CLINICAL INVESTIGATION

A new SAS macro for flexible parametric survival modeling: applications to clinical trials and surveillance data

Ron Dewar[1] & Iftekhar Khan*,[2]

[1]Surveillance & Epidemiology Unit, Cancer Care Nova Scotia 1278 Tower Road, Halifax B3H 2Y9, Canada

[2]Department of Applied Health Research, University College London, UK

*Author for correspondence: iftekhar.khan@ucl.ac.uk

Survival analysis is often performed using the Cox proportional hazards model. Parametric models are useful in several applications, including health economic evaluation, cancer surveillance and event prediction. Flexible parametric models extend standard parametric models (e.g., Weibull) to increase the flexibility of the shape of the hazard function. We present a new SAS® macro for implementing flexible parametric models with a similar functionality to that of Stata®, with examples using data from cancer surveillance and clinical trials. Results from SAS® were identical with similar computational time to Stata®. The flexible parametric approach to modeling survival data is shown to be superior to standard parametric methods. This SAS® macro will facilitate an increase in the use of flexible parametric models.

Keywords: cancer surveillance • clinical trials • economic evaluation • parametric survival analysis • Royston–Parmar • splines

The semiparametric Cox proportional hazards (PH) model has continued to dominate the analysis and reporting of survival data for over 40 years [1]. One reason is the simplicity of estimating the relationship between covariates and the hazard rate while not having to make (sometimes unjustifiable) assumptions about the baseline hazard rate. However, despite the widespread use of the Cox PH model, there are nevertheless limitations, particularly when the PH assumptions are violated [2].

Where modeling covariates whose effects may vary over time, or prediction of survival rates are of importance, the Cox PH model may have some limitations [3]. In practice the Cox PH model is used for estimating hazard ratios and little else. Where more information is required, the baseline hazard is required. For example, the behavior of the hazard function itself might be of medical interest [2,4,5]. Differences in practices between hospital sites may result in hazard functions whose shape varies and might explain differences in mortality rates. In health economic evaluation, prediction of survival proportions beyond the observed follow-up period of a clinical trial are often required (extrapolation) so that future health benefits over the entire life time of patients can be estimated [6,7]. Flexible parametric models can be useful to predict the target number of (death) events [8]. In clinical trials, sample sizes are estimated based on the target number of events required. Follow-up of death events is often an ongoing process until trial completion or when the target number of events required are reached. Current patterns of (death) events are used to predict when the trial is likely to be stopped or further recruitment closed. This is sometimes achieved using parametric models assuming Weibull distributions. It is plausible that flexible parametric methods can improve on the pattern of death events and consequently estimate when the target event rates might be achieved [8]. This way clinical trial logistics (e.g., reporting and staff recruitment) can be planned more accurately. Standard parametric models (e.g., Weibull) could be used in these applications, but not all will retain the useful assumption of PH or even fit the

FUTURE SCIENCE part of fsg

observed survival pattern well. Efficient estimation of a smooth survival curve in situations where there are non-PH or modeling using proportional odds can therefore be implemented using more flexible parametric methods.

Applications in cancer surveillance

In PH regression, the relationship between a prognostic factor and patient survival is summarized in the hazard ratio. The PH model is particularly useful for assessing the impact of covariates on patient survival experience at an 'average' level, but has no simple interpretation at an individual level. Parametric modeling of survival data offers a way of presenting results using understandable metrics and visual displays that can be easily interpreted for a wider audience.

Cancer registries increasingly collect important prognostic factors useful for oncologists, policy makers and others in order to make decisions on expected survival proportions for subgroups of patients. Calculating patient- and stratum-specific survival is not easily accomplished using Cox PH modeling. For example, calculating the survival probability at a specific time point after diagnosis, for an individual with a specific set of prognostic covariates may be particularly useful for policy makers and planners, especially as far as assessing the future costing of cancer treatments; this is difficult with the Cox PH model, especially when interactions with the time scale are present. The parametric modeling described here does facilitate such computations.

Currently, software for implementing flexible parametric models is restricted to the more commonly used Stata [2] software. R code is also available for implementing extensions of the Cox model is also available using the 'flexsurv' program from the comprehensive R archive network archives [9].

The SAS® software [10] is widely used in academia and industry and currently does not have a specific option to implement these useful models. We therefore introduce a suite of SAS macros with similar functionality to that available in the Stata program 'stpm2' [2] and postestimation command 'predict' that can be used to fit these models.

This paper will first discuss briefly aspects of parametric modeling, then, outline flexible parametric methods, followed by details of the technical notation. Following this, the computational algorithm used in SAS code will be provided and finally examples of using the SAS macro. Supplementary Data 1 describes the input SAS dataset and Supplementary Data 2 details parameters and macro call set up using the SAS code with further descriptions to help the user. The SAS code for the macro suite is available from the authors on request and/or is also available for download from details in the references [11] (Supplementary Data 3).

Parametric modeling

Survival analysis is often reported using three commonly used methods: Kaplan–Meier (KM) methods, Cox PH models and parametric modeling. The KM method does not involve modeling and estimates of survival probabilities are based on using a nonparametric method (product limit estimator) to estimate the survival function [12]. When it is of interest to relate survival with covariates, a semiparametric method using the Cox PH model is often used. The Cox PH model is often adopted when the probability distribution of the sampled survival times is unknown or it might be complicated to fit a model to the data.

The third method is the parametric approach where it is tentatively assumed that the probability distribution of survival times is known (e.g., assuming a Weibull or Gompertz distribution).

In the Cox PH model, the hazard function for patient i, is:

$$h_i(t) = \phi(X_i)\, h_0(t)$$

where $h_0(t)$ is the baseline hazard function and $\varphi(x_i)$ is a function of explanatory variables, x_i. The baseline hazard function is the hazard function with the covariates all held at a reference level. The baseline hazard is not estimated when fitting a Cox model. In the case of a clinical trial with a single explanatory variable (treatment) with two treatments (active and control), $h_i(t)$ is:

$$h_i(t) = h_0(t) \exp\,(X_i\beta) *$$

where x_i= 1 for the active treatment and x_i = 0 for the control. When the treatment effect is estimated, the $h_0(t)$ terms cancel out (through the partial likelihood).

When a parametric model is required and the PH model is appropriate, the Weibull function is one commonly employed model. In this case, the baseline hazard function for a Weibull model is given by $h_0(t)$:

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

where λ and γ are scale and shape parameters, respectively. Therefore, the hazard rate for an individual patient i, consist of two parts: the baseline hazard (i.e., $h_0[t]$) and the covariate function $\varphi(x_i)$. In the case where covariates effect the baseline hazard multiplicatively, as in the Cox PH model, the covariate function is:

$$\phi(X_i) = \exp\,(\Sigma\beta_i X_i)$$

Therefore, the Weibull hazard function for an individual i, is [11]:

$$h_i(t) = \lambda\gamma t^{\gamma-1} * \exp\,(X_i\beta_i + X_2\beta_2 + \ldots X_k\beta_k$$

Parametric models of the type above can be a useful starting point to model survival data and in many cases one of the standard parametric forms might be adequate.

However, not all parametric models are suited in situations when PH does not hold such as some accelerated failure time models (e.g., log-logistic). Moreover, when the PH assumption is reasonable, the simple parametric models may not capture the underlying shape of the hazard function. Where PH is not a reasonable assumption, but models on other scales (e.g., AFT/proportional odds) are adequate, time-dependent hazard ratios may be still of direct interest and the flexible parametric framework offers a means to estimate them. Therefore, a more flexible approach to modeling survival time might be required.

Flexible parametric modeling

Extensions to the Cox model have been proposed earlier. Abrahamowicz *et al.* [13] and Wyant and Abrahamowicz [14] used splines to model the baseline survival, including linear and nonlinear effects of covariates. Flexible parametric modeling methodology as expounded by Beck and Jackman [15] and Royston and Parmar [16] is based on using a ‘flexible’ polynomial function for the hazard, the Royston–Parmar (RP) model. This consists of several functions which are joined together at ‘knots’ in such a way that the overall fitted function is smooth. The idea can be compared with linear spline fitting, but instead of linear splines (which are polynomial functions in the first degree), third-degree polynomial curves are joined together to fit the observed data.

The general idea behind flexible parametric modeling involves joining pairs of data points using higher degrees of polynomial functions (e.g., quadratic for order two and cubic for order three, or fractional polynomials). One useful property of using splines is that the underlying functional form does not need to be known (Kruger) [17]. Splines can be used in the context of a Cox PH model to smooth the hazard function (Sleeper and Harrington) [18], however, in this paper we present the use of splines for fully parametric models. Although the Weibull model is a useful alternative to the Cox PH model, especially when the assumptions around distributions are reasonable, a more flexible approach (Royston and Lambert; Royston; Lambert *et al.*) [2,19,20] by directly modeling the (log cumulative) baseline hazard function $h_0(t)$ as a polynomial function has shown to be a versatile approach to fitting smooth survival functions.

Applications to cancer surveillance

In our application of the SAS code in cancer surveillance, we give examples of estimating relative survival and crude probability of death (CPD). These will be briefly discussed.

Relative survival

Relative survival addresses a question such as ‘what are the chances of surviving 5 years after diagnosis, in the hypothetical world where cancer is the only possible cause of death?’ Relative survival methods make use of population life tables, rather than cause of death data, which is often used in the estimation of ‘net’ survival, although both measures address the same question. Relative survival is of great importance when comparing patient outcomes between differing jurisdictions, or over time, when background mortality rates differ.

Crude probability of death

This measure addresses a slightly different question: ‘what are the chances of surviving 5 years after diagnosis in the real world where a patient may die of some other cause first?’. This implies a competing risks framework, where death due to cancer and death due to other causes are considered to be independent.

Both the above measures imply an excess hazard model, where the user specifies a background probability of the event derived from local life tables, matched on attained age (at the time of the death), sex, time period and any other determinants of general population mortality that are of interest and available. The estimation of CPD requires numerical integration, which can be accomplished in the post-estimation step that follows fitting a relative survival model. Cronin and Feuer [21] described the computations from relative survival estimated in a life table framework. Technical details of the estimation in the parametric modeling framework described here have been described by Lambert [20].

Notation & methods

The following notation based on Royston *et al.* [2] and Royston [19] is used for a brief exposition of the flexible parametric modeling approach. This is also the description provided in the use of stmp2 command in Stata.

The survival function S(t) for a Weibull distribution is:

$$S(t) = \exp(-\lambda t^{\gamma})$$

The hazard function is:

$$h(t) = h_0(t)\exp\left(X\beta\right)$$

The cumulative hazard function is therefore:

$$H(t) = H_0(t)\ \exp(X\beta) = \left(\int_0^t h_0(u)\,du\right)\exp(X\beta)$$

Transforming to the log cumulative hazard scale gives so that we get a linear function of log (natural logarithm) time:

$$\mathrm{Ln}\left\{H(t)\right\} = \mathrm{Ln}\left[-\mathrm{Ln}\left\{S(t)\right\}\right] = \mathrm{Ln}(\lambda) + \gamma \mathrm{Ln}(t)$$

Adding covariates gives: Ln $\{H(t|x_i)\}$ = Ln(λ) + γLn(t) +$X_i\beta$, where Ln(λ) + γLn(t) is the baseline log cumulative hazard function (covariates on an additive scale). A covariate function that is proportional on the hazard scale is also proportional on the cumulative hazard scale.

The hazard and survival functions are required for likelihood estimation of model parameters. Using these estimates, predictions can be made through the parametric functions. These predictions can be improved through the use of splines which are more flexible.

Restricted cubic splines

Splines are flexible parametric functions defined through piecewise polynomials. The points at which polynomials are joined together are called 'knots' which 'forces' the fitted function to have zero-, first- and second-order derivatives [2,19].

Restricted cubic splines can be fitted by creating K-1 derived variables for K nots, k_1, k_2,…. k_k. A restricted cubic spline can therefore be noted as:

$$S(x) = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \ldots + \gamma_{k-1} Z_{k-1}$$

With derivatives (termed basis functions) compute as:

$$Z_1 = x \text{ and } Z_j = (X - K_j)^3_+ - \Psi_j (X - K_j)^3 - (1 - \Psi_j)(X - K_j)^3$$
$$_+ \text{..for } j = 2, \ldots, K - 1$$
$$\text{and } \psi_j = (k_k - k_j)/(k_k - k_1)$$

Incorporating splines in flexible parametric models

A PH model on the log cumulative scale can be written (Royston) [19] as:

$$\mathrm{Ln}(H(t| X_1)) = \mathrm{Ln}\ (H_0(t)) + x_1\beta$$

The PH model using a (restricted) spline function of Ln(t), with knots k_0 is also: S{Ln(t)|γ,k_0).

Hence,

$$\mathrm{Ln}(H(tX_i)) = \eta_i = \gamma_0 + \gamma_1 + Z_{1i} + \gamma_2 + Z_{2i} + \gamma_3 + Z_{3i} + X_i\beta$$

$$= M(\chi, \gamma) + X_i\beta, \text{ where } M(\chi, \gamma) = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \gamma_3 Z_{3i}$$

Royston *et al.*^{2} in a similar notation, using [5] define the basis functions as:

$$z_j = x$$
$$z_j = (x - k_j)^3_+ - \lambda_j (x - k_{min})^3_+ - (1 - \lambda_j)(x - k_{max})^3$$
$$_+ \text{.. for } j = 2, \ldots, m + 1$$

Therefore, λ_j = (k_{max} - k_j)/(k_{max} - k_{min}), where k_{max} and k_{min} are called the external knots and k_1….k_m are internal knots stated prior to the final analysis.

The likelihood function for an uncensored observation is:

$$L = \frac{1}{t}\frac{dM(\chi, \gamma)\exp(q - \exp(q))}{d\chi}$$

For a censored observation, L = exp(-exp(q)), where q = Ln H(t|x).

Royston and Parmar [16] suggest that the starting values for solving [6] can be established by a fitting a Cox model that accounts for censoring, transforming the estimated survival function to the log cumulative hazard scale, and deriving estimates finally using linear regression. For time-dependent effects, interactions with the time scale and the covariates of interest can be specified.

The SAS macro estimates the parameters in [6] using a maximization of the full log (natural log) likelihood using a Newton–Raphson optimization search algorithm with ridging in the PROC NLMIXED procedure.

Computational algorithms in SAS

The computer program has been tested with SAS version 9.3 and 9.4. No previous SAS macro is available as far as we are aware for this type of analysis. The only example of SAS code we identified which might be in anyway similar was in a recent online PhD thesis where the objective was simulation of data in the context of a single knot spline model [22] The SAS code used in the thesis was clearly not meant to be generalized and moreover does not fit covariates nor is there evidence the output from the code is comparable to Stata.

The SAS code we present consists of four main macro programs: %sas_stset, %sas_stpm2, %predict and %rcsgen**.** The macros were written to mirror usage of the corresponding Stata commands [2] of the same name. The macro calls are reasonably straight forward to use, while programmed checking for consistency of parameter specifications will help users avoid many errors. The focus of the work on this macro was to facilitate fitting RP models in SAS. The stset analogue we introduce is only an introduction, and not intended to replicate all of the functionality in the Stata command. The SAS macro %sas_stpm2 is designed to allow for estimation in situations where there is late entry (for example, in so-called 'period' analysis, or an analysis where age at diagnosis is the time scale), but the programmer would have to manually code the appropriate structures in the dataset to do so. Example code in

Supplementary Data 1 is given. In this SAS program the only RP model available at present is on the log cumulative hazard scale. RP models are strictly a class of model with different link functions of the survival function. Further development to other link functions can be built in the current macro to extend to other link functions. The details of the macro calls are presented in Supplementary Data I and II.

%sas_stpm2 defines and fits the RP model using the standard dataset created by %sas_stset. This part of the macro fits models on the log cumulative hazard scale. The user supplies parameters to the macro in order to specify covariates, the number of degrees of freedom for knots in baseline hazard, baseline risk (for relative survival models), any variables which can be considered as time-varying covariates (TVC), degrees of freedom for TVC, indicators to turn off the intercept in a model, an indicator to turn off orthogonalization of spline variables and an indicator to turn off computation of baseline splines. As an alternative to specifying degrees of freedom for splines, knot locations (for both baseline and TVC splines) can also be specified. The term 'TVC' refers to interactions between the time scale and specified covariates. This is the nomenclature presented in the documentation around Stata's stpm2, and it is used here in the same vein. Definitions of variables whose values change over the course of follow-up are not encompassed in this software.

%predict computes estimates and confidence intervals for a variety of survival functions for a specified covariate pattern using estimates from the previous run of %sas_stpm2. The time points used for estimation can be either the actual time points in the standard dataset, or a user-supplied set of time points. This latter option is most useful when computing survival functions from large datasets. The function to be predicted at each point can be a cumulative measure (cumulative hazard or survival) or one of hazard, hazard difference or hazard ratio. Confidence intervals for the cumulative functions (cumulative hazard, survival) may be computed using direct analytical integration (as used in Stata and replicated in the SAS macro, using the %predict call), the methods suggested by Carstensen [4] or using bootstrap methods.

%rcsgen is used by %sas_stpm2 and %predict, but can also be called to generate restricted cubic splines for a continuous covariate for purposes of analysis. An example would be to analyze the effect of age as a continuous variable allowing for nonlinear effects.

Example applications

We now provide some examples of how to use the SAS macro. For each example, we explain the background, data used, macro call and results.

Example 1: application to a Phase III randomized clinical trial in lung cancer patients

Background

The data are from a published Phase III clinical trial in 670 UK non-small-cell lung cancer patients treated with erlotinib compared with placebo [23]. Nearly all, (98%) of the patients had died by the time of analyses (658 deaths). The objective is to fit a parametric survival curve to each treatment group to predict the survival pattern and compare the results from Stata and SAS software.

Data

In this analysis we use overall survival as the time to event variable OS_event as the censoring variable (using a value of 1 for deceased), trt as the treatment variable (coded as 1 or 0) and patient identifier (patientid). Hence only four variables were used as inputs for the SAS macro call.

SAS macro call

Starting with a dataset (named 'topic') that contains variables identified above, the following sequence of macros are called:

- %sas_stset(topic, os_event(1), os, patientid); **data set up**;
- %sas_stpm2(trt, scale = hazard, df = n); ** estimation of RP model*;
- %predict(surv, survival, at = trt:xx); **predicted survival rate*.

The first macro generates a standard dataset called '_events_' in the user's 'work' library. The second call describes the RP model to be implemented and performs estimation. 'df = n' refers to the number of spline variables to be computed. Setting df = 1, is considered be equivalent to a Weibull model [2]. Note that df = 1 implies 1 derived variable. The number of knots generated is df + 1, since default knots are placed at the minimum and maximum of the noncensored survival times (also called boundary knots). If df is greater than 1, then df - 1 internal knots are placed at locations such that approximately equal numbers of noncensored events are in each interval between knots.

The%predict macro in the third step estimates and saves the predicted survival proportions based on estimates from the previous execution of %sas_stpm2. The 'at = trt*:* xx'is the syntax for specifying a particular covariate pattern at which to predict survival rates ('at = trt:1'for erlotinib for example).

Table 1. Estimates of coefficients and hazard ratios from analysis programs run in Stata and SAS.

Estimated parameter	Cox PH	Weibull[†]	RP(1)[‡]	RP(3)[§]
HR(SE) [Stata]	0.95 (0.08) [0.95 (0.08)]	0.93 (0.09) [0.93 (0.09)]	0.93 (0.09) [0.93 (0.9)]	0.95 (0.08) [0.94 (0.08)]
Lower 95% [Stata]	0.82 [0.82]	0.78 [0.78]	0.78 [0.78]	0.81 [0.81]
Upper 95% [Stata]	1.11 [1.11]	1.11 [1.11]	1.11 [1.11]	1.09 [1.09]
Predicted 6 month survival[¶] [Stata]		40.1 vs 38.4 [40.1 vs 38.4]	39.5 vs 37.2 [39.5 vs 37.2]	34.0 vs 32.2 [34.0 vs 32.2]
AIC	7340	2238	2238	2176

Stata results are presented in square brackets.
[†] Using PROC LIFEREG in SAS.
[‡] RP(1): flexible parametric using one knot.
[§] RP(3): flexible parametric, Royston–Parmar (RP) model using three knots.
[¶] Observed 6 month survival rates (Kaplan–Meier) for erlotinib versus placebo were 36.3 versus 31.9%.
AIC: Aikakes Information Criteria; HR: Hazard ratio; PH: Proportional hazard; RP:Royston–Parmar; SE: Standard error.

Results

Results from SAS are identical (Table 1) to the Stata output, which confirms the SAS macro is performing as expected. The empirical survival curve (Figures 1 & 2) is also plotted along with the Weibull and RP(3) model. The KM (black solid line) is approximated well by the RP (dotted line) with three knots. The Weibull (dashed line) is a slightly worse fit.

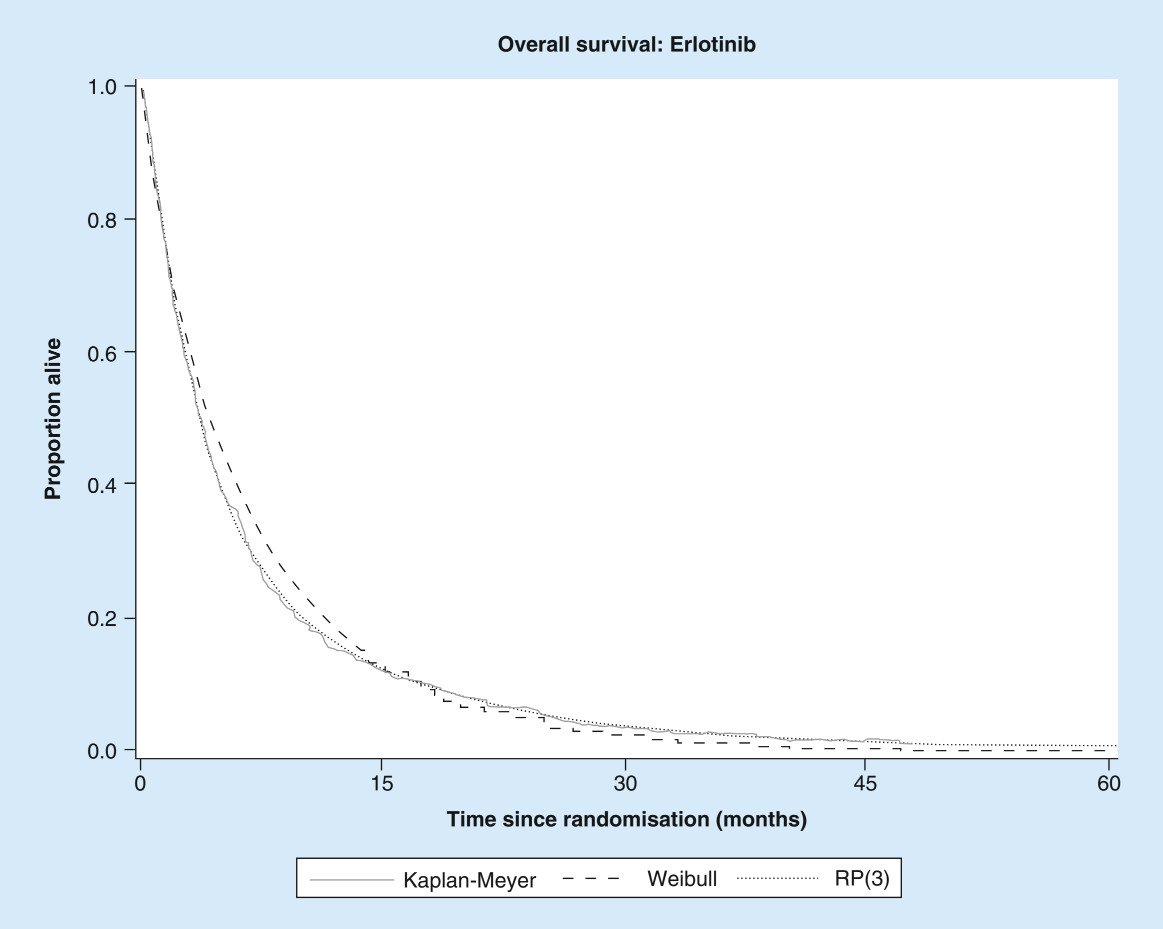

Figure 1. Comparison of predicted survival rates from Weibull and flexible parametric model applied to data from the TOPICAL trial using SAS macro (treatment arm).
RP:Royston–Parmar.

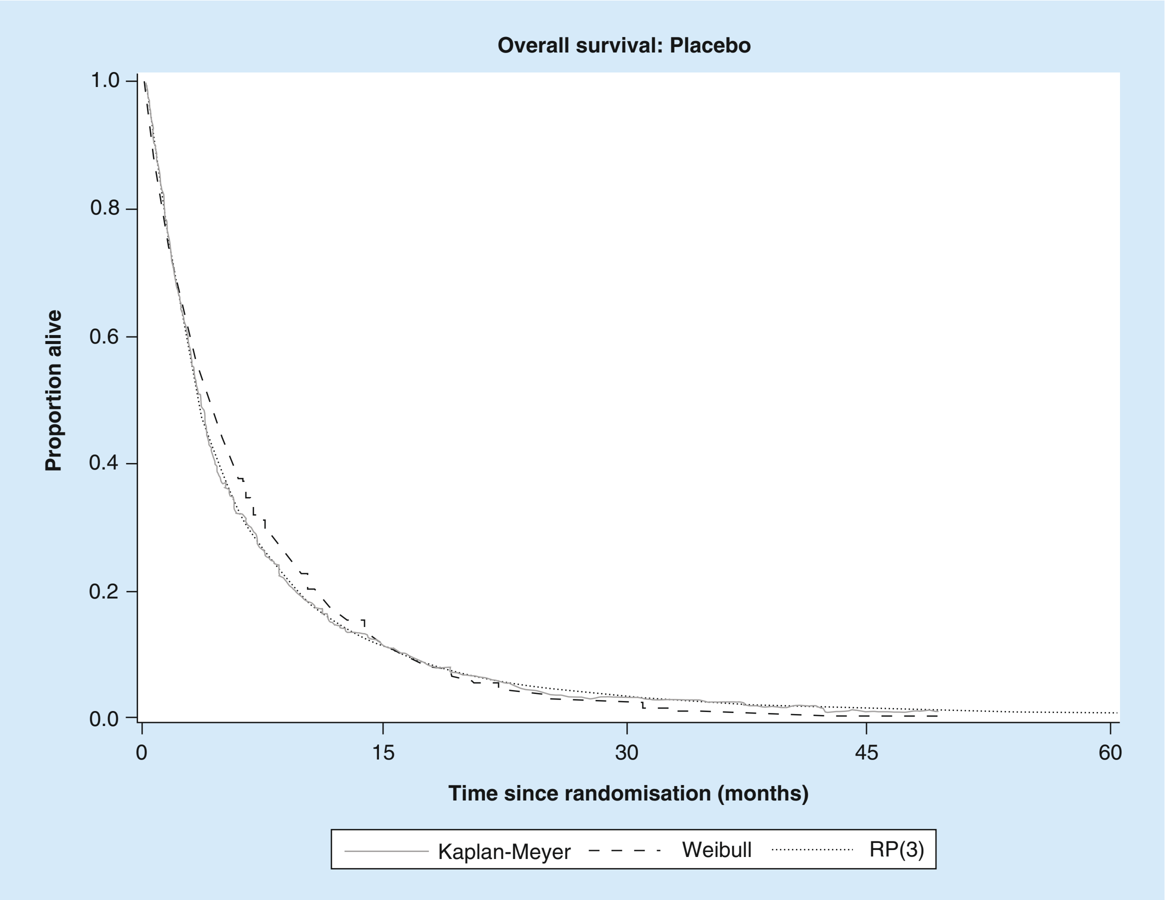

Figure 2. Comparison of predicted survival rates from Weibull and flexible parametric model applied to data from the TOPICAL trial using SAS macro (placebo arm).
RP:Royston–Parmar.

Interestingly, the RP(3) offers a 'better' fit than RP(1) (AIC smaller). The Weibull is a commonly employed model to compute the mean survival time (area under the survival curve) in economic evaluations for calculating quality adjusted life years [7,16]; in this example the mean survival time for erlotinib versus placebo were 6.95 versus 6.53, 6.96 versus 6.47 and 7.05 versus 6.62 months for the KM Weibull and RP(3) models, respectively. The Weibull model therefore overestimates mean survival time and hence quality adjusted life years (assuming quality of life is the same between groups) in this example (difference of 0.49 vs 0.43 for Weibull and RP(3), respectively.

Example 2: application to cancer surveillance

Background

In the field of cancer surveillance, policy makers and planners are often interested in measures of the impact of a diagnosis of cancer. Here, we apply the SAS macro to a Canadian cancer population (size under one million) to compute examples of RS and CPD.

SAS macro calls

The SAS code used to fit a relative survival model starts by appending the expected mortality rate to each subject, given that subject's sex, attained age and year of death (or censoring).

Relative survival

The input dataset CRC_deaths is first created with the variables required for analysis (age, sex and stage at diagnosis, details not shown), plus the general population probability of death (given attained age, sex and year of death):

- Data CRC_deaths:
- Merge CRC;
- Life_1991_2011 (keep = age sex death_year rate);
- By age sex death_year; run
- We then create the standard dataset that is required for the model fitting and estimation steps:

Table 2. Estimates of net and crude probability of death from SAS and Stata.

Estimates for	Stage	NPD (%)		CPD (%)			
				Cancer		Other causes	
		SAS	Stata	SAS	Stata	SAS	Stata
Male (55 years)	Stage I	5.8	5.8	5.7	5.7	3.3	3.3
	Stage II	9.7	9.7	9.5	9.4	3.2	3.2
	Stage III	24.1	24.1	23.6	23.6	3.0	3.0
	Stage IV	87.6	87.6	86.7	86.7	1.2	1.2
Male (85 years)	Stage I	20.7	20.7	18.3	18.2	39.9	39.9
	Stage II	32.7	32.7	26.7	26.5	37.5	37.4
	Stage III	65.7	65.7	52.0	51.9	29.6	29.6
	Stage IV	100.0	100.0	95.3	95.1	4.3	4.3

CPD: Crude probability of death; NPD: Net probability of death;

- %sas_stset(CRC_deaths, censor(0), surv, patient_ID);

- This is followed by executing the macro:

- %sas_stpm2 (sex stage2 stage3 stage4 agercs1 agercs2 agercs3, scale = hazard, df = 3, tvc = sex stage2 stage3 stage4, dftvc = 2, bhazard = rate);

which specifies covariates for sex, stage (as three indicator variables) and age. The above call requests to fit a model with sex, age (as a set of three spline variables) and stage at diagnosis, with stage I as a reference level. The TVC option has also been used, allowing the shape of the hazard functions for sex and stage to vary over time in a nonproportional way. The bhazard parameter is used to specify the life table probability, making this a relative survival or excess hazard model.

To compute the estimated survival curve for age 55 years, with sex and stage at the reference levels, we first compute the specific values of the age splines by a call to%rcsgen:

%rcsgen(age_yrs, gen = agercs, knots = &age_knots., tmatrix = age_mat, scalar = 55);

The age, knots and matrix to orthogonalize the computed splines have been saved from an earlier call to %rcsgen. The use of the %predict macro now provides survival estimates at the specified covariate values:

- %predict(Rt, survival, at = agercs1:. agercs2:. agercs3:. zero);

- The other covariates are held at their reference values (males, stage) by the use of the 'zero' keyword.

Results from the use of the %predict macro are interpreted as relative survival, excess hazard, excess hazard ratios, etc. and can be displayed graphically. Table 2 shows the estimated net probability of death (NPD, i.e., 1-relative survival) at 5 years, for male patients diagnosed at either 55 or 85 years of age, with colorectal cancer in Nova Scotia. A note of caution should be added that that relative survival is not an estimation of net survival even if on most situations the difference is small.

Estimates by stage of disease at diagnosis are presented. The SAS and Stata estimates are essentially identical.

Crude probability of death

Estimation of crude probability of death implies a competing risks framework, and requires knowledge of the background risk of death over the follow-up time. The cumulative background risk of death is estimated from local complete life tables, and the excess risk of

Table 3. Estimates of model fits and survival probabilities from Stata and SAS for SOCCAR for non-proportional hazards.

Survival (month)	Concurrent (%)			Sequential (%)		
	RP(2)	KM	Weibull	RP(5)	KM	Weibull
12	71	70	78	80	83	76
24	47	50	53	46	46	47
36	33	33	39	26	26	39

RP: Royston–Parmar; KM: Kaplan–Meyer.

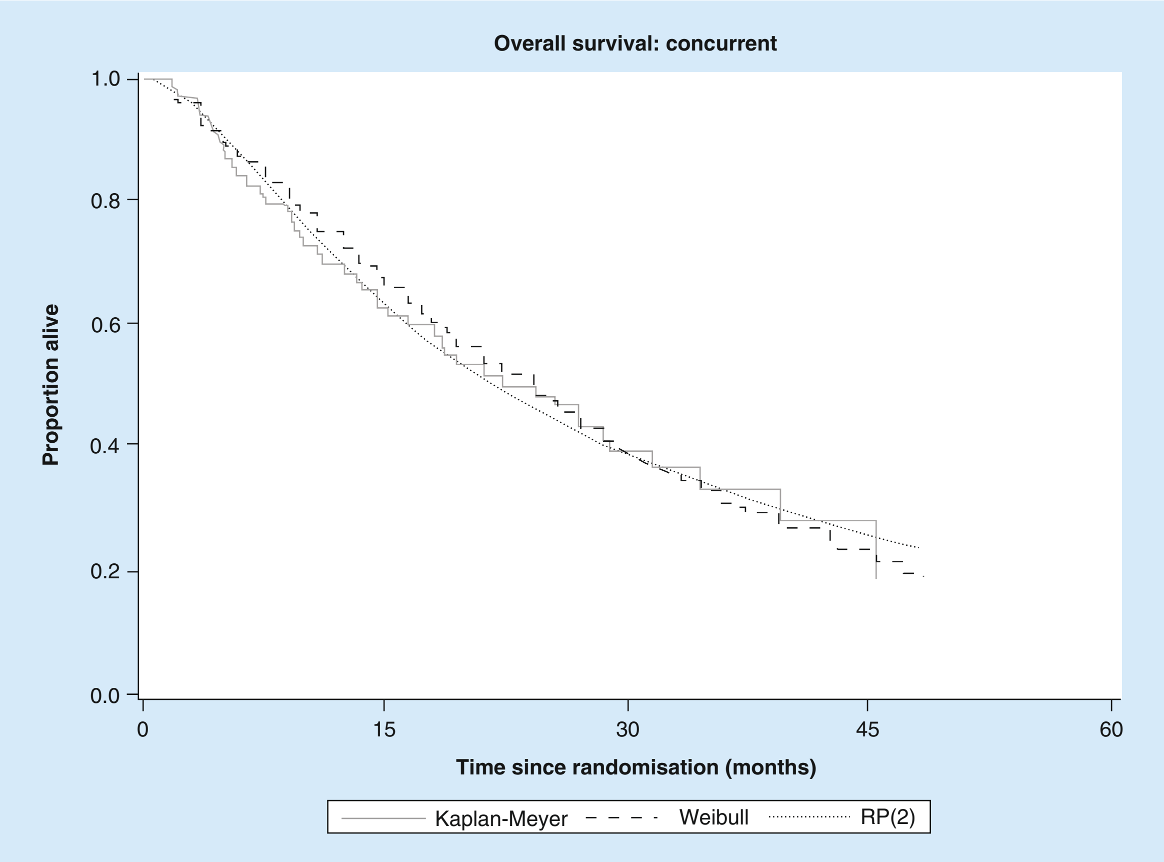

Figure 3. Royston–Parmar models fitted to SOCCAR data using SAS Macro.
RP:Royston–Parmar.

death is estimated from a net survival model, as above. Thus, the information needed to compute crude survival probabilities is the same as is required for relative survival. The details of the calculations are given in Lambert [20] and involve numerical integration using methods suggested by Carstensen [4] Estimates of the crude probability of death (due to cancer and due to all other causes) are presented in Table 2. Again, the SAS and Stata estimates are essentially identical.

For a male aged 55 years with stage III cancer, the NPD was about 24%; the risk of death from cancer alone was about 24% and from other causes about 3%. For an older male, the corresponding risks of death were 66, 52 and 30% for NPD and the two crude probabilities of death, respectively, highlighting the greater estimated impact of noncancer causes of death at older ages.

Example 3: application to a randomized Phase II cancer trial in lung cancer with non-proprotional hazards

The SOCCAR trial (Maguire *et al.*) [24] compared survival outcomes in non-small-cell lung cancer patients receiving concurrent versus sequential chemotherapy. In this example we show the model effect and fit when the KM curves cross and the PH assumption is violated. Table 3 shows the results from this analysis and compares with Stata.

The SAS macro was used to generate the following survival probabilities at each of 12, 24, 26 and 48 months. The results from SAS and Stata are identical.

Table 3 compares estimates of survival rates and model fits between SAS and Stata. RP(5) and RP(2) when fitted separately to each treatment group showed the closest estimates to the empirical KM curve (Figures 3 & 4). The AIC were identical between SAS and Stata (e.g., AIC = 325.1 and 139.5 for RP[2] and RP[5], respectively). Note that parameter estimates (not shown) cannot be compared between SAS and Stata because SAS and Stata orthogonalize splines in different ways. The Log Likelihood (-2LL) were also identical (-2LL = 312.6 and 127.8 for RP[2] and RP[5]), respectively. The above is an example of the TVC option to allow for a non-PH model to be evaluated.

Event prediction

One potentially useful feature of parametric survival modeling is the ability to predict (death) events by modeling the survival (hazard) time. This can be very

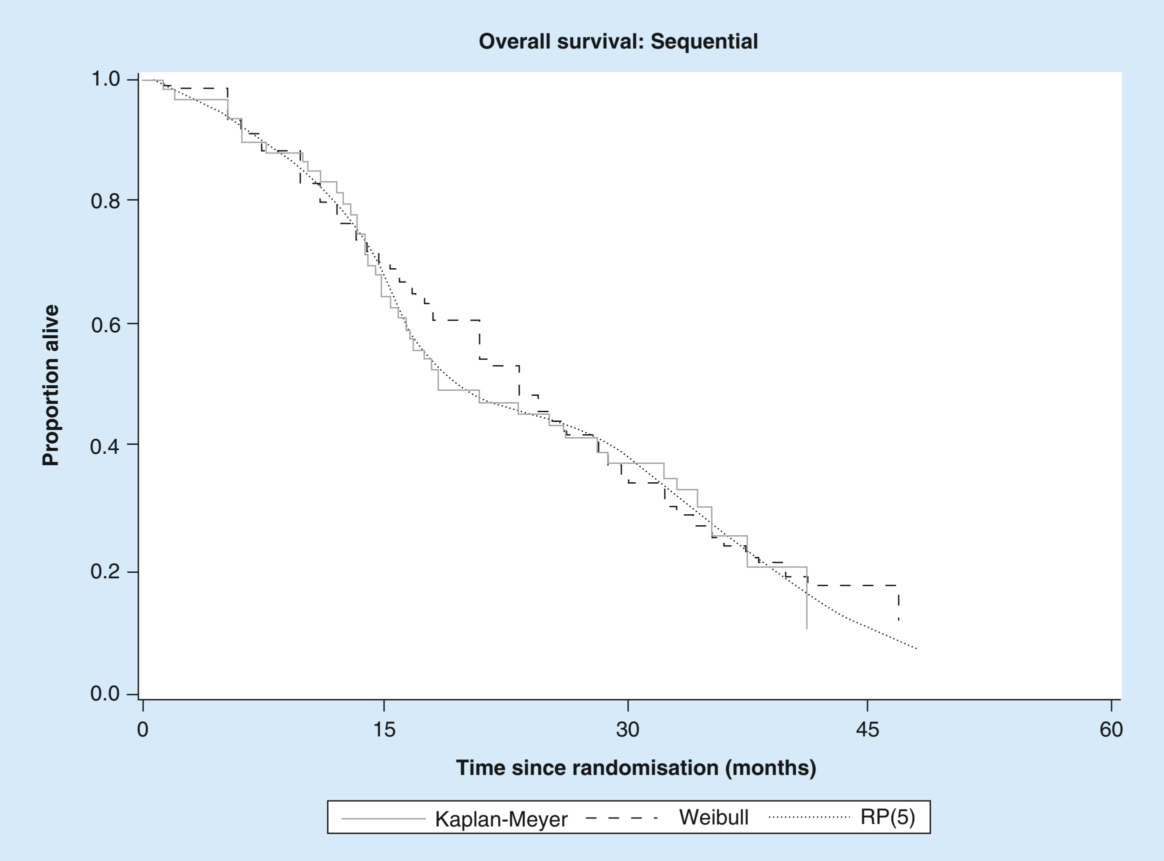

Figure 4. Royston–Parmar models fitted to SOCCAR data using SAS Macro.
RP: Royston–Parmar

useful in clinical trials because often the timing of when the last desired event occurs is a trigger for preparing for trial closure, reporting results and several other trial operating procedures such as when funding is likely to cease (as trial staff are needed for a longer period if event rates are slow).

Recent approaches to event prediction have used Weibull type models [8], however the potential of flexible parametric model to improve predictions of the survival more closely can be realized with the use of this macro for interested researchers. The use of extrapolation to estimate future health benefits (and costs) and subgroup analyses is supported by the National Institute of Clinical and Health Excellence as evidenced by the Decision Support Unit's technical document on approaches to extrapolation.

Conclusion

Availability of this SAS program will increase the use of applying flexible parametric models. We have shown results from SAS which are identical to that of Stata using various examples. In addition, we discussed how the code can be used for event prediction which can be very useful for clinical trial logistics and planning, particularly when the event rate is low or in rare cancers. A limitation to our code is that it is long (although the Stata codes are even longer) and might take some time for a user to become fluent with it. However, the actual macro call is short and reasonably straight forward. It is hoped that the availability of this SAS code will result in more widespread use and publication of flexible survival methods. The SAS code is available to download on request from the authors [10].

Future perspective

The field of flexible parametric modeling will increase in the next years especially in applications in oncology. As the cost of cancer care becomes more expensive, there will be a need to understand the benefits and costs associated with such treatments not just over the duration of a clinical trial, but over the survival period outside the trial. In addition, estimation of survival rates for specific groups of patients is essential for determining future policy direction. Flexible parametric models will also become more important as registry data and cancer surveillance data become more important for assessing real world evidence for new treatments. Consequently, this paper, through the use of an

SAS macro will facilitate further use and development in this area.

Acknowledgments

We acknowledge permission to use the data from Siow-Ming (University College Hospital, London) and J Maguire for use of the TOPICAL trial data. We also thank anonymous reviewers whose comments helped to improve the manuscript. Finally, we thank P Lambert for encouragement and explanations in an earlier form of this work.

Financial & competing interests disclosure

Some data used in this research were from the TOPICAL trial sponsored by CRUK Grant no. C1438/A4147 and from the SOCCAR trial, Grant no. C11922/A4558. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- The Cox proportional hazards model has dominated survival analyses methods for over 40 years.
- Parametric methods are being used increasingly.
- Flexible parametric models are shown to improve upon standard parametric methods for modeling survival type data. Such methods can have a wide area of application such as in clinical trials, health economics and cancer surveillance.
- A new SAS code is provided that will facilitate use of such methods.
- Data from two randomized trials and a cancer surveillance study were used as examples.
- The SAS code and output has been validated against Stata code and output and both provide identical results.
- This article should result in a wider use of flexible parametric models being used.

References

1 Laine T, Reyes Eric M. Survival estimation for cox regression models with time-varying coefficients using SAS. www.jstatsoft.org/

2 Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model.* Stata Press, TX, USA (2011).

3 Bellera CA, MacGrogan G, Debled M, Tunon de Lara C, Brouste V, Mathoulin-Pélissier S. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med. Res. Methodol.* 10(20), (2010).

4 Carstensen B. Demography and epidemiology: practical use of the Lexis diagram in the computer age or: who needs the Cox-model anyway? Annual meeting of Finnish Statistical Society 23–24 May 2005. http://publichealth.ku.dk/sections

5 Stocken D, Hassan AB, Altman DG *et al.* Modelling prognostic factors in advanced pancreatic cancer. *Br. J. Cancer.* 99(6), 883–893 (2008).

6 Khan I. *Design and analysis of Clinical trials for Cost–effectiveness and Reimbursement: an Applied Approach using SAS and Stata.* Chapman & Hall Press, Abingdon, Oxford, UK (2015).

7 Latimer N. NICE DSU Technical Support Document 14: Survival Analysis for Economic Evaluations Alongside Clinical Trials – Extrapolation with Patient – Level Data, report by the Decision Support Unit (2013). www.nicedsu.org.uk/NICE

8 Ying GS, Heitjan DF. Prediction of event times in the REMATCH trial. *Clin. Trials* 10(2), 197–206 (2013).

9 Documentation and installation instructions can be downloaded from https://cran.r-project.org/web/packages/flexsurv/flexsurv.pdf

10 SAS; SAS institute, Cary 100 Cary, NC 27513, United States

11 Full SAS code can be downloaded free from. www.researchgate.net/profile/Iftekhar_Khan5

12 Collett D. *Modelling Survival Data in Medical research (2nd Edition).* Chapman & Hall, Abingdon, Oxford, UK (2003).

13 Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat. Med.* 26, 392–408 (2007).

14 Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat. Med.* 33(19), 3318–3337 (2014).

15 Beck N, Jackman, S. Beyond linearity by default: generalized additive models. *Am. J. Pol. Sci.* 42(2), 596–627 (1998).

16 Royston P, Parmar MKB. Flexible proportional-hazards and proportionalodds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* 21, 2175–2197 (2002).

17 Kruger CJC. Constrained cubic spline interpolation for chemical engineering Applications (2004). www.korf.co.uk/spline.pdf

18 Sleeper LA & Harrington DP. Regression Splines in the Cox Model with application to covariate effects in liver disease. *J. Am. Stat. Assoc.* 85, 941–949 (1990).

19 Royston P. Flexible parametric alternatives to the Cox model. *Stata J.* 1(1), 1–28 (2001).

20 Lambert PC, Dickman PW, Nelson CP & Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statist. Med.* 29, 885–895 (2010).

21 Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Stat. Med.* 19(13), 1729–1740 (2000).

22 Hamid AH. *Flexible Parametric Survival Models with Time dependent covariates for right censored data [PhD thesis].* University of Southampton, UK (2012).

23 Lee SM, Khan I, Upadhyay S *et al.* First-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy (TOPICAL): a double-blind, placebo-controlled, Phase 3 trial. *Lancet Oncol.* 13(11), 1161–1170 (2012).

24 Maguire J, Khan I, McMenemin R *et al.* SOCCAR: a randomised Phase II trial comparing sequential versus concurrent chemotherapy and radical hypofractionated radiotherapy in patients with inoperable stage III non-small cell lung cancer and good performance status. *Eur. J. Cancer* 50(17), 2939–2949 (2014).