Research Article

Statistics in Medicine

Received 28 June 2010, Accepted 28 March 2011 Published online 25 May 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4274

The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt

Patrick Royston*† and Mahesh K. B. Parmar

In most randomized clinical trials (RCTs) with a right-censored time-to-event outcome, the hazard ratio is taken as an appropriate measure of the effectiveness of a new treatment compared with a standard-of-care or control treatment. However, it has long been known that the hazard ratio is valid only under the proportional hazards (PH) assumption. This assumption is formally checked only rarely. Some recent trials, particularly the IPASS trial in lung cancer and the ICON7 trial in ovarian cancer, have alerted researchers to the possibility of gross non-PH, raising the critical question of how such data should be analyzed. Here, we propose the use of the restricted mean survival time at a prespecified, fixed time point as a useful general measure to report the difference between two survival curves. We describe different methods of estimating it and we illustrate its application to three RCTs in cancer. The examples are graded from a trial in kidney cancer in which there is no evidence of non-PH, to IPASS, where the opposite is clearly the case. We propose a simple, general scheme for the analysis of data from such RCTs. Key elements of our approach are Andersen's method of 'pseudo-observations,' which is based on the Kaplan–Meier estimate of the survival function, and Royston and Parmar's class of flexible parametric survival models, which may be used for analyzing data in the presence or in the absence of PH of the treatment effect. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: time-to-event data; randomized controlled trials; hazard ratio; non-proportional hazards; restricted mean survival time; flexible parametric survival models

1. Introduction

Most researchers in the world of randomized clinical trials (RCTs) with a right-censored time-to-event outcome are accustomed to thinking of the hazard ratio (HR) as the most appropriate measure of the effectiveness of a new treatment compared with a standard-of-care or control treatment. Following the seminal papers of Freedman [1] and Schoenfeld [2], such trials are usually planned and powered with a target HR in mind. In cancer, for example, we often see a target HR of 0.75 for a new treatment, to be detected with 80 or 90 per cent power at a two-sided significance level (α) of 5 per cent. The null hypothesis of HR = 1 is commonly assessed using the logrank test. A likelihood ratio test of $\beta = 0$ in a Cox proportional hazards (PH) model, which is asymptotically equivalent to the logrank test, is an alternative; under PH, this 'Cox test' is a good choice when we wish to adjust for covariates. The logrank and Cox tests are known to be robust to non-PH in the sense that they retain some power to distinguish between treatments for which the hazard functions are not proportional. The logrank test is a tool to compare two distribution functions, allowing for censoring. It makes no assumption about

Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, 222 Euston Road, London NW1 2DA, U.K.

**Correspondence to: Patrick Royston, Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, 222 Euston Road, London NW1 2DA, U.K.*

†E-mail: pr@ctu.mrc.ac.uk

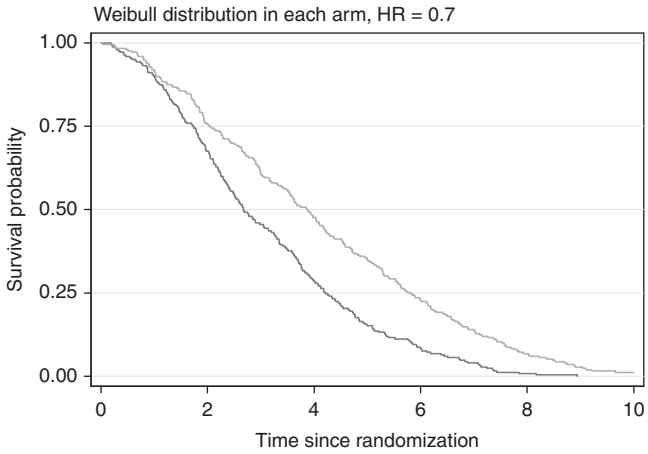

Figure 1. Kaplan–Meier survival curves for data ($n=500$) simulated from Weibull distributions with HR 0.7.

the shape of the underlying distribution functions. However, due to the way the test is constructed, it will have reduced power in extreme cases (e.g. when the survival curves cross).

Analysis of the primary outcome measure in an RCT with a time-to-event outcome typically has two key components: (i) a test of the null hypothesis and (ii) a summary of the treatment effect (with a 95 per cent confidence interval). As we discuss later, if the HR is the target measure of the treatment effect, the analysis ought to include (iii) a test of the PH assumption, but this is rarely done. It is quite natural that within the PH framework, (i) and (ii) are the logrank test and a suitable estimate of the HR, respectively. Almost all reports of RCTs include plots of the Kaplan–Meier estimates of the survival functions in each arm.

Summary statistics such as the HR are only useful if they have a relevant interpretation. In many RCTs in which a significant treatment effect is reported, the Kaplan–Meier curves bear some general resemblance to Figure 1. The data were simulated from Weibull distributions with an HR of 0.7. The two survival curves are vertically separated at all values of time, t, with no evidence of crossing. Perhaps, the most obvious example of non-PH is when the survival curves cross. The presence of non-PH throws into doubt the interpretation of a single reported HR, since the HR must now vary with time. (It must in fact cross 1, but not at the time point the survival curves cross.) We are not surprised, therefore, to learn that in letters commenting on Mok *et al.*'s paper [3], two correspondents commented on the evident breach of the PH assumption and felt that it cast doubt on the validity of the authors' analysis and the value they reported for the estimated treatment effect. Mok *et al.* [3] presented a highly significant ($P<0.001$) HR of 0.74 (95 per cent CI 0.65, 0.85) in favour of gefitinib, and estimated progression-free survival at 12 months of 24.9 and 6.7 per cent for the two arms, respectively. The critics noted that gefitinib's advantage at 12 months is reversed at 4 months (being about 75–60 per cent in favour of the non-gefitinib arm), and that the median progression-free survival times are almost identical in the two arms. The key question is, when the HR is an inappropriate summary statistic because non-PH is present, what (if anything) is an appropriate summary? How should the statistical analysis of such data be conducted?

When non-PH is present, biostatisticians may interpret a single HR for a treatment effect as some kind of average HR (AHR) over the observed follow-up time. A recent paper by Schemper *et al.* [4] clarified and explored several possible definitions of 'average' in this context. They concluded in favour of a definition proposed by Kalbfleisch and Prentice [5], and estimated this average HR by weighted Cox regression. Their definition of the 'average ' (AHR) is as follows:

$$AHR=\frac{\int[h_1(t)/h(t)]w(t)f(t)\,\mathrm{d}t}{\int[h_0(t)/h(t)]w(t)f(t)\,\mathrm{d}t}$$

where $h_0(t)$ and $h_1(t)$ are the hazard functions in the two treatment groups, $h(t)=h_0(t)+h_1(t)$. AHR is not a direct estimate of an AHR. The function $w(t)$ is a weight function, to be chosen by the user. $f(t)$ is a density function, but it is unclear (to us) what distribution $f(t)$ is the density function of. More importantly, although we accept that an AHR can be a well-defined concept, we are far from convinced that it is a useful summary statistic when non-PH is present. As we shall see in one of our examples, it can be seriously misleading.

In this paper, we propose to use the restricted mean survival time [6] (see also Reference [7]) as an appropriate outcome measure in a time-to-event trial, as the primary measure when non-PH is observed and as a useful secondary measure when the PH assumption appears to be satisfied. In Section 2, we describe restricted mean survival time, its interpretation and several methods for estimating it. In Section 3 we introduce and reanalyze three randomized trials to be used as examples. In Section 4, we outline a general strategy for analyzing data from time-to-event trials, distinguishing between situations in which the PH assumption cannot reliably be asserted to hold and those where it can. Section 5 is a discussion.

2. Methods

2.1. Restricted mean survival time

The restricted mean survival time, $\mu(t^*)$ say, of a random variable T is the mean of $\min(T, t^*)$. It may be evaluated as the area under the survival curve $S(t)$ up to t^* [6]:

$$\begin{aligned}\mu(t^*) &= E(\min(T, t^*)) \\ &= \int_0^{t^*} S(t)\,\mathrm{d}t\end{aligned}$$

When T is time to death, we may think of $\mu(t^*)$ as the 't^*-year life expectancy.' For example, a patient might be told that 'your life expectancy with X treatment and Z disease over the next 18 months is 9 months', or 'treatment A increases your life expectancy during the next 18 months by 2 months, compared with treatment B'. We may also explain it as a mean score, the score being created by assigning a value equal to the survival time T if $T \leqslant t^*$ and t^* otherwise. It has been used to summarize survival outcomes when non-PH has been observed, see e.g. Reference [8].

The (unrestricted) mean survival time, e.g. the life expectancy if birth is the time origin, is the limit of $\mu(t^*)$ as $t^* \to \infty$ (or to some appropriate finite upper limit). Because $\int_0^{t_2} S(t)\,\mathrm{d}t > \int_0^{t_1} S(t)\,\mathrm{d}t$ when $t_2 > t_1$, the mean exceeds the restricted mean for any t^*, and $\mu(t^*)$ is a monotonically increasing function of t^*. In survival analysis, we usually have right censoring of event times. We then do not observe the crucial upper tail of the survival distribution, making estimation of the (unrestricted) mean impossible unless we are willing to assume some statistical model for the distribution. This amounts to extrapolation.

2.2. Pseudovalues

Andersen *et al.* [9] described the use of 'pseudo-observations' (or pseudovalues, as we shall call them) as a route to assessing the effects of covariates on restricted mean survival time. Pseudovalues are leave-one-out (i.e. jackknife) estimates of a parameter of interest, here, restricted mean survival time. They are constructed such that their mean is an estimate of the restricted mean survival time at t^* for the entire sample. They are approximately unbiased for each individual observation and hence also unbiased for the overall mean. They are computed from the Kaplan–Meier estimate of the survival curve for the sample. We may model covariate effects on restricted mean survival time with the pseudovalues as the response variable in generalized linear models (GLMs). Standard errors of parameters must employ the robust 'sandwich' estimator.

One advantage of pseudovalues is that since they are based on the Kaplan–Meier estimates, they provide distribution-free estimates of restricted mean survival time. We can meaningfully apply techniques such as smoothed scatter plots of pseudovalues against continuous covariates. Pseudovalues may also be used in model diagnostics [10]. We note that since the link function in the GLM (typically chosen as identity or log) differs from that used in, for example, PH models for censored survival data, covariate adjustment may result in somewhat different estimates of restricted mean survival time. This is similar to the situation of estimating an AHR in a standard analysis of an RCT.

Pseudovalues may be calculated in Stata [11], or in SAS or R [12].

2.3. Cox model

Restricted mean survival time is easily calculated after fitting a Cox model for the treatment effect. The predicted survival function for treatment j ($j=0, 1, \ldots$) is $S_j(t)=S_0(t)^{\exp(\widehat{\beta}_j)}$, where $S_0(t)$ is the baseline survival function (which can be estimated by standard methods) and $\widehat{\beta}_j$ is the log HR for treatment j compared with the control group. To obtain the restricted mean survival time, we numerically integrate $S(t;\mathbf{x})$ over the observed failure and censoring times up to the chosen time point, t^*.

However, we do not recommend the above approach, because it is clearly biased under non-PH. A simple alternative is to integrate the Kaplan–Meier estimate $\widehat{S}_j(t)$ of $S_j(t)$ on $(0, t^*)$ in each treatment group separately, which amounts to using survival estimates from a Cox model stratified by treatment group. Although we use this Kaplan–Meier method for the purpose of illustration, we do not in general recommend it. For example, it is known that $\widehat{S}_j(t)$ is unstable when the risk set size is small.

2.4. Flexible parametric survival models

Royston and Parmar [13] described extensions of standard parametric survival models to accommodate a wide range of baseline distributions in a flexible way. Their hazard-scaled family of models with covariate vector $\mathbf{x}$ is defined through the log cumulative hazard function as

$$\ln H(t;\mathbf{x})=\ln H_0(t)+\mathbf{x}'\boldsymbol{\beta}=s(\ln t)+\mathbf{x}'\boldsymbol{\beta} \tag{1}$$

where the baseline log cumulative hazard function, $\ln H_0(t)=s(\ln t)$, is modelled as a restricted cubic spline in log time. The spline $s(\ln t)$ comprises a linear combination of basis functions and regression parameters $\boldsymbol{\gamma}$:

$$s(\ln t)=\gamma_0+\gamma_1 \ln t+\gamma_2 v_1(\ln t)+\cdots+\gamma_{K+1} v_K(\ln t)$$

Each of the $K+1$ basis functions except the first ($\ln t$) depends on an *interior knot*, which is the join-point in log time of a pair of contiguous cubic polynomial segments. The basis functions are constructed such that their polynomial segments are joined at the knots, and the same holds for the first and second derivatives. The spline function is further constrained to be linear in $\ln t$ in the tails beyond two predefined boundary knots placed at extremes of the observed event times. Note that if $K=0$ the model simplifies to the Weibull model, for which the baseline log cumulative hazard function is $\gamma_0+\gamma_1 \ln t$.

As well as providing flexible baseline distribution functions, Royston–Parmar models have the major advantage of being easy to extend for non-PH. This is done by incorporating a different spline function at the different values of a variable modelled to have a non-proportional effect on the (cumulative) hazard. Statistically, this is simply an interaction between the spline function and the covariate. In the present context, we can include non-PH for the treatment variable.

We estimate the restricted mean survival time at $\mathbf{x}$ by predicting the log cumulative hazard function from equation (1) at $\mathbf{x}$ over a suitably fine grid of time values, transforming it into the survival function and integrating the latter over $(0, t^*)$. Since the model is fully parametric, the survival function is completely specified and smooth. We estimate standard errors using the bootstrap or the delta method.

So far, flexible parametric survival models are available only in Stata. They were originally implemented as the `stpm` command [14], now superseded by `stpm2` [15].

2.5. Adjusting for covariates

Often, we have a model for a randomized treatment effect, and we may wish to adjust for one or more covariates (typically, prognostic factors) measured at baseline. In a PH model, the HR is an interpretable measure of the treatment effect, whether or not covariates are included. When we have non-PH of the treatment effect, things are trickier, since we need an alternative measure of the treatment effect.

Clearly, the restricted mean survival time in each treatment group is not independent of the values of covariates, such as disease severity. How should we produce a meaningful estimate of the treatment effect from an adjusted analysis? One approach is to use the model to estimate the treatment effect at clinically relevant values of the covariate(s). For example, we could present the restricted mean survival times and their difference for the most and least sick patients (e.g. in cancer, according to their 'performance status'). Another approach, similar in spirit to that of Shen and Fleming [16] for estimating the integrated weighted difference in Kaplan–Meier survival probabilities in the context of

Cox model(s) for the covariates, is to average the preferred measure over the empirical distribution of the covariates in each treatment group, and estimate the resulting treatment difference. However, exploring these themes and their variations is beyond the scope of the present paper.

3. Examples

We consider the analysis of three data sets, all from RCTs in advanced cancer. They have been chosen to exemplify different cases: first, a treatment effect that is plausibly described by an HR, with no evidence that non-PH is present; second, a treatment effect whose survival curves do not cross and that traditionally (lacking a test of non-PH) would be described by a single HR, although non-PH is present; and third, a trial with crossing survival curves whose treatment effect clearly breaches PH, making an appropriate analysis more challenging. We aim to show that in all three situations, restricted mean survival time is straightforward to estimate, gives interpretable results and seems to yield P-values for testing the null hypothesis that are comparable with those from the logrank test. In each case, we chose the time point, t^*, for calculating the restricted mean survival time to be near the last observed event time. In choosing t^*, we implicitly assumed that the period of clinical interest in the survival experience was the whole observed follow-up time for the trial.

3.1. Example 1. RE01 trial in advanced kidney cancer

The MRC RE01 trial compared the effects of interferon-α (IFN) with medroxyprogesterone acetate (MPA) on the overall survival of patients with advanced kidney cancer that had metastasized (i.e. spread from the original cancer site to other organs) [17]. The study was a randomized trial recruiting patients between 1992 and 1997. A 28 per cent reduction in the mortality rate in the interferon-α group was reported. We reanalyzed data updated to June 2001, including 322 deaths in 347 patients.

Kaplan–Meier curves for the overall survival by treatment arm are shown in Figure 2(a). The survival curves in the RE01 trial show a clear (if modest) difference between the arms. The P-value from the logrank test is 0.009. The estimated HR in a Cox model for treatment is 0.75 (95 per cent CI 0.60, 0.93) in favour of the experimental arm, again with $P = 0.009$. Median survival (SE) is 0.57 (0.07) and

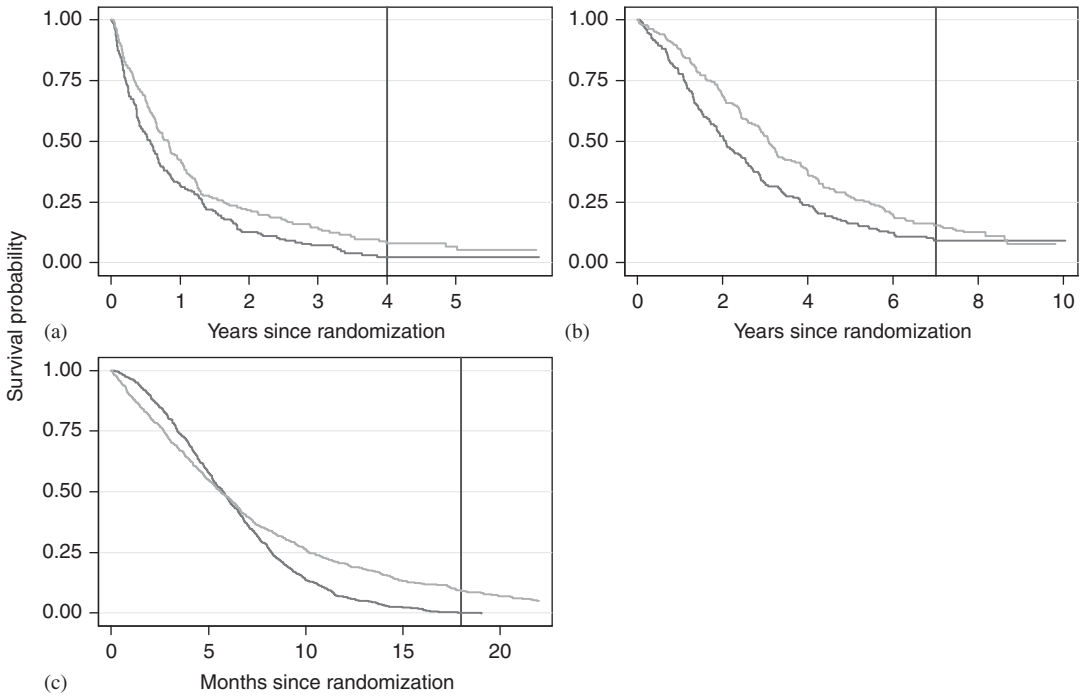

Figure 2. Kaplan–Meier survival curves by treatment group for the three trial data sets. Dark lines: control group; pale lines: experimental group. The vertical lines show the time point t^* for which restricted mean survival time was estimated: (a) RE01 trial; (b) GOG111 trial; and (c) IPASS trial (simulated data).

Table I. Analysis of restricted mean survival time for three randomized controlled trials in advanced cancer. Estimates are given in the same units as t^*. Exp. = experimental.

			Control arm		Exp. arm		Exp. minus control		
Trial	Method	t^*	Est.	SE	Est.	SE	Est.	SE	P
(a) RE01	Pseudovalues	4 y	0.929	0.076	1.243	0.094	0.314	0.121	0.009
	Kaplan–Meier		0.909	0.072	1.199	0.095	0.290	0.120	0.02
	Flex. parametric		0.932	0.078	1.256	0.095	0.324	0.123	0.008
(b) GOG111	Pseudovalues	7 y	2.68	0.15	3.49	0.16	0.81	0.22	<0.001
	Kaplan–Meier		2.60	0.15	3.44	0.17	0.84	0.23	<0.001
	Flex. parametric		2.72	0.15	3.46	0.17	0.74	0.23	0.001
(c) IPASS	Pseudovalues	18 months	6.18	0.14	7.08	0.22	0.90	0.27	0.001
	Kaplan–Meier		6.02	0.15	7.04	0.22	1.03	0.27	<0.001
	Flex. parametric		6.17	0.14	7.08	0.21	0.91	0.25	<0.001

0.82 (0.07) y in the control and experimental arms, respectively, a difference (95 per cent CI) of 0.25 (0.05, 0.46) y, or about 3 months.

According to the Grambsch–Therneau test, there is no evidence of non-PH of the treatment effect in RE01 ($P=0.5$). In this particular trial, therefore, the HR provides an adequate estimate of the treatment effect. However, even under PH, the difference in restricted mean survival times provides a useful perspective. The results of analyses of restricted mean survival time at $t^*=4$ y are given in Table I(a). According to the flexible parametric model, the restricted mean survival time difference (95 per cent CI) is 0.324 (0.083, 0.565) y. Thus, interferon-α treatment extended 4-y mean survival time by about one-third of a year, from 0.93 to 1.26 y. The results are similar (and significant at the 1 per cent level) according to the pseudovalues and flexible parametric estimation methods, but slightly lower (and significant at the 2 per cent level) with the Kaplan–Meier method.

3.2. Example 2. GOG111 trial in advanced ovarian cancer

In the GOG111 trial [18], 410 women with advanced ovarian cancer and residual masses larger than 1 cm after initial surgery were randomized to receive cisplatin (75 mg per square metre of body-surface area) with either cyclophosphamide (750 mg per square metre, control arm) or paclitaxel (135 mg per square metre over a period of 24 h, experimental arm). The eligibility criteria were met by 386 women, with 202 (180 deaths) assigned to the control arm and 184 (163 deaths) to the experimental arm.

Kaplan–Meier curves for overall survival by treatment arm are shown in Figure 2(b). There is a clear improvement in survival on the experimental arm (cisplatin/taxol). Median survival (95 per cent CI) is 2.1 (1.8, 2.5) and 3.1 (2.7, 3.4) y in the control and experimental arms, respectively, a difference (95 per cent CI) of 1.0 (0.6, 1.4) y. The estimated HR (Cox model) is 0.73 (95 per cent CI 0.59, 0.90; $P=0.004$). However, there is clear evidence of non-PH ($P=0.006$, Grambsch–Therneau test), so the appropriateness of the HR as an overall summary statistic is less clear. Further analysis (not shown) suggests that the hazard functions for the two treatment arms converge by about 8 y. The levelling-off of the Kaplan–Meier curves hints that a small fraction (perhaps $\sim$10 per cent) of the trial patients may be cured of the disease, a possible contributor towards the non-PH. For example, if the data are censored at 5 y, the test of non-PH is no longer significant at the 5 per cent level. Because the follow-up time in the trial report [18] was limited to about 4 y, non-PH was not evident in the data originally analyzed.

The results of the analyses of restricted mean survival time at $t^*=7$ y are given in Table I(b). The restricted means and their SEs according to the pseudovalues and flexible parametric modelling methods agree fairly closely, whereas the results from the Kaplan–Meier method are again slightly at variance. According to the flexible parametric model, the restricted mean survival time difference (95 per cent CI) is 0.74 (0.29, 1.19) y, or about 9 months.

3.3. Example 3. IPASS trial in lung cancer

IPASS (Iressa Pan-ASia Study) is a phase 3, open-label trial of previously untreated patients in East Asia who had advanced pulmonary adenocarcinoma (lung cancer) [3]. Patients, who were nonsmokers or former light smokers, were randomly assigned to receive 250 mg per day of the 'targeted' agent gefitinib (experimental arm, 609 patients) or carboplatin (at a dose calculated to produce an area under

the curve of 5 or 6 mg per millilitre per minute) plus paclitaxel (200 mg per square metre of body-surface area) (control arm, 608 patients). The primary outcome measure was progression-free survival.

Individual patient data for the trial were not available for further analysis. The main results are summarized in Mok *et al.* [3]'s Figure 2, which shows the distribution of time-to-event in each arm as Kaplan–Meier curves. Their Figure 2A (i.e. Figure 2, Panel A) shows that the progression-free survival curves cross at approximately 6 months, thus showing extreme non-PH. To enable us to simulate a plausible representation of the data for demonstration purposes, we manually read off progression-free survival probabilities at several times from Mok *et al.* [3]'s Figure 2A. We transformed these survival probabilities into log cumulative hazard values and regressed the latter on the log times, enabling us to estimate the scale and shape parameters of Weibull distributions in each arm separately. We used Monte–Carlo simulation to create a data set of the same size as the original ($n=1217$ patients). We truncated the time to event at 22 months to resemble the follow-up pattern in Mok *et al.* [3]'s Figure 2A. The resulting synthetic data set had 608 and 579 events in the control and experimental arms, respectively.

Kaplan–Meier curves for progression-free survival by treatment arm are shown in Figure 2(c). The Kaplan–Meier curves cross at approximately the median survival time, which is about 5.7 months in each arm. According to a Cox model, the estimated HR is 0.73 (95 per cent CI 0.65, 0.82; $P<0.001$). Since the hazards are markedly non-proportional ($P<0.001$), the HR is a poor summary of the treatment effect.

The results of the analyses of restricted mean survival time at $t^*=18$ months are given in Table I(c). The difference (SE) in restricted mean survival times is about 0.9 (0.3) months by the pseudovalues and flexible parametric methods, a small (5 per cent of t^*) but statistically highly significant treatment effect. The Kaplan–Meier method again gives slightly lower restricted mean survival times, although a larger treatment difference than the other two methods.

3.4. Model selection

When working with flexible parametric models for survival time, we must make several choices: the complexity (d.f.) of the baseline distribution function, whether or not to include a time-dependent treatment effect in the model, and if so, how complex should the treatment/spline interaction be.

We explored the complexity issue in the example data sets by systematically varying the d.f. for the baseline distribution between 1 (no knots) and 6 (5 interior knots). Also, we either assumed PH (i.e. no time-dependent treatment effect) or modelled the interaction between treatment and the spline to have 1 d.f., the most parsimonious such model. Previous experience with similar data sets has suggested that we gain little by adding further d.f. for the interaction. Restriction to 1 d.f. forces the shape of the relationship between the log cumulative hazard function and log time to be the same in each arm of the trial, but not to be parallel, which seems a reasonable compromise.

Values of the Akaike information criteria (AIC) and the Bayesian information criterion (BIC) are plotted against the d.f. in Figure 3. The 'best' (selected) model would be that which minimizes the preferred information criterion. For RE01, the AIC is approximately equal for d.f. $\geqslant 2$ and much higher (worse) for 1 d.f. The BIC shows a clear minimum for 2 d.f. For GOG111, the AIC is minimized by the model with 3 d.f. and a time-dependent effect. We obtain a similar result with BIC, except that the support for a time-dependent effect is weaker. For IPASS, the AIC is minimized by the non-PH model with 1 d.f. Both information criteria indicate a strong time-dependent effect, and both select the same model (which is in fact the model from which the data were simulated).

The corresponding estimates of restricted mean survival time are plotted in Figure 4. For RE01, estimates are stable for $\geqslant 2$ d.f. The difference between estimates from PH and non-PH models is negligible, which indirectly supports the PH assumption. For GOG111, 3 d.f. seems enough, since the estimate hardly changes for larger d.f. There is a difference between the results for the PH and non-PH models, the treatment effect being consistently larger for the latter. For IPASS, the estimates are markedly different between the PH and non-PH models, as we would expect. The treatment effect is smaller with the non-PH model, and is nearly independent of the d.f.

Our initial conclusion from these analyses is that a sensible default strategy when estimating the restricted mean survival time with flexible parametric models is to assign 3 d.f. to the baseline distribution and 1 d.f. to a time-dependent treatment effect to account for possible non-PH. While the model will sometimes overfit the data, the effect is likely to be mild, at worst to introduce a small amount of 'noise' into the estimate of restricted mean survival time. Underfitting is a more serious issue and is

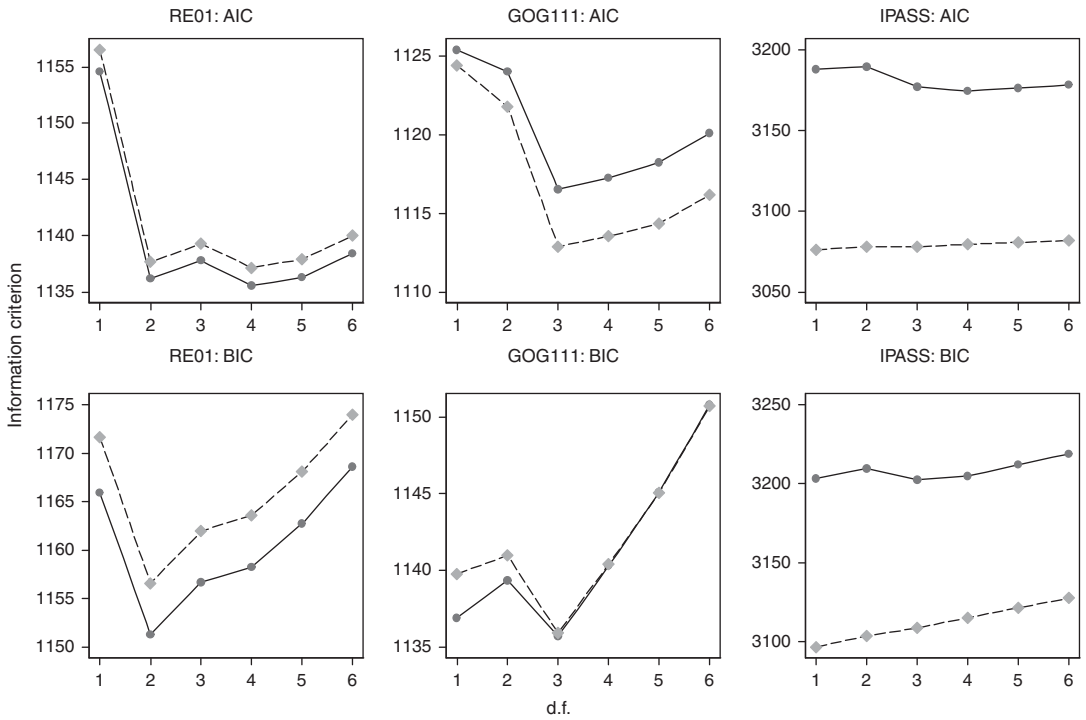

Figure 3. Akaike (AIC, upper panels) and Bayesian (BIC, lower panels) information criteria versus the number of d.f. for the baseline distribution function in flexible parametric survival models for the three example data sets. Solid lines: with PH treatment effect; dashed lines, with time-dependent treatment effect.

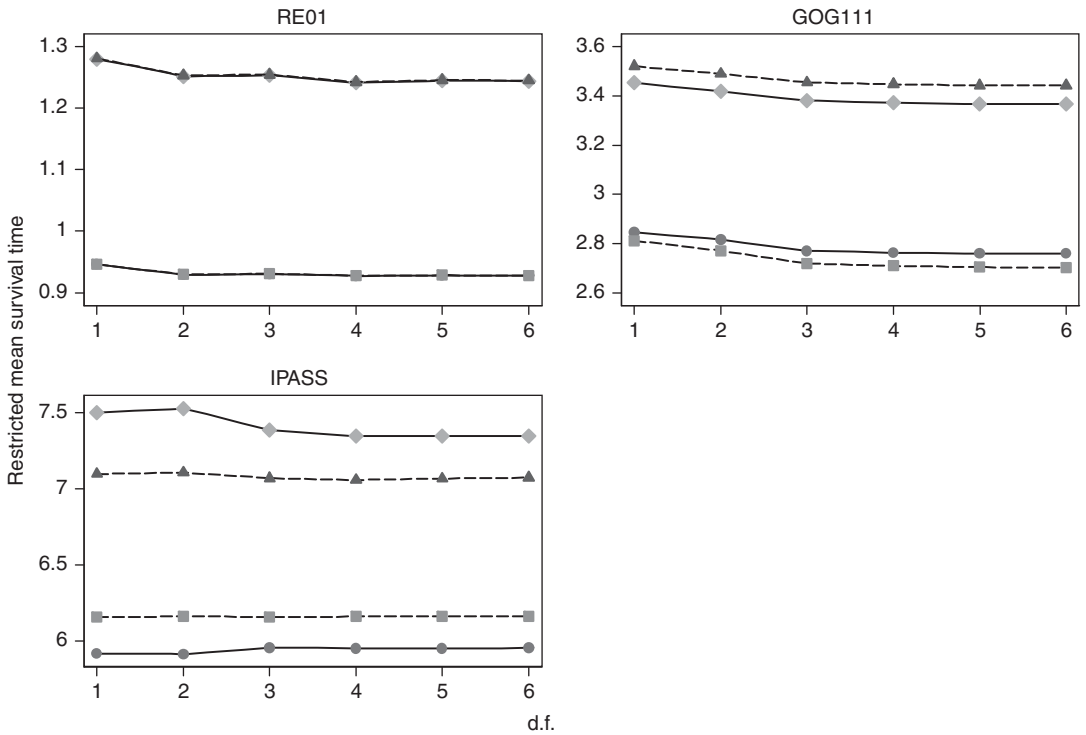

Figure 4. Estimated restricted mean survival time versus the number of d.f. for the baseline distribution function in flexible parametric survival models for the three example data sets. Lower pairs of lines: control group; upper pairs of lines: experimental group. Solid lines: with PH treatment effect; dashed lines, with time-dependent treatment effect.

best avoided, since it may bias the restricted mean. Figure 4 demonstrates both effects; a presumed bias is seen if the baseline d.f. are too few, whereas the estimate is remarkably stable if more baseline d.f. are included.

Since it is important in a statistical analysis plan and a trial protocol to specify precisely which analyses will be carried out, we suggest the flexible parametric model with 3 d.f./1 d.f. as a good overall choice. If we wish to estimate a treatment effect with a constant HR, the 3 d.f. PH model may be used.

3.5. Time-dependent HR

One of the advantages of the flexible parametric family of models is that it is straightforward to estimate the HR as a function of time. With the 3 d.f./1 d.f. model discussed above, we calculated the HR and its 95 per cent CI for the treatment effect in each of the example trials. The HR varies with time. We suggest that such a plot is a useful routine diagnostic of how the HR may depend on time, i.e. the extent to which non-PH is present. The results are shown in Figure 5. In the RE01 trial, there is no evidence of non-PH, since the HR is nearly constant over time. In the GOG111 and IPASS trials, the HR crosses 1 (i.e. no treatment effect) at about 2 y and 5 months, respectively, suggesting non-PH. In the IPASS trial, the overall HR of 0.73 is a particularly poor summary, since the time-dependent estimate ranges between about 0.2 and 30. In the GOG111 trial, the overall value, again 0.73, is also misleading. The HR starts much lower than 0.73, but may even be >1 at long follow-up times.

3.6. Variation of restricted mean survival time with t^*

The restricted mean survival time is a function of t^*. As already remarked, $\mu(t^*)$ increases monotonically with t^*, and for large t^*, tends to the unrestricted mean survival time, μ.

It is of interest to see how the treatment effect, i.e. the estimated difference in $\mu(t^*)$ between trial arms, evolves as t^* increases. If desired, the results for $\mu(t^*)$ could be presented in relative terms as a percentage of t^*, by multiplying each $\mu(t^*)$ and its confidence interval by $100/t^*$. Both the absolute

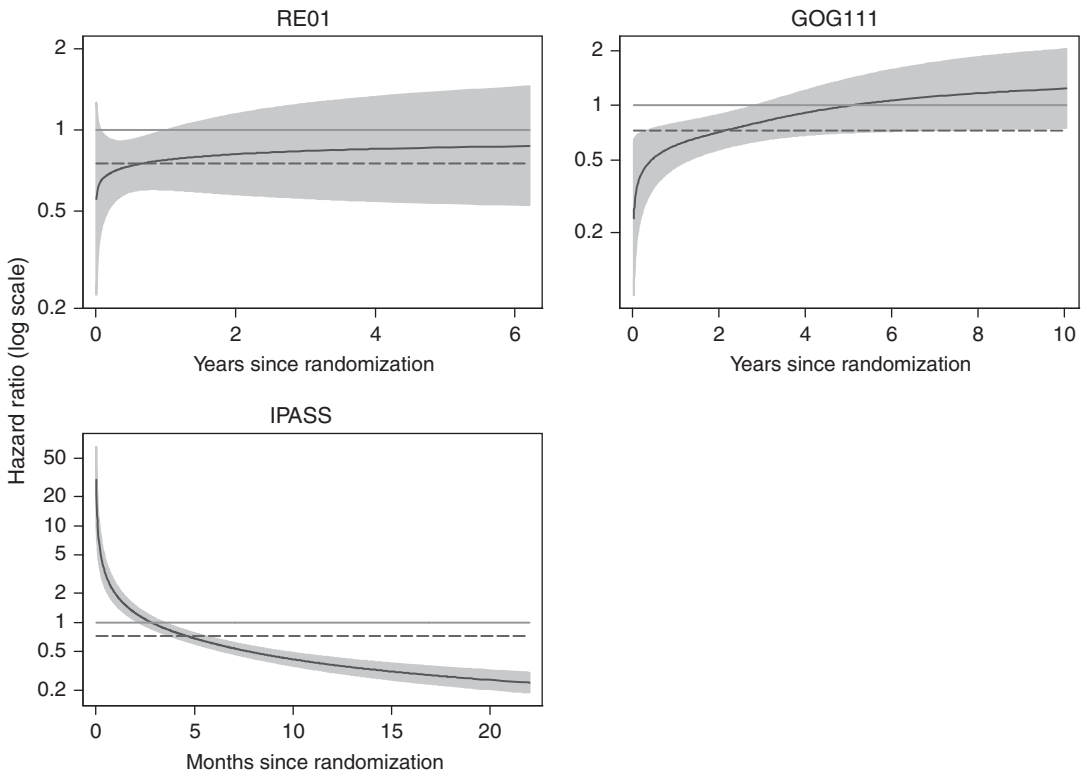

Figure 5. HR and 95 per cent confidence intervals estimated from a flexible parametric model with time-dependent treatment effect in three randomized trials. HRS <1 favour the new treatment, HRS $\geqslant 1$ favour the control treatment. Horizonal lines: solid, HR $= 1$; dashed, HR for comparing the arms using a Cox PH model.

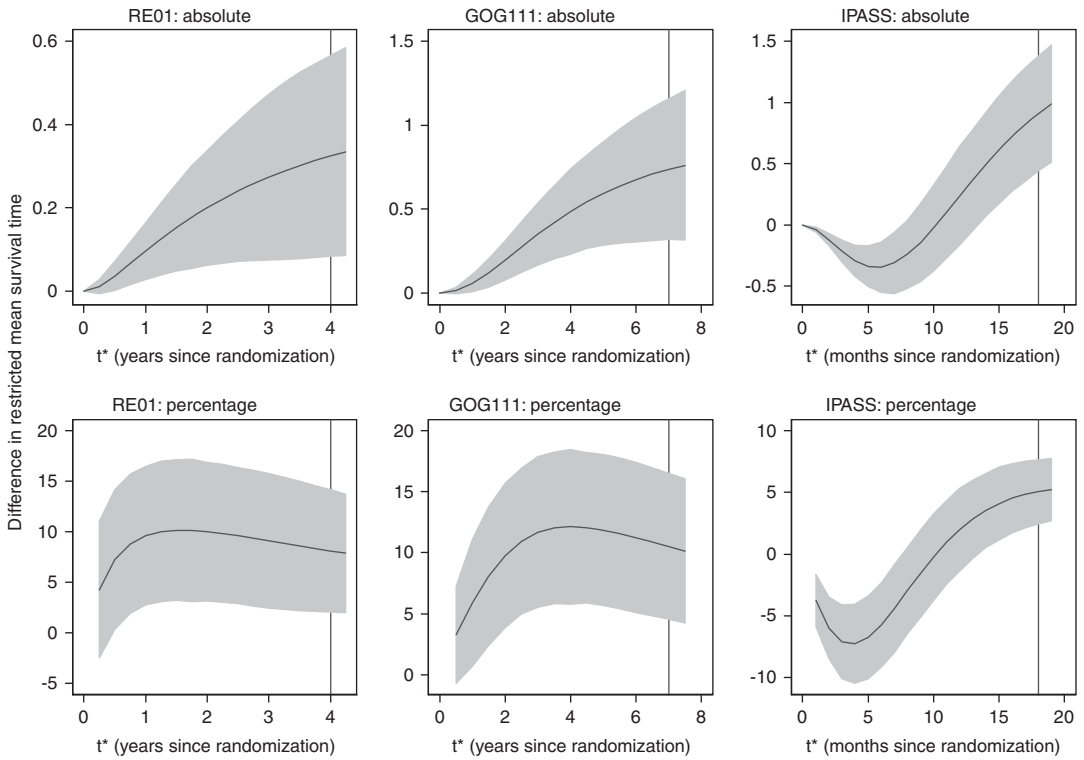

Figure 6. Effect of varying t^* on the treatment difference in $\mu(t^*)$, the restricted mean survival time. Curves labelled 'absolute' are difference in $\mu(t^*)$ with 95 per cent CI as a function of t^*. Curves labelled 'percentage' are the absolute values multiplied by $100/t^*$. The original t^* is shown by a vertical line. Values were estimated by a flexible parametric survival model with 3 d.f./1 d.f. (see text for details).

and the relative measure of difference in restricted mean survival time are shown for the three example data sets in Figure 6. As previously, we estimated $\mu(t^*)$ using flexible parametric models, in each case varying t^* up to slightly above the value chosen in the earlier analyses. The results using the pseudovalues method (not shown) are extremely similar to those in Figure 6.

Both the absolute and the relative treatment effects change with t^*. In the RE01 and GOG111 trials, the relative effect is less variable than the absolute effect. In the latter trials, the relative effect near t^* is approximately 10 percent, whereas for IPASS, it is about 5 percent. In the simulated IPASS trial, we observe a treatment effect that significantly favours the control arm at about 6 months, but that subsequently establishes itself in the opposite direction.

4. Implications for trial statistical analysis plans

The sample size and timelines of most trial designs with a time-to-event outcome are predicated on a constant HR between the treatment arms. The statistical analysis plan for such designs should therefore naturally include an estimate and a confidence interval for the HR. But how should the primary analysis be performed, and what statistics should be reported, if the data indicate non-PH?

We argue that if non-PH is present, the HR, which now depends on time, is a less useful summary of the trial data. In the IPASS trial, for example, it is obvious that the HR is uninterpretable. Even in GOG111, in which we see moderate non-PH, Figure 5 shows that the HR may vary quite considerably over the follow-up time, bringing into question the interpretability of a single estimated HR.

An important question for the trialist is, what is the primary analysis in such a situation? We suggest the following four-step approach:

(1) The treatment effect should be tested using a logrank test and conclusions regarding the null hypothesis drawn accordingly. The logrank test is known to have good power under PH. Our experience is that it retains reasonable power under mild departures from PH.

Table II. Results of tests of non-PH (Grambsch–Therneau (G-T) test and likelihood ratio test from flexible parametric survival models). Also, estimates of differences in restricted mean survival time (computed from flexible parametric models and pseudovalues) in three example trials. Flex. = flexible parametric; Pseudo. = pseudovalues.

		Tests of non-PH					Restricted mean survival time		
Trial	P-value (logrank)	G-T	Flex.	HR	95 per cent CI	t^*	Method	Diff.	95 per cent CI
RE01	0.009	0.5	0.5	0.75	(0.60, 0.93)	4 y	Flex.	0.32	(0.08, 0.57)
							Pseudo.	0.31	(0.08, 0.55)
GOG111	0.004	0.006	0.02	0.73	(0.59, 0.90)	7 y	Flex.	0.74	(0.29, 1.18)
							Pseudo.	0.81	(0.38, 1.23)
IPASS	<0.001	<0.001	<0.001	0.73	(0.65, 0.82)	18 months	Flex.	0.90	(0.38, 1.42)
							Pseudo.	0.91	(0.42, 1.40)

(2) Regardless of whether the logrank test in step 1 is 'significant', the treatment effect should be tested for non-PH. The best-known approach is the Grambsch–Therneau test based on scaled Schoenfeld residuals from a Cox model. Since a bare P-value from a hypothesis test is not particularly informative, we also recommend producing a graphical diagnostic. One possibility is shown in Figure 5, where the general purpose choice of 3 d.f./1 d.f. may be used in the flexible parametric model. An alternative is a scatterplot smooth of the scaled Schoenfeld residuals; however, the plot reveals only the pattern of the log relative hazards over time, not their absolute magnitude.

(3) If there is no evidence of non-PH, the primary summary of the treatment effect is the HR and its confidence interval.

(4) If there is evidence of non-PH, the primary measure of the treatment effect switches to the difference in restricted mean survival time at t^* and its confidence interval. Either the flexible parametric or the pseudovalues method may be used. The value of t^* should be used as specified *a priori* in the trial protocol. A reasonable choice for t^* should be clinically motivated (see further comments in the Discussion). In the absence of such a choice, a default t^* may be taken to be slightly below the maximum expected follow-up time. Although a single estimate of the HR is no longer meaningful, it is of scientific interest in a secondary analysis to estimate and plot the HR as a function of time, as mentioned at step 1.

We stress that even when the validity of the PH assumption is not in doubt, the restricted mean survival time is still a useful measure that can be presented along with other measures and analyses. Here, we are mainly concerned with what should be reported as the *primary* analysis of the trial data. See further comments in the Discussion.

Finally, Table II summarizes the above approach when applied to the three trials. The two tests of non-PH are in reasonable agreement. Although, as Figure 5 shows, the HR in the GOG111 trial progresses from a rather low value to approximately 1 at longer survival times, the overall value of 0.73 gives some feel for the magnitude of the treatment effect. This is not the case in the IPASS trial, where extreme non-PH is present; we would not normally report the HR as a primary outcome measure here.

5. Discussion

Restricted mean survival time has great potential as a meaningful and sensitive outcome measure in the analysis of trial data with a time-to-event outcome. We recommend it as the primary outcome measure when we cannot be confident that the PH assumption holds and therefore doubt that a single HR is appropriate. Flexible parametric models give appropriate estimates of treatment effects under PH or with extension to non-PH, and estimates of restricted mean survival time can easily be obtained from them. Although we do not discuss extensions here, the models can if required include adjustment for covariates, even with time-dependent (i.e. non-PH) effects of the latter if needed. Pseudovalues can also be calculated with appropriate software. We note that adjustment for covariates can be made using generalized linear modelling with robust standard errors, which are a standard feature of most modern statistical software packages. Flexible parametric models, although perhaps more complex,

provide additional useful quantities, such as smooth estimates of the time-varying HR (see Figure 5, for example) and smooth estimates of the hazard function for any desired combination of covariate values [13]. We recommend using either flexible parametric models or pseudovalues in the primary analysis of time-to-event data in RCTs.

Some advantages of restricted mean survival time as an outcome measure are (i) its interpretation is straightforward, (ii) through integration of the survival function, the entire survival distribution up to t^* is taken into account, not just a snapshot at a single time point, (iii) with pseudovalues, analysis of the data with linear regression models is available, and (iv) structural assumptions (e.g. PH of a treatment effect) are minimal. We have seen in examples that a test comparing restricted means between groups gives P-values comparable to those from the logrank test, although we have not studied this comparison systematically.

We see the sensitivity of the restricted mean survival time (RMST) to crossing survival curves and other departures from PH as one of its advantages. In the IPASS trial, an advantage of the control treatment is apparent at 6 months; this is plain from the Kaplan–Meier curves, and the statistically significant RMST difference reflects it. In the longer term comparison at 18 months the situation is reversed, and again the RMST faithfully reflects that. Therefore, the choice of t^* is important and that is why we recommend that its value for the primary analysis (ideally, clinically motivated; see further comments below) should be specified in the trial protocol. There is nothing to stop one specifying other values of t^* in secondary analyses.

It is clear (e.g. from Figure 6) that satisfactory results of an analysis of RMST depend on an appropriate choice of t^*. In a trial of treatments for a metastatic cancer, for example, three-year survival may be an appropriate measure, so $t^* = 3$ y would be reasonable. In the setting of a less lethal primary cancer, a longer follow-up time is typically required to evaluate treatments; a sensible value of t^* would certainly be larger than 3 y. In primary breast cancer, for example, a horizon of at least 5 y, and possibly longer, would be appropriate when the outcome was recurrence-free survival time.

Rather than pre-specifying the primary analysis, we suggest the somewhat broader approach of pre-specifying the primary analysis *procedure*. Our procedure comprises a logrank test, a test of non-PH, and the calculation and reporting either of a HR or of an RMST difference, depending on the result of the test of non-PH. The main reason is that we do not believe that 'one size fits all' in this situation; in the face of non-PH, we think it makes more sense to report the RMST difference than the HR. Clearly, the test of non-PH carries the chance of a type 1 error, in which cases we would erroneously reject PH. The only damage would be that we would report the RMST difference instead of the HR. When PH has been rejected, we recommend producing a plot of the estimated time-dependent HR against t. Under an incorrect rejection of the null hypothesis of PH, the (presumably small) variation of the estimated HR over time would likely be visible in the plot. A variant of the proposed strategy is to carry out the preferred test of non-PH at a significance level, such as 1 percent, more stringent than the conventional 5 percent. There would then be a negligible type 1 error probability for the test of non-PH, but at the cost of lower power to pick up non-PH. We do not favour this approach as the power to detect non-PH is often limited.

We acknowledge that restricted mean survival time also has limitations. Since it depends on t^*, a value of t^* must be selected (see above), but an inappropriate choice may give misleading results. To avoid misinterpretation, the Kaplan–Meier plots of survival in the treatment groups are obviously an important adjunct. Calculation of pseudovalues makes the key assumption that right censoring of survival times is random. See Section 2.2 of Reference [19] for a discussion and simulation study of bias caused by dependent censoring.

Restricted mean survival time might seem an attractive option for designing trials, since standard tools for determining sample size when comparing the means of a continuous variable in two groups are available. The stumbling block with such an approach is the question of within-group variance. A sample size calculation for a continuous outcome variable in two groups requires hypothesized values of the means and the SDs in each treatment group. The variance of the pseudovalues in a sample of patients increases with the proportion of censored observations. However, the precise relationship between the amount (and pattern) of censoring and the variance of restricted mean survival time has not to our knowledge been worked out. In a typical trial with censoring induced by staggered entry of patients, for example, it is unclear how to determine realistic within-group variances, even when loss to follow-up is not an issue. In practice, analysis of an existing data set might be needed to inform the choice of parameter values, but it might not be sufficiently relevant to the real trial data.

In summary, we believe that restricted mean survival time is a technique that should be included in the trial statistician's toolkit. The difference in restricted mean survival time should be seriously

considered as a standard outcome measure in the analysis of censored survival data under non-PH of the treatment effect. In particular, it provides a straightforward and easily communicated way to analyze such data and report the results in relatively familiar terms. It gives a useful summary estimate of the difference between the two survival curves when the PH assumption is untenable, and is informative whether or not the PH assumption holds. Choice of an appropriate analysis of trial data when the PH assumption is clearly wrong provoked controversy in the letters pages of the journal that published the IPASS lung cancer trial. Our reanalysis of simulated data broadly replicating the IPASS data supports the trial originators' claim [3] that gefitinib extends progression-free survival in the population of lung cancer patients studied over about a year and a half. It quantifies the treatment effect as a small (approximately 1 month) improvement in progression-free survival time between 0 and 18 months after randomization.

Acknowledgements

We thank the Gynecologic Oncology Group for permission to use updated individual patient data from the GOG111 trial as an example. We are grateful to two reviewers and an associate editor for comments which helped us to strengthen the paper.

References

1. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1982; **1**:121–129.
2. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**:499–503.
3. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, Nishiwak Y, Ohe Y, Yang J-J, Chewaskulyong B, Jiang H, Duffield EL, Watkins CL, Armour AA, Fukuoka M. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine* 2009; **361**:947–957.
4. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine* 2009; **28**:2473–2489.
5. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981; **68**:105–112.
6. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene* 1949; **47**:188–189.
7. Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association* 1998; **93**:702–709.
8. Yusuf S, Zucker D, Peduzzi P, Fisher LD, Takaro T, Kennedy JW, Davis K, Killip T, Passamani E, Norris R, Morris C, Mathur V, Varnauskas E, Chalmers TC. Effect of coronary artery bypass graft on survival: overview of ten-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet* 1994; **344**:563–570.
9. Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 2004; **10**:335–350.
10. Perme MP, Andersen PK. Checking hazard regression models using pseudo-observations. *Statistics in Medicine* 2008; **27**:5309–5328.
11. Parner ET, Andersen PK. Regression analysis of censored data using pseudo-observations. *Stata Journal* 2010; **10**:408–422.
12. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine* 2008; **89**:289–300.
13. 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.
14. Royston P. Flexible parametric alternatives to the Cox model, and more. *Stata Journal* 2001; **1**:1–28.
15. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal* 2009; **9**:265–290.
16. Shen Y, Fleming TR. Weighted mean survival test statistics: a class of distance tests for censored survival data. *Journal of the Royal Statistical Society* (*Series B*) 1997; **59**:269–280.
17. Medical Research Council Renal Cancer Collaborators, Interferon-α and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *Lancet* 1999; **353**:14–17.
18. McGuire WP, Hoskins WJ, Brady MF, Kucera PR, Partridge EE, Look KY, Clarke-Pearson DL, Davidson M. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. *New England Journal of Medicine* 1996; **334**:1–6.
19. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 2010; **19**:71–99.