The University Of Sheffield.

School of Health And Related Research

Health Economics & Decision Science (HEDS)

Discussion Paper Series

Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times?

Authors: Nicholas Latimer, Ian White, Keith Abrams, Uwe Siebert

Corresponding author: Nicholas Latimer

ScHARR, University of Sheffield,

Regent Court, 30 Regent Street,

Sheffield, S1 4DA,

Tel: +44 (0) 114 222 0821, Email: n.latimer@sheffield.ac.uk

No. 17.09

Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times?

Latimer NR[1], White IR[2], Abrams KR[3], Siebert U[4]

[1] School of Health and Related Research, University of Sheffield, UK

[2] MRC Clinical Trials Unit, University College London, UK

[3] Department of Health Sciences, University of Leicester, UK

[4] Department of Public Health, Health Services Research and Health Technology Assessment, UMIT - University for Health Sciences, Medical Informatics and Technology, Hall i.T., Austria
Oncotyrol – Center for Personalized Cancer Medicine, Innsbruck, Austria
Harvard T.H. Chan School of Public Health and Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Corresponding author: Nicholas Latimer, ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, Tel: +44 (0) 114 222 0821, Email: n.latimer@shef.ac.uk

ABSTRACT

Treatment switching often has a crucial impact on estimates of effectiveness and cost-effectiveness of new oncology treatments. Rank preserving structural failure time models (RPSFTM) and two-stage estimation (TSE) methods estimate 'counterfactual' (i.e. had there been no switching) survival times and incorporate re-censoring to guard against informative censoring in the counterfactual dataset. However, re-censoring causes a loss of longer term survival information which is problematic when estimates of long-term survival effects are required, as is often the case for health technology assessment decision making. We present a simulation study designed to investigate applications of the RPSFTM and TSE with and without re-censoring, to determine whether re-censoring should always be recommended within adjustment analyses. We investigate a context where switching is from the control group onto the experimental treatment in scenarios with varying switch proportions, treatment effect sizes and time-dependencies, disease severity and switcher prognosis. Methods were assessed according to their estimation of control group restricted mean survival (that would be observed in the absence of switching) at the end of the simulated trial follow-up. We found that RPSFTM and TSE analyses which incorporated re-censoring usually produced negative bias (i.e. under-estimating control group restricted mean survival and therefore over-estimating the treatment effect). RPSFTM and TSE analyses that did not incorporate re-censoring consistently produced positive bias (i.e. under-estimating the treatment effect) which was often smaller in magnitude than the bias associated with the re-censored analyses. We believe that analyses should be conducted with and without re-censoring, as this may provide decision makers with useful information on where the true treatment effect is likely to lie. Analyses that incorporate re-censoring should not always represent the default approach when the objective is to estimate long-term survival times and treatment effects on long-term survival.

Key words: Treatment switching; treatment crossover; survival analysis; overall survival; oncology; health technology assessment; time-to-event outcomes; prediction; re-censoring

INTRODUCTION

Treatment switching commonly occurs in randomised controlled trials (RCTs), whereby patients randomised to the control group are permitted to switch onto the experimental treatment during trial follow-up. Switching is permitted primarily due to ethical considerations, and the rationale for switching, its implications and analytical methods for adjusting for it has been the focus of much discussion in the literature.[1-4] Given that switching in trials is likely to continue to occur and often has a large impact on estimates of the effectiveness of new treatments, it is important for regulators and health technology assessors to engage with methods that attempt to adjust for switching. Several statistical adjustment methods are available, but all make strong assumptions that are not possible to test perfectly. In addition, each of these methods can be applied in a multitude of ways and seemingly innocuous choices around how a particular method is applied can importantly affect the results they produce. This is sure to influence the thinking of decision makers when they seek to interpret the results of adjustment analyses, and may lead to a lack of trust in adjustment methods. It has been suggested that decision makers require manufacturers to describe and justify adjustment analyses in detail – including rationale for each application decision made – in order that robust and informed decisions can be made.[5,6]

Whether or not to apply re-censoring represents an application decision that can have a substantial impact on the results of Rank Preserving Structural Failure Time Model (RPSFTM) and two-stage adjustment analyses. In a recently published study, Latimer *et al.* presented a series of adjustment analyses applied to a trial analysing the effect of trametinib compared to chemotherapy in patients with metastatic melanoma.[7] A standard intention-to-treat (ITT) analysis resulted in a hazard ratio (HR) of 0.72 (95% confidence interval (CI) 0.52 to 0.98). However, 67% of control group patients had switched onto the experimental treatment. An RPSFTM analysis designed to adjust for the treatment switching gave a HR of 0.38 (95% CI 0.15 to 0.95) when re-censoring was applied, and an HR of 0.49 (95% CI 0.25 to 0.96) when re-censoring was not applied. The HRs for a two-stage analysis to adjust for the treatment switching were 0.43 (95% CI 0.20 to 0.96) with re-censoring and 0.53 (95% CI 0.29 to 0.97) without re-censoring. Such substantial differences in the point-estimate of the treatment effect can be critical particularly for estimates of the expected cost-effectiveness of new interventions – overall survival benefit estimates are often the most influential parameters within cost-effectiveness models of cancer interventions.[8] Cost-effectiveness analyses are key factors in reimbursement decisions made on new healthcare interventions around the world.[9-12]

It is generally recommended to apply re-censoring – which will be described in the next section – when applying RPSFTM and two-stage adjustment methods.[13-15] However, it is recognised that whilst re-censoring helps avoid one type of bias – informative censoring – it can result in a type of missing information bias when the treatment effect changes over time, because longer-term information is lost.[1,15-18] It may therefore be possible that in some situations analyses which do not re-censor are preferable to analyses which do. Currently, little is known about the impact of re-censoring in realistic scenarios, or how results should be interpreted when the choice of whether or not to apply re-censoring has a large impact on the estimated treatment effect. Simulation studies have shown that adjustment methods produce varying levels of bias depending upon factors such as the switch proportion and the treatment effect size, but have only considered applications of adjustment methods that include re-censoring.[19-21] In this paper we conduct a new simulation study to investigate the performance of adjustment methods with and without applying re-censoring. Our objective is to determine whether it is possible to discern the likely impact of re-censoring in various scenarios, in order that expectations over the likely bias associated with analyses that do or do not apply re-censoring can be informed. This should allow analysts and decision-makers to better interpret the results of adjustment analyses, enabling more constructive use of adjustment methods.

METHODS

Statistical adjustment methods

The RPSFTM [22] and two-stage adjustment methods [20] can be used to estimate counterfactual survival times in the presence of treatment switching in RCTs – that is, they estimate survival times that would have been observed if treatment switching had not occurred.

The simple one-parameter version of the RPSFTM splits the observed event time, T_i, for each patient into time spent on the control treatment, T_{A_i}, and time spent on the intervention treatment, T_{B_i}. For patients who are randomised to the intervention treatment, and who do not switch onto the control treatment (that is, when there is full compliance in the treatment group), T_{A_i} is equal to zero. For patients randomised to the control group who do not switch onto the intervention (i.e. compliance is full in the control group) T_{B_i} is equal to zero. However, for patients who switch treatments (for whom compliance is therefore only partial) both T_{A_i} and T_{B_i} will be greater than zero. The RPSFTM method relates T_i to the counterfactual survival time (U_i) with the following causal model:

$$U_i = T_{A_i} + e^{\psi_0} T_{B_i} \quad (1)$$

$e^{-\psi_0}$ represents the acceleration factor (AF) associated with the intervention – the amount by which an individual's expected survival time is increased by treatment. The RPSFTM assumes that there is a common treatment effect associated with the experimental treatment (i.e. that the treatment effect, e^{ψ_0}, is the same no matter when the treatment is received) and that if no patients received the experimental treatment average survival times in the randomised groups would be equal. Given these assumptions, g-estimation is used to estimate ψ_0, with the true value being that for which counterfactual survival times (U_i) are independent of randomised group.[22] This is done by computing U_i for a range of values of ψ and each time testing whether the U_i are independent of randomised group.

The two-stage adjustment method also involves estimating counterfactual survival times. The counterfactual survival model (1) is again used, but the two-stage adjustment method estimates ψ based upon an assumption of no unmeasured confounding. Under the simple two-stage adjustment method, it is assumed that treatment switching only occurs after a disease-related secondary baseline, such as disease progression. Then (assuming switching is only from the control group onto the experimental treatment), post-secondary baseline survival times in control group patients who switch onto the experimental treatment are compared to those in control group patients who do not switch, using a parametric accelerated failure time model (e.g. Weibull or Generalised Gamma), controlling for prognostic covariates measured at the secondary baseline time-point and including the switch indicator as a time-dependent variable which equals '1' after the time of switch. A treatment effect (ψ) associated with switching is then obtained, and is incorporated into (1) to estimate counterfactual survival times in switching patients.

Censoring is problematic for the RPSFTM and two-stage method due to an association between treatment received, counterfactual censoring time, and prognosis. For ease of exposition, we assume the experimental treatment is beneficial, though similar arguments apply if it is harmful. The counterfactual survival model then involves shrinking survival times for all patients who receive the experimental treatment. For some patients, the event time (usually death) may not be observed – instead it is censored. For these patients, the RPSFTM and two-stage methods estimate shrunken censoring times. The amount by which survival or censoring times are shrunk depends upon the size of the treatment effect and the duration for which the experimental treatment is

received. Counterfactual censoring times will be prone to informative censoring bias if either/both of the two following criteria are met:

- If treatment discontinuation/initiation decisions are related to prognostic factors;
- If the duration of treatment is related to prognostic factors.

It has been suggested that possible bias associated with informative censoring can be avoided by breaking the dependence between the counterfactual censoring time and treatment received by re-censoring the counterfactual survival time associated with a given value of ψ (that is, $U_i(\psi)$) for all patients at the minimum of the administrative censoring time C_i and $C_i \exp \psi$, representing the earliest possible censoring time over all possible treatment trajectories, $D_i^*(\psi)$. $U_i(\psi)$ is then replaced by $D_i^*(\psi)$ if $D_i^*(\psi) < U_i(\psi)$.[13-15]

Unfortunately, re-censoring involves a loss of information as the survival data are artificially censored at a time-point earlier than the follow-up times observed in the trial. A treatment effect calculated by comparing counterfactual control group survival times and observed experimental group survival times is therefore based upon shorter-term data for the control group (see Figure 1, which presents counterfactual survival curves with and without re-censoring from the trametinib example mentioned previously). If the treatment effect is not constant over time, using the re-censored survival data would result in bias if the objective is to estimate the overall longer-term treatment effect.[15] It is common for regulatory and health technology assessment (HTA) agencies to attempt to estimate longer-term treatment effects for interventions that affect survival, with HTA agencies typically requiring estimates of lifetime treatment effects.[9-12] There has recently been considerable interest in moving away from the hazard ratio as a summary of the treatment effect, partly because treatment effects are often observed to change over time.[23,24] Therefore, there is a legitimate question as to whether re-censoring or *not* re-censoring is likely to produce least bias in an adjustment analysis, given an objective of estimating long-term survival times and treatment effects.

We aim to investigate whether re-censoring or not re-censoring is likely to produce least bias in a range of realistic scenarios.

Simulation study design

We simulated independent datasets in which treatment switching was permitted, and in which the true survival times for each treatment option were known. We then applied each of the switching adjustment methods with

and without re-censoring, and compared the bias in their estimation of restricted mean survival time (RMST) in the control group. We focussed on control group RMST because we simulated scenarios where switching was only in the control group and therefore the objective of the adjustment analysis was to estimate counterfactual survival times for the control group. For each method we also calculated the empirical standard error, root mean squared error and coverage associated with estimates of control group RMST. The study was designed such that the data simulated reflected data typically observed in clinical trials in the advanced/metastatic cancer disease area. The simulation study was conducted using Stata software, version 13.1.[25]

Underlying survival times

A joint survival and longitudinal model was used to simultaneously generate a continuous time-dependent covariate (referred to as 'biomarker') and survival times,[26] similar to the approach taken in a previous simulation study.[19] The underlying biomarker level influenced survival and was influenced by treatment received, and observed values of the biomarker (which were subject to an error term) influenced the probability of treatment switching. Within the data-generating joint model, the longitudinal model for the underlying biomarker value for the i^{th} patient at time t was:

$$biomarker_i(t) = \beta_{0_i} + \beta_1 + \beta_2 \times trt_i + \beta_3 badprog_i \qquad (2)$$

where,

$\beta_{0_i} \sim N(\beta_0, \sigma_0^2)$.

Here β_{0_i} is the random intercept, β_1 is the average rate of change of the biomarker for a patient in the control group, and $\beta_1 + \beta_2$ is the average rate of change of the biomarker for a patient in the experimental treatment group. trt_i is a binary covariate that equals 1 when the patient is in the experimental group and 0 otherwise, $badprog_i$ is a binary covariate that equals 1 when a patient has poor prognosis at baseline and 0 otherwise, and β_3 is the change in the intercept for a patient with a poor prognosis compared to a patient with a good prognosis. We simulated data in which biomarker observations were made at randomisation, and at 21 day intervals after randomisation. Biomarker observations were subject to an error term with a standard normal distribution with mean 0 and variance σ^2.

We used a 2-component mixture Weibull baseline survival function and the general survival simulation framework described by Crowther and Lambert (2013)[26] to simulate survival dependent on a time-varying biomarker. Simulating using a mixture model allows us to simulate complex hazard functions, which is important given the recognition that real-world survival data frequently does not follow standard parametric distributions.[27] The model can be written as:

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2}) \quad (3)$$

where $\lambda_1, \lambda_2 > 0$ and $\gamma_1, \gamma_2 > 0$ are scale and shape parameters, respectively. The mixture parameter, p, with $0 \leq p \leq 1$, represents the contribution of the first Weibull to the overall survival model, and $1-p$ represents the contribution of the second Weibull. The related baseline hazard function is:

$$h_0(t) = \frac{\lambda_1\gamma_1 p t^{\gamma_1-1}\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2(1-p)t^{\gamma_2-1}\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})} \quad (4)$$

The linear predictor of the survival model was incorporated as follows:

$$h_i(t) = h_0(t)\exp[\delta_1 trt_i + \eta\, t \times trt_i + \delta_2\, badprog_i + \alpha\, biomarker_i(t)] \quad (5)$$

where δ_1 is the direct effect of treatment at time 0, η is the rate at which the direct effect of treatment changes with time, δ_2 is the impact of poor prognosis, and α is the coefficient of the underlying biomarker level.

Disease progression times were simulated to equal survival times multiplied by a value from a beta distribution with shape parameters (5,10). We assumed that patients had consultations with their clinician every 21 days, and that disease progression was observed to have occurred at the first consultation following the actual progression event.

We simulated random entry into the study. The maximum administrative censoring time was set at 548 days (1.5 years), and patients in the control group had a random uniform entry time from 0 to 183 days – hence their administrative censoring times ranged from 365 to 548 days.

In the 'base case' (Scenario 1) simulation, the parameter values for the mixture Weibull survival model and the longitudinal biomarker model were:

$\beta_0 = 20$, $\sigma_0^2 = 1$, $\beta_1 = 0.04$, $\beta_2 = -0.02$, $\beta_3 = 2.5$, $\sigma^2 = 1$, $\delta_1 = -1.30$, $\delta_2 = 0.3$, $\alpha = 0.01$, $\lambda_1 = 0.00001$, $\gamma_1 = 2.0$, $\lambda_2 = 0.00001$, $\gamma_2 = 0.8$, $p = 0.5$, $\eta = 0.003$.

An example of the Kaplan-Meier curves and hazard function produced by the simulation model (in the absence of treatment switching) from a single simulated data set using the base case parameter values is presented in Figure 2. We simulated a hazard function that was initially low, then steadily increased before decreasing towards the end of the trial follow-up. This is similar to the data simulated in our previous study,[19] and we believe that this is typical of the types of hazards observed in a metastatic oncology RCT setting: initial hazards are likely to be low, because trial inclusion criteria dictate that trial participants usually have relatively good prognosis. The seriousness of the disease dictates that hazards are likely to rise, before falling in the longer term as those who remain alive are of relatively better prognosis. The resulting Kaplan Meier curves are also reminiscent of those observed in the trametinib in metastatic melanoma example presented in Figure 1.

Treatment effect in the experimental group

For the majority of scenarios investigated we cannot summarise the treatment effect experienced in the experimental group using a single value, because our hazard function includes '*t*' terms. The treatment effect initially increases during the period of greatest hazard, before falling in the longer-term. We believe that this is representative of a realistic treatment effect, which falls in the longer-term when the initial treatment effect may have worn off, or when only better prognosis patients remain alive.

In one set of scenarios we excluded the '*t*' terms from the data generating mechanism and did not use a mixture survival model, in order to test the different methods in instances with a constant treatment effect (i.e. with proportional hazards) over time. In these scenarios re-censoring should not produce bias and comparing results from scenarios with a time-dependent treatment effect to scenarios with a constant treatment effect should show how sensitive methods that apply re-censoring are to a time-dependent treatment effect. In this set of scenarios the true treatment effect was known, with δ_1 representing the log hazard ratio. In scenarios that incorporated a time-dependent treatment effect, to give an idea of the size of the treatment effect we calculated the 'average' HR and AF by generating scenario-specific survival data for a large number of patients (1 000 000) without applying switching, and by fitting Cox and accelerated failure time models to this.

The switching mechanism

Only patients in the control group could switch treatments, and switching could only occur during the three consultations immediately following disease progression – switching was not permitted before disease progression, to reflect the treatment switching typically seen in metastatic cancer trials.[1] During this 'at risk' period, the probability of switching declined for each individual patient with each simulated consultation, which were assumed to occur every 21 days. The probability of switching during the 'at risk' period was calculated using a logistic function and depended upon the time of observed disease progression, and the observed biomarker value at that time-point. In reality, switching is highly likely to be related to prognosis and therefore in half of our simulated scenarios patients with relatively good prognosis were more likely to switch, and in the other half switching was more likely in patients with relatively poor prognosis. Switching probabilities were varied to test different switching proportions. Further details on the probability of switching in different simulated groups are presented in Appendix A.

Treatment effect in switchers

For patients who were simulated to switch from the control treatment onto the experimental treatment, the period after switching was multiplied by a factor (ω) to estimate survival times incorporating the impact of switching (T_{z_i}), using the following approach:

$$T_{z_i} = T_{A_i} + \omega \times T_{B_i} \quad (6)$$

where T_{A_i} represents the time of switching and T_{B_i} represents the survival time after the switch point that was simulated to occur in the absence of switching. This is the same as the accelerated failure time model presented in (1), but here we denote the treatment effect as ω rather than $e^{-\psi_0}$.

The magnitude of ω was varied across scenarios to represent relative reductions in the average treatment effect (in terms of an AF) of 0% and 20%. This allowed us to test scenarios in which the 'common treatment effect' assumption did and did not hold. For instance, in scenarios where the common treatment effect assumption held, the scenario-specific survival data were generated for 1 000 000 patients without applying switching and the RPSFTM was applied to estimate ψ_0, with ω then set to equal $e^{-\psi_0}$. In scenarios where a 20% treatment effect reduction was simulated ω was set to equal $((e^{-\psi_0} - 1) \times 0.8) + 1$. In scenarios where there was a time-dependent treatment effect, the common treatment effect assumption did not hold in the truest sense even when the treatment effect received by switchers was the same as the average treatment effect in the experimental

group, because the treatment effect in the experimental group was time-dependent. However, in the set of scenarios that did not incorporate a time-dependent treatment effect the common treatment effect assumption did truly hold.

Scenarios investigated

The simulated data generating mechanism had several variables for which values had to be assumed. These are listed in Appendix B, together with details on how they were altered in the 'base' scenarios. Scenarios were devised in order to cover key variables that were likely to change in trials in the real world, and also to test the sensitivities of the different adjustment methods with respect to their key assumptions. Scenarios were run varying the following characteristics:

- Severity of disease: moderate severity (restricted mean survival in control group approximately 357 days, administrative censoring proportion approximately 40-50%); severe (restricted mean survival in control group approximately 228 days, administrative censoring proportion approximately 17-25%)
- Relative treatment effect reduction received by switchers: 20%; 0%
- Switch proportion: moderate (approximately 55% of control group patients who experienced disease progression) ; low (approximately 25% of control group patients who experienced disease progression)
- Treatment effect: high (average HR under the incorrect assumption of proportional hazards, approximately 0.56); moderate (average HR approximately 0.80)
- Switcher prognosis: good prognosis more likely to switch; poor prognosis more likely to switch;
- Time dependency of treatment effect: moderate ($\alpha = 0.01, \eta = 0.003$); zero ($\alpha = 0.00, \eta = 0.000$); strong ($\alpha = 0.01, \eta = 0.006$)

Using a 2x2x2x2x2x3 factorial design resulted in a total of 96 scenarios. The scenarios were numbered 1-96 with all levels of one factor nested inside one level of the next factor, following the order listed above. The first 16 scenarios were regarded as the 'base' scenarios, varying the first four factors, and holding switcher prognosis as "good prognosis more likely to switch" and the time dependency of the treatment effect as "moderate". One thousand simulations were run for each scenario. Scenario settings are detailed in Appendix C.

Adjustment methods compared

To provide context on the performance of the adjustment methods, we present results from a 'No Switching' analysis, representing the results of a standard ITT analysis undertaken on the simulated dataset before switching was applied. This does not represent a feasible estimator, but provides a useful upper bound for adjustment method performance which may be considered a 'gold standard'. We also present a standard ITT analysis after switching has been applied.

For the RPSFTM we used a log-rank test within the g-estimation procedure using the Stata command strbee.[28] We included the RPSFTM with and without re-censoring (referred to as RPSFTM and RPSFTMnr respectively).

We applied the two-stage method using a Weibull model, used disease progression as the secondary baseline time-point, and included covariates for switching, baseline prognosis group, observed biomarker value at time 0, observed time-to-disease progression, and observed biomarker value at disease progression. We included the two-stage method with and without re-censoring (referred to as TSE and TSEnr respectively).

Performance measures

We used control group restricted mean survival time (RMST) as our true value, or estimand, upon which to base our performance measures. We did not focus on estimated treatment effects because in the majority of scenarios we did not simulate proportional hazards. Our simulated survival function was not analytically tractable so for each scenario we simulated data for 1 000 000 patients without incorporating treatment switching, and we estimated the RMST at 548 days (the maximum administrative censoring time in the simulated datasets). Because this value is the product of a simulation rather than a calculation it is prone to error, but this is likely to be extremely minimal given the large number of patients simulated.

To estimate RMST at 548 days for each of the adjustment methods, we could not simply calculate the area under the counterfactual Kaplan-Meier curve because this may restrict the mean estimation to too short a time period, particularly for methods that apply re-censoring. Instead, we used what we believe is the most appropriate approach given the context that these methods are usually used in – that is, for health technology assessment. TSE, TSEnr, RPSFTM and RPSFTMnr each provide counterfactual datasets, to which we fitted flexible parametric models in order to obtain the survivor function extrapolated to 548 days. The Stata command stpm2 was used to fit the models on the log cumulative hazard scale, with 3 knots placed at equally spaced

centiles of the distribution of the log survival times [29] Where the final observed survival time was less than 548 days, the RMST at 548 days was estimated through a linear extrapolation from the last knot. This is in line with recommendations made in the UK for undertaking survival modelling in the absence of proportional hazards.[30,31] To estimate confidence intervals (CIs), counterfactual datasets were derived for the lower and upper 95% CIs of the estimated treatment effect (ψ) for each of the adjustment methods. Then flexible parametric models were fitted as described above to estimate 95% CIs for RMST at 548 days.

We evaluated the performance of methods according to the percentage bias in their estimate of control group RMST at 548 days. Percentage bias was estimated by taking the difference between the mean estimated RMST and the true RMST and expressing this as a percentage of true RMST.[32] The root mean squared error (RMSE) of the percentage bias was calculated to provide information on the variability of estimates in combination with percentage bias. The empirical standard error (SE) of the RMST estimate was also calculated for each method, as was coverage, defined as the proportion of simulations where the 95% confidence interval of the RMST contained the true RMST. Convergence was measured, defined as the proportion of times that each method resulted in an estimate of control group RMST. Percentage bias, RMSE, empirical SE and coverage were calculated based upon simulations in which convergence occurred. Monte Carlo standard errors were also calculated for each performance measure, for each method.

RESULTS

We present detailed results from 8 of the base scenarios that illustrate the key findings. First we report key results in scenarios that involved moderate (approximately 55%) and low (approximately 25%) switching proportions, before summarising the extent to which these reflect the results of the other scenarios simulated.

A summary table describing the characteristics of each scenario is presented in Appendix D. Appendices E, F and G present the percentage bias, empirical standard error and RMSE respectively across all scenarios for each method.

Scenarios with moderate switching proportions

Tables 1 and 2 present detailed results from Scenarios 1, 2, 3 and 4, in which the switching proportion was approximately 57 – 58% of at-risk patients (40 – 55% of all control group patients).

The characteristics of Scenario 1, with regard to survival times, switch proportion, treatment effect and censoring proportion, are described in Table 1. To summarise, this scenario incorporated a moderate switch proportion, a large treatment effect, a high censoring proportion, and violated the common treatment effect assumption. The average HR and AF are included in Tables 1, 2, 3 and 4 for illustrative purposes, to give an idea of the size of the treatment effect. Given that neither the proportional hazards nor constant acceleration factor assumptions held in our simulations, these estimates are prone to error.

As expected, in Scenario 1 the ITT analysis estimated a higher control group RMST than would have been observed in the absence of treatment switching, equivalent to a percentage bias of 6.5%. The RPSFTM and TSE analyses that applied re-censoring both under-estimated control group RMST, with the level of bias more appreciable for the RPSFTM (percentage bias -5.3%, compared to -1.9% for TSE). In contrast, RPSFTMnr and TSEnr analyses over-estimated control group RMST (percentage bias 2.1% for the RPSFTMnr and 3.0% for the TSEnr).

The only substantive difference between Scenario 1 and Scenario 2 was that disease severity was greater in Scenario 2, leading to the censoring proportion being approximately halved. The TSE, TSEnr and RPSFTMnr methods were relatively unaffected by this change (percentage bias -1.5%, 3.5% and 1.5% respectively), but the percentage bias associated with the RPSFTM increased (percentage bias -8.3%).

Table 2 presents detailed results of Scenario 3 and Scenario 4. Scenario 3 was approximately equivalent to Scenario 1 and Scenario 4 was approximately equivalent to Scenario 2, except the common treatment effect assumption held. This had little impact on the TSE and TSEnr analyses. However, in comparison to Scenarios 1 and 2, in Scenarios 3 and 4 the percentage bias associated with the RPSFTM reduced (percentage bias -2.8% and -5.5% in Scenarios 3 and 4 respectively), and the percentage bias associated with the RPSFTMnr increased (percentage bias 3.5% and 4.2% respectively).

Tables 1 and 2 show that coverage was poor for all the adjustment methods, although methods that applied re-censoring provided better coverage than those that did not. RMSE results demonstrate that the levels of variability associated with the different adjustment methods differed importantly. Higher levels of bias were not always associated with higher RMSEs. For instance, in all four scenarios TSEnr produced least RMSE aside from the gold standard ‘no switching’ analysis, and both the TSEnr and RPSFTMnr produced appreciably lower RMSE than TSE and RPSFTM, even when the applications that applied re-censoring resulted in lower

percentage bias. This reflects the fact that the empirical standard errors of the percentage bias differed substantially between methods. Two-stage and RPSFTM methods that applied re-censoring produced empirical standard errors that were close to double the size of those associated with methods that did not apply re-censoring. Amongst the adjustment methods, the empirical standard error was consistently lowest for the TSEnr. This was always higher than for the gold standard 'no switching' analysis, but the difference was not substantial.

Successful estimation was achieved with all of the adjustment methods across Scenarios 1, 2, 3 and 4.

Scenarios with low switching proportions

Tables 3 and 4 present detailed results from Scenarios 5, 6, 7 and 8, in which the switching proportion was approximately 25% of at-risk patients (17 – 24% of all control group patients). Scenarios 5, 6, 7 and 8 were similar to Scenarios 1, 2, 3 and 4, with the only substantive difference the switching proportion.

The reduced switching proportion had an important impact on the adjustment methods that did not apply re-censoring, with percentage bias reducing substantially for TSEnr and RPSFTMnr. In Scenarios 5-8 the RPSFTMnr and TSEnr always led to lower percentage bias than RPSFTM and TSE, whereas in Scenarios 1-4 TSE always produced lower percentage bias than TSEnr, and RPSFTM produced lower percentage bias than RPSFTMnr in Scenario 3. The direction of the bias remained the same – applications that included re-censoring resulted in negative bias, and those that did not apply re-censoring resulted in positive bias. In these scenarios, RMSE and empirical standard errors remained substantially lower for TSEnr and RPSFTMnr compared with TSE and RPSFTM. TSEnr consistently produced the lowest empirical standard errors and RMSE of the adjustment methods, and these were only marginally higher than those produced by the gold standard 'no switching' analysis.

Other base case scenarios

The results presented above provide a good overview of our key findings. RPSFTMnr and TSEnr produced positive bias in all 16 base case scenarios, over-estimating control group mean survival. RPSFTM always produced negative bias (under-estimating control group mean survival) whereas TSE produced negative bias when the treatment effect was high, but occasionally produced positive bias when the treatment effect was low. Percentage bias was increased for both methods that applied re-censoring when the treatment effect was high. TSE, TSEnr and RPSFTMnr generally produced similar levels of bias – though TSE usually produced bias in

the opposite direction. The TSE produced least percentage bias most often (see Table 5) but TSEnr produced the lowest RMSE across all scenarios. TSE always produced lower RMSE than RPSFTM, TSEnr always produced lower RMSE than RPSFTMnr, and methods that did not apply re-censoring always produced lower RMSE than those that did apply re-censoring. The RPSFTM generally performed relatively poorly, only rarely producing lower percentage bias than the RPSFTMnr.

Results according to prognosis of switchers

Scenarios 17-32 repeated Scenarios 1-16, but patients with a relatively poor prognosis were more likely to switch, rather than patients with a relatively good prognosis. The performance of RPSFTM and TSE methods was very similar to that observed in Scenarios 1-16. The impact on TSEnr and RPSFTMnr was larger – both produced reduced bias and whilst TSEnr continued to consistently produce positive bias, RPSFTMnr produced low negative bias in 4 of the 16 scenarios (although the estimated RMST always remained higher than that estimated by the RPSFTM). This occurred in scenarios in which the switching proportion was low and the common treatment effect assumption was violated.

Across Scenarios 17-32, RPSFTMnr produced least percentage bias in the most (10) scenarios, followed by TSE (5 scenarios, see Table 5). Again, often several methods resulted in similarly low levels of bias. TSEnr produced the lowest RMSE in all scenarios, and this was often only marginally higher than the RMSE associated with the gold standard 'no switching' analysis.

Scenarios with zero and strong time-dependent treatment effects

Scenarios 1-32 were repeated in Scenarios 33-64 with the treatment-related HR constant over time, and in Scenarios 65-96 with the treatment effect reducing more substantially over time.

In most of Scenarios 33-64 TSE, RPSFTM, TSEnr and RPSFTMnr produced percentage bias that was approximately half the size of that produced in Scenarios 1-32. For the RPSFTM and TSEnr, patterns in results were similar to those observed in Scenarios 1-32 – the RPSFTM produced negative bias in all but one scenario and TSEnr produced positive bias in all but 5 scenarios. In contrast, the TSE produced bias in varying directions in Scenarios 33-64, whereas it consistently produced negative bias in Scenarios 1-32. In particular, TSE produced positive bias when good prognosis patients were more likely to switch in Scenarios 33-64. The pattern in the direction of bias also altered for the RPSFTMnr in Scenarios 33-64 – positive bias was less consistently

produced when the common treatment effect assumption held, and negative bias was often produced when the common treatment effect assumption was violated. All methods generally produced very low levels of bias across Scenarios 33-64, with the only exception being the RPSFTM when there was a high, non-common treatment effect and a high switching proportion.

No single method consistently produced least bias in Scenarios 33-64 and often several methods produced similarly low levels of percentage bias (Table 5). The range of RMSE produced by the different adjustment methods was much narrower in these scenarios, but TSEnr continued to consistently produce the lowest values with these again often only marginally higher than those produced by the 'no switching' analysis.

The stronger time-dependency of the treatment effect in Scenarios 65-96 led to marginal increases in the percentage bias and RMSE associated with the adjustment methods, and patterns and directions of bias closely mimicked those observed in Scenarios 1-32. Again, no single method consistently produced least bias (Table 5). TSE, TSEnr and RPSFTMnr were the most consistent, being least affected by scenario characteristics, and TSEnr continued to consistently produce the lowest RMSE values.

DISCUSSION

Our study demonstrates the value in conducting adjustment analyses with and without re-censoring. Re-censored and non-re-censored analyses are likely to often produce bias in opposing directions, potentially providing additional information on where the true treatment effect is likely to lie.

In many of the scenarios tested – particularly those with a low and common treatment effect, and a low censoring proportion – all adjustment methods produced low percentage bias. TSE, TSEnr and RPSFTMnr produced low levels of bias across all scenarios, never performing appreciably worse than other adjustment methods. There was a trend towards non-re-censored analyses performing relatively better than re-censored analyses when the switching proportion was low and when the treatment effect was high. This is likely to be because small switching proportions mean fewer patients become informatively censored, and because re-censoring leads to a greater loss of information when the treatment effect is high. Perhaps most importantly, the direction of bias differed consistently between re-censored and non-re-censored analyses, and therefore the choice of method remains important.

Our intention was primarily to compare re-censored and non-re-censored analyses within the RPSFTM and TSE classes. However, our results also provide new information allowing us to update the comparison between these two classes. Whilst no single method produced least percentage bias consistently across all scenarios, TSEnr produced the lowest RMSE in all 96 scenarios, suggesting that when bias and variability are considered together, it consistently represented the optimal method. However, it remains important to carefully consider trial and switching characteristics to assess the likely performance of the different methods. In addition, the RPSFTM (with re-censoring) performed substantially worse than other adjustment methods in a subset of scenarios, allowing scenarios to be identified in which this method should not be relied upon.

The RPSFTM with re-censoring consistently produced negative bias (over-estimating the treatment effect) and performed substantially worse than other adjustment methods in scenarios with a high, time-dependent treatment effect (irrespective of whether there was a common treatment effect) and also when there was not a common treatment effect (irrespective of the size of the treatment effect). When the treatment effect decreases over time, re-censoring causes a negative bias (more substantially so when the treatment effect is high). In addition, when switchers receive a decreased treatment effect the RPSFTM – which assumes that the treatment effect is the same in all patients who receive it – will over-adjust survival times for switchers, again causing negative bias. Hence, the RPSFTM with re-censoring is clearly prone to negative bias in scenarios such as those investigated in this study – two-thirds of scenarios incorporated a treatment effect that reduced over time, and half incorporated a violation of the common treatment effect assumption, where switchers received a reduced treatment effect. There were only 16 scenarios with a constant, common treatment effect – and one of these scenarios represented the only instance in which the RPSFTM produced positive bias. If, in reality, the treatment effect is expected to decline over time, or the treatment effect in switchers is expected to be lower than that received by patients in the experimental group, an RPSFTM with re-censoring is highly likely to over-estimate the treatment effect. If both of these characteristics are expected, an over-estimated treatment effect is even more likely.

The TSE with re-censoring is prone to the same negative bias as the RPSFTM when the treatment effect falls over time, but not the negative bias associated with violations of the common treatment effect assumption. This explains why the TSE consistently produced more conservative estimates of restricted mean survival than the RPSFTM. When there is not a time-dependent treatment effect, the TSE with re-censoring should not result in systematic negative bias, and indeed it produced a mixture of positive and negative bias in scenarios that met

this criteria. The TSE produced low levels of positive bias in some scenarios with a decreasing treatment effect over time, when the treatment effect was low and disease severity was high. This is likely to be because re-censoring has a smaller impact in these scenarios. The TSE appears to represent a better method than the RPSFTM for adjusting for treatment switching unless the treatment effect is common and constant, provided the switching mechanism matches the requirements of the TSE method.

When re-censoring is not applied within RPSFTM and TSE adjustment methods, they are no longer exposed to the negative bias associated with a loss of longer-term information in the presence of a treatment effect that decreases over time. However, they become exposed to bias associated with informative censoring. The RPSFTMnr remains exposed to the negative bias associated with a non-common treatment effect. Originally we hypothesised that informative censoring would be associated with positive bias (over-estimates of control group survival) when poor prognosis patients were more likely to switch treatments – because more poor prognosis patients would be censored at earlier time-points than good prognosis patients. Conversely, when good prognosis patients were more likely to switch we expected that not re-censoring would lead to negative bias (under-estimates of control group survival), because good prognosis patients would generally be censored at earlier time-points. In fact, in scenarios where there was a time-dependent treatment effect, RPSFTMnr and TSEnr almost always produced positive bias, irrespective of the prognosis of switchers.

After thorough investigation, we conclude that this will occur when there are any non-switching long-term survivors (see Appendix H for more details). These patients most influence the impact of informative censoring, because re-censoring primarily affects the right-hand-side of the Kaplan-Meier curve. The implication is that TSE and TSEnr are likely to result in biases in opposing directions when the treatment effect decreases over time. It is difficult to conclude which of these biases will be greater – in our simulations TSE and TSEnr often produced similar levels of bias, in opposite directions, although as previously mentioned TSEnr always produced lower RMSE and therefore may be preferred when the aim is to estimate long-term treatment effects. Opposing directions of bias can also be expected with the RPSFTM and RPSFTMnr, provided there is a common treatment effect. This is less clear-cut when there is not a common treatment effect, but more confident conclusions may be made about which analysis is likely to produce least bias. In these scenarios the RPSFTMnr is subject to opposing forces of bias – violation of the common treatment effect assumption induces negative bias, whereas informative censoring is likely to cause positive bias. Conversely, the RPSFTM is prone to the dual negative biases associated with re-censoring and a non-common treatment effect. Whilst the RPSFTM is

likely to result in appreciable negative bias in these scenarios, the direction of bias associated with the RPSFTMnr depends upon the extent to which negative and positive biases cancel out. Given that these biases are likely to cancel out to some extent, it seems reasonable to conclude that the RPSFTMnr is likely to produce lower bias than the RPSFTM in these scenarios – this was almost exclusively the case in our simulations. The RPSFTMnr also consistently produced lower RMSE than RPSFTM and therefore may be preferred when the aim is to estimate long-term treatment effects.

We are aware of three studies that have presented analyses adjusting for treatment switching both with and without re-censoring, or which have investigated the impact of re-censoring. White *et al.* (1999) presented RPSFTM analyses undertaken on the Concorde trial of immediate versus deferred zidovudine for patients with HIV. The analysis without re-censoring led to more conservative estimates of the treatment effect and the authors observed that the treatment effect appeared to decrease over time. They concluded that their re-censored analysis may have over-estimated the treatment effect, whilst their non-re-censored analysis may have produced an under-estimate because switchers appeared to have a better prognosis than non-switchers.[15] Latimer *et al.* reported an adjustment analysis applied to an RCT comparing trametinib and chemotherapy in patients with metastatic melanoma.[7] RSPFTM and two-stage analyses which excluded re-censoring produced the most conservative estimates of the treatment effect. The authors found evidence of a decreasing treatment effect over time, and concluded that the analyses that excluded re-censoring were likely to be least biased. The pattern in these results is identical to that seen in our study. This was not the case in White and Goetghebeur's analysis of an RCT comparing two anti-hypertensive treatments affected by treatment switching. Heavily re-censored analyses resulted in less optimistic estimates of the treatment effect, because the treatment effect only became apparent in the long-term.[18] It is possible that in some situations the treatment effect may rise and then fall over time – in fact this was the pattern simulated in our scenarios (see Figure2). If re-censoring leads to analyses being based on data observed before the treatment effect becomes apparent, under-estimates of the long-term treatment effect may result. This was not the case in our simulations, but with a more delayed treatment effect it is conceivable.

Previous authors have noted that failing to re-censor may result in a small bias but a large gain in precision.[15] We found that failing to re-censor often led to a large gain in precision and reduced bias. RMSE and empirical standard errors were substantially reduced when re-censoring was excluded from RPSFTM and TSE analyses, highlighting important advantages associated with not re-censoring.

Our study has limitations. We sought to investigate many realistic scenarios, but a simulation study can never be exhaustive. Our choice of endpoint could also be questioned – we used restricted means to limit the impact of extrapolation on our results. When extrapolation is required to estimate unrestricted means (for HTA purposes) bias associated with all adjustment methods could increase, but methods that re-censor may be most seriously affected owing to the associated loss of information. Extrapolation should always be undertaken with care, incorporating external information where possible to provide credible projections.[30,31,33] We could instead have chosen to use a mean restricted to a shorter time-period, to prevent the results of re-censored analyses from being affected by extrapolation. However, given that our intention is to help inform the choice of adjustment method used primarily within HTA analyses we deemed it of little value to assess the performance of the different adjustment methods in estimating short-term treatment effects. Also, we recognise that results of survival analyses are usually summarised as hazard ratios. The majority of our scenarios had non-proportional hazards so HRs were inappropriate for measuring performance. Despite this, we did calculate 'average' HRs to allow assessment of adjusted HRs. This is presented in Appendix I. We found that estimates of HRs were prone to higher levels of bias than estimates of restricted mean survival – particularly if there is a time-dependent treatment effect and re-censoring is used. This should be borne in mind if adjustment analyses are summarised using HRs.

Also, as with previous simulation studies on switching adjustment methods, we did not incorporate bootstrapping for confidence intervals.[19,20] Coverage levels associated with the adjustment analyses are correspondingly poor because confidence intervals only took into account uncertainty in the treatment effect – not the uncertainty in the underlying survival distribution. In reality, the entire adjustment process should be bootstrapped to obtain appropriate confidence intervals.

Both re-censored and non-re-censored adjustment analyses are prone to bias, depending upon scenario characteristics. Our study provides valuable information on the likely direction and extent of these biases, and on their variability. Analyses that exclude re-censoring are likely to produce under-estimates of the treatment effect, irrespective of the perceived prognosis of switchers. Re-censored analyses are likely to produce over-estimates of the treatment effect if the treatment effect decreases over time, especially RPSFTM analyses if switchers receive a reduced treatment effect. Our results can be used to enable better interpretation of treatment switching adjustment analyses, by helping determine a range in which the true treatment effect is likely to lie. We suggest that analyses should be conducted with and without re-censoring, and that analyses that incorporate

re-censoring should not always represent the default approach when the objective is to estimate long-term survival times and treatment effects. This could be pivotal in the context of HTA, where accurate estimates of long-term treatment effects are critical to evaluations of the cost-effectiveness of novel treatments.

Acknowledgements

Parts of this work have been presented at the following conferences:

International Society for Pharmacoeconomics and Outcomes Research, 22nd Annual International Meeting, Boston, United States, May 2017

Funding

NRL is supported by the National Institute for Health Research (NIHR Post Doctoral Fellowship, Dr Nicholas Latimer, PDF-2015-08-022)

IRW is supported by the Medical Research Council (Programme number MC_UU_12023/21)

KRA was partially supported by the National Institute for Health Research (NIHR) as a Senior Investigator [NF-SI-0512-10159] & is a NIHR Senior Investigator Emeritus

US was in part supported by the COMET Center ONCOTYROL (Grant no. 2073085), which is funded by the Austrian Federal Ministries BMVIT/BMWFJ (via FFG) and the Tiroler Zukunftsstiftung/Standortagentur Tirol (SAT)

Conflict of interests statement

This report is independent research partly supported by the National Institute for Health Research and the Medical Research Council. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

REFERENCES

[1]. Latimer NR, Abrams KR, Lambert PC et al. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – an economic evaluation context: Methods, limitations and recommendations. Med Decis Making. DOI:10.1177/0272989X13520192 (2014).

[2]. Jonsson L, Sandin R, Ekman M et al. Analyzing overall survival in randomized controlled trials with crossover and implications for economic evaluation. Value Health. 17(6),707-713 (2014).

[3]. Ishak KJ, Proskorovsky I, Korytowsky B, Sandin R, Faivre S, Valle J. Methods for adjusting for bias due to crossover in oncology trials. Pharmacoeconomics. 32(6),533-46 (2014).

[4]. Watkins C, Huang X, Latimer N, Tang Y, Wright EJ. Adjusting overall survival for treatment switches: commonly used methods and practical application. Pharm Stat. 12(6),348-57 (2013).

[5]. Latimer NR, Henshall C, Siebert U, Bell H. Treatment Switching: statistical and decision making challenges and approaches. International Journal of Technology Assessment in Health Care 32(3):160-166 14 Sep 2016

[6]. Henshall C, Latimer NR, Sansom L, Ward RL. Treatment switching in cancer trials: Issues and proposals. International Journal of Technology Assessment in Health Care 32(3):167-174 01 Jan 2016

[7]. Latimer NR, Bell H, Abrams KR, Amonkar MM, Casey M. Adjusting for treatment switching in the METRIC study shows further improved overall survival with trametinib compared with chemotherapy. Cancer Medicine 2016; 5(5):806–815

[8]. Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. European Journal of Cancer 2006;42;17:2867-2875

[9]. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. London: NICE, 2013. nice.org.uk/process/pmg9 (Accessed 2 June 2017).

[10]. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. Oxford University Press Inc., New York, 2006

[11]. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M. et al. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. JAMA 2016; 316(10):1093-1103

[12]. Canadian Agency for Drugs and Technologies in Health, Guidelines for the economic evaluation of health technologies: Canada, 4th Edition, 2017

[13]. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Health Service Research Methodology: A Focus on AIDS. Eds: Sechrest L., Freeman H., Mulley A. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research.1989;113-159.

[14]. Robins JM. Analytic methods for estimating HIV treatment and cofactor effects. Methodological Issues of AIDS Mental Health Research. Eds: Ostrow D.G., Kessler R. New York: Plenum Publishing. 1993;213-290.

[15]. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. Stat Med. 1999; 18(19):2617-2634.

[16]. Latimer, N. R., K. R. Abrams, M. M. Amonkar, C. Stapelkamp, and R. S. Swann. 2015. Adjusting for the confounding effects of treatment switching – the Break-3 trial: dabrafenib versus dacarbazine. Oncologist20:798–805.

[17]. Walker, A. S., I. R. White, and A. G. Babiker. 2004. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. Stat. Med. 23:571–590.

[18]. White IR and Goetghebeur EJT. Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? Stat Med. 1998;17(3):319-339

[19]. Latimer NR, Abrams KR, Lambert PC, Morden JP, Crowther MJ. Assessing methods for dealing with treatment switching in clinical trials: A follow-up simulation study. Stat Methods Med Res. 25 Apr 2016

[20]. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting for treatment switching in randomised controlled trials – A simulation study and a simplified two-stage method. Stat Methods Med Res. 21 Nov 2014

[21]. Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ. Assessing statistical methods for dealing with treatment switching in randomised controlled trials: A simulation study. BMC Methodology 2011;11(4).

[22]. Robins JM, Tsiatis AA. Correcting for Noncompliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. Commun Stat Theory Methods. 1991; 20(8):2609-2631.

[23]. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. Journal of Clinical Oncology 2014;32;22:2380-2385

[24]. Hernan MA. The hazards of hazard ratios. Epidemiology 21:13-15, 2010

[25]. StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

[26]. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Statistics in Medicine 2013;32(23):4118-4134.

[27]. Crowther MJ, Lambert PC. A general framework for parametric survival analysis. Statistics in Medicine 2014;33(30):5280-5297.

[28]. White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. The STATA Journal 2002; 2(2):140-150.

[29]. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. The Stata Journal 9[2],265-290. 2009

[30]. Latimer NR. Survival analysis for economic evaluations alongside clinical trials – Extrapolation with patient-level data: Inconsistencies, limitations, and a practical guide. Medical Decision Making 2013;33(6):743-754.

[31]. Latimer N. NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data, Report by the Decision Support Unit, June 2011.

[32]. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25:4279-4292

[33]. Guyot P, Ades AE, Beasley M, Lueza B, Pignon J-P, Welton NJ. Extrapolation of survival curves from cancer trials using external information. Med Decis Making 2016;37(4):353-366.

Table 1: Scenarios 1 and 2 – performance measures for estimation of control arm RMST

Scenario details	Method	Bias (% of true RMST)	Empirical SE (% of true RMST)	RMSE (% of true RMST)	Coverage (%)	Successful estimation (%)
Scenario number: 1 True RMST: Control: 357 Experimental: 430 Mean switch: 57% True ave. HR: 0.57 True ave. AF: 1.53 Mean censored: 49% Switcher treatment effect: 20% reduction	No switching	0.2	3.8	3.8	94.3	100
	ITT	6.5	3.6	7.5	58.8	100
	TSE	-1.9	6.6	6.9	66.8	100
	TSEnr	3.0	4.0	5.0	33.6	100
	RPSFTM	-5.3	9.3	10.7	56.8	100
	RPSFTMnr	2.1	5.1	5.5	36.0	100
	min/max MC error	0.1/0.3	0.1/0.2	0.1/0.3	0.7/1.6	-
Scenario number: 2 True RMST: Control: 228 Experimental: 322 Mean switch: 58% True ave. HR: 0.57 True ave. AF: 1.85 Mean censored: 26% Switcher treatment effect: 20% reduction	No switching	0.2	6.1	6.1	93.4	100
	ITT	12.7	6.0	14.1	45.0	100
	TSE	-1.5	8.8	8.9	63.4	100
	TSEnr	3.5	6.8	7.6	36.6	100
	RPSFTM	-8.3	12.9	15.3	53.9	100
	RPSFTMnr	1.6	8.6	8.8	41.4	100
	min/max MC error	0.2/0.4	0.1/0.3	0.2/0.3	0.8/1.6	-

Note: RMST: restricted mean survival time; HR: hazard ratio; AF: acceleration factor; SE: standard error; RMSE: root mean squared error; MC: Monte-Carlo; ITT: intention to treat; TSE: two-stage estimation; TSEnr: two-stage estimation without re-censoring; RPSFTM: rank preserving structural failure time model; RPSFTMnr: rank preserving structural failure time model without re-censoring

Table 2: Scenarios 3 and 4 – performance measures for estimation of control arm RMST

Scenario details	Method	Percent bias	Empirical SE of % bias	RMSE of % bias	Coverage (%)	Successful estimation (%)
Scenario number: 3 True RMST: Control: 357 Experimental: 430 Mean switch: 57% True ave. HR: 0.57 True ave. AF: 1.53 Mean censored: 50% Switcher treatment effect: 0% reduction	No switching	-0.0	3.7	3.7	94.4	100
	ITT	7.5	3.4	8.3	46.7	100
	TSE	-2.3	6.9	7.3	63.1	100
	TSEnr	3.5	3.9	5.3	29.4	100
	RPSFTM	-2.8	8.9	9.3	62.1	100
	RPSFTMnr	3.5	4.9	6.1	30.7	100
	min/max MC error	0.1/0.3	0.1/0.2	0.1/0.3	0.7/1.6	-
Scenario number: 4 True RMST: Control: 228 Experimental: 322 Mean switch: 57% True ave. HR: 0.57 True ave. AF: 1.85 Mean censored: 26% Switcher treatment effect: 0% reduction	No switching	-0.1	5.7	5.7	94.7	100
	ITT	15.1	5.5	16.0	29.1	100
	TSE	-3.5	9.1	9.8	64.3	100
	TSEnr	4.0	6.5	7.6	33.0	100
	RPSFTM	-5.5	11.8	13.0	64.9	100
	RPSFTMnr	4.2	8.2	9.2	40.5	100
	min/max MC error	0.2/0.4	0.1/0.3	0.1/0.3	0.7/1.6	-

Note: RMST: restricted mean survival time; HR: hazard ratio; AF: acceleration factor; SE: standard error; RMSE: root mean squared error; MC: Monte-Carlo; ITT: intention to treat; TSE: two-stage estimation; TSEnr: two-stage estimation without re-censoring; RPSFTM: rank preserving structural failure time model; RPSFTMnr: rank preserving structural failure time model without re-censoring

Table 3: Scenarios 5 and 6 – performance measures for estimation of control arm RMST

Scenario details	Method	Percent bias	Empirical SE of % bias	RMSE of % bias	Coverage (%)	Successful estimation (%)
Scenario number: 5 True RMST: Control: 357 Experimental: 430 Mean switch: 25% True ave. HR: 0.57 True ave. AF: 1.53 Mean censored: 48% Switcher treatment effect: 20% reduction	No switching	0.1	3.7	3.7	94.6	100
	ITT	2.8	3.6	4.5	90.0	100
	TSE	-1.6	5.7	5.9	53.5	100
	TSEnr	1.4	3.7	4.0	24.7	100
	RPSFTM	-3.8	7.2	8.2	36.3	100
	RPSFTMnr	0.9	4.1	4.2	17.2	100
	min/max MC error	0.1/0.2	0.1/0.2	0.1/0.2	0.7/1.6	-
Scenario number: 6 True RMST: Control: 228 Experimental: 322 Mean switch: 25% True ave. HR: 0.57 True ave. AF: 1.85 Mean censored: 25% Switcher treatment effect: 20% reduction	No switching	0.3	6.1	6.1	93.4	100
	ITT	5.5	5.9	8.0	87.2	100
	TSE	-2.0	8.0	8.2	57.2	100
	TSEnr	1.7	6.2	6.4	26.9	100
	RPSFTM	-6.5	10.4	12.3	40.3	100
	RPSFTMnr	0.8	6.9	6.9	21.8	100
	min/max MC error	0.2/0.3	0.1/0.2	0.1/0.3	0.8/1.6	-

Note: RMST: restricted mean survival time; HR: hazard ratio; AF: acceleration factor; SE: standard error; RMSE: root mean squared error; MC: Monte-Carlo; ITT: intention to treat; TSE: two-stage estimation; TSEnr: two-stage estimation without re-censoring; RPSFTM: rank preserving structural failure time model; RPSFTMnr: rank preserving structural failure time model without re-censoring

Table 4: Scenarios 7 and 8 – performance measures for estimation of control arm RMST

Scenario details	Method	Percent bias	Empirical SE of % bias	RMSE of % bias	Coverage (%)	Successful estimation (%)
Scenario number: 7 True RMST: Control: 357 Experimental: 430 Mean switch: 57% True ave. HR: 0.57 True ave. AF: 1.53 Mean censored: 25% Switcher treatment effect: 0% reduction	No switching	-0.2	3.8	3.8	94.5	100
	ITT	3.0	3.7	4.8	87.4	100
	TSE	-2.3	6.7	7.1	48.6	100
	TSEnr	1.3	3.9	4.1	25.1	100
	RPSFTM	-3.0	7.3	7.9	35.5	100
	RPSFTMnr	1.2	4.2	4.3	16.5	100
	min/max MC error	0.1/0.2	0.1/0.2	0.1/0.2	0.7/1.6	-
Scenario number: 8 True RMST: Control: 228 Experimental: 322 Mean switch: 25% True ave. HR: 0.57 True ave. AF: 1.85 Mean censored: 25% Switcher treatment effect: 0% reduction	No switching	0.1	5.7	5.7	95.4	100
	ITT	6.2	5.6	8.3	86.4	100
	TSE	-3.3	8.4	9.0	59.9	100
	TSEnr	1.8	5.9	6.1	28.7	100
	RPSFTM	-5.5	10.0	11.4	44.9	100
	RPSFTMnr	1.5	6.6	6.7	21.9	100
	min/max MC error	0.2/0.3	0.1/0.2	0.1/0.2	0.7/1.6	-

Note: RMST: restricted mean survival time; HR: hazard ratio; AF: acceleration factor; SE: standard error; RMSE: root mean squared error; MC: Monte-Carlo; ITT: intention to treat; TSE: two-stage estimation; TSEnr: two-stage estimation without re-censoring; RPSFTM: rank preserving structural failure time model; RPSFTMnr: rank preserving structural failure time model without re-censoring

Table 5: Methods producing least bias

Method	Scenarios 1-16	Scenarios 17-32	Scenarios 33-48	Scenarios 49-64	Scenarios 64-80	Scenarios 81-96	Total
ITT	0	0	0	0	0	0	0
TSE	10	5	6	4	8	3	36
TSEnr	0	1	1	4	3	4	13
RPSFTM	1	0	4	3	3	1	12
RPSFTMnr	5	10	5	5	2	8	35

Note: ITT: intention to treat; TSE: two-stage estimation; TSEnr: two-stage estimation without re-censoring; RPSFTM: rank preserving structural failure time model; RPSFTMnr: rank preserving structural failure time model without re-censoring

Figure 1. (a) Overall survival in primary efficacy population. (A). Rank-preserving structural failure time models (RPSFTM) with re-censoring. (B). RPSFTM without re-censoring. (C). Two-stage method with re-censoring. (D). Two-stage method without re-censoring. Reproduced from Latimer *et al*, 2016 [7]

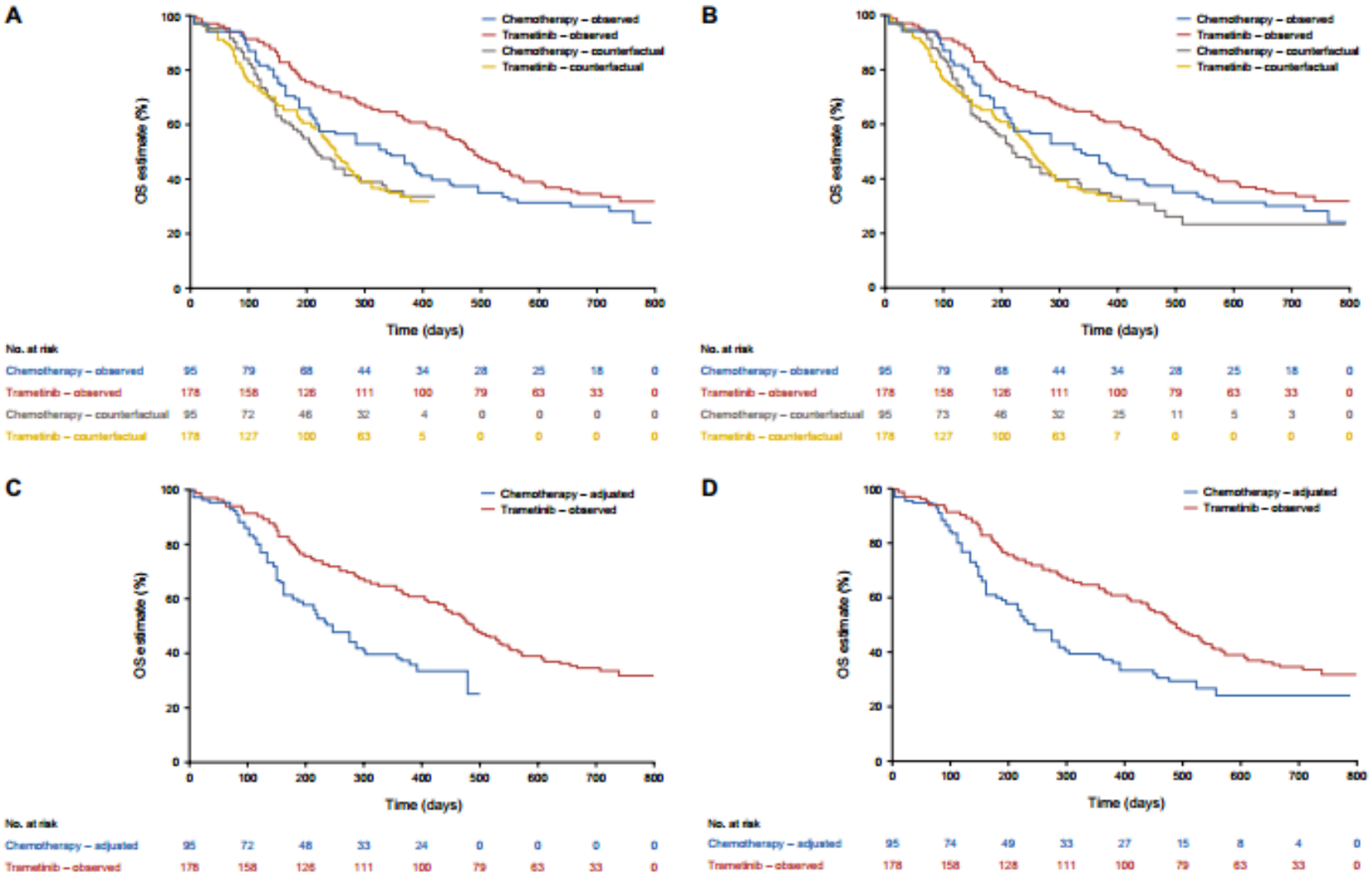
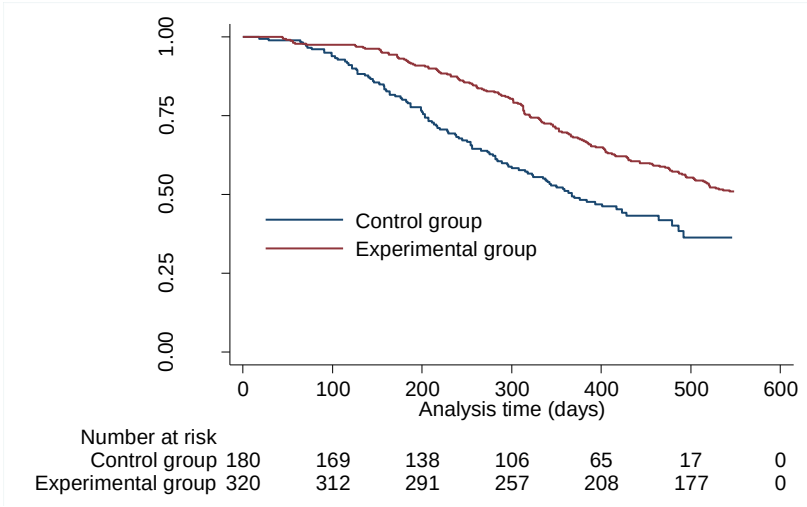


Figure 2. One simulated dataset from Scenario 1 with no switching: (a) Overall survival Kaplan–Meier (b) Smoothed hazard rate

(a)



(b)