Research Article

Statistics in Medicine

Received 1 December 2014, Accepted 18 October 2015 Published online 17 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6801

Gaining power and precision by using model–based weights in the analysis of late stage cancer trials with substantial treatment switching

Jack Bowden,[a,b*†] Shaun Seaman,[b] Xin Huang[c] and Ian R White[b]

In randomised controlled trials of treatments for late-stage cancer, it is common for control arm patients to receive the experimental treatment around the point of disease progression. This treatment switching can dilute the estimated treatment effect on overall survival and impact the assessment of a treatment's benefit on health economic evaluations. The rank-preserving structural failure time model of Robins and Tsiatis (*Comm. Stat.*, 20:2609–2631) offers a potential solution to this problem and is typically implemented using the logrank test. However, in the presence of substantial switching, this test can have low power because the hazard ratio is not constant over time. Schoenfeld (*Biometrika*, 68:316–319) showed that when the hazard ratio is not constant, weighted versions of the logrank test become optimal. We present a weighted logrank test statistic for the late stage cancer trial context given the treatment switching pattern and working assumptions about the underlying hazard function in the population. Simulations suggest that the weighted approach can lead to large efficiency gains in either an intention-to-treat or a causal rank-preserving structural failure time model analysis compared with the unweighted approach. Furthermore, violation of the working assumptions used in the derivation of the weights only affects the efficiency of the estimates and does not induce bias or inflate the type I error rate. The weighted logrank test statistic should therefore be considered for use as part of a careful secondary, exploratory analysis of trial data affected by substantial treatment switching. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: late-stage cancer; logrank test; RPSFTM; treatment switching

1. Introduction

In randomised controlled trials (RCTs) of late-stage cancer therapies, it is common to give the experimental treatment to placebo arm patients at the point of disease progression. This could occur for several reasons: an individual clinician may feel it the best course of action for their patient; it may be pre-specified in the trial protocol as part a dynamic treatment strategy; or emerging evidence (internally or externally to the specific RCT) of the active treatment's benefit may have broken the original trial equipoise. Regardless of the reason, treatment switching (also called contamination or cross-over) does not generally affect the assessment of early outcome measures such as progression free survival but can substantially dilute the estimated treatment effect on overall survival. For example, Demetri *et al.* [1] reported the results of a randomised controlled trial into the use of Sunitinib for the treatment of advanced gastrointestinal stromal tumours in patients for whom conventional therapy (Imatinib) had failed because of resistance or intolerance. Early trial results were unequivocal; randomisation of patients to either placebo or Sunitinib stopped early after a planned interim analysis showed a strong benefit in favour of

[a] *MRC Integrative Epidemiology Unit, University of Bristol, Bristol, U.K.*
[b] *MRC Biostatistics Unit, Cambridge, U.K.*
[c] *Pfizer, La Jolla, San Diego CA, U.S.A.*
[*] *Correspondence to: Jack Bowden, MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Bristol, BS8 2BN, U.K.*
[†] *E-mail: jack.bowden@bristol.ac.uk*

the new treatment in terms of time to tumour progression (hazard ratio 0.33, $p < 0.001$). However, subsequent analysis and interpretation of the data were made harder by the understandable decision to change to an open label protocol and make Sunitinib available to the patients in the placebo arm, which led to the vast majority of eligible patients in the placebo arm switching to receive Sunitinib. By the end of follow-up, the intention-to-treat (ITT) hazard ratio between treatment and control groups for overall survival — the original primary outcome — had weakened from 0.49 ($p = 0.007$) at the interim analysis to 0.88 ($p = 0.31$).

Another example of this phenomenon is described by Motzer *et al.* [2], who reported the results of a double blind Phase III randomised trial into the use of Everolimus for patients with advanced renal cell carcinoma. In order to address ethical concerns and improve recruitment rates, the trial protocol stipulated that, upon disease progression or unacceptable toxicity, patients were to be unblinded and given the option to switch to an open-label Everolimus. After 8 months of follow up, the Everolimus arm patients were experiencing significantly fewer disease progressions (progression–free survival hazard ratio 0.3, $p < 0.001$) but after 10 months, any early difference in overall survival had disappeared (OS hazard ratio 0.83, $p = 0.23$).

In both cases, a strong and early treatment benefit in progression free survival was enough for licence approval by industry regulators. So why, one may ask, does the subsequent artificial dilution in the estimate of overall survival matter? One reason is that it makes it far harder for authorities, such as the UK's National Institute of Clinical Exellence (NICE), to subsequently assess whether these treatments are cost-effective. This is because their calculations rely heavily on an accurate, unbiased measure of overall survival, which trials of this nature do not readily provide [3].

This issue does not just affect industry sponsored trials that seek to gain approval for new treatments. The Concorde trial [4] evaluated the relative effectiveness of two pragmatic approaches to the management of patients with HIV with a proven therapy (Zidovudine): one arm (the Imm group) received Zidovudine immediately, whilst the other arm — the deferred (or Def) group — received it when their condition deteriorated to AIDS Related Complex (ARC). However, whilst the study was ongoing, a change was made to the protocol to allow those in the Def arm to receive Zidovudine if they experienced a persistently low CD4 count, in order to be in line with current clinical practise of the time. At the end of the trial, only a small non-significant difference could be detected between the two arms in terms of overall survival. Several authors subsequently queried whether a larger difference would have been observed if the original protocol had been followed (see [5] for example).

The rank preserving structural failure time model (RPSFTM) of Robins and Tsiatis [6] can be used to estimate the causal effect of a randomised treatment allowing for contamination. The method is based on the initial randomisation of patients and requires specification of an ITT test (typically, the logrank test). The procedure for testing the null hypothesis of no causal effect of treatment in the RPSFTM is *exactly* the same as testing for no effect of treatment assignment in an ITT analysis. *P*-values resulting from the RPSFTM and ITT analyses are consequently identical. For this reason, and also because it does not require an assumption of no unmeasured confounders, the RPSFTM has been acknowledged by NICE as the most statistically principled and robust analysis procedure in this setting [7]. It has been applied to the above trials in subsequent analyses, for example, in the case of the Concorde trial by White *et al.* [8]. Unfortunately, the logrank test can suffer from a loss of power — so that one consistently fails to reject the ITT null hypothesis even when the treatment is strongly effective — if a substantial portion of the participants switch treatments during the course of the trial. This is because the logrank test is optimal when the hazard ratio is constant, whilst treatment switching causes the hazard ratio to change over time (when the treatment effect is non-zero and constant). Our aim in this paper is to find methods of analysis that adjust for treatment switches whilst respecting the randomisation, but that give greater power to the ITT test and consequently greater precision to the causal effect estimator than the standard implementation of the RPSFTM. This is achieved by building on the work of Schoenfeld [9] and Lagakos *et al.* [10], who developed weighted versions of the logrank ITT test. The weights in our logrank ITT test are largely driven by the proportion of people on the active treatment in each arm over time, which we feel is simple and heuristically appealing.

In Section 2, we review the standard and weighted forms of the logrank test and use working assumptions to derive approximately 'optimal' weights for testing the ITT null hypothesis using the weighted logrank test. In Section 3, we review the RPSFTM framework in [6] and discuss how it can be implemented using our weighted logrank test approaches. In Section 4, we use Monte-Carlo simulation to compare the performance of the standard and weighted logrank tests when used for hypothesis testing within an ITT analysis (in terms of power and type–I error) and within an RPSFTM analysis (in terms of bias and mean-squared error of estimates for the causal effect parameter). In Section 5, we apply our new methodology to a re-analysis of the Sunitinib and Concorde trials. We conclude in Section 6 with a discussion of the issues raised and point to future research.

2. Logrank tests for intention-to-treat analysis

2.1. *The standard and weighted logrank tests*

Consider a clinical trial assessing whether a new 'experimental' treatment is superior to standard (control) therapy. Let $T_1, \ldots, T_n$ denote the failure or censoring times of n participants randomised into the trial. Let $T_{(1)} \leqslant \ldots \leqslant T_{(J)}$ $(J \leqslant n)$ denote the J ordered observed event times. The logrank test statistic for the null hypothesis, H_0, that there is no difference between the distribution of survival times in the experimental treatment and control arms can be written as follows:

$$Z = \frac{\sum_{j=1}^{J} \left(O_j - E_j\right)}{\sqrt{\sum_{j=1}^{J} V_j}}. \tag{1}$$

O_j is the observed number of failures in the treatment arm at time $T_{(j)}$. E_j is the expected number of failures at time $T_{(j)}$ in the treatment arm given the total number of failures at time $T_{(j)}$, and V_j is its variance, both under H_0.

The weighted logrank test statistic is

$$Z^W = \frac{\sum_{j=1}^{J} W_j \left(O_j - E_j\right)}{\sqrt{\sum_{j=1}^{J} W_j^2 V_j}}, \tag{2}$$

where $W_1, \ldots, W_J$ are a set of weights. Schoenfeld [9] derived the optimal weights for the logrank test, $W_1^{\text{opt}}, \ldots, W_J^{\text{opt}}$, that is. the weights that asymptotically maximise its power to reject H_0 under the alternative hypothesis. He showed that $W_{(j)}^{\text{opt}}$ is proportional to the true log hazard ratio at time $T_{(j)}$. Hence, if the true log hazard ratio is constant, the standard (unweighted) logrank test is optimal. However, when patients are permitted to switch treatment over the course of follow-up, the log hazard ratio will, in general, not be constant, and so a weighted logrank test will be optimal.

Lagakos *et al.* [10] explored the relative efficiency of the weighted logrank test using the optimal weights compared with the unweighted logrank test in a two-arm RCT in which there is non-compliance in the treatment arm but perfect compliance in the control arm. They assumed that censoring times were independent of both failure times and treatment switching times, as would be the case if censoring were administrative. In the next section, we modify their approach to derive optimal weights in the general situation where treatment switching can occur in the treatment and control arms. We continue to make the working assumption that censoring is independent of failure and treatment switching.

2.2. *An 'optimal' weighted logrank intention-to-treat test*

Let $X_i(t)$ be a binary random variable with $X_i(t) = 1$ if person i is on the experimental treatment at time t; $X_i(t) = 0$ if they are off the experimental treatment (and therefore on the control treatment), and let $x_i(t)$ be its observed value. We assume throughout that it is only possible, at each time point, to take the experimental treatment/control treatment fully or not at all (although this could be relaxed easily). We write $\bar{x}_i(t) = \{x_i(s) : 0 \leqslant s \leqslant t\}$ to represent person i's full treatment history up to time t. Further, let $R_i = r$ be their randomisation indicator so that, at $t = 0$, subject i receives either the experimental treatment (if $R_i = 1$) or control (if $R_i = 0$), and $x_i(0) = R_i$. For $t \in (0, T_i]$, a person's treatment may, however, depart from the original assignment so that $x_i(t)$ is not necessarily equal to R_i. The reason for this departure may be associated with the underlying health of the patient.

Suppressing the subscript i for convenience, let

$$h^r(t) = h(t|R = r) = \lim_{\Delta \to 0} \frac{P(T \leqslant t + \Delta | T \geqslant t, R = r)}{\Delta}$$

denote the conditional hazard of failure at time t given randomisation to arm r. Let $h^r_{\text{on}}(t) = h^r(t|x(t) = 1)$ and $h^r_{\text{off}}(t) = h^r(t|x(t) = 0)$ denote the conditional hazards of failure at time t given randomisation to arm r and given that at time t treatment is, respectively, being received and not being received. Note that these hazards depend not only on the distribution of failure times in the absence of treatment and on the true treatment effect but also on the treatment switching process. In order to derive optimal weights, we make the following 'working' assumptions. It is important to first stress that WA1 and WA2 are not necessary for the weighted logrank test asymptotically to maintain its nominal type-I error rate. Only the non-informative censoring assumption is required, and this is also required by the standard (unweighted) logrank test.

- Working assumption 1 (WA1): $h^0_{\text{off}}(t) = h^1_{\text{off}}(t)$ and $h^0_{\text{on}}(t) = h^1_{\text{on}}(t)$.

Given WA1, we can ignore the randomisation superscript and denote the hazard functions simply as $h_{\text{on}}(t)$ and $h_{\text{off}}(t)$.

- Working assumption 2 (WA2): $\log\{h_{\text{on}}(t)/h_{\text{off}}(t)\} = \theta_0 \quad \forall t$.

In a slight modification of the words used by Lagakos *et al.*, these assumptions would hold if, at any given time, patients switch treatment randomly and independently of their underlying health at that time. We stress that this is an implausible assumption which is not required for the validity of the analysis. Let $\gamma^r(t) = Pr(X_i(t) = 1|T \geqslant t, R_i = r)$ denote the conditional probability of being on treatment at time t given randomisation to arm r and given that failure has not occurred prior to time t. Then, using WA1

$$h^r(t) = \gamma^r(t)h_{\text{on}}(t) + \{1 - \gamma^r(t)\}\, h_{\text{off}}(t)$$

and so using WA2, the log hazard ratio at time t comparing the two randomisation arms is

$$\log\left(\frac{h^1(t)}{h^0(t)}\right) = \log\left(\frac{1 + \gamma^1(t)e^{\theta_0} - \gamma^1(t)}{1 + \gamma^0(t)e^{\theta_0} - \gamma^0(t)}\right).$$

Therefore, under our working assumptions, the optimal weight for the jth failure is

$$W_j^{\text{opt}} = \log\left\{1 + \left[\gamma^1\left(t_{(j)}\right)\right]\left(e^{\theta_0} - 1\right)\right\} - \log\left\{1 + \left[\gamma^0\left(t_{(j)}\right)\right]\left(e^{\theta_0} - 1\right)\right\} \tag{3}$$

If, moreover, θ_0 is close to, but not equal to, zero, we can use the fact that $\log(1 + x) \approx x$ for small x to derive from equation (3) that

$$W_j^{\text{opt}} \approx \left\{\gamma^1\left(t_{(j)}\right) - \gamma^0\left(t_{(j)}\right)\right\}\left(e^{\theta_0} - 1\right)$$

The $(e^{\theta_0} - 1)$ term cancels in Equation (2) and so we can write

$$W_j^{\text{opt}} \approx \gamma^1\left(t_{(j)}\right) - \gamma^0\left(t_{(j)}\right). \tag{4}$$

We will refer to the weights defined in Equation (4) as 'simple ITT weights'. If WA1 and WA2 are violated, they could lose some of their efficiency to detect a treatment effect when one truly exists. Of course, there may be very good reasons to doubt the validity of WA1 and WA2 in the late-stage cancer context, when treatment switching is often initiated by the progression of a patient's disease. We address this issue in detail in Section 4 and in the Discussion.

2.3. Is it acceptable to weight by treatment usage in an intention-to-treat analysis?

In this paper, we follow Lagakos *et al.* [10] in referring to our approach as a weighted ITT test. We do so because it adheres to the ITT principles of basing the analysis on a comparison between the original randomised groupings and including all patients in the analysis. However, some may disapprove of our use of this terminology. Indeed, a reviewer has expressed serious concern that, whilst ITT tests are perfectly entitled to weight patients differently (for example, the Fleming-Harrington family of tests [11, 12] allows the weight given to the jth patient's failure to be a pre-specified function of the survival proportion at time $t_{(j)}$) treatment usage rates should *not* be used for this purpose. We understand the reviewer's concerns and stress that this weight is not a function of patient j's individual treatment usage, as in a *per-protocol* or *as treated* analysis say but is based on the treatment usage of the entire study population at time $t_{(j)}$. Of course, it would be unacceptable if the type I error rate of the weighted logrank test was inflated because of the estimation of $\gamma^0(t)$ and $\gamma^1(t)$. Lagakos *et al*. noted that in the simpler scenario, they considered, where there was non-compliance only in one arm, the nominal type-I error rate was achieved (despite estimation of their weights). Using simulation studies, we shall investigate whether this is also true in our setting. It would also be a concern if the weighted test encouraged erroneous interpretations of the data under the alternative hypothesis. We shall return to this important issue within the context of a hypothetical scenario in the Discussion section.

3. Logrank tests for causal inference

3.1. The rank-preserving structural failure time model

If a substantial proportion of patients depart from their originally allocated treatment, then an ITT analysis, which would provide an estimate of the causal effect of treatment *assignment*, could differ substantially from the causal effect of the treatment itself. The RPSFTM provides a statistical framework for estimating this causal effect, by linking each patient's event time (T) to their (possibly counterfactual) event time had they had not received any treatment (T_0) via the formula:

$$T_0 = T_0(\beta) = \int_0^T \exp\{\beta x(t)\}\, dt \tag{5}$$

The single parameter β, indexing the RPSFTM, can be interpreted in the following manner. The rate at which a person's lifetime is used up is e^β times greater when on treatment than when off treatment. For example, if $\beta = \log(0.5) \approx$-0.7 then treatment slows this rate by a half. If $\beta = 0$, then $T_0 = T$, meaning that treatment has no causal effect on survival time. If we write T as $T = T_{\text{on}} + T_{\text{off}}$, where T_{on} and T_{off} denote times spent on and off treatment, respectively, then (5) can be written as $T_0(\beta) = e^\beta T_{\text{on}} + T_{\text{off}}$. We shall refer to the $T_0(\beta)$'s and any statistic that is a function of them as being on the 'β-transformed' timescale and denote the true value of the causal parameter β as β_0.

In order to estimate β_0, we use the fact that, under the RPSFTM, $T_0(\beta_0)$ is independent of R. For each value of β in a range of possible values, the null hypothesis $H_{0,\beta}$ that $T_0(\beta) \perp\!\!\!\perp R$ is tested using an appropriate test statistic. We first consider the commonly used logrank test statistic (and variations thereof). If $Z(\beta)$ denotes the logrank test statistic calculated using Equation (1) with the β-transformed failure times, then $E\{Z(\beta)\} = 0$ when $T_0(\beta) \perp\!\!\!\perp R$. So β_0 can be estimated as the value $\hat{\beta}$ of β for which $Z(\hat{\beta})$ crosses zero (that is, where $\text{sign}(Z(\hat{\beta}-\epsilon)) \neq \text{sign}(Z(\hat{\beta}+\epsilon))$ for some small value of ϵ). This method is often referred to as 'g–estimation'. A 100(1-α)% confidence interval for β_0 can be obtained by finding the range of values of β for which $|Z(\beta)| \leqslant z_{1-\alpha/2}$. This is the set of values for which $H_{0,\beta}$ can not be rejected at significance level α. Note this is typically not symmetrical around $\hat{\beta}$.

3.2. Weighted logrank tests for the rank-preserving structural failure time model

In this section, we consider the use of the weighted log-rank test statistic in a RPSFTM analysis to estimate the causal effect of treatment in the presence of treatment switching. Intuitively, the idea is to apply the RPSFTM not only to the event time but also to the treatment change times, and hence to apply the ideas of Section 2.2 on the $T_0(\beta)$ scale. In particular, we derive, for any given β, optimal weights for the test of the null hypothesis that $T_0(\beta) \perp\!\!\!\perp R$ under the following assumptions:

- Basic causal assumption (BCA): RPSFTM (5) holds.
- Causal working assumption 1 (CWA1): Patients switch treatment at random independently of their treatment-free failure times $T_0(\beta_0)$.
- Causal working assumption 2 (CWA2): $T_0(\beta_0) \sim exp(\lambda)$ for some rate parameter λ.

Assumption CWA1 differs from WA1 and WA2 because using the RPSFTM requires us to condition on the complete treatment history, not just on current treatment. Assumption CWA2 is used to make the RPSFTM yield a simple model for the hazard. BCA is needed for consistency of the g–estimator of β even when the unweighted log rank test statistic is used. CWA1 and CWA2 are only needed for optimality of the weights. Thus, if CWA1 and CWA2 are violated but BCA holds then, we still obtain a consistent estimate for β_0.

Just as in Section 2.2, the optimal weights we derive are functions of the ratio of the hazard of failure in the treatment arm relative to the hazard in the control arm. As before, they depend on the proportion of patients in each arm that are on treatment at a given time. However, unlike in Section 2.2, the hazards of interest, here, are the hazards of the β-transformed time $T_0(\beta)$ in the two arms, rather than those of the untransformed time T. This is because we are testing $T_0(\beta) \perp\!\!\!\perp R$ when fitting the RPSFTM, as opposed to testing the ITT null hypothesis that $T \perp\!\!\!\perp R$. We now derive the hazard for $T_0(\beta)$.

Let an individual's treatment switching times be denoted $S_1 < S_2 < \ldots < S_K < T$, where K denotes the number of switches before failure, and let $S_0 = 0$. Let X_k denote his treatment (0 if off treatment and 1 if on treatment) during time interval $[S_{k-1}, S_k)$. In order to make things more concrete, we will illustrate the general setup by referring to the special case where $\beta_0 = \log\frac{3}{4}$ and to a single hypothetical patient, Peter. For Peter,

$$K = 3, (S_1, S_2, S_3) = (1, 2, 3), T = 4, X_1 = X_3 = 0, X_2 = X_4 = 1$$

(for convenience, we assume that time is measured in years). Peter starts the trial off treatment (at year $S_0 =$ 0) and proceeds to have three treatment switches (at the start of years 1,2 and 3, respectively) before failing at the start of year 4. According to the RPSFTM, Peter's treatment-free lifetime is $T_0(\beta_0) = exp(\beta_0)T_{\text{on}} + T_{\text{off}}$. Therefore, his treatment free lifetime $T_0(\beta_0)$ is expended at a rate of 1 year per year of observed (or untransformed) study time whilst he is off treatment, but at a rate of $\exp(\beta_0) = 3/4$ years per year of observed study time whilst he is on treatment. In general, a unit of untransformed time corresponds to either $\exp(\beta_0)$ units or one unit of β_0-transformed time.

We derive the hazard of $T_0(\beta)$, in general, that is, for any switching pattern and any β_0 and β, and then for the hypothetical individual, Peter, using $\beta_0 = \log\frac{3}{4}$ (the true value) and β=log $\frac{1}{2}$ (our guess at the true value). In order to do this, we first derive the hazard function $h^r\{t \mid \bar{x}(t)\}$ of T in the same (general, then specific) manner.

3.3. *Step 1: Derivation of the hazard of T*

Our assumptions state that the β_0-transformed failure time $T_0(\beta_0)$ is independent of randomisation arm (BCA) and treatment switching times (CWA1) and has a constant hazard λ (CWA2). It follows that the hazard of T at time t given randomisation arm and treatment history up to time t depends only on treatment at time t and equals $\lambda\exp(\beta_0)$ whilst on treatment and λ whilst off treatment, that is,

$$\begin{aligned} h^r\{t \mid \bar{x}(t)\} &= h^r\{t \mid \bar{x}(t)\}\exp\{\beta_0 x(t)\} \\ \text{which, given CWA1} &= h(t)\exp\{\beta_0 x(t)\} \\ \text{and CWA2} &= \lambda\exp\{\beta_0 x(t)\}. \end{aligned} \tag{6}$$

This is illustrated in Figure 1(a) for Peter. His T is the outcome of a failure process in which the hazard is λ during the first year, $\lambda\exp(\beta_0) = \frac{3\lambda}{4}$ during years 1–2, λ during years 2–3 and $\lambda\exp(\beta_0)$=$\frac{3\lambda}{4}$ from years 3–4.

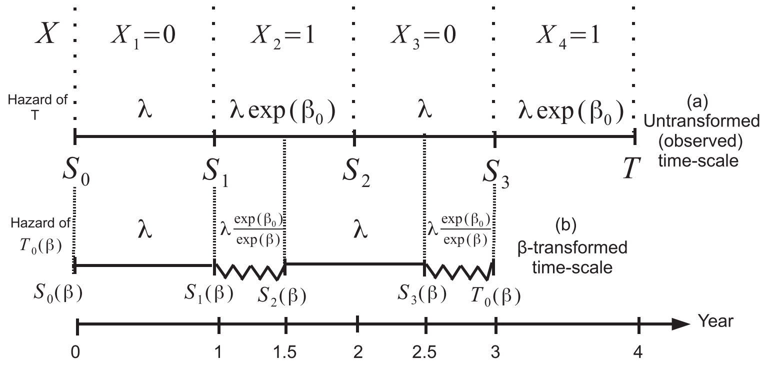

Figure 1. An illustration, for a hypothetical patient, Peter, of the relative switching times and corresponding hazard rates on: (a) the observed timescale and (b) the β-transformed timescale.

3.4. Step 2: Derivation of the hazard of $T_0(\beta)$

We can now derive the hazard of the β-transformed failure time $T_0(\beta)$ for any switching pattern and for any given β. When $T_0(\beta)$ is calculated from T using the RPSFTM (Equation (5)), each year of observed (untransformed) study time on treatment becomes $\exp(\beta)$ years of β-transformed time. So the length of each period on treatment is multiplied by $\exp(\beta)$ and, crucially, the hazard of $T_0(\beta)$ during this period is divided by $\exp(\beta)$. The length of, and hazard during, periods off treatment remain unchanged. In general, we transform the kth switching time on the observed time scale, S_k, to the switching time on the β-transformed timescale, $S_k(\beta)$, via the formula

$$S_k(\beta) = \sum_{j=1}^{k} \left(S_j - S_{j-1}\right) \exp\left\{\beta X_j\right\}, \tag{7}$$

where $S_0(\beta) = 0$. Using Equation (6), and dividing through by $\exp(\beta)$, the hazard of $T_0(\beta)$ is given by

$$\lambda \exp\left\{(\beta_0 - \beta)X_k\right\}, \tag{8}$$

during interval $[S_{k-1}(\beta), S_k(\beta))$. We now calculate the β-transformed hazard for our hypothetical individual, Peter, using the specific value $\beta = \log\frac{1}{2}$. This is illustrated in Figure 1(b). From Equation (7), $S_k(\beta)$, k=1,2,3 and $T_0(\beta)$ equal 1, 1.5, 2.5 and 3, respectively. For Peter, $T_0(\beta)$ is the outcome of a failure process in which the hazard is λ during the first year, $\lambda\exp(\beta_0)/\exp(\beta) = \frac{3\lambda}{2}$ during years 1–1.5, λ during years 1.5–2.5 and $\frac{3\lambda}{2}$ during years 2.5–3. Note that, if β were *actually* equal to β_0, then the hazard of $T_0(\beta)$ would equal a constant value of λ, in line with CWA2.

3.5. Step 3: Derivation of the optimal weights

We now derive the optimal weights for the RPSTM using the hazard of the β-transformed failure time, $T_0(\beta)$. The argument is similar to that in Section 2.2. Let $x(t;\beta) = X_k$ for $t \in [S_{k-1}(\beta), S_k(\beta))$ represent the treatment indicator on the β–transformed timescale, and let $\bar{x}(t;\beta) = \{x(s;\beta) : 0 \leqslant s \leqslant t\}$. So the hazard of $T_0(\beta)$ at time t given randomisation arm $R=r$ and $x(t;\beta)$, that is, $\lim_{\Delta t\to 0} P\{t \leqslant T_0(\beta) < t + \Delta t \mid \bar{x}(t;\beta), R = r, T_0(\beta) \geqslant t\}/\Delta t$, is equal to $\lambda\exp\{(\beta_0 - \beta)x(t;\beta)\}$ from Equation (8). It follows that the hazard of $T_0(\beta)$ at time t given $R = r$ equals $\lambda\exp\{(\beta_0 - \beta)\}\gamma^r(t;\ \beta) + \lambda\{1 - \gamma^r(t;\ \beta)\}$, where $\gamma^r(t;\ \beta) = P(x(t;\beta) = 1 \mid T_0(\beta) \geqslant t, R = r)$.

Let $t_{(j),\beta}$ denote the jth of the ordered β-transformed failure times $T_0(\beta)$ $(j = 1,\ldots,J)$. Straightforward application of the results of Section 2.2 shows that the optimal weights for the weighted log rank test of the null hypothesis that $T_0(\beta) \perp\!\!\!\perp R$ are given by

$$\begin{aligned} W_j^{\text{opt}}(\beta) = &\log\left\{1 + \gamma^1\left(t_{(j),\beta};\beta\right)\left(e^{\beta_0-\beta} - 1\right)\right\} \\ &- \log\left\{1 + \gamma^0\left(t_{(j),\beta};\beta\right)\left(e^{\beta_0-\beta} - 1\right)\right\}, \end{aligned} \tag{9}$$

$(j = 1,\ldots,J)$. Moreover, if β is close, but not equal to, β_0, then

$$W_j^{\text{opt}}(\beta) \approx \gamma^1\left(t_{(j),\beta};\beta\right) - \gamma^0\left(t_{(j),\beta},\beta\right) \tag{10}$$

Typically, $\gamma^0(t;\beta)$ and $\gamma^1(t;\beta)$ will be unknown, but they can be estimated from the observed data: $\gamma^r(t;\ \beta)$ is estimated as the number of patients with $R = r$, $T_0(\beta) \geqslant t$ and $x(t;\beta) = 1$ divided by the number of patients with $R = r$ and $T_0(\beta) \geqslant t$. We will refer to the weights defined in Equation (10) as ‘simple causal weights’.

3.6. Censored observations

For ease of explanation, we have so far assumed in this section that every subject’s event time is observed, but real trial data will be inevitably be affected to some degree by censoring. Let C_i denote patient i’s (potential) censoring time. If $C_i < T_i$, then T_i is censored. For those with $C_i > T_i$, C_i should represent the end of planned follow up. In the context of an ITT analysis, provided censoring is non-informative on the observed event timescale, that is $C_i \perp\!\!\!\perp T_i$, the type-I error rate of the logrank test is asymptotically equal to its nominal value.

In a RPSFTM analysis, in order that the type-I error rate is maintained for the null hypothesis H_{0,β_0} (and so the nominal coverage of confidence intervals for β is maintained), we need to ensure that the censoring remains independent of failure times on the β-transformed scale. This can be achieved in the following manner, as in [6]. Let $\Delta_i(0) = I(T_i < C_i)$ be a failure indicator on the original timescale denoting whether T_i is observed and assume that treatment switching is possible in both arms of the trial. Let $C_i(\beta) = \min\{C_i, C_i \exp(\beta)\}$ be patient i's censoring time on the β–transformed scale. Their corresponding failure indicator on this time-scale, $\Delta_i(\beta)$, is given by

$$\Delta_i(\beta) = \begin{cases} 1 \text{ if } T_{0,i}(\beta) < C_i(\beta) \\ 0 \text{ otherwise.} \end{cases} \tag{11}$$

This scheme makes it possible for originally uncensored individuals to become censored on the β-transformed scale but does not allow originally censored individuals to become uncensored. Thus, we define $J(\beta)$ ($J(\beta) \leqslant J$) to be the number of unique event times on the β-transformed timescale. To test $H_{0,\beta}$, we re-calculate logrank test statistic (1) using the $J(\beta)$ ordered failure times on the β-transformed timescale, $T_{(1),\beta}, \ldots, T_{(J(\beta)),\beta}$.

4. Simulation study

In this section, we use a simulation study to compare the performances of the standard and weighted log-rank test statistics: firstly, for testing the null hypothesis of no treatment effect in an ITT analysis (using the simple ITT weights from Equation (4) in Section 2.2) ; and secondly, for estimating the causal parameter β using a RPSFTM (using the simple causal weights from Equation (10) in Section 3.2). We do not consider the optimal weights defined in Equations (3) (for the ITT analysis) and (9) (for the causal RPSFTM analysis) any further. In a preliminary investigation, their performance was found to be highly similar to the simple weights, even when implemented in the most favourable circumstances (that is, by correctly plugging in the true values for either θ_0 or β_0). Section 4.1 describes our framework for simulating data using a RPSFTM. In Section 4.2, we specify four simulation scenarios by choosing which causal working assumptions to meet and which to violate. Results are presented in Sections 4.3 and 4.4.

4.1. Framework for data generating process

Patients randomised to the experimental arm ($R = 1$) start on the active treatment. If their disease progresses ('progression 1'), they are switched to standard (control) therapy for the remainder of the trial. Patients randomised to the control arm ($R = 0$) start on standard therapy. If they experience a disease progression (also 'progression 1'), they are then switched to the active drug. However, if they suffer a further progression ('progression 2'), they are switched back to the control treatment for the remainder of the trial. Thus, regardless of the arm to which a patient is randomised to, there is a good chance that, before death, he/she will spend some time on and off the active treatment.

Time-to-event data following this treatment switching pattern were generated. As a first step, three independent random variables were simulated for each person i from the following exponential model:

$$D_{ij} \sim \exp\left(\lambda_{ij}^{-1}\right) \quad j = 1, 2, 3, \quad \lambda_{ij} = \lambda_i \exp\left(\mu_j + \eta_i\right), \tag{12}$$

where λ_{ij}^{-1} is the rate parameter (hence $\mathrm{E}[D_{ij}] = \lambda_{ij}$). The 'treatment–free' times to disease progression '1' , disease progression '2' and death are then given, respectively, by D_{i1}, $(D_{i1} + D_{i2})$ and D_{i3}. It is imagined that these times are on the month scale. The μ_j represent fixed parameters, whereas λ_i and η_i represent (potentially) random frailty terms for patient i. In all simulations $\mu_1 = \mu_2 = \mu_3 = 2.5$. Before discussing the assumed distributions of λ_i and η_i — and its implications for the distribution of D_{ij} — we describe the rest of the data generating process.

Whilst progression 2 can not occur before progression 1, death can occur before, in between or after the two progressions, as illustrated in Figure 2. Person i's D_{i1}, $(D_{i1} + D_{i2})$ and D_{i3} are temporally ordered and relabeled as either $S_{i,1}(\beta_0)$, $S_{i,2}(\beta_0)$ or $S_{i,3}(\beta_0)$ according to the following principles: Death is the final event for person i, so events occurring after D_{i3} are subsequently ignored. Person i, therefore, has k_i pertinent potential events $(S_{i,1}(\beta_0), \ldots, S_{i,k_i}(\beta_0))$, where $k_i \in \{1, 2, 3\}$. For example, if D_{i3} is in position (c) then $k_i = 3$, $S_{i,1}(\beta_0) = D_{i1}$, $S_{i,2}(\beta_0) = D_{i1} + D_{i2}$ and $S_{i,3}(\beta_0) = D_{i3}$. Or, if D_{i3} is instead in position (a),

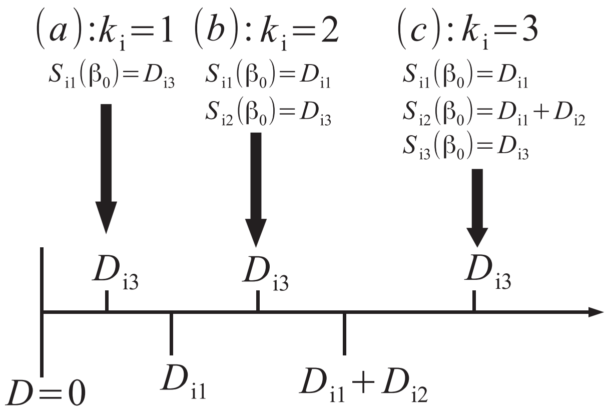

Figure 2. Illustration of the possible temporal orderings of person i's treatment free events (switching and failure times). D_{i3} can occur before (position (a)), in between (position (b)) or after (position (c)) both disease progressions.

then $k_i = 1$ and $S_{i,1}(\beta_0) = D_{i3}$. Note that, for convenience, we are now using a common notation, S, to denote switching times *and* failure times, unlike in Section 3.2.

For a given value of the true causal treatment effect, β_0, observed event times are then calculated using

$$S_{i,j} = \sum_{l=1}^{j}(S_{i,l}(\beta_0) - S_{i,l-1}(\beta_0))\exp\{-\beta_0 X_{i,l}\}, \quad \text{for } j = 1, \ldots, k_i. \tag{13}$$

where $X_{i,j}$ are the treatment indicator variables defined in Section 3 and $S_{i,0} = S_{i,0}(\beta_0) = 0$ for all i. Note that Equation (13) is the inverse of Equation (7). The general treatment switching pattern observed over time in the two arms is

$$R = 1 : \text{On} \rightarrow \text{Off}$$
$$R = 0 : \text{Off} \rightarrow \text{On} \rightarrow \text{Off},$$

and so the implied full vector for $(X_{i,1}, X_{i,2}, X_{i,3})$ is (1,0,0) for the $R = 1$ group and (0,1,0) for the $R = 0$ group, although only the first k_i entries of this vector are relevant to person i.

4.2. Data generation and our causal working assumptions

When we wish to satisfy the working assumptions used to derive the form of the simple causal weights, that is, CWA1 and CWA2, we set $\lambda_i = 0.75$ and $\eta_i = 0$ for all i. In that case, D_{ij} are i.i.d exponential with rate parameter $\lambda_j = \lambda\exp(\mu_j)$. When we wish to violate these working assumptions, we set $\lambda_i \sim$ Uniform(0.6 and 0.9) and $\eta_i \sim$ N(0,1). The D_{ij}'s are not then i.i.d exponential, because of the patient–specific frailty terms. This means that: (i) when the $D_{i,j}$'s are transformed to $S_{i,j}(\beta_0)$'s, $S_{i,k_i}(\beta_0)$ (which equals person i's $T_{0,i}(\beta_0)$) is no longer i.i.d exponential, a violation of CWA2; (ii) when the $S_{i,j}(\beta_0)$'s are transformed to the $S_{i,j}$'s via formula (13), the observed treatment switching times are correlated with the $S_{i,k_i}(\beta_0)$'s (or $T_{0,i}(\beta_0)$'s), a violation of CWA1. In other words, the timing of a patient's treatment switches is predictive of their treatment—free failure time (i.e. their underlying prognosis). This in turn means that the hazard of a patient at observed time t will not only depend on $x(t)$ but also on their full treatment history $\bar{x}(t)$, thus invalidating Equation (6).

In addition to generating the trial data as described earlier (from Equations (12) and (13)), we also considered two data–generating mechanisms that violated the BCA. Firstly, we assume that the treatment has a delayed effect of 3 months, that is, a person's lifetime is used up at the 'off treatment' rate (1 day per day) for the first 3 months they take the treatment and thenceforth at rate $\exp(\beta_0)$ until they stop treatment. They must, therefore, stay alive and on treatment for at least 3 months to realise any potential benefit the treatment may hold. Secondly, we assume that the treatment's effectiveness is lessened or degraded after the first disease progression; specifically, the treatment effect is β_0 before disease progression 1 and $\beta_0/\sqrt{2}$ afterwards.

To summarise, the four data generating mechanisms were as follows:

- Scenario 1: CWA1 & CWA2 true (λ_i=0.75, η_i=0), BCA true;

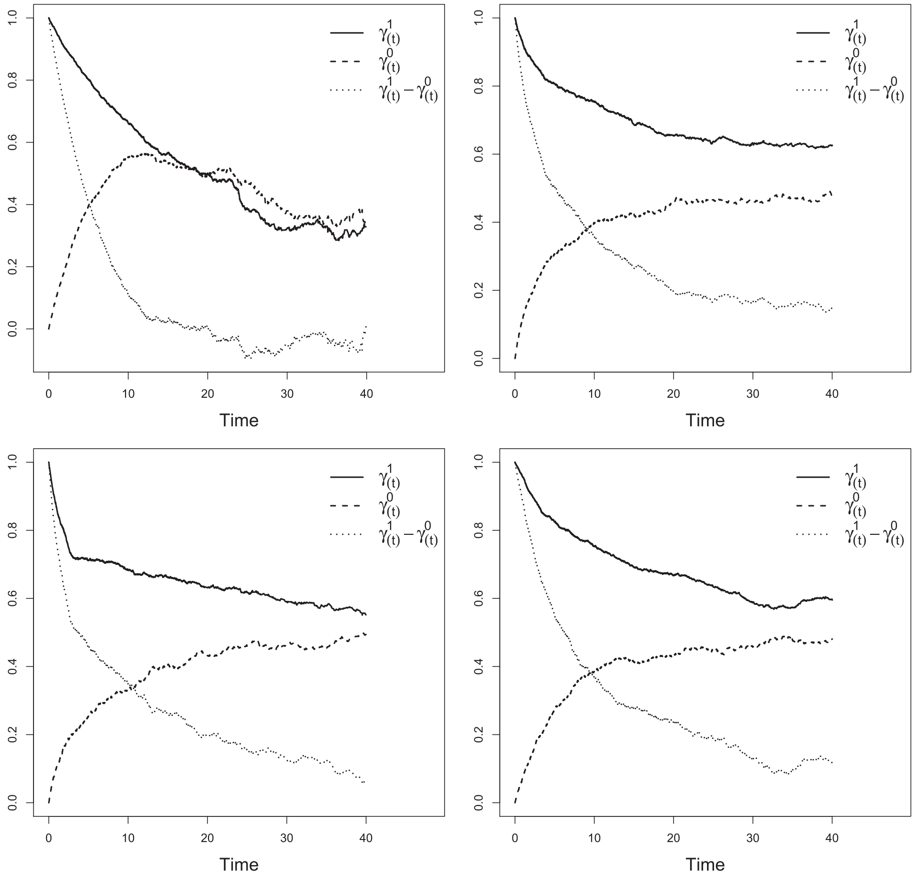

Figure 3. Treatment switching pattern for scenarios 1 (top-left), 2 (top-right), 3 (bottom-left), 4 (bottom-right), under H_1, $\beta = \log(0.5)$. Results based on 20 000 subjects.

- Scenario 2: CWA1 & CWA2 false ($\lambda_i \sim$ Uniform(0.6,0.9), $\eta_i \sim$ N(0,1)), BCA true;
- Scenario 3: CWA1 & CWA2 false ($\lambda_i \sim$ Uniform(0.6,0.9), $\eta_i \sim$ N(0,1)), BCA false (3 month delay to treatment effect), and
- Scenario 4: CWA1 & CWA2 false ($\lambda_i \sim$ Uniform(0.6,0.9), $\eta_i \sim$ N(0,1)), BCA false ($1/\sqrt{2}$ degrading of treatment effect after progression 1).

For each scenario, data were simulated both under $\beta_0 = 0$ (so that the ITT null hypothesis H_0 was true) and under $\beta_0 = \log(0.5)$ (so that H_0 was false — this is referred to as H_1). Two forms of right censoring were also introduced. For each subject, we generated a censoring time C from an exponential distribution with mean 250 months. Additionally, everyone with a failure time above 40 months (the assumed length of the trial follow up) was censored. This led on average to a censoring proportion of 10–15% across the distinct simulation scenarios. Figure 3 shows the basic pattern of treatment switching in the two groups across scenarios 1–4 under H_1, and Figure 4 shows the Kaplan Meier survival curves for the two treatment groups across scenarios 1–4 under H_1. These curves are shown for large trials of 20 000 subjects, in order to reveal their true (large sample) shape.

4.3. Results: intention-to-treat analyses

Table I (column 3) shows the type I error rate of the standard and simple weighted logrank test when used within an ITT analysis to test H_0 at the two-sided 5% significance level. All results in Table I are based on 5000 simulated trials of $n = 250$ patients. We see that, across all scenarios, the type I error rate remains close to the nominal level for both the standard and simple weighted ITT test. Figure 5 (top) shows the power to reject H_0 under H_1 using the standard and simple weighted ITT test, for scenarios 1–4 and for varying sample size ($n = 100$ to 650). Solid lines indicate power for the standard test, dashed lines indicate power for the weighted test, and scenarios are differentiated by colour. We see that the simple ITT weights furnish a substantially more powerful test in Scenarios 1, 2 and 4. To understand why

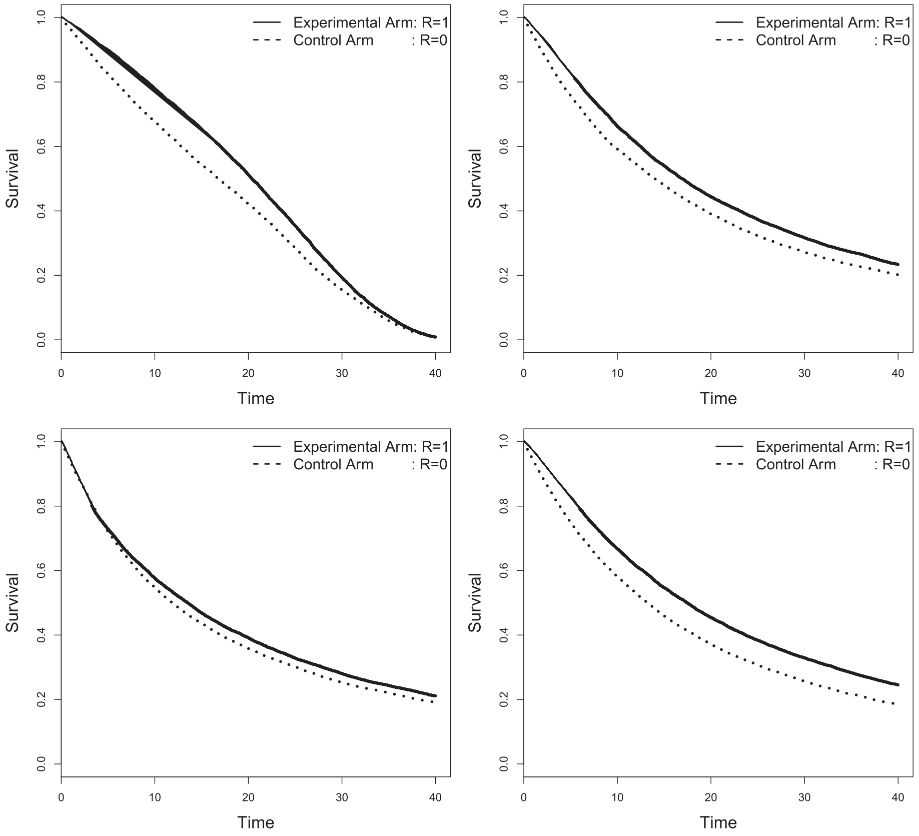

Figure 4. Kaplan Meier survival functions for scenarios 1 (top-left), 2 (top-right), 3 (bottom-left) and 4 (bottom-right), under H_1, $\beta = \log(0.5)$. Results based on 20 000 subjects.

this occurs, it is helpful to look at the relevant treatment switching patterns and Kaplan Meier curves in Figures 3 and 4, respectively. In each of these scenarios, large values of the weight function $\gamma^1(t) - \gamma^0(t)$ coincide with a strong rate of separation in the Kaplan Meier curves during the first half of the trial ($t = (0, 20)$). In Scenario 3, the standard logrank ITT test is more powerful than the simple weighted ITT test, although the power of both tests is fairly low. From Figure 4 (bottom-left), we see that this is due to the delayed treatment effect preventing early separation of the survival curves, which is precisely when $\gamma^1(t) - \gamma^0(t)$ is large. In order to better understand this result, we conduct a further simulation for a set of subcases of Scenario 3. Figure 5 (bottom-left) shows the power of the standard and simple weighted ITT tests under H_1 when the delay for treatment to take effect is varied between 0 and 3 months. We see that the weighted ITT test is substantially more powerful for small delayed effects and remains more powerful than the standard test for up to a $1\frac{1}{2}$ month delay.

Figure 5 (bottom-right) reports the results of a more in-depth investigation of scenario 4. It shows the power of the standard and simple weighted ITT tests under H_1 when the degrading factor is equal to $1/(\mathrm{M}\times\sqrt{2})$, for M in the interval from 1 to 4 (where $M = 1$ is Scenario 4). We see that the power of both tests reduces towards zero as M increases, but the weighted test always retains an advantage over the unweighted test.

4.4. Results: causal rank-preserving structural failure time model analysis

Table I (columns 4 and 5) shows the bias and mean squared error (MSE) of estimates obtained for β within a RPSFTM analysis, using the standard (un-weighted) logrank test and the logrank test with simple causal weights, for Scenarios 1–4. Unsurprisingly, both methods return unbiased estimates for the causal parameter in Scenario 1, but the MSE of the estimate obtained via the weighted analysis is only 44% of that obtained via an unweighted analysis. In Scenario 2, both tests returned estimates with a small amount of bias, with the weighted test being more severe. This bias was seen to diminish for larger trial sizes, however (results not shown). Again, MSE for the weighted analysis was considerably reduced compared with the standard unweighted analysis (approximately 61% of its magnitude).

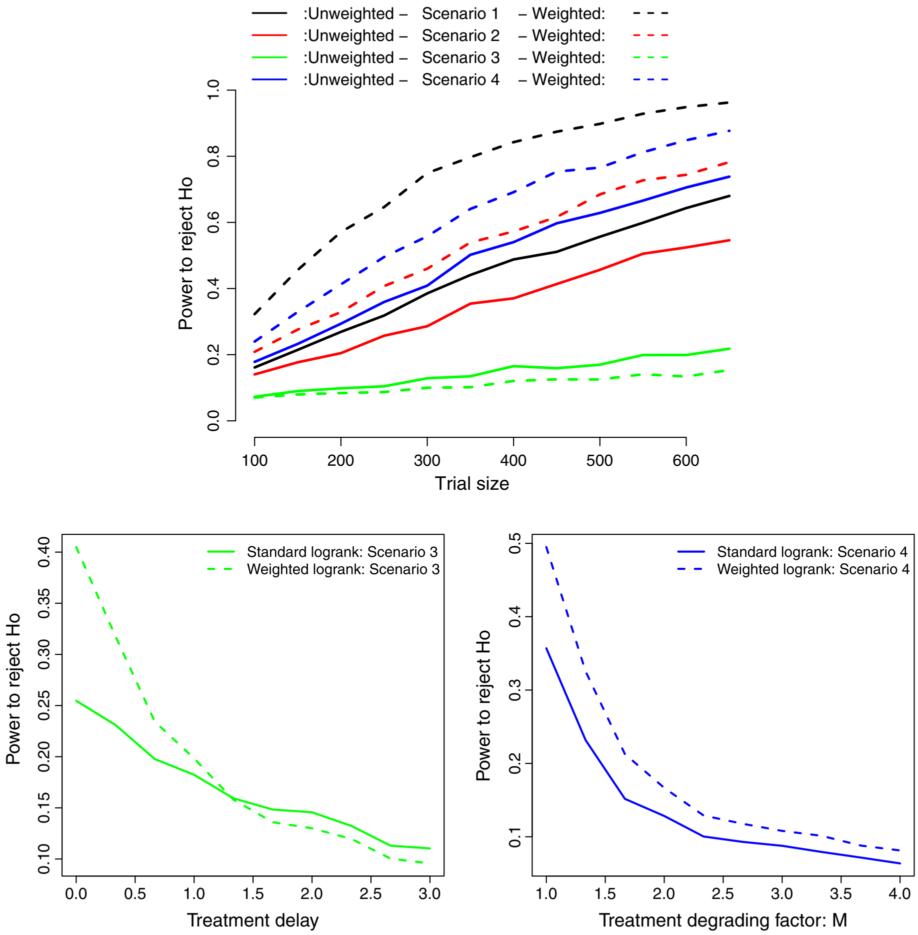

Figure 5. Top: Power to reject H_0 under H_1, $\beta = \log(0.5)$ for scenarios 1–4. Standard logrank test shown via solid lines; simple weighted logrank test shown via dashed lines. Bottom-left: Power of the weighted and unweighted logrank test to reject H_0 under H_1 for scenario 3 and for a varying delay (trial size $n = 250$). Bottom-right: Power of the weighted and unweighted logrank test to reject H_0 under H_1 for scenario 4 and for a varying degradation factor (trial size $n = 250$).

Table I. Type I error of the standard and simple weighted ITT tests under H_0 and bias/MSE of the standard and simple weighted RPSFTM analysis under H_1 for scenarios 1–4.

Scenario	Logrank	ITT analysis	Causal RPSFTM analysis	
	Method	H_0: $\beta_0 = 0$	H_1: $\beta_0 = \log(0.5)$	
		Type I error	Bias	MSE
1.	Standard	0.052	−0.003	0.204
	Simple Weighted	0.051	−0.009	0.089
2.	Standard	0.054	−0.029	0.277
	Simple Weighted	0.048	−0.037	0.170
3.	Standard	0.052	0.270	0.411
	Simple Weighted	0.051	0.417	0.381
4.	Standard	0.052	−0.117	0.266
	Simple Weighted	0.049	−0.094	0.176

ITT, intention-to-treatl; RPSFTM, rank-preserving structural failure time model; MSE, mean squared error.

Results for Scenarios 3 and 4 are more mixed, and harder to interpret, because the RPSFTM model is misspecified in both cases. For example, in Scenario 3 the treatment effect is (in truth) zero for 3 months and then $\log(0.5) = -0.693$ afterwards. Therefore, the 'average' treatment effect is somewhere between 0 and -0.693. The weighted logrank test estimates a treatment effect of around -0.28 $(0.417 + \log(0.5))$ whereas the unweighted test's estimate is -0.42 $(0.270 + \log(0.5))$. In Scenario 4, the bias is not so easily explained. The true effect is -0.693 before progression 1 and $-0.693/\sqrt{2} = -0.49$ afterwards, and one might therefore expect the mean estimates to lie somewhere between these values. However, the mean estimates were approximately -0.81 and -0.79 for the unweighted and weighted approaches, respectively. By comparing the Kaplan Meier curves in Figure 4 (top-right, no degradation) with Figure 4 (bottom-right, degradation present), one can begin to see why the unweighted ITT logrank test statistic (and therefore the ITT hazard ratio) would be larger in Scenario 4 than in Scenario 2. In Scenario 4, the $(R = 1)$ treatment group come off treatment at the point of disease progression 1 and do not take the treatment when its effect becomes degraded. The $(R = 0)$ control group, by contrast, start treatment only after disease progression 1 has occurred, so only ever experience the degraded effect. The Kaplan Meier curve of the control group is not pulled towards that of the treatment group as quickly as it would have been had no degrading occurred (Scenario 2). Moving from Scenario 2 to Scenario 4 therefore reduces the benefit of post-progression treatment for the placebo arm, thereby increasing the treatment effect estimated via an ITT or RPSFTM analysis.

In conclusion, hypothesis testing via an ITT analysis remains valid when the assumptions CWA1, CWA2 and BCA are violated, but the RPSFTM parameter will only be meaningful in general when the BCA is true. When the BCA is violated, the standard and weighted RPSFTM will identify different causal parameters depending on the treatment switching pattern and the type of violation.

5. Real data examples

In this section, we apply our weighted approach to re-analyses of the Sunitinib trial [1, 13] and the Concorde trial [4, 8].

5.1. *The Concorde trial*

Figure 6 (left) shows the Kaplan–Meier survival curves (for the earliest time to either AIDS, ARC or death) of the 1745 patients with HIV who were randomised to the Imm $(R = 1)$ or Def $(R = 0)$ arms of the Concorde trial. The logrank test statistic, for a comparison of the two survival curves, was -1.35, with an associated hazard ratio estimate of 0.89 and a p-value of 0.18. It was subsequently questioned whether a departure from the original protocol, which enabled those in the Def arm to receive Zidovudine before a diagnosis of AIDS or ARC, had diluted the difference between the two treatment strategies. For this reason, White *et al.* [8] applied the RPSFTM (with the standard logrank test) to ask what would have happened if the original trial protocol had been followed with 100% adherence. To this end, they defined the treatment indicator variable for patient i at time t, $X_i(t)$, as:

$$X_i(t) = \begin{cases} 1 & \text{if } R_i = 1 \text{ or } (R_i{=}0 \text{ and } AZT_i \leqslant t) \\ 0 & \text{Otherwise.} \end{cases}$$

where AZT_i was the time at which patient i initiated Zidovudine treatment. The causal parameter β is then defined using $X_i(t)$ via Equation (5). We aim to repeat their analysis using our simple weighted approach. Figure 6 (right) shows $\gamma^r(t)$, $(l = 0, 1)$ for the Concorde trial.

The function $\gamma^1(t)$ is fixed at 1 throughout; not because all patients in the Imm arm were on Zidovudine for their total participation in the trial (many would have stopped treatment earlier on their doctor's advice) but because all satisfied the original protocol of beginning Zidovudine treatment immediately, however briefly. Our simple weight from Equation (4), for use within an ITT analysis, is therefore $1 - \gamma^0(t)$. A further consequence of $\gamma^1(t)$ being fixed at 1 is that, in a RPSFTM analysis, no patient in the treatment arm who was originally uncensored can be 're-censored', regardless of the proposed value of β. This can be guaranteed by setting person i's censoring indicator on the β–transformed scale, $C_i(\beta)$, equal to $C_i \exp(\beta)$ (see [8] for further explanation). Before proceeding with an analysis of the Concorde trial, we first introduce our second real data example.

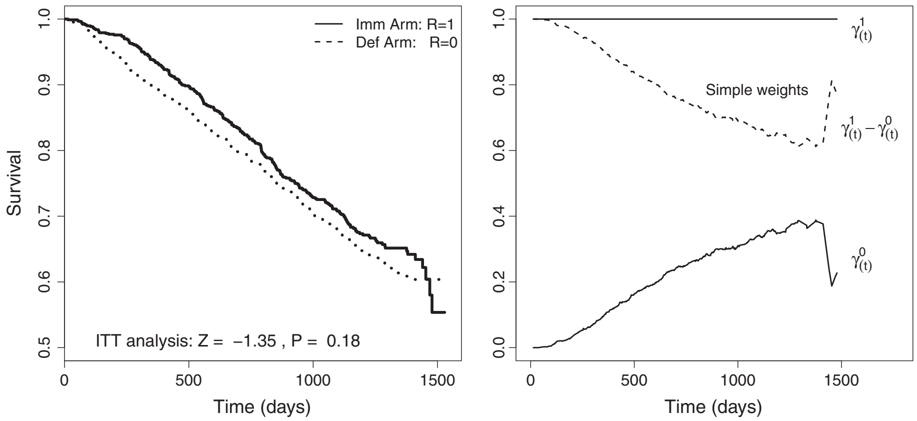

Figure 6. Concorde trial left: survival curves for the Imm and Def arms. Right: Proportion of people in the Imm and Def Concorde arms on treatment as a function of survival time.

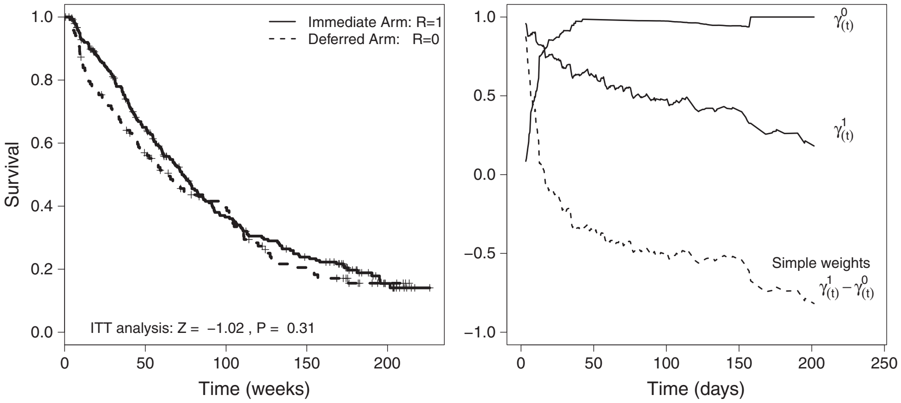

Figure 7. Sunitinib trial left: survival curves for the immediate and deferred arms. Bottom right: proportion of people in the immediate and deferred arms on treatment as a function of survival time.

5.2. The Sunitinib trial

Figure 7 (left) shows the Kaplan–Meier survival curves (time to death) for the 312 patients with gastrointestinal stromal tumours who were randomised to the Sunitinib arm ($R = 1$) and placebo arm ($R = 0$) as part of a multi-centre phase III randomised controlled trial. The trial was unblinded approximately 50 weeks after study initiation because of large observed differences in tumour progression rates between the two groups (in favour of Sunitinib).

At this point, patients in the placebo group were allowed to switch treatments and receive the new drug, which they did *en masse*. We label the two arms as ‘immediate’ or ‘deferred’ (Sunitinib) for convenience. By the end of the trial, the highly significant difference in tumour progression rates was not mirrored in the overall survival measure between the two treatment groups. The final hazard ratio estimate reduced from 0.49 (p-value 0.007) at the interim to 0.88 (p-value 0.31). In order to investigate whether the disappearance of any treatment effect was due to the treatment switching, Demetri *et al*. [14] conducted a causal analysis using the RPSFTM with the standard logrank test, which was later discussed in [13]. Here, we follow their example by defining the treatment indicator variable for patient i at time t, $X_i(t)$, as 1 if patient i was on Sunitinib at time t and 0 otherwise. This defines the parameter β as the causal effect that would have been observed if all patients on the immediate arm had received Sunitinib for the entire duration, and all patients on the deferred arm had received placebo

Figure 7 (right) shows the simple weight function for the Sunitinib trial. $\gamma^1(t)$ decreased over the follow-up period because a sizable proportion of patients in the immediate arm discontinued treatment at the point of tumour progression if their physician thought they would no longer benefit. $\gamma_0(t)$ is seen to increase rapidly from 0 to near 1 after 50 weeks of follow up. The switching in the Sunitinib trial was so extreme that, from 25 weeks of follow-up onwards, the simple ITT weights $\gamma^1(t) - \gamma^0(t)$ are negative.

Some patients who were randomised to the control arm, and who switched over to Sunitinib, would have additionally come off treatment before death. Although our method could account for this pattern (indeed, the design of the simulation study did just that), this extra information was not available for these data. Our analysis therefore assumed that the treatment indicator variable for patients in the control group was fixed to 1 from the time they started to take Sunitinib to the end of follow-up. It should therefore be treated as illustrative and conditional on this fact.

Under the assumptions used in their derivation, we would expect a negative weight at time t to coincide with a negative value for the difference between observed and expected numbers of events at time t, because the control group becomes the *de facto* treatment group. This would mean the numerator of the weighted logrank statistic in Equation (2) would still be positive, and the test would remain powerful. However, some may not be comfortable with this interpretation or the idea of negative weights at all. For this reason, we perform additional weighted ITT and causal analyses by truncating the simple weights in Equations (4) and (10) at 0 if they are negative. The truncated weights were also thought to provide a closer approximation to the true weights that would have been estimated and had full information on the treatment status of control group patients in the latter stages of the trial been available (as this would have probably have acted to restrict the weights from becoming negative in the first place).

5.3. Intention-to-treat and rank-preserving structural failure time model analyses

An ITT analysis of the Concorde data using the simple weighted logrank test yielded a test statistic Z^W of −1.64 with a corresponding p-value of 0.10. Figure 8 (left) and Table II shows the results of a causal RPSFTM analysis using both the unweighted logrank test and the weighted test with the simple weights from Equation (10).

A 95% confidence interval for β under each test can be read off from Figure 8 as the range of β values for which test statistics lie between $\pm$ 1.96. The point estimate for β (where the test statistic equals 0) is slightly larger using the weighted test and its 95% confidence interval is around 2% narrower.

An ITT analysis of the Sunitinib data using the simple weighted logrank test yields a test statistic of −1.49 and a p-value of 0.137. An ITT analysis using the zero-truncated simple weights yields a test statistic of −2.48 with a p-value of 0.013. Figure 8 (right) and Table II shows the results of a causal RPSFTM analysis using the unweighted logrank test, the simple weights from Equation (10) and the zero-truncated simple weights.

There is considerable uncertainty as to the point estimate for β based on the standard and simple weighted tests — in both cases the test statistic is not a monotonically decreasing function of β, crossing zero several times. The simple weighted statistic yields a 95% confidence interval for β, which is approximately 43% narrower than that obtained using the standard logrank test. However, the zero-truncated simple weights furnish the most precise estimate for β ($\hat{\beta}$=−1.25, 95% C.I −2.21,−0.32).

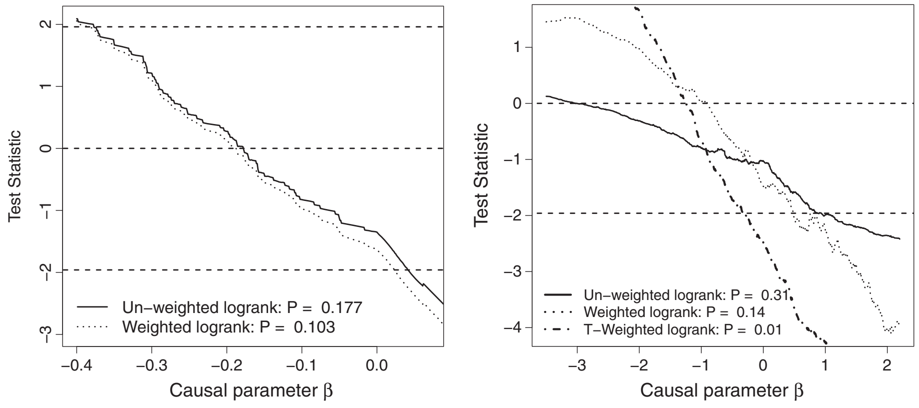

Figure 8. Estimation of the rank-preserving structural failure time model causal parameter β using the standard and weighted logrank tests for the the Concorde trial (left) and the Sunitinib trial (right). 'T-weighted' refers to the zero-truncated weighted analysis.

Table II. ITT and RPSFTM causal analyses of the Sunitinib and Concorde trials using the standard and weighted logrank test. Z^{w*} and $\hat{\beta}^{*}$ refer to the test statistic and estimate, respectively, based on zero truncated weights (Sunitinib trial only).

Trial	Switching pattern		Analysis	Standard	Simple weighted
example	R=0	R = 1	type	logrank test	logrank test
Concorde	Off → On	On	ITT	Z = −1.35 (0.177)	Z^{w} = −1.63 (0.103)
			RPSFTM	$\hat{\beta}$ = −0.178 (−0.378, 0.0411)	$\hat{\beta}$ =-0.188 (−0.386,0.0234)
Sunitinib	Off → On	On → Off	ITT	Z = −1.03 (0.306)	Z^{w} = −1.49 (0.137) Z^{w*} = −2.48 (0.013)
			RPSFTM	$\hat{\beta}$ = −3.01 (−3.95, 0.88)	$\hat{\beta}$ = −0.89(−3.06,0.56) $\hat{\beta}^{*}$ = −1.25(−2.21,−0.32)

ITT, intention-to-treatl; RPSFTM, rank-preserving structural failure time model.

6. Discussion

Given the severity of the disease, it is of paramount importance that clinical trials in late–stage cancer meet the needs of the patients who take part, as well as future patients. For this reason, the emphasis is usually, and rightly, on the pragmatic evaluation of different treatment strategies, rather than of simple all–or–nothing comparisons. Amendments to the original protocol are also therefore likely, if felt to be in the patients' interest. In this paper, we have shown that a simple weighted logrank test can be far more powerful than the standard logrank test for testing the ITT null hypothesis of no difference in the survival distributions between randomised groups when (a) substantial treatment switching occurs and (b) large rates of separation between the two group's survival functions over time coincides with large differences in the proportion of patients on treatment in each arm. This remained true even when the assumptions used to derive the weights were violated, and, when this was the case, the weighted test's type I error rate was maintained at its nominal level.

The evidence gathered from late-stage cancer trials often needs to be subsequently interpreted by external bodies who are tasked with making decisions about a single treatment's suitability for future patient populations. This may require an assessment of its effectiveness compared with another therapy under different conditions than those originally followed in the trial. The RPSFTM has been used extensively to provide quantitative answers to such 'what if' questions. In addition to showing its use for ITT analyses, we have also shown that the weighted logrank test can furnish a more powerful method for comparing counterfactual treatment assignments within the RPSFTM framework, assuming that the assumptions of the RPSFTM hold.

It is important to stress that rejection of the null hypothesis by the weighted test should not be automatically equated with a conclusion of treatment superiority. For example, a reviewer highlighted the scenario (not explicitly explored in this paper) where being treated with an experimental drug at time t has the effect of reducing the individual's hazard at time t when $t < A$, but increasing it when $t > A$, for some time A. That is, the treatment causes benefit early on, but harm later. Suppose that this change in the direction of treatment effect causes the survival curves of the two treatment arms eventually to cross over. In this scenario, the early benefit and later harm of the treatment might cancel one another out in the test statistic of the standard unweighted log rank test, so that the null hypothesis of no treatment effect were not rejected. Any test statistic that gave more weight to earlier event times, on the other hand, might be more likely to reject the null hypothesis in favour of the experimental drug. In particular, this would be true of the weighted log rank test in the presence of treatment switching. In a very real sense, it is correct to reject the null, because the survival curves are truly different, and so the failure of the standard log rank test to reject the null constitutes a type-II error [15], but it is important not to interpret this falsity of the null as being the same as treatment benefit.

If one believed that the treatment effect may not be constant over time, even in the absence of any treatment switching, then an extended RPSFTM could easily be used to probe this possibility via an additional sensitivity analysis. For example, to explore the case where the experimental treatment offered a reduced (or even negative) benefit over standard therapy after time A, then a two-parameter RPSFTM of the form:

$$T_0(\beta_0, \psi_0) = \exp\{\beta_0\} T_{\text{on}}^{\text{Pre–A}} + \exp\{\psi_0\} T_{\text{on}}^{\text{Post–A}} + T_{\text{off}}, \tag{14}$$

could be fitted instead. Here, $T_{\text{on}}^{\text{Pre–A}}$ and $T_{\text{on}}^{\text{Post–A}}$ refer to time on treatment before and after time A, respectively, and the inclusion of a second parameter ψ_0 allows for a different treatment effect either side of A. Time A could be common to all patients, or one could use patient—specific information such as time to disease progression. If Equation (14) is correctly specified, then β_0 and ψ_0 could be estimated consistently by g–estimation using this two–parameter RPSFTM. Similar models were investigated in the Concorde trial to explore the hypothesis that Zidovudine is more beneficial to patients with full-blown AIDS than those who were only HIV-positive [8]. However, this was shown to be challenging because of the increased dimensionality of the parameter space. A simpler approach would be to define ψ_0 (or perhaps even ψ_0/β_0) as a sensitivity parameter in Equation (14), and then find the values of ψ_0 for which evidence of an apparent beneficial treatment effect would be brought into doubt.

Estimation of the RPSFTM causal parameter under the weighted logrank test is more of a computational challenge than when using the standard unweighted statistic. This is because each new proposal of β necessitates re-censoring of the data, which in turn requires a re-calculation of the weights. The parametric accelerated failure time modelling approach of Branson and Whitehead [16] uses a less stringent re-censoring mechanism, and the additive hazard modelling approach of Martinussen [17] avoids the need for 'artificial' re-censoring altogether. It would be interesting to see if these approaches could be extended to incorporate time-varying treatment weights as outlined in this paper, to subsequently enable more precise causal estimates to be obtained with less computational effort.

7. Software

R code is available from the corresponding author to perform the analyses set out in this paper.

Acknowledgements

In unpublished work produced, specifically for a NICE appraisal that has now been seen by the authors (personal communication), Robins [18] derives locally efficient weights for the logrank test, for both ITT and causal RPSFTM analyses. In the ITT case, his weights are identical to the 'simple' weights given in Equation (4) of this paper, although his derivation is different. For the RPSFTM, his weights take a more complex form than those in given Equation (10). They additionally depend on the cumulative amount of treatment an individual has taken and assume that survival times follow a Weibull distribution. We wish to make the reader aware of this work in advance of its publication and to acknowledge its precedence to our own.

This work was supported by the Medical Research Council [Unit Programme number U105260558 and Methodology grant MR/N501906/1]. We also thank Pfizer and the Concorde Coordinating Committee for allowing us to use their data.

References

1. Demetri GD, van Oosterom AT, Garrett CR, Blackstein ME, Shah MH, Verweij J, McArthur G, Judson IR, Heinrich MC, Morgan JA, Desai J, Fletcher CD, George S, Bello CL, Huang X, Baum CM, Casali PG. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of Imatinib: a randomised controlled trial. *Lancet* 2006; **368**:1329–1338.
2. Motzer RJ, Escudier B, Oudard S, Hutson TE, Porta C, Bracarda S, Grünwald V, Thompson JA, Figlin RA, Hollaender N, Urbanowitz G, Berg WJ, Kay A, Lebwohl D, Ravaud A. Efficacy of everolimus in advanced renal cell carcinoma: a double-blind, randomised, placebo-controlled phase III trial. *Lancet* 2008; **372**:449–456.
3. NICE. Technology appraisal 179: Sunitinib for the treatment of gastrointestinal stromal tumours. Technical Report, National Institute for Health and Clinical Excellence, 2009.
4. Concorde Coordinating Committee. Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred Zidovudine in symptom-free HIV infection. *Lancet* 1994; **343**:871–881.
5. Gore SM, Bird AG. Concorde trial of immediate versus deferred Zidovudine (letter). *Lancet* 1994; **343**:1357–1358.
6. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank-preserving structural failure time models. *Communications in Statistics* 1991; **20**:2609–2631.
7. NICE. Renal cell carcinoma (second-line metastatic): Everolimus final appraisal determination. Technical Report, National Institute for Health & Clinical Excellence, 2010.
8. White IR, Babiker AG, Walker S, Darbyshire J. Randomization-based methods for correcting for treatment changes: examples from the Concorde trial. *Statistics in Medicine* 1999; **18**:2617–2634.
9. Schoenfeld D. The asymptotic properties of non-parametric tests for comparing survival distributions. *Biometrika* 1981; **68**:316–319.

10. Lagakos SW, Lim LL-Y, Robins JM. Adjusting for early treatment termination in comparative clinical trials. *Statistics in Medicine* 1990; **9**:1417–1424.
11. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**:553–566.
12. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 2005.
13. Blay JY. Pharmacological management of gastrointestinal stromal tumours: an update on the role of Sunitinib. *Annals of Oncology* 2010; **21**:208–215.
14. Demetri GD, Huang X, Garrett CR, Schoffski P, Blackstein M, Shah M, Verweij J, Tassell V, Baum C, Casali P. Novel statistical analysis of long-term survival to account for crossover in a phase III trial of Sunitinib (su) vs. placebo (pl) in advanced GIST after Imatinib (im) failure [poster discussion] (abstr 10524). *Journal of Clinical Oncology* 2008; **15S**:26.
15. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One* 2015; **10**(1):e0116774.
16. Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Statistics in Medicine* 2002; **21**:2449–2463.
17. Martinussen T, Vansteelandt S, Gerster M, Hjelmborg JVB. Estimation of direct effects for survival data by using the Aalen additive hazards model. *Journal of the Royal Statistical Society (Series B)* 2011; **73**:773–788.
18. Robins JM. Updated algorithm for locally efficient weighted log rank test. Technical Report, Harvard University, 2010.