

```
. gen age_entry = rnormal(30,3)
. //Simulate times from a user defined log-hazard function, with delayed entry
. survsim time1 event, loghazard(0.01:*#t:^(-2) :- 8:*#t:^(-0.5))
> covariates(treatment -0.5) enter(age_entry) maxtime(50)
Warning: 16 survival times were above the upper limit of 50
         They have been set to 50 and can be considered censored
         You can identify them by _survsim_rc = 3
```

Acknowledgement

We are very grateful to an anonymous reviewer for comments which greatly improved the manuscript. Michael Crowther is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (DRF-2012-05-409). Paul Lambert performed part of this work while on study leave from the University of Leicester.

References

1. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med* 2006; **25**(24):4279–4292.
2. Leemis LM. Variate generation for accelerated life and proportional hazards models. *Operations Research* 1987; **35**(6):pp. 892–894. URL http://www.jstor.org/stable/171438.
3. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005; **24**(11):1713–1723.
4. Rashid I, Marcheselli L, Federico M. Estimating survival in newly diagnosed cancer patients: use of computer simulations to evaluate performances of different approaches in a wide range of scenarios. *Stat Med* 2008; **27**(12):2145–2158, doi:10.1002/sim.3178. URL http://dx.doi.org/10.1002/sim.3178.
5. Belot A, Abrahamowicz M, Remontet L, Giorgi R. Flexible modeling of competing risks in survival analysis. *Stat Med* 2010; **29**(23):2453–2468, doi:10.1002/sim.4005. URL http://dx.doi.org/10.1002/sim.4005.
6. Murtaugh P, Dickson E, Van Dam M G Malincho, Grambsch P. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology* 1994; **20**:126–134.
7. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata J* 2009; **9**:265–290.
8. Eloranta S, Lambert PC, Andersson TM, Czene K, Hall P, Bjorkholm M, Dickman PW. Partitioning of excess mortality in population-based cancer patient survival studies using flexible parametric survival models. *BMC Med Res Methodol* 2012; **12**(86), doi:10.1186/1471-2288-12-86. URL http://dx.doi.org/10.1186/1471-2288-12-86.
9. Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. Stata Press, 2011.
10. King NB, Harper S, Young ME. Use of relative and absolute effect measures in reporting health inequalities: structured review. *BMJ* 2012; **345**:e5774.
11. Siannis F, Copas J, Lu G. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics* 2005; **6**(1):77–91, doi:10.1093/biostatistics/kxh019. URL http://dx.doi.org/10.1093/biostatistics/kxh019.
12. Crowther MJ, Lambert PC. Simulating complex survival data. *Stata J* 2012; **12**(4):674–687.
13. Crowther MJ. SURVSIM: Stata module to simulate complex survival data. Statistical Software Components, Boston College Department of Economics 2011. URL http://ideas.repec.org/c/boc/bocode/s457317.html.
14. Schumacher M, Bastert G, Bojar H, Hübner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RL, Rauschecker HF. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *J Clin Oncol* 1994; **12**(10):2086–2093.
15. McLachlan GJ, McGiffin DC. On the role of finite mixture models in survival analysis. *Stat Methods Med Res* 1994; **3**(3):211–226.
16. Lambert PC, Dickman PW, Weston CL, Thompson JR. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *J Roy Statist Soc Ser C* 2010; **59**(1):35–55, doi:10.1111/j.1467-9876.2009.00677.x. URL http://dx.doi.org/10.1111/j.1467-9876.2009.00677.x.
17. Anderson PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New York, Springer, 1993.
18. Jann B. MOREMATA: Stata module (Mata) to provide various functions. Statistical Software Components, Boston College Department of Economics 2005. URL http://ideas.repec.org/c/boc/bocode/s455001.html.
19. Crowther MJ, Lambert PC. STMIX: Stata module to fit two-component parametric mixture survival models. Statistical Software Components, Boston College Department of Economics 2011. URL http://ideas.repec.org/c/boc/bocode/s457339.html.
20. Royston P, Sauerbrei W. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, 2008.
21. Stoer J, Burlirsch R. *Introduction to Numerical Analysis*. 3rd edn., Springer, 2002.
22. Jatoi I, Anderson WF, Jeong JH, Redmond CK. Breast cancer adjuvant therapy: time to consider its time-dependent effects. *J Clin Oncol* 2011; **29**(17):2301–2304, doi:10.1200/JCO.2010.32.3550. URL http://dx.doi.org/10.1200/JCO.2010.32.3550.
23. Lambert PC, Holmberg L, Sandin F, Bray F, Linklater KM, Purushotham A, Robinson D, Møller H. Quantifying differences in breast cancer survival between England and Norway. *Cancer Epidemiol* 2011; **35**(6):526–533, doi:10.1016/j.canep.2011.04.003. URL http://dx.doi.org/10.1016/j.canep.2011.04.003.
24. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, *et al.*. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009; **361**(10):947–957, doi:10.1056/NEJMoa0810699. URL

http://dx.doi.org/10.1056/NEJMoa0810699.
25. Austin PC. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Stat Med* 2012; **31**(29):3946–3958, doi:10.1002/sim.5452. URL http://dx.doi.org/10.1002/sim.5452.
26. Crowther MJ, Abrams KR, Lambert PC. Flexible parametric joint modelling of longitudinal and survival data. *Stat Med* 2012; **31**(30):4456–4471, doi:10.1002/sim.5644. URL http://dx.doi.org/10.1002/sim.5644.
27. Thiebaut ACM, Benichou J. Choice of time-scale in cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine* 2004; **23**(24):3803–3820, doi:10.1002/sim.2098. URL http://dx.doi.org/10.1002/sim.2098.
28. Mackenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping. *Stat Comp* 2002; **12**:245–252. URL http://dx.doi.org/10.1023/A:1020750810409.
29. Sylvestre MP, Abrahamowicz M. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Stat Med* 2008; **27**(14):2618–2634, doi:10.1002/sim.3092. URL http://dx.doi.org/10.1002/sim.3092.
30. Royston P. Tools to simulate realistic censored survival-time distributions. *Stata J* 2012; **12**(4):639–654.
31. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med* 2009; **28**(6):956–971, doi:10.1002/sim.3516. URL http://dx.doi.org/10.1002/sim.3516.
32. Cox DR. Regression models and life-tables. *J Roy Statist Soc Ser B* 1972; **34**(2):187–220.