Estimating a smooth baseline hazard function for the Cox model [1]

Patrick Royston
Hub for Trials Methodology Research
MRC Clinical Trials Unit and University College London
Aviation House
125 Kingsway
LONDON WC2B 6NH
UK.

13 September 2011

Summary

The aim of fitting a Cox model to time-to-event data is to estimate the effect of covariates on the baseline hazard function. The baseline hazard function, not itself estimated within the model, is the hazard function obtained when all covariate are set to zero. In several applications, it is important to have an explicit, preferably smooth, estimate of the baseline hazard function, or more generally the baseline distribution function. A key application is model-based prediction of survival probabilities in independent data. In particular, external validation of a Cox model should include an assessment of the calibration of prediction, and this requires an explicit baseline. We propose using simple parametric models (fractional polynomials and restricted cubic splines) to approximate the baseline log cumulative hazard function to a sufficient degree of accuracy. Other functions, including hazard and survival, are easily derived from the log cumulative hazard. We derive and compare estimated functions in two real datasets, one of prognostic factors in breast cancer, the other a randomized controlled trial in oesophageal cancer.

1 Introduction

The aim of fitting a Cox model to time-to-event data is to estimate the effect of covariates on the baseline hazard function. The baseline hazard function is not itself estimated. Depending on one's perspective, the approach can be viewed as an advantage—since there is no risk of misspecifying the baseline distribution—or a drawback—since the model is incompletely specified, with consequences for the interpretation of results and for subsequent users of the model.

As Royston and Altman [1] discuss, an explicit estimate of the baseline hazard function is useful in evaluating the performance of a Cox model in independent data, a process often known

[1] Research report No. 314, Department of Statistical Science, University College London. Date: Sept. 2011.

as external validation [2, 3]. An explicit estimate is essential when assesing the calibration of a Cox model in independent data. Calibration concerns prediction accuracy; a well-calibrated model accurately predicts survival probabilities. Van Houwelingen [4], when discussing external validation of a prognostic model, nicely states the case for an explicit baseline hazard for a Cox model as follows:

> It is the duty of the bio-statistician involved in reporting the prognostic model to give all the information needed to build further on their model. For Cox models that should also include the baseline hazard or survival rate, if possible smoothed somehow or given in an approximate functional form using (fractional) polynomials, exponentials, rational functions or something similar.

As van Houwelingen implies, a second important reason for wanting the baseline hazard is in reporting prognostic models in the literature. In general, the standard of reporting such models is low [5]. Without the baseline, an important component of the model is missing, and as a result it is not possible to report the entire model or to use the model to reproduce or predict survival curves for given covariate patterns.

We start from van Houwelingen's viewpoint [4] that it is useful to find a parametric estimate of the baseline hazard function, or more generally, the baseline distribution function, after fitting a Cox model. More specifically, we assume we have the prognostic index (linear predictor), that is the regression coefficients and a clear definition of the participating covariates and their coding and/or transformations. When reporting a baseline distribution, for example the baseline survival function, practicality and transportability to other settings clearly demand that the function be simple to express and to present. For example, a cubic smoothing spline [6] or a penalized spline function with many knots and many regression parameters [7] are too complex to present as mathematical formulae. As van Houwelingen [4] notes, this leaves us with some simple parametric functions. In the present paper, we consider linear functions, fractional polynomials and restricted cubic regression splines, all of which have relatively simple mathematical forms.

The question, then, is how to approximate the baseline. Well-known non-parametric estimates [8] of the baseline survival function, $S_0(t)$, and the baseline cumulative hazard function, $H_0(t)$, are available after fitting a Cox model in most software packages. It is in principle easy to differentiate the cumulative hazard function to obtain an estimate of the baseline hazard function, i.e. $h_0(t) = H_0'(t)$, but non-parametric estimates of $H_0(t)$ is too 'noisy' for this to be feasible. We first need a smooth estimate of $H_0(t)$. According to the reasoning given in References [9, 10], a practical approach is to model the log cumulative hazard as a flexible function of log time, starting with the simplest parametric model, the Weibull distribution, for which the linear relationship $\ln H_0(t) = \gamma_0 + \gamma_1 \ln t$ holds. Fruitful generalizations of the Weibull model involve replacing or supplementing the term in $\ln t$ with a non-linear function, such as a fractional polynomial (FP) of suitable degree or a restricted cubic regression spline (RCS) of suitable complexity. We consider both approaches here.

The structure of the paper is as follows. Section 2 briefly describes two datasets we use in examples. The first dataset is from a two-arm randomized controlled trial in oesophageal cancer. One motivation is to encapsulate the baseline distribution function in a simple form which can be used in planning future trials. The second dataset has been used to develop a prognostic model for recurrence of primary operable breast cancer [11] and to illustrate methodological issues in prognostic model development [12]. A transportable estimate of the baseline survival function is sought. Section 3 discusses several methods for estimating a smooth baseline function, and section 4 gives the results of applying the methods to the two example datasets. Section 5 is a discussion.

2 Data

OE02 is a randomized, controlled trial of preoperative chemotherapy in patients undergoing radical surgery for esophageal cancer [13]. Random assignment was to surgery alone (control arm) or to two cycles of combination cisplatin and fluorouracil therapy before surgery (experimental arm). There were 655 deaths in 802 patients recruited, with a hazard ratio (HR) favouring the experimental arm of 0.84 (95% CI, 0.72 to 0.98; $P = 0.03$). We seek a smooth estimate of the baseline distribution function in a Cox model including only the effect of treatment.

From July 1984 to December 1989, the German Breast Cancer Study Group (GBSG) recruited 720 patients with primary node positive breast cancer into a factorial 2 × 2 trial investigating the effectiveness of three versus six cycles of chemotherapy and of additional hormonal treatment with tamoxifen [14]. The dataset comprises recurrence-free survival (RFS) time of the 686 patients (with 299 events) who had complete data on several standard prognostic variables. The Cox model we use is prognostic model III proposed in Reference [12]. We refer the reader to the latter article for further details.

3 Methods

Our main aim is to produce a smooth, parsimonious parametric estimate of the baseline hazard, survival and cumulative hazard functions for a dataset on which a 'plain' Cox model, i.e. including neither stratification nor time-dependent effects, has been fitted. We assume that the model has covariates $\mathbf{x} = (x_1, \ldots, x_k)$ and regression parameters $\beta_1, \ldots, \beta_k$. We write the Cox model for a typical observation as

$$h(t; \mathbf{x}) = h_0(t) \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \tag{1}$$

The intercept, β_0, is non-standard and is discussed below. For a clinical trial with two arms ('control' and 'experimental'), unadjusted for covariates, we set $\beta_0 = 0$ and take the control arm (coded $x_1 = 0$) to be the baseline group.

For a prognostic model, we manipulate β_0 to ensure that the baseline hazard function represents, loosely speaking, an 'average risk' curve by centring the linear predictor or prognostic index, $\widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$, to have mean zero over the n individuals in the dataset:

$$\widehat{\beta}_0 = -n^{-1} \sum_{i=1}^{n} \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_k x_{ki} = -\left(\widehat{\beta}_1 \overline{x}_1 + \cdots + \widehat{\beta}_k \overline{x}_k\right)$$

Let $\eta = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$ be the centred prognostic index.

3.1 Ordinary least squares estimation of baseline distribution functions

The usual estimates of the cumulative hazard and related functions obtained after fitting model (1) are high-dimensional and unsuitable for presentation as a simple formula. A rough initial parametric approximation to the log baseline cumulative hazard function can be found by assuming the times-to-event follow a Weibull distribution, conditional on covariates. Let $H(.)$ denote the cumulative hazard function and $z_0(t) = \ln H_0(t) = \ln H(t; \mathbf{x} = \mathbf{0})$ be the log baseline cumulative hazard function. For the Weibull model, we have

$$z_0(t) = \ln H_0(t) = \gamma_0 + \gamma_1 \ln t \qquad (2)$$

The parameters γ_0 and γ_1 are transformations of the scale and shape parameters of the distribution. Ordinary least squares (OLS) estimates, $\widetilde{\gamma}_0$ and $\widetilde{\gamma}_1$, may be obtained by linear regression of the 'noisy' estimate of $\ln H_0(t)$ on $\ln t$. Let $\widetilde{z}_0(t) = \widetilde{\gamma}_0 + \widetilde{\gamma}_1 \ln t$.

For any continuous distribution with support on $t > 0$, the baseline hazard function, $h_0(t)$, may be obtained from $z_0(t)$ and its derivative with respect to $\ln t$ as follows:

$$\begin{aligned} h_0(t) &= \frac{dH_0(t)}{dt} = \frac{d \ln H_0(t)}{d \ln t} \frac{d \ln t}{dt} \frac{dH_0(t)}{d \ln H_0(t)} \\ &= t^{-1} \frac{dz_0(t)}{d \ln t} \exp\left[z_0(t)\right] \end{aligned}$$

For the Weibull baseline hazard function with parameters fitted by OLS, we have

$$\widetilde{h}_0(t) = t^{-1} \widetilde{\gamma}_1 \exp\left(\widetilde{\gamma}_0 + \widetilde{\gamma}_1 \ln t\right)$$

There is of course no particular reason why a Weibull model should fit the data. If it does not, a better-fitting model may be sought by extending the linear function (2) for the log cumulative baseline hazard to include non-linear terms in t and/or $\ln t$. This is the starting-point for the class of flexible parametric proportional hazards models described by Royston and Parmar [9, 10]. Obviously many extensions are possible. We consider FPs in t (which includes

$\ln t$ as a special case) and RCS functions in $\ln t$ as being simple enough for practical purposes, yet sufficiently flexible to represent many distributional shapes.

For FPs, we consider models in t of degree 1, 2 or 3, with (approximate) dimension of 2, 4 or 6 d.f. respectively. Spline functions with broadly comparable flexibility are restricted cubic splines with 1, 2 or 3 interior knots, corresponding to 2, 3 or 4 d.f. respectively. Brief descriptions of FP and RCS functions are given in the Appendix.

3.2 Maximum likelihood estimation of baseline distribution functions

Consider a model of the type just discussed. We generalize the Weibull model such that the linear function $z_0(t) = \gamma_0 + \gamma_1 \ln t$ is replaced with a more flexible form, either an FP of degree m, $z_0(t) =$ FP$m(t)$ or an RCS, $z_0(t) = s(\ln t)$. The linear predictor, η, estimated by the Cox model is regarded as given. The log cumulative hazard function $z(t; \mathbf{x})$ is

$$z(t; \mathbf{x}) = z_0(t) + \eta \tag{3}$$

Thus η is an 'offset' from $z_0(t)$. The log-likelihood contribution l_i of the ith observation in model (3) is [9]

$$l_i = \begin{cases} -\ln t_i + \ln \frac{dz_0(\ln t_i)}{d\ln t_i} + z(t_i; \mathbf{x}_i) - \exp z(t_i; \mathbf{x}_i) & \text{for an uncensored observation,} \\ -\exp z(t_i; \mathbf{x}_i) & \text{otherwise.} \end{cases} \tag{4}$$

The total log likelihood for the model is $l = \sum_{i=1}^{n} l_i$. For a given dataset, the term $\sum(-\ln t_i)$, where the sum is over the uncensored observations, is independent of the parameters to be estimated. For consistency with log likelihoods reported by standard software, e.g. the `streg` command in Stata, we omit the constant $\sum(-\ln t_i)$ from l.

The parameters of $z_0(t)$ for the different formulations of this function may be estimated by maximum likelihood using standard optimisation tools. For the RCS formulation, the Stata command `stpm2` [15], either with the `offset()` option or by constraining the regression coefficient of η to 1, is suitable. MLEs of the exponents $p_1, \ldots, p_m$ in an FPm function for $z_0(t)$ are found by systematic search among all combinations of the exponents in the restricted set S (see Appendix). Conditional on the exponents, $z_0(t)$ is linear in FP transformation(s) of t. A Stata program (`stpmfp`) for the FP case is available from the author on request.

3.3 Selecting a model for the log baseline cumulative hazard

When $z_0(t)$ is to be modelled as an RCS function, Royston and Parmar [9] suggested selecting the number of knots for $z_0(t)$ that minimized the Akaike Information Criterion (AIC) for model (3). The AIC is defined as $-2l + 2\times$ d.f., where d.f. is the number of degrees of freedom of the RCS. A more stringent alternative tending to give more parsimonious models, the Bayesian Information Criterion (BIC), is defined for time-to-event data as $-2l+\ln(\text{number of events})\times$ d.f.

Model	d.f.	Fit by OLS			Fit by MLE		
		Deviance	**AIC**	**BIC**	**Deviance**	**AIC**	**BIC**
Weibull	1	2848.0	2850.0	2854.4	2817.7	2819.7	2824.2
FP1	2	2848.0	2852.0	2860.9	2785.4	2789.4	2798.3
FP2	4	2722.7	2730.7	$\boxed{2748.6}$	2709.5	2717.5	$\boxed{2735.4}$
FP3	6	2709.7	$\boxed{2721.7}$	2748.6	2703.8	$\boxed{2715.8}$	2742.7
Spline	2	2719.8	2723.8	2732.8	2714.7	2718.7	2727.6
Spline	3	2706.7	2712.7	$\boxed{\mathbf{2726.1}}$	2706.2	2712.2	$\boxed{\mathbf{2725.7}}$
Spline	4	2704.3	$\boxed{\mathbf{2712.3}}$	2730.2	2703.8	$\boxed{\mathbf{2711.8}}$	2729.8

Table 1: Oesophageal cancer data. Deviances, AIC and BIC for the baseline models considered. Boxed values represent the optimal models in their class. Additionally, bolding indicates the overall best models.

[16]. Practical experience suggests that BIC gives more satisfying results than AIC in the sense that the number of likely artefacts (‘wiggles’) in the fitted hazard function is lower. We report both criteria for a variety of d.f., with the models ranging from simple (1 or 2 d.f.) to relatively complex (4 to 6 d.f.).

A model for $z_0\,(t)$ estimated by OLS can be ascribed a log likelihood value via (4), hence its AIC and BIC can be calculated. By definition, such AIC and BIC values can never be smaller than those from the MLE of the same model.

4 Results

Before proceeding to detailed modelling, we examine graphically the fit of Weibull distributions for the baseline cumulative hazard and survival functions in the example datasets (see Figure 1). [FIGURE 1 NEAR HERE] A Weibull model evidently fits the breast cancer data slightly better than the oesophageal cancer data, but is not completely satisfactory. The fit to the oesophageal cancer data looks quite poor.

4.1 Oesophageal cancer trial

The deviances and related statistics for the various models fitted to the OE02 data, the only covariate being randomized treatment arm, are shown in Table 1. [TABLE 1 NEAR HERE] If we take minimal AIC as the optimization criterion, the best FP model is FP3 and the best spline model has 4 d.f. (3 interior knots). If minimizing by BIC, the selected models are FP2 and an RCS with 3 d.f. The discrepancy between the OLS and MLE fits is much larger for the

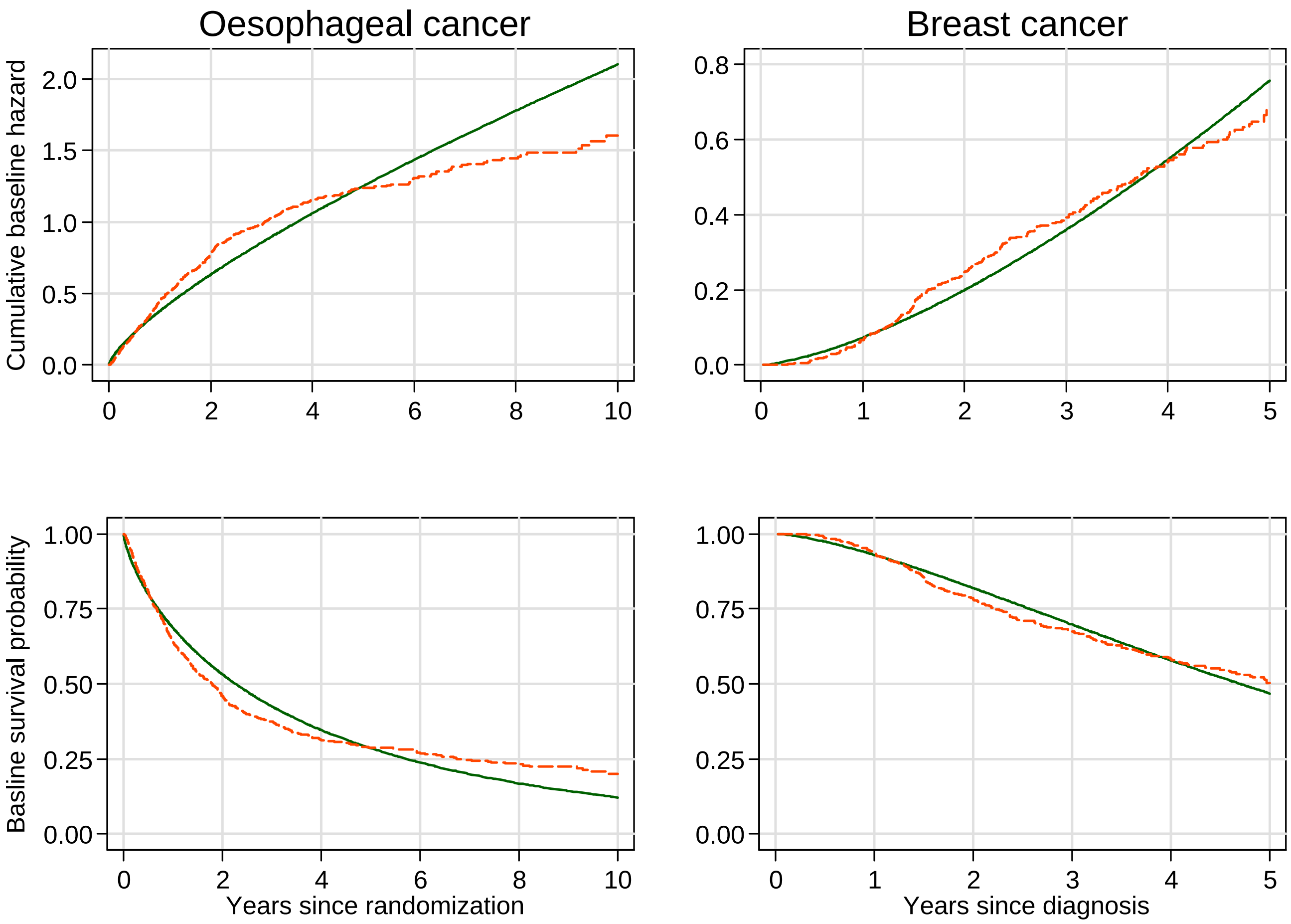

Figure 1: Oesophageal cancer and breast cancer datasets. Comparison of baseline cumulative hazard and survival functions for a Weibull model using the linear predictor estimated from a Cox model (solid lines) and non-parametrically after fitting the same Cox model (dashed lines).

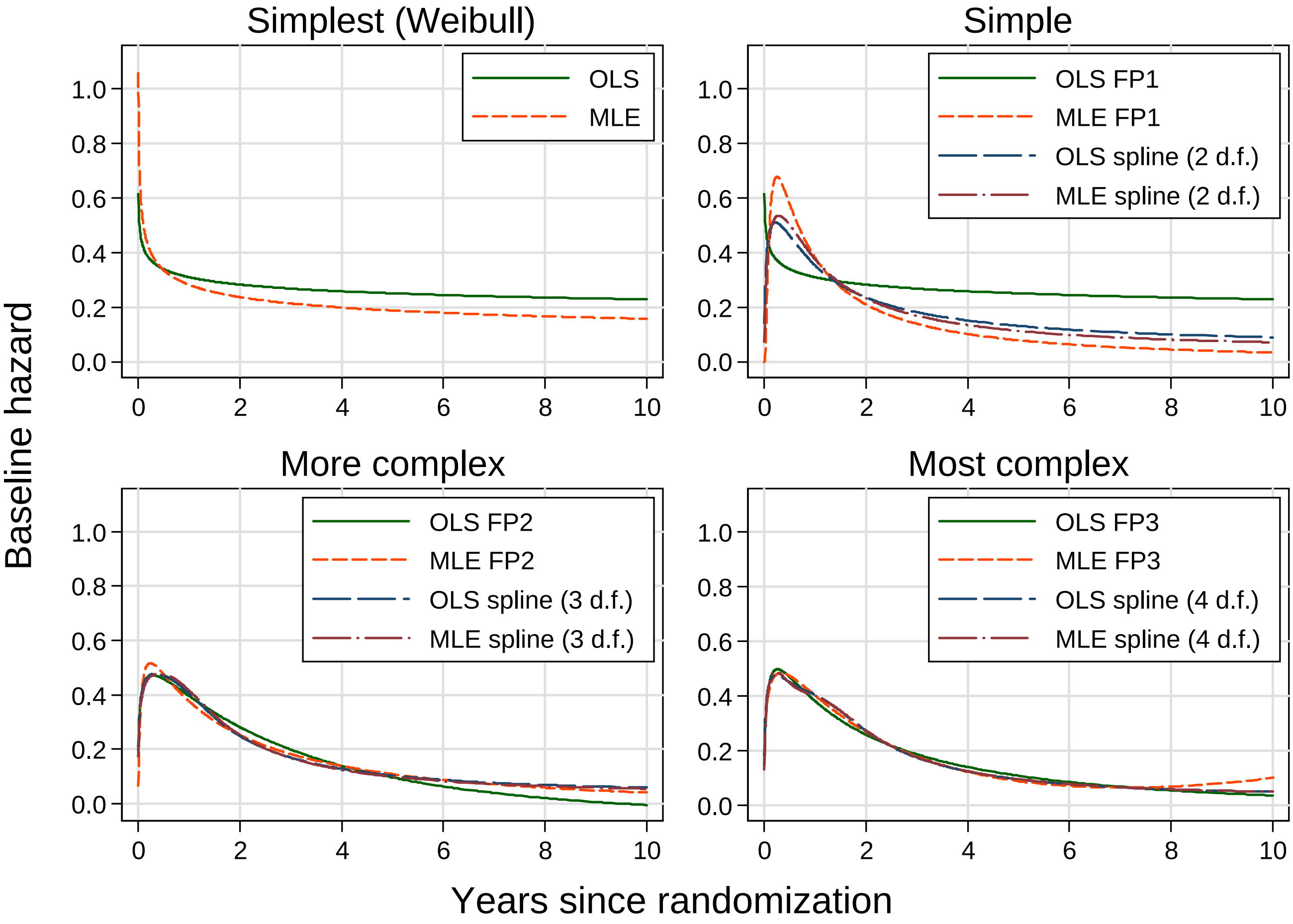

Figure 2: Oesophageal cancer data. Plots of the estimated baseline hazard function derived from all models considered.

FP models than for the spline models. Indeed, the OLS method sometimes selects the ‘wrong’ powers for the FP representing $z_0(t)$, particularly when the model is underfitted (FP1).

Estimates of the baseline hazard function derived from the models in Table 1 are shown in Figure 2. [FIGURE 2 NEAR HERE] Since the Weibull hazard function is monotonic in t, the model cannot capture the early peak in the hazard function seen with almost all the other models. The only other model that fails to display this feature is the FP1 curve fitted by OLS. The results for the simple models vary, whereas those for the more and most complex models agree closely.

Plots of various baseline functions from the BIC/MLE-selected RCS and FP models are shown in Figure 3. [FIGURE 3 NEAR HERE] The 3 d.f. RCS function provides an excellent fit. The FP2 function appears to fit rather poorly for very low hazards, but the effect is exaggerated by the log-log axis scaling used in the upper left graph. The remaining three plots

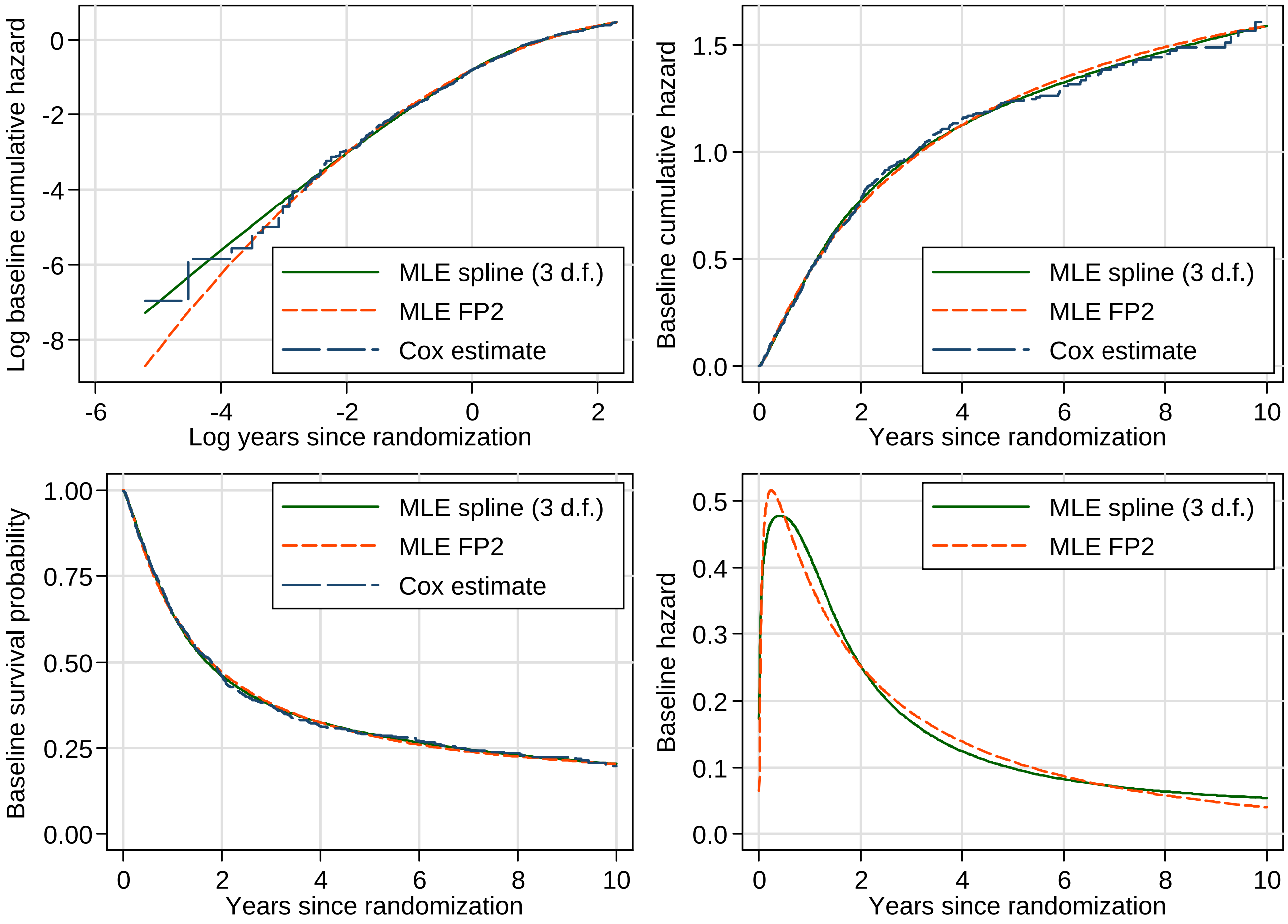

Figure 3: Oesophageal cancer data. Plots of various functions from the BIC/MLE-selected FP and spline models.

Model	d.f.	Fit by OLS			Fit by MLE		
		Deviance	AIC	BIC	Deviance	AIC	BIC
Weibull	1	1214.4	1216.4	1220.0	1210.7	1212.7	1216.4
FP1	2	1170.8	$\boxed{1174.8}$	$\boxed{1182.1}$	1170.8	$\boxed{1174.8}$	$\boxed{1182.1}$
FP2	4	1170.7	1178.7	1193.3	1170.0	1178.0	1192.6
FP3	6	1167.6	1179.6	1201.5	1167.1	1179.1	1201.1
Spline	2	1168.1	$\boxed{\mathbf{1172.1}}$	$\boxed{\mathbf{1179.4}}$	1168.0	$\boxed{\mathbf{1172.0}}$	$\boxed{\mathbf{1179.3}}$
Spline	3	1168.7	1174.7	1185.7	1168.0	1174.0	1185.0
Spline	4	1166.6	1174.6	1189.2	1166.5	1174.5	1189.1

Table 2: Breast cancer data. Deviances, AIC and BIC for the models considered. Boxed values represent the optimal models in their class. Additionally, bolding indicates the overall best models.

suggest there is little difference between the fits. However, the RCS model has a BIC some 10 less than the FP2 model (see Table 1) and may therefore be preferred.

4.2 Breast cancer data

The fit statistics for the various baseline models for the breast cancer data, using the centred prognostic index from Sauerbrei and Royston's [12] model III, are shown in Table 2. [TABLE 2 NEAR HERE] The minimal AIC and BIC criteria select the same RCS and the same FP model for the OLS and for the MLE fitting methods, namely an RCS with 2 d.f. and an FP1. As Figure 4 shows, the estimated hazard functions are similar for all models with d.f. > 1 and with both fitting methods. [FIGURE 4 NEAR HERE] The Weibull model clearly fits the data poorly, having a deviance more than 40 higher than the models with 2 d.f.

Plots of baseline functions from the BIC/MLE-selected RCS and FP models are shown in Figure 5. [FIGURE 5 NEAR HERE] The fitted FP1 and RCS functions plotted on the log-log scale (upper left panel) have been truncated below $t = 0.2$ years, the first non-censored observation. The estimate of $H_0(t)$ from the Cox model is 0 for $t < 0.2$ years, and observations for which $H_0(t) = 0$ do not contribute to the likelihood. Such observations were not included in the OLS regressions.

The fits of both the FP1 function and the 2 d.f. RCS function appear excellent.

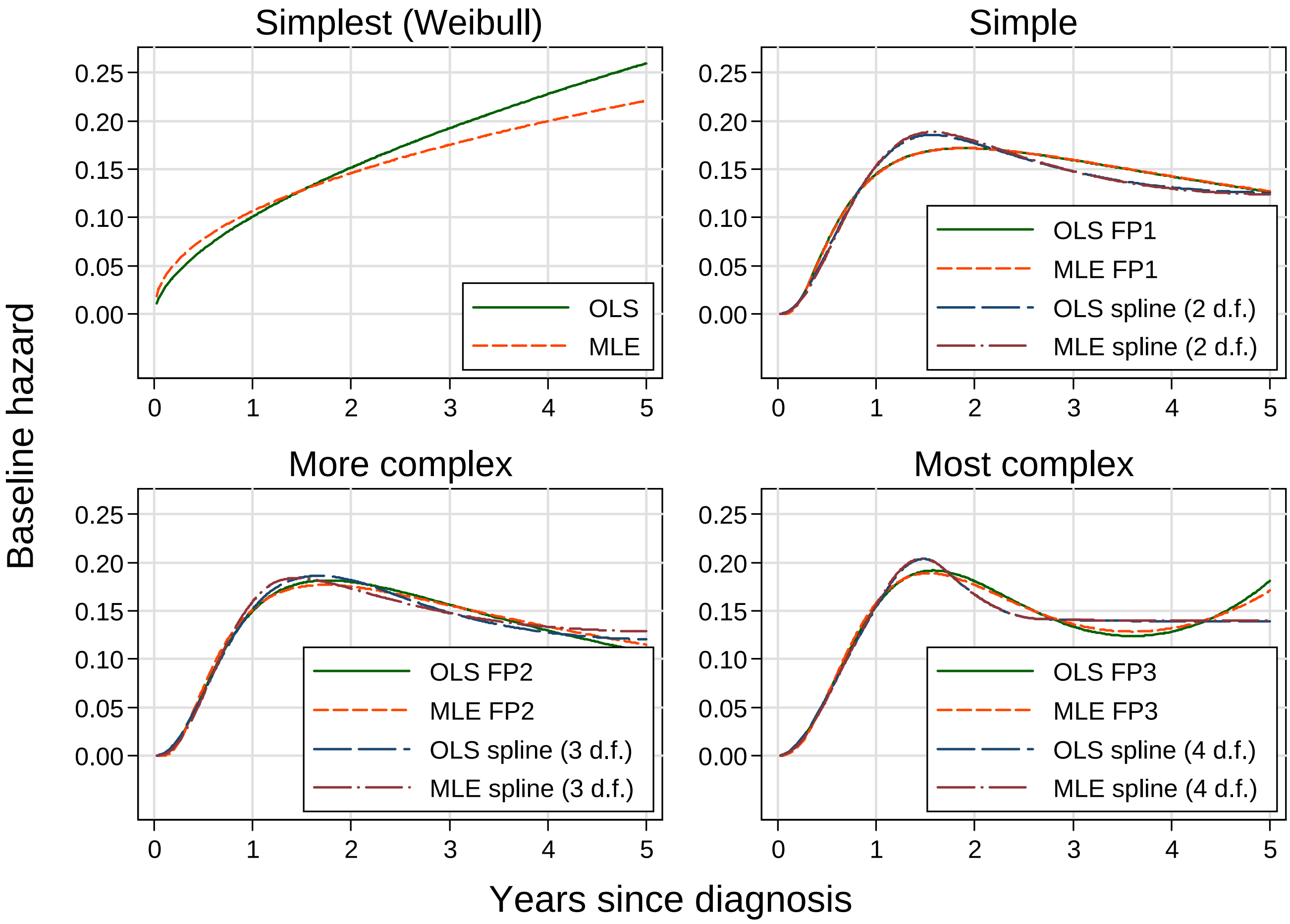

Figure 4: Breast cancer data. Plots of the estimated baseline hazard function for all models considered.

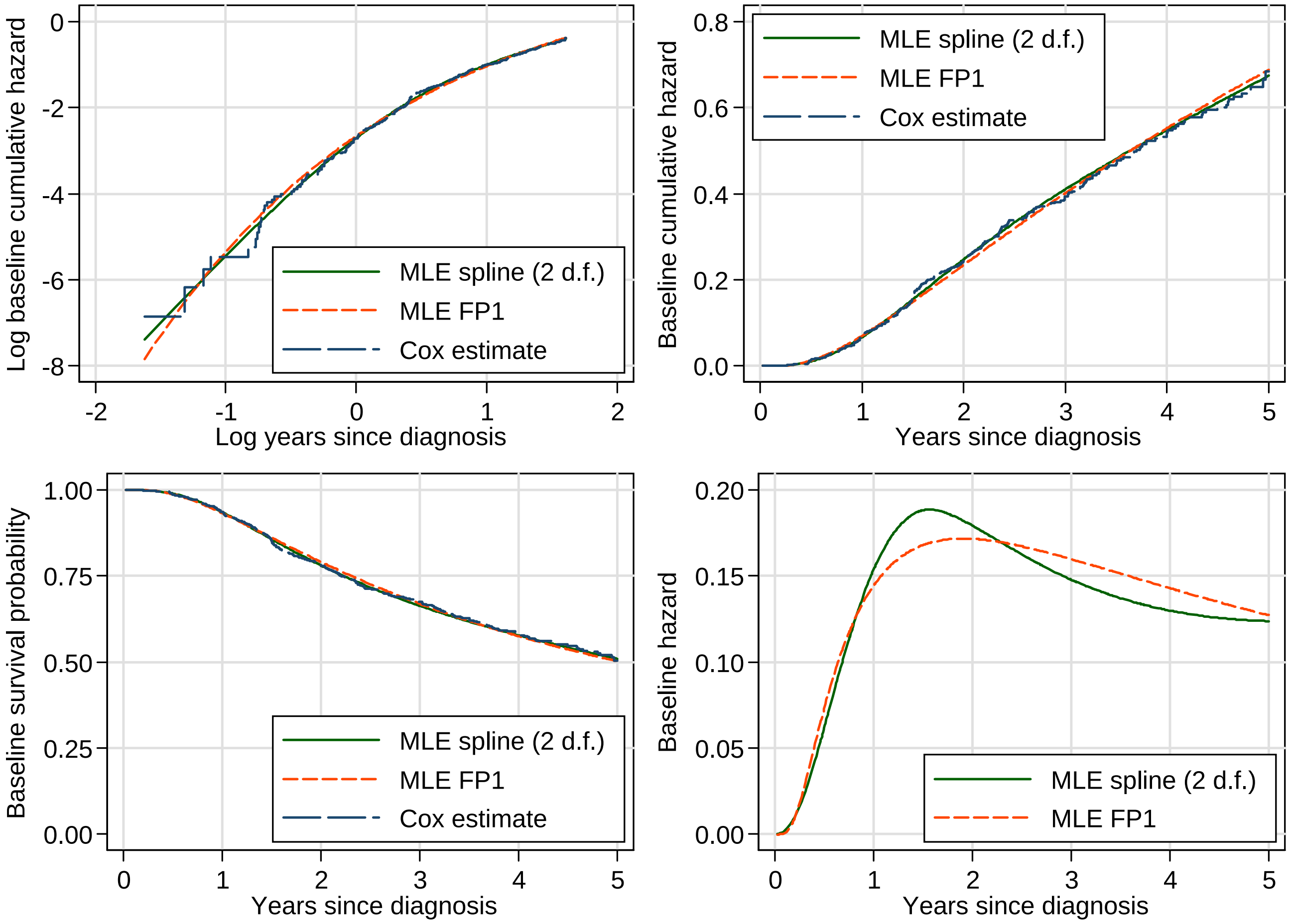

Figure 5: Breast cancer data. Plots of various functions from the BIC/MLE-selected FP and spline models.

4.3 Example of presenting the baseline function

A numerical worked example showing how to calculate the spline basis functions $v_j\left(.\right)$ in eqn. (5) is provided in section 5 of Royston and Parmar [9] and on pp. 109–110 of Royston and Lambert [10]. We do not repeat it here. Estimates of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are from the Cox model. Estimates of $\gamma_0, \gamma_1, \ldots, \gamma_{K+1}$ come either from OLS regression of $z_0\left(t\right)$ on the spline basis functions or from the ML estimation of model (3).

As an example of the FP approach, we present the numerical results for the MLE/BIC-based FP2 model for the oesophageal cancer example (see Table 1). The selected exponents are $p_1 = p_2 = 0$ (a repeated powers model, in fact a quadratic in $\ln t$). Let $x = 0$ for the control arm and $x = 1$ for the experimental arm. The fitted equations are

$$\begin{aligned} \eta &= 0.172\,(0.078)\,x \\ H_0\left(t\right) &= \exp\left[z_0\left(t\right)\right] \\ &= \exp\left[-0.812\,(0.047) + 0.849\,(0.029)\ln t - 0.128\,(0.013)\,(\ln t)^2\right] \end{aligned}$$

Standard errors are given in parentheses, although SEs are not strictly required when reporting the baseline function. It is important to remember that the fitted $z_0\left(t\right)$ function represents only the observed time interval, in this example up to 10 years after randomization. Values of $z_0\left(t\right)$ predicted for $t > 10$ years would be extrapolations and should not be trusted.

5 Discussion

We have suggested practical ways to obtain simple estimates of the baseline log cumulative hazard function, and hence the baseline survival and hazard function, following the fitting of a Cox model. The routine availability of such estimates in published reports of randomized trials and prognostic studies would allow users to apply the model to predict survival and other quantities of interest in their own data. In particular, it would improve the evaluation of the predictive accuracy (calibration) of Cox models in independent data, a process which at present is difficult if not impossible. In clinical trials, the researcher often needs an estimate of the survival function in the experimental arm of the current trial to give a rough estimate of the control-arm survival function expected in a successor trial. This is available as $S_0\left(t\right)^{\exp \eta}$ where η is the log relative hazard for the experimental treatment.

The cumulative hazard function is by definition monotonic increasing (or at least non-decreasing) in t. It is therefore important to check whether the fitted approximation is also monotonic within the range of values of t studied. All FP1 and some FP2 functions are globally monotonic. Whether the latter holds can be determined from the powers and regression coefficients of a selected FP2 model. An FP2 function $\beta_1 x^{p_1} + \beta_2 x^{p_2}$ is monotonic when sign$(\beta_1\beta_2)$sign(p_2) =sign(p_1) (Reference [17], p. 75). The monotonicity of spline approximations always needs to be checked. Our experience, however, is that it is extremely uncommon to find non-monotonic spline approximations to the cumulative hazard function.

A technical issue is, why would one want to use the 'inferior' OLS method for estimating the parameters of FP or RCS models when the 'superior' MLE method gives better results? The main reason is that OLS does not need special software. As we have demonstrated in our examples, in most cases, provided underfitting is avoided, OLS gives perfectly acceptable results. Given the prognostic index, it is quite easy to evaluate the AIC or BIC of any of the proposed parametric models for the baseline using the supplied formula (4) for the likelihood contributions. Hence, comparing the fit of different models estimated by OLS is straightforward. Plots resembling figures 3 and 5 can help one to assess the quality of the fits.

RCS functions may fit the log cumulative hazard function rather better than FPs, but they are more cumbersome to present and to use than FPs. In any case, it is probably unnecessary to seek a very accurate fit. In practice, members of either family of curves are likely to give acceptable results.

In conclusion, we hope that the proposed methods will encourage researchers to report both the regression coefficients and the baseline survival function for their Cox model. We believe that prognostic research and clinical trials with a time-to-event outcome will benefit if this recommendation is adopted.

6 Appendix

6.0.1 Fractional polynomials

A (univariate) fractional polynomial function FP$m\,(x)$ with degree $m > 0$ of an argument $x > 0$ is concisely defined as follows. Let $h_0\,(x) = 1$ and $p_0 = 0$. Then

$$\mathrm{FP}m\,(x) = \gamma_0 + \sum_{j=1}^{m} \gamma_j h_j\,(x)$$

where

$$h_j\,(x) = \begin{cases} x^{p_j} & \text{if } p_j \neq p_{j-1} \\ h_{j-1}\,(x) \ln x & \text{if } p_j = p_{j-1} \end{cases}$$

Cases with $p_j = p_{j-1}$ define the so-called repeated powers models. For example, with $m = 2$, $p_1 = p_2 = -1$, we have $\mathrm{FP2}(x) = \gamma_0 + \gamma_1 x^{-1} + \gamma_2 x^{-1} \ln x$. Cases with $p_j \neq p_{j-1}$ are straightforward extensions of standard polynomials, replacing exponent (power) j with p_j for $j = 1, \ldots, m$. The p_j are taken from a restricted ordered set $S = \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\}$, where by convention x^0 denotes $\ln x$. Full details of univariate FP models are given in chapters 4 and 5 of Reference [17].

6.0.2 Restricted cubic splines

A restricted cubic spline function $s(x)$ of an argument x with $K \geq 1$ interior knots $k_1, \ldots, k_K$ and two boundary knots, $k_{\min} < k_1$ and $k_{\max} > k_K$, is weighted sum of so-called basis functions $x, v_1(x), \ldots, v_K(x)$. The basis functions $v_j(x)$ are derived from cubic polynomial segments defined on the intervals between the knots. Constraints are imposed such that the spline function and its first two derivatives are continuous at the knots, and, to restrain tail behaviour, such that the function is linear for $x \leq k_{\min}$ and for $x \geq k_{\max}$. The spline is defined as

$$s(x) = \gamma_0 + \gamma_1 x + \sum_{j=1}^{K} \gamma_{j+1} v_j(x)$$

Let $\lambda_j = (k_{\max} - k_j) / (k_{\max} - k_{\min})$ be the relative position of the jth knot $(j = 1, \ldots, K)$ on the interval $[0, 1]$. Then

$$v_j(x) = \max\left[0, (x - k_j)^3\right] - \lambda_j \max\left[0, (x - k_{\min})^3\right] - (1 - \lambda_j) \max\left[0, (x - k_{\max})^3\right] \quad (5)$$

The number of knots determines the complexity of the function and its dimension (d.f.). Excluding γ_0, the d.f. equals $K + 1$.

To reflect local data density, knot positions are often chosen according to fixed percentiles of the distribution of the argument x. For example, with equal percentile spacing, 3 knots would be placed at the 25th, 50th and 75th percentiles of x. Harrell [18] proposes a scheme with unequal spacing that includes knots further into the tails of the distribution of x.

Acknowledgments: I thank the investigators and trial management group of the MRC OE02 trial for permission to use the data. I am grateful to Hans van Houwelingen for comments that helped me to strengthen the manuscript.

References

[1] Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. *Statistics in Medicine* 2011; Submitted.

[2] Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**:453–473.

[3] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal* 2009; **338**:b605.

[4] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**:3401–3415.

[5] Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. *BMC Medicine* 2010; **8**:21.

[6] Green PJ, Silverman BW. *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman & Hall: London, 1994.

[7] Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression.* Cambridge University Press: Cambridge, 2003.

[8] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* second edn., Wiley: New York, 2002.

[9] Royston P, Parmar MKB. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; **21**:2175–2197.

[10] Royston P, Lambert PC. *Flexible parametric survival analysis using Stata: Beyond the Cox model.* StataPress: College Station, TX, 2011.

[11] Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M, the German Breast Cancer Study Group. Modelling the effects of standard prognostic factors in node positive breast cancer. *British Journal of Cancer* 1999; **79**:1752–1760.

[12] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 1999; **162**:71–94.

[13] Allum WH, Stenning SP, Bancewicz J, Clark PI, Langley RE. Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in esophageal cancer. *Journal of Clinical Oncology* 2009; **27**:5062–5067.

[14] Schumacher M, Bastert G, Bojar H, Hübner K, Olschweski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RLA, Rauschecker HF. Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* 1994; **12**:2086–2093.

[15] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal* 2009; **9**:265–290.

[16] Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics* 2000; **56**:256–262.

[17] Royston P, Sauerbrei W. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* Wiley: Chichester, 2008.

[18] Harrell FE. *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis.* Springer: New York, 2001.