Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2017, Vol. 26(2) 724–751
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280214557578
journals.sagepub.com/home/smm
SAGE

Adjusting for treatment switching in randomised controlled trials – A simulation study and a simplified two-stage method

Nicholas R Latimer,[1] KR Abrams,[2] PC Lambert,[2,3] MJ Crowther,[2] AJ Wailoo,[1] JP Morden,[4] RL Akehurst[1] and MJ Campbell[1]

Abstract

Estimates of the overall survival benefit of new cancer treatments are often confounded by treatment switching in randomised controlled trials (RCTs) – whereby patients randomised to the control group are permitted to switch onto the experimental treatment upon disease progression. In health technology assessment, estimates of the unconfounded overall survival benefit associated with the new treatment are needed. Several switching adjustment methods have been advocated in the literature, some of which have been used in health technology assessment. However, it is unclear which methods are likely to produce least bias in realistic RCT-based scenarios. We simulated RCTs in which switching, associated with patient prognosis, was permitted. Treatment effect size and time dependency, switching proportions and disease severity were varied across scenarios. We assessed the performance of alternative adjustment methods based upon bias, coverage and mean squared error, related to the estimation of true restricted mean survival in the absence of switching in the control group. We found that when the treatment effect was not time-dependent, rank preserving structural failure time models (RPSFTM) and iterative parameter estimation methods produced low levels of bias. However, in the presence of a time-dependent treatment effect, these methods produced higher levels of bias, similar to those produced by an inverse probability of censoring weights method. The inverse probability of censoring weights and structural nested models produced high levels of bias when switching proportions exceeded 85%. A simplified two-stage Weibull method produced low bias across all scenarios and provided the treatment switching mechanism is suitable, represents an appropriate adjustment method.

Keywords

survival analysis, prediction, treatment switching, treatment crossover, health technology assessment, inverse probability of censoring weights, time-to-event outcomes

[1]School of Health and Related Research, University of Sheffield, Sheffield, UK
[2]Department of Health Sciences, University of Leicester, Leicester, UK
[3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[4]Clinical Trials and Statistics Unit (ICR-CTSU), Division of Clinical Studies, The Institute of Cancer Research, London, UK

Corresponding author:
Nicholas R Latimer, ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK.
Email: n.latimer@shef.ac.uk

1 Introduction

It is commonplace for new drugs to be assessed formally by health technology assessment (HTA) agencies for their effectiveness and value for money before approval is given for their reimbursement. Typically, the evidence to support the effectiveness of the drug comes from randomised controlled trials (RCTs) from which the effect size for the intervention is estimated. Clearly, for a fair assessment of the drug, estimating the effect size is of central importance. For treatments that affect survival, it is recommended that economic evaluations take a lifetime time horizon, and thus appropriate estimates of overall survival (OS) are key.[1–4] However, treatment switching – where patients randomised to the control group of a clinical trial are permitted to switch onto the experimental treatment at some point during follow-up – is common in trials of oncology treatments and causes problems for HTA.[5–8] RCTs allow a comparison of effects between the novel drug and a comparator, used in separate arms of the trial. When treatment switching occurs, the separation of the treatment arms is lost. If control group patients switch and benefit from the experimental treatment, an intention-to-treat (ITT) analysis (a comparison of treatment groups as randomised) will underestimate the 'true' survival benefit associated with the new treatment – that is, the benefit that would have been observed had treatment switching not been allowed.

Treatment switching may occur for a number of reasons, both ethical and practical. Ethically, when there are no other non-palliative treatments available, it may be deemed inappropriate to deny control group patients the new treatment if interim analyses indicate a positive treatment effect. Practically, including the possibility of treatment switching within a trial protocol is likely to significantly help enrolment as patients (and their clinicians) know that they are likely to receive the novel treatment at some point whichever trial group they are randomised to. In addition, clinical trials of cancer treatments are often powered to investigate differences in progression-free survival as a primary endpoint, rather than OS, because drug regulatory agencies such as the United States Food and Drug Administration and the European Medicines Agency accept that this represents an acceptable primary endpoint for drug approval.[9,10] Hence, there is less motivation for pharmaceutical companies to ensure that randomised groups are maintained beyond disease progression for registration purposes.

Simple methods for adjusting for treatment switching, such as excluding or censoring patients who switch, will clearly lead to substantial bias when switching is associated with prognosis. More complex switching adjustment methods have been described in the literature, and previous research has shown that some of these, such as the rank preserving structural failure time model (RPSFTM),[11] perform very well when their key methodological assumptions are satisfied.[5] However, a full comparison of all relevant methods across a range of realistic scenarios – including scenarios where key methodological assumptions are *not* satisfied – has not previously been undertaken. The aim of this paper is to evaluate the relative performance (in terms of bias, coverage and mean squared error (MSE) related to the estimation of true mean survival in the control arm of a trial) of alternative switching adjustment methods in a range of realistic scenarios. Our objective is to identify which adjustment methods allow us to most accurately estimate what mean survival in the control group would have been in the absence of treatment switching because means are what are used in economic evaluation, in order that appropriate decisions on resource allocation can be made.

Section 2 provides a brief overview of simple and more complex switching adjustment methods. Section 3 describes the simulation study undertaken to assess the performance of the alternative methods. Section 4 presents key results of the simulation study. Section 5 discusses the results and their implications, while Section 6 considers the limitations of our study and suggests future research priorities.

2 Switching adjustment methods

The different switching adjustment methods considered in this paper can be grouped into simple methods (those which are currently widely used) and more complex methods.[6] Further, the more complex methods can be classified as 'observational-based' methods and 'randomisation-based' methods (referred to as 'randomisation-based efficacy estimators' by White et al.[12]). An additional method not before considered in the literature is described in Section 2.3.

2.1 Simple methods

2.1.1 Intention-to-treat

An ITT analysis does not attempt to adjust for treatment switching but represents the standard analysis undertaken on an RCT. Groups are compared as initially randomised, and thus the randomisation balance of the trial is respected. The ITT analysis represents a valid comparison of randomised groups, but in the presence of treatment switching, this may not be what is required for an HTA because it is not a comparison of patients treated with versus without the new therapy.[12]

2.1.2 Per protocol – Excluding and censoring switchers

Where treatment received in an RCT differs from what was planned, a common approach to analysing the resulting data is to conduct a per protocol (PP) analysis. In the case of treatment switching, data from patients who switch would either be excluded entirely from the analysis, or would be censored at the point of the switch. Such analyses are prone to selection bias because the randomisation balance between groups may be broken, particularly if switching is associated with prognostic patient characteristics.[13,14,15]

2.2 Complex methods

2.2.1 Observational-based complex methods

2.2.1.1 Inverse probability of censoring weights. The inverse probability of censoring weights (IPCW) method represents a proportional hazards approach to adjusting estimates of a treatment effect in the presence of informative censoring. Data are sorted into a panel format, with observations for individuals recorded at regular intervals through time until death or censoring. In the context of treatment switching from the control onto the experimental treatment, patients are artificially censored at the time of switch. Remaining observations for control group patients are weighted based upon baseline and time-dependent values of prognostic covariates in an attempt to remove the selection bias caused by the censoring – patients who do not switch and have similar characteristics to patients who did switch receive higher weights. In effect, the IPCW formulates a 'pseudo population' for the control group, through applying weights to control group patients who did not switch.

Stabilised weights applied to each individual for time interval (t), as specified by Hernan et al.[16] are

$$\hat{W}(t) = \prod\nolimits_{k=0}^{t} \frac{Pr[C(k) = 0|\bar{C}(k-1) = 0, \bar{A}(k-1), V, T > k]}{Pr[C(k) = 0|\bar{C}(k-1) = 0, \bar{A}(k-1), \bar{L}(k)V, T > k]} \tag{1}$$

where $C(k)$ is an indicator function demonstrating whether informative censoring (switching) had occurred at the end of interval k, and $\bar{C}(k-1)$ denotes censoring history up to the end of the previous interval $(k-1)$. $\bar{A}(k-1)$ denotes an individual's treatment history up until the end of the previous interval ($k - 1$), and V is an array of an individual's baseline covariates. $\bar{L}(k)$ denotes the history of an individual's time-dependent covariates measured at or prior to the beginning of interval k. Hence, the numerator of equation (1) represents the probability of an individual remaining uncensored (i.e. not having switched) at the end of interval k given that that individual was uncensored at the end of the previous interval ($k - 1$), conditional on baseline characteristics and past treatment history. The denominator represents that same probability conditional on baseline characteristics, time-dependent characteristics and past treatment history. When the cause of informative censoring is treatment switching, past treatment history is removed from the model because as soon as switching occurs the individual is censored. Artificially censoring patients at the point of switch is associated with selection bias if switching is related to baseline and time-dependent prognostic factors. Stabilised weights attempt to account for any time-dependent confounding, removing this selection bias.

The IPCW-adjusted Cox hazard ratio (HR) can be estimated by fitting a time-dependent Cox model to a dataset in which switching patients are artificially censored at the point of switching.[17] The model includes baseline covariates and uses the time-varying stabilised weights for each patient and each time interval. Similarly, the Kaplan–Meier estimator and log-rank test can be replaced with their respective IPCW versions.[18]

The IPCW method is reliant on the 'no unmeasured confounders' assumption – only if there are data on all time-dependent prognostic factors for mortality that independently predict informative censoring (switching) can the method produce unbiased results – hence, data collection in the clinical trial in question must be comprehensive. This assumption cannot be tested using the observed data.[19,20] Models for switching and survival must be correctly specified,[21] and the method fails if there are any covariates which ensure (that is, the probability equals 1) that treatment switching will occur.[17,20,22]

2.2.1.2. Structural nested models. Structural nested failure time models (SNMs) are causal models which estimate the effect of a time-dependent treatment on a survival time outcome in the presence of time-dependent confounding. They were developed for use on observational and RCT datasets.[23,24] Here, we describe a simple, one-parameter SNM. Counterfactual survival times – that is, the survival times that would have been observed if no treatment had been given – are fundamental to SNM methodology. An accelerated failure time (AFT) model structure is used, such as that presented by Robins[23]

$$U = \int_0^T \exp[\psi A_i(t)]\mathrm{d}t \tag{2}$$

where U is the counterfactual survival time for each patient, which is a known function of observed survival time, T, observed treatment A, (where A is a binary time-dependent variable equal to 1 or 0 over time) and the unknown treatment effect parameter ψ. It is assumed that exposure to treatment accelerates the time-to-event (such as death) by a factor $\exp(-\psi)$, and that exposure to treatment is independent of counterfactual survival times, conditional on a 'no unmeasured confounders' assumption. To identify the SNM estimate of the treatment effect, the model presented in equation (2) is used to estimate counterfactual survival times for a range of possible values of ψ. Then, g-estimation is used to determine a value ψ_0 for which treatment exposure at each time-point is independent of counterfactual survival. Provided that the 'no unmeasured confounders' assumption holds, it is this value of ψ_0 that provides us with the average treatment effect associated with exposure to the treatment. The model used for the g-test, as specified by Robins,[23] is a time-dependent Cox proportional hazards model for the hazard of treatment change

$$\lambda_0(t)\exp[\alpha' W(t)] \tag{3}$$

where $W(t)$ is a known vector-valued function of treatment history and covariate history up until time t, α is an unknown parameter vector and $\lambda_0(t)$ is an unspecified baseline hazard function. To conduct the g-test the term $\theta Q(t,\psi)$ is added to $\alpha' W(t)$ in the model, where $Q(t,\psi)$ is a function of treatment and covariate history up until time t and the estimated counterfactual survival time for a given value of ψ. The value of ψ that results in a Cox partial likelihood score test (g-test) statistic of zero for the hypothesis $\theta = 0$ in this model provides a consistent and asymptotically normal estimator of ψ_0, given the 'no unobserved confounders' assumptions holds, the Cox model of the hazard of treatment change is correct, and the SNM is correct. When these requirements are satisfied, an estimate of the treatment effect has been found that is adjusted to take into account any differences in covariate history between patients who received treatment and patients who did not. The confidence interval (CI) for ψ_0 is given by the values of ψ that result in the g-test not being rejected at the 0.05 level.[23]

In this paper, we apply the SNM to control group patients after the point of disease progression. Hence, the time of progression becomes time zero, and we compare control group patients who switched onto the experimental treatment to patients who did not. In this instance, the SNM should provide an estimate of the treatment effect specifically associated with switchers compared to non-switchers, adjusted for any time-dependent prognostic differences between these patient groups. The method provides us with counterfactual survival times for switchers, and thus we can compare observed survival times in the experimental group to counterfactual survival times in the control group.

Like the IPCW, the SNM method is reliant upon the untestable 'no unmeasured confounders' assumption, which requires that all variables that contribute to the process that determines whether a patient switches treatment are measured.[25]

2.2.2 Randomisation-based efficacy estimators

2.2.2.1. Rank preserving structural failure time model. The RPSFTM uses the same causal model for counterfactual survival as the SNM introduced in equation (2).[11] The only difference is that the RPSFTM bases estimation of the treatment effect on the randomisation of the trial, treatment history and observed survival times and does not make the 'no unmeasured confounders' assumption. In this paper, we consider the simple one-parameter version of this model. The method splits the observed event time, T_i, for each patient into two, that is the event time when the patient is on the control treatment, T_{A_i}, and the event time when the patient is 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 total 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 imperfect), both T_{A_i} and T_{B_i} will be greater than zero.

The RPSFTM method relates T_i to the counterfactual event time (U_i) with the following causal model

$$U_i = T_{A_i} + e^{\psi 0} T_{B_i} \tag{4}$$

$e^{-\psi 0}$ represents the acceleration factor (AF) associated with the intervention – the amount by which an individual's expected survival time is increased by treatment. By defining a binary process $X_i(t)$ which equals 1 when a patient is on the intervention treatment and equals zero when the patient is on control treatment, the causal model can be rewritten as

$$U_i = \int_0^{T_i} \exp[\psi X_i(t)]\mathrm{d}t \tag{5}$$

which is identical to the SNM introduced in equation (2). The value of ψ is estimated using a grid search. For each value of ψ, equation (4) is used to estimate U_i, and the true value of ψ is that for which $U(\psi)$ is independent of randomised groups. A log-rank or Wilcoxon test can be used for the RPSFTM g-test in a non-parametric setting, testing the hypothesis that the baseline survival curves are identical in the two treatment groups, or a Wald test could be used for parametric models.[26] The point estimate of ψ is that for which the test (z) statistic equals zero.

Censoring is problematic for the RPSFTM due to an association between treatment received, counterfactual censoring time and prognosis.[24,27,28] It is suggested that possible bias be avoided by breaking the dependence between censoring time and treatment received by re-censoring $U_i(\psi)$ at the minimum of the administrative censoring time C_i and $C_i \exp\psi$. $U_i(\psi)$ is then replaced by the censoring time of the counterfactual event time $D_i^*(\psi)$ if $D_i^*(\psi) < U_i(\psi)$.

The one-parameter version of the RPSFTM assumes that the relative treatment effect is equal for all patients no matter when the treatment is received (the 'common treatment effect' assumption), and that the randomisation of the trial means that there are no differences between the treatment groups, apart from treatment allocated.[11] The 'common treatment effect' assumption is particularly important in the context of treatment switching in oncology RCTs, because if switching is only permitted after disease progression, it may be expected that the capacity for benefit amongst switchers is lower than in patients initially randomised to the experimental group. The IPCW and SNM methods previously described are not reliant upon this assumption. In contrast, the RPSFTM is not reliant upon the 'no unmeasured confounders' assumption.

2.2.2.2. IPE algorithm. Branson and Whitehead[29] extended the RPSFTM method by developing a novel iterative parameter estimation (IPE) procedure. A parametric failure time model is fitted to the original, unadjusted ITT data to obtain an initial estimate of ψ. The observed failure times of switching patients are then re-estimated using the counterfactual survival time model presented in equation (5), and the treatment groups are then compared again using a parametric failure time model. This will give an updated estimate of ψ, and the process of re-estimating the observed survival times of switching patients is repeated. This iterative process continues until the new estimate for $\exp(\psi)$ is very close to the previous estimate (the authors suggest within 10^{-5} of the previous estimate but offer no particular rationale for this), at which point the process is said to have converged.[29] Bootstrapping is recommended to obtain standard errors and CIs for the treatment effect.[29]

The IPE procedure makes similar assumptions to the RPSFTM method – for example, the randomisation assumption is made, as is the 'common treatment effect' assumption, whereas the 'no unmeasured confounders' assumption is not required. An additional assumption is that survival times follow a specific parametric failure time distribution, and thus the method uses a parametric estimation procedure rather than g-estimation which may allow the method to converge more quickly.

2.3 Two-stage estimation – A simplified method

We believe that a simplified adjustment method not previously used in an HTA context is worthy of consideration, driven by a consideration of the type of treatment switching typically seen in oncology RCTs. Usually switching is only permitted after disease progression but is likely to happen soon after this time-point. In this case, if we assume that all patients are at a similar stage of disease at the point of disease progression, we can estimate the effect of the new treatment on extending survival from the point of disease progression to death, specifically for control group patients who switch. Disease progression can be used as a secondary baseline – that is, a time-point from which analysis time can be reset to zero – and data for patients in the control group can be treated as an observational dataset. By fitting an AFT model (such as a Weibull model) to this data (excluding patients in the experimental group) including covariates measured at the secondary baseline and including a time-varying covariate indicating treatment switch, we can estimate the treatment effect received by patients who switched compared to control group patients who did not switch. This would be expected to produce a reasonable estimate of the treatment effect associated with switching, provided the model fits the data, there are 'no unmeasured confounders' at the point of the secondary baseline and provided switching occurs soon after the secondary baseline. Counterfactual survival times could then be obtained using

$$U_i = T_{A_i} + \frac{T_{B_i}}{\mu_B} \tag{6}$$

where T_{A_i} represents the time spent on control treatment, T_{B_i} represents the time spent on the new intervention and μ_B is the treatment effect (AF) in switching patients.

Essentially, this represents a simplification of the approach used by Robins and Greenland[19] and Yamaguchi and Ohashi[20] to adjust for treatment switches, in which an SNM was utilised to estimate the treatment effect in the control group, rather than a less complex AFT model as suggested here. The simplified approach suggested here makes no attempt to adjust for time-dependent confounding beyond disease progression, and thus requires the strong assumption that there is no time-dependent confounding between the time of disease progression and the time of treatment switch. It also makes additional parametric assumptions according to which parametric AFT model is used to estimate the treatment effect in switchers. The benefits of making these assumptions are that the method requires fewer data (the 'no unmeasured confounders' assumption is only required at the secondary baseline timepoint) and does not require modelling of the treatment switching process. In addition, it may be argued that if switching occurs soon after the secondary baseline bias caused by time-dependent confounding may be minimal, although as long as there is some difference between secondary baseline and switch the method will be prone to bias. Unlike the RPSFTM and IPE methods, the simple two-stage method suggested here does not require the 'common treatment effect' assumption because the initial step of the approach involves estimating a treatment effect specifically for switchers in the control group. However, this method is only possible to apply if a suitable secondary baseline exists, and if the required covariate data are available at this time-point.

2.4 Methods summary

It is clear that the 'simple' switching adjustment methods described in Section 2.1 are highly prone to bias in certain situations. The complex methods described in Section 2.2 may improve upon these, but they too are associated with important limitations. While IPCW, SNM, RPSFTM and IPE methods will be unbiased in scenarios in which their methodological assumptions hold, their assumptions are limiting and may not be plausible in particular RCT contexts. For instance, observational-based methods are reliant upon the 'no unmeasured confounders' assumption and require sufficient data availability to allow the treatment and survival processes to be accurately modelled. Randomisation-based methods are less data-reliant, but depend upon the 'common treatment effect' assumption, which may not be clinically plausible. The two-stage estimation method is potentially useful, since it is designed specifically with the treatment switching mechanism often observed in oncology RCTs in mind. In order to inform the practical use of these methods, we will compare the bias associated with them in realistic scenarios – particularly in situations in which key methodological assumptions do not hold.

3 Simulation study design

We simulated independent datasets in which the true survival differences between treatment options were known, and in which treatment switching was allowed. We then applied each of the switching adjustment methods and compared their bias, MSE and coverage. We designed our study such that the data simulated reflected data typically observed in clinical trials in the advanced/metastatic cancer disease area, based upon input from pharmaceutical companies and our own knowledge. The simulation study was conducted using STATA software, version 11.0.[30]

3.1 Underlying survival times

We used a joint survival and longitudinal model to simultaneously generate a continuous time-dependent covariate (referred to as 'biomarker') and survival times, using the *survsim* STATA command.[31–33] We incorporated a time-dependent covariate that influenced both survival and the probability of treatment switching and was influenced by treatment received. Within the data-generating joint model, the longitudinal model for the biomarker value for the ith patient at time t was

$$biomarker_i(t) = \beta_{0_i} + \beta_1 \log(t) + \beta_2 \log(t) trt_i + \beta_4 badprog_i \tag{7}$$

where,

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

β_{0_i} is the random intercept, β_1 the slope against time for a patient in the control group, $\beta_1 + \beta_2$ the slope against time for a patient in the experimental treatment group. β_4 is the change in the intercept for a patient with a poor prognosis (referred to as 'badprog') compared to a patient with a good prognosis, trt_i is a binary covariate that equals 1 when the patient is in the experimental group and 0 otherwise, and $badprog_i$ is a binary covariate that equals 1 when a patient has poor prognosis at baseline, and 0 otherwise.

Based on methods described in detail by Bender et al.[34], the biomarker level was incorporated into the survival simulating process based upon the Weibull model hazard function

$$h(t) = \lambda \gamma t^{\gamma - 1} \exp(X\beta) \tag{8}$$

where

$$X\beta_i = \delta_1(trt_i) + (\eta \log(t)) trt_i + \delta_2 badprog_i + \alpha(biomarker_i(t)) \tag{9}$$

δ_1 is the baseline log hazard ratio intercept, η the rate at which the treatment effect changes with time, δ_2 is the impact of poor prognosis and α is the coefficient of the biomarker level. Given this, the survivor function was used to simulate survival times for the control group and the experimental group. The survivor function is

$$S(t) = \exp\left(\frac{-\lambda\gamma}{\gamma + \alpha(\beta_1 + \beta_2(trt_i)) + \eta * trt_i} \exp\big((\delta_1(trt_i)) + \delta_2 badprog_i + \alpha(\beta_{0_i} + \beta_4 badprog_i)\big) \times \left(t^{\gamma+\alpha(\beta_1+\beta_2(trt_i))+\eta(trt_i)}\right)\right) \tag{10}$$

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

$$\beta_0 = 20, \quad \sigma_0^2 = 1, \quad \beta_1 = 15, \quad \beta_2 = -4, \quad \beta_4 = 5, \quad \delta_1 = -0.7, \quad \delta_2 = 0.5,$$
$$\alpha = 0.02, \quad \lambda = 0.0005, \quad \gamma = 0.9, \quad \eta = 15$$

One example of the Kaplan–Meier curves produced by the simulation model (in the absence of treatment switching) using these parameter values are presented in Figure 1. Note that trtrand = 0 represents the control group, and trtrand = 1 represents the experimental group.

These Kaplan–Meier curves provided a close match to Kaplan–Meier curves for OS presented in two recent NICE technology appraisals of treatments for metastatic cancer in which treatment switching was an issue.[35,36]

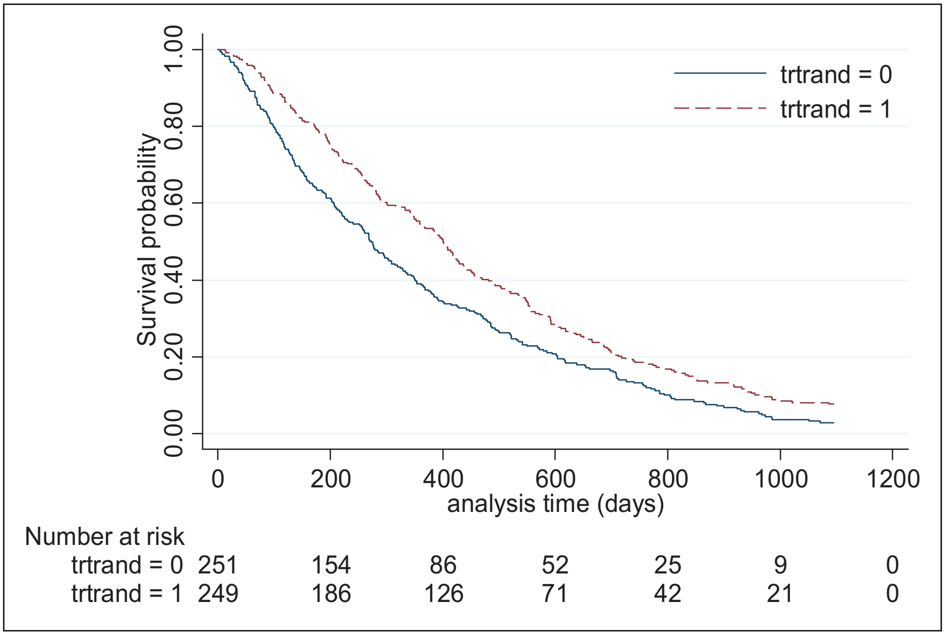

Figure 1. Overall survival Kaplan–Meier from one simulated dataset. Scenario 1: no switching.

3.2 Treatment effect in the experimental group

Equations (9) and (10) show how the treatment effect received by patients in the experimental group was incorporated into the survivor function, and that this was related to biomarker level and time. Collett[37] showed the AF φ relates the survivor function in the experimental S_E and control S_C group by

$$S_E = S_C\left(\frac{t}{\varphi}\right) \tag{11}$$

And thus, given the survivor function presented in equation (10), our AF over time can be estimated

$$\varphi = \exp\left[\frac{-\log\left(\frac{(\gamma+\alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2+n}}{(\gamma+\alpha\beta_1+\alpha\beta_2+n}\right)}{(\gamma+\alpha\beta_1)}\right] \tag{12}$$

Using this equation, we were able to ensure that we selected values for model parameters to produce a reasonable treatment effect over time. The overall treatment effect was assumed to reduce over time as disease steadily progresses. To achieve this, suitable values for α and η were chosen. It was necessary to include parameter η because otherwise a treatment that reduced the progression of the biomarker level would have a treatment effect that increased over time.

3.3 Treatment effect in switchers

The treatment effect applied to patients who switched from the control group to the experimental treatment was not calculated using the time-dependent AF shown in equation (12), since this equation estimates the treatment effect at a certain time-point assuming the patient had been receiving the treatment since baseline. Instead, the baseline treatment effect was applied to switchers but was multiplied by a factor (ω) such that the effect received was lower than the average effect received by experimental group patients. The magnitude of ω was varied across scenarios to represent reductions in the average treatment effect of 0%, 15% and 25%. This allowed us to test scenarios in which the 'common treatment effect' assumption did not hold.

3.4 The switching mechanism

In the base case scenario, the probability of treatment switching was allowed to depend upon the biomarker value at the time of disease progression and the time of progression itself. Disease progression times were simulated as a proportion of OS times dictated by a beta distribution with alpha and beta parameters (5,5). This implies the assumption that time-to-progression was approximately half of the OS time. The primary reason for generating disease progression times was to allow us to simulate switching that could only occur after a certain survival-related time-point. Switching was only allowed from the control group on to the experimental treatment and was not allowed prior to disease progression, to reflect the treatment switching typically seen in metastatic cancer trials. In addition, switching was only allowed to occur at one of the three visits immediately following disease progression (including the visit at which progression was first observed), and the probability of switching declined in each of these visits. Visits were assumed to occur every 21 days (also in line with metastatic cancer trials), and hence the earliest that switching could occur was 21 days after randomisation, and the latest that switching could occur was 42 days after the first visit at which disease progression was observed.

The probability of switching was calculated for each control group patient using a logistic function. In the base case, the probability of switching increased if the biomarker value was low at the time of disease progression, and if time-to-progression was high. The relationship between the probability of switching and prognostic measures was altered across scenarios, as specified in Section 3.5. Further details on the probability of switching in the different scenarios are presented in Online Appendix A (Available at http://smm.sagepub.com/).

3.5 Scenarios investigated

The simulated data generating mechanism had several variables for which values had to be assumed. These are listed in Online Appendix B, Table B1. The variables altered within the simulations related to:

- Treatment effect decrement received by switchers: 0% (zero time-dependency); 15%; 25%
- Switch proportion: moderate (approximately 65–70% of control group); high (approximately 85–95% of control group)
- Treatment effect: moderate (average HR approximately 0.75); high (average HR approximately 0.50)
- Disease severity: moderate (mean survival in control group approximately 340–375 days); high (mean survival in control group approximately 200–220 days)

This resulted in 24 scenarios. In addition, we tested the impact of alternative switching mechanisms. All 24 of the base scenarios were tested again in simulations in which the treatment switching mechanism was based only upon the simulated biomarker level at the time of disease progression. Scenarios 25–48 simulated a situation where patients with poor prognosis (based upon biomarker level at progression) were more likely to switch treatments, and Scenarios 49–72 simulated a situation where patients with good prognosis were more likely to switch.

In total, 72 scenarios were run. In each scenario, 500 patients were simulated to reflect study sizes commonly found in oncology clinical trials, and 1000 simulations were run for each scenario.

3.6 Performance measures

The time-dependency of the simulated treatment effect meant that it was not possible to produce a single ‘true’ HR or AF to which the results of the switching adjustment methods could be compared, apart from in the scenarios where the treatment effect was not time-dependent. Instead, the true mean survival time in the control group was calculated from the survivor function given in equation (10), and restricted mean survival at 1095 days (the administrative censoring time in the simulated dataset) was used as the ‘truth’ upon which performance measures were based. Hence, methods were assessed based upon how well they could estimate counterfactual survival times in the control group, in the presence of treatment switching from the control group to the experimental treatment. This is particularly relevant given the economic evaluation context of our work, since NICE require estimates of mean survival in order that appropriate resource allocation decisions can be made.[1] The restricted mean was convenient to use as it avoided potential additional biases associated with having to extrapolate survival times and placed the focus of the simulation study on switching adjustment methods, rather than extrapolation methods.

For the ITT, PP exclude switchers and PP censor switchers analyses mean survival time restricted to the longest follow-up time (that is, 1095 days) was computed by calculating the area under the Kaplan–Meier survivor function using Statas *stci* command for the specified group. For the IPCW analysis-adjusted control group, survival was derived by first applying the inverse of the estimated treatment effect (the IPCW hazard ratio) to the experimental group hazard function (modelled using a Weibull model) to obtain the adjusted control group hazard function, from which the control group survivor function was derived. Mean survival at 1095 days was then calculated. CIs were derived in the same way, except the inverse of the 95% CIs of the estimated treatment effect were applied to the experimental group hazard function. We refer to this approach for estimating mean survival as the 'survivor function' approach. While the IPCW method can provide a weighted Kaplan–Meier curve, generating this for each simulation would be extremely computationally intensive. Using the IPCW adjusted HR within a 'survivor function' approach should represent a close approximation of the IPCW-weighted Kaplan–Meier.

A 'survivor function' approach was also used to estimate mean survival at 1095 days for the RPSFTM and IPE methods, except that the estimated AFs generated by the methods were applied to the time-points associated with the experimental group survivor function, since AFs work directly on the time-scale. 95% CIs were estimated by using the 95% CIs of the estimated treatment effects within the 'survivor function' process.

The SNM and two-stage estimation methods involved estimating treatment effects specific to switching patients. For these, a counterfactual dataset was derived using equation (6), and Stata's *stci* command was used to estimate mean survival at 1095 days.

Bias (δ) was measured by the difference between the true restricted mean (β) and the estimated restricted mean ($\hat{\beta}$). Percentage bias was calculated as $\frac{\delta}{\beta} \times 100$.[38] The MSE was also calculated, to provide information on the variability of the estimates obtained using the different adjustment methods in combination with their bias. Variability is useful to consider, because if methods produce similar levels of bias, but one produces much more variable estimates, the method that produces less variability may be preferred. The standard error presented and included within the MSE calculation was that associated with the mean restricted mean estimated by each adjustment method, reflecting the variation in estimated mean survival across the 1000 replications for each scenario.

The coverage of each method was also calculated, defined as the proportion of simulations where the 95% CIs of the restricted mean estimated by each method contained the true restricted mean. Finally, we recorded the proportion of estimations in which the adjustment methods converged to give an estimate of the adjusted treatment effect.

3.7 Methods included

All methods described in Section 2 were included in the simulation study. For the RPSFTM method, we used a log-rank test within the g-estimation procedure, as this has proven most reliable in previous studies.[5] For the IPE algorithm, we tested alternative methods using exponential and Weibull models within the estimation procedure, in order to assess whether the performance of the method is sensitive to this. The STATA command *strbee* was used for the application of RPSFTM and IPE methods.[39] For both the IPCW and SNM methods, we included two versions – one in which all covariates were included in the relevant models, and one in which key covariates were excluded – in order to test their sensitivity to the 'no unmeasured confounders' assumption. The STATA command *stgest* was used for the application of the SNM method.[25] The two-stage estimation method was applied using a Weibull model.

4 Results

The performance of each method differed importantly depending upon the scenario investigated. We present detailed results from eight scenarios that clearly illustrate the key findings. In Section 4.1, we report key results in scenarios that involved moderate (approximately 65–70%) switching proportions, and in Section 4.2, we report key results in scenarios that involved high (approximately 85–95%) switching proportions. In Section 4.3, we summarise the extent to which the eight scenarios focussed upon reflect the results of the other 64 scenarios completed. A summary table (Table C1) describing the characteristics of each scenario is presented in Online Appendix C, and Figure D1 to D6 depict percentage bias across all scenarios, presented in Online Appendix D. In this section, method names are abbreviated as follows: ITT, Exclude switchers (PPexc), Censor at switch (PPcens), IPCW, IPCW excluding the time-dependent prognostic covariate (IPCWn), structural nested failure time model (SNM), SNM excluding the baseline prognostic covariate (SNMn), RPSFTM, Iterative Parameter Estimation (applied using a Weibull model) (IPE), iterative parameter estimation with an exponential model (IPEexp), two-stage Weibull estimation (Weib2m).

4.1 Scenarios with moderate switching proportions

Tables 1 and 2 present detailed results from Scenarios 25–28. These are illustrative of the results of scenarios in which the switching proportion simulated was approximately 65–70%. In Scenario 25, mean true survival in the control group (in the absence of treatment switching) was 372 days, and in the experimental group was 462 days, which was associated with an average HR of 0.75. This average HR is included only for illustrative purposes, to give an idea of the size of the treatment effect – it was estimated by generating the scenario data for a large number of patients (1,000,000) without applying switching, and by applying a Cox model to this. However, given that the proportional hazards assumption does not apply in our simulated datasets, this estimate of the treatment effect is prone to error. The mean switching proportion was 70% of control group patients, which equated to 72% of those who became at risk of switching (that is, those who lived for 21 days or longer). Seven percent of patients were administratively censored at 1095 days, and the treatment effect applied to switching patients was on average 15% lower than the average treatment effect received by patients in the experimental group – hence, the common treatment effect assumption did not hold. The probability of switching was related to the time-dependent biomarker covariate value at the time of disease progression, and patients with a lower value were more likely to switch. Because the biomarker score increased over time, patients who had a low score at the time of disease progression tended to be of relatively poor prognosis. The 95% CIs presented represent the mean of the lower and upper 95% CIs calculated for each adjustment method across the 1000 simulations.

As expected, the ITT analysis overestimated the true, unconfounded control group mean survival time in this scenario. The absolute bias was 9.05 days, equivalent to a percentage bias of 2.43%. This bias was low relative to several other scenarios, due to the moderate true average HR of 0.75 and the 15% decrement in the effect received by switchers. Simple adjustment methods (PPexc, PPcens) all produced substantially higher percentage bias than the ITT analysis, ranging from 10.46% to 65.14%. The IPCW and the IPCWn both underestimated mean survival in the control group (hence overestimated the treatment effect) and produced higher bias than the ITT analysis, with the version that included the biomarker covariate resulting in marginally less bias than the version that excluded this covariate (percentage bias of −7.36% compared to −7.76%). The SNM and SNMn overestimated mean survival in the control group and produced higher bias than the ITT analysis, with the version that included the baseline prognostic covariate resulting in marginally

Table 1. Estimates of the true mean survival (days) in the control group – Scenarios 25 and 26, 65–75% switch proportion, reduced effect in switchers.

Scenario details	Method	Mean estimate	SE of mean[a]	95% Confidence interval		Bias	Percentage bias	MSE	Coverage (%)	Convergence (%)
				Lower	Upper					
Scenario number: 25	ITT	381.11	18.32	343.88	418.33	9.05	2.43	417.38	93.70	100.00
True mean survival:	PPexc	410.98	37.84	334.43	487.53	38.92	10.46	2946.93	85.80	100.00
Control: 372.06	PPcens	614.43	35.10	545.17	683.70	242.38	65.14	59977.59	100.00	100.00
Experimental: 462.27	IPCW	344.68	36.73	262.54	443.19	−27.38	−7.36	2098.90	97.00	100.00
	IPCWn	343.20	33.47	263.45	438.25	−28.86	−7.76	1952.72	97.00	100.00
Mean switch %: 69.65%	Weib2m	371.97	22.30	351.24	401.28	−0.09	−0.02	497.27	73.10	100.00
True average HR: 0.75	SNM	388.00	34.09	374.78	402.78	15.95	4.29	1416.18	31.20	100.00
Mean censored: 7.02%	SNMn	386.51	34.46	374.01	401.34	14.45	3.88	1396.05	31.30	100.00
	RPSFTM	351.62	26.27	284.94	428.98	−20.44	−5.49	1108.18	96.60	100.00
Treatment effect:	IPE	352.02	26.24	304.76	403.82	−20.04	−5.39	1090.01	86.70	100.00
15% decrement	IPEexp	351.32	26.12	295.52	413.60	−20.74	−5.57	1112.30	92.30	100.00
Scenario number: 26	ITT	419.11	20.45	379.47	458.75	47.05	12.65	2631.94	35.10	100.00
True mean survival:	PPexc	403.31	41.00	323.27	483.36	31.26	8.40	2658.12	89.90	100.00
Control: 372.06	PPcens	629.26	39.58	557.30	701.23	257.20	69.13	67720.38	100.00	100.00
Experimental: 579.21	IPCW	338.07	35.37	255.73	437.48	−33.99	−9.13	2405.83	95.60	100.00
	IPCWn	335.91	32.82	256.15	431.73	−36.15	−9.72	2383.60	94.70	100.00
Mean switch %: 71.62%	Weib2m	372.10	22.83	351.46	397.84	0.04	0.01	521.43	68.60	100.00
True average HR: 0.52	SNM	399.64	34.31	386.63	415.47	27.58	7.41	1938.33	23.10	100.00
Mean censored: 13.10%	SNMn	397.95	34.05	384.99	414.29	25.89	6.96	1829.76	24.00	100.00
	RPSFTM	341.72	28.70	271.69	424.16	−30.34	−8.15	1743.74	93.80	100.00
Treatment effect:	IPE	342.01	28.69	293.49	395.72	−30.05	−8.08	1725.70	77.40	100.00
15% decrement	IPEexp	341.80	28.59	284.16	406.95	−30.26	−8.13	1732.54	86.30	100.00

MSE: mean squared error; HR: hazard ratio; ITT: intention-to-treat; PPexc: exclude switchers; PPcens: censor at switch; IPCW: inverse probability of censoring weights; IPCWn: IPCW excluding the time-dependent prognostic covariate; Weib2m: two-stage Weibull estimation; SNM: structural nested model; SNMn: SNM excluding the baseline prognostic covariate; RPSFTM: rank preserving structural failure time model; IPE: iterative parameter estimation with a Weibull model; IPEexp: iterative parameter estimation with an exponential model.

[a] The standard error presented and included within the MSE calculation is that associated with the mean restricted mean estimated by each adjustment method, reflecting the variation in estimated mean survival across the 1000 replications for each scenario.

Table 2. Estimates of the true mean survival (days) in the control group – Scenarios 27 and 28, 65–75% switch proportion, common treatment effect – Results.

				95% Confidence interval						
Scenario details	Method	Mean estimate	SE of mean[a]	Lower	Upper	Bias	Percentage bias	MSE	Coverage (%)	Convergence (%)
Scenario number: 27	ITT	429.23	24.51	383.51	474.96	84.77	24.61	7785.97	4.30	100.00
True mean survival:	PPexc	348.30	41.41	267.73	428.88	3.84	1.11	1729.52	94.10	100.00
Control: 344.47	PPcens	628.90	42.38	554.01	703.79	284.43	82.57	82698.36	100.00	100.00
Experimental: 568.12	IPCW	315.32	38.62	224.90	422.86	−29.14	−8.46	2340.69	97.50	100.00
	IPCWn	312.23	36.02	224.80	415.84	−32.24	−9.36	2336.33	97.10	100.00
Mean switch %: 65.86%	Weib2m	351.76	24.45	335.49	372.26	7.29	2.12	651.16	51.50	100.00
True average HR: 0.51	SNM	383.21	31.31	372.44	395.86	38.74	11.25	2481.10	10.30	100.00
Mean censored: 19.08%	SNMn	381.87	31.56	371.03	394.17	37.40	10.86	2394.85	11.20	100.00
	RPSFTM	344.62	36.81	258.27	446.44	0.16	0.05	1355.09	98.70	100.00
Treatment effect:	IPE	344.72	36.80	285.71	408.80	0.25	0.07	1354.38	90.90	100.00
0% decrement	IPEexp	344.76	36.85	292.07	401.49	0.29	0.09	1357.67	87.40	100.00
Scenario number: 28	ITT	374.25	22.31	331.89	416.61	29.78	8.65	1384.65	74.00	100.00
True mean survival:	PPexc	353.29	40.33	275.98	430.61	8.83	2.56	1704.20	94.00	100.00
Control: 344.47	PPcens	609.83	38.45	536.99	682.68	265.37	77.04	71898.60	100.00	100.00
Experimental: 437.88	IPCW	324.01	35.19	234.54	428.85	−20.46	−5.94	1657.21	98.70	100.00
	IPCWn	320.96	33.30	234.53	421.90	−23.51	−6.82	1661.48	98.50	100.00
Mean switch %: 64.03%	Weib2m	344.47	23.38	327.49	367.04	0.00	0.00	546.48	59.60	100.00
True average HR: 0.75	SNM	369.84	34.29	358.31	382.45	25.37	7.37	1819.56	18.50	99.90
Mean censored: 11.15%	SNMn	368.46	34.43	357.04	381.09	23.99	6.96	1760.91	19.30	99.90
	RPSFTM	344.16	32.90	265.40	436.61	−0.31	−0.09	1082.66	98.60	100.00
Treatment effect:	IPE	344.20	32.80	288.90	403.90	−0.26	−0.08	1076.20	91.90	100.00
0% decrement	IPEexp	344.17	32.80	295.16	396.63	−0.29	−0.08	1076.14	88.10	100.00

MSE: mean squared error; HR: hazard ratio; ITT: intention-to-treat; PPexc: exclude switchers; PPcens: censor at switch; IPCW: inverse probability of censoring weights; IPCWn: IPCW excluding the time-dependent prognostic covariate; Weib2m: two-stage Weibull estimation; SNM: structural nested model; SNMn: SNM excluding the baseline prognostic covariate; RPSFTM: rank preserving structural failure time model; IPE: iterative parameter estimation with a Weibull model; IPEexp: iterative parameter estimation with an exponential model.

[a] The standard error presented and included within the MSE calculation is that associated with the mean restricted mean estimated by each adjustment method, reflecting the variation in estimated mean survival across the 1000 replications for each scenario.

higher bias than the version that excluded the covariate (percentage bias 4.29% compared to 3.88%). In this scenario, the RPSFTM, IPE and IPEexp all produced very similar levels of bias (percentage biases of −5.49%, −5.39% and −5.57%, respectively), underestimating mean survival in the control group and producing higher bias than the ITT analysis. The IPE using a Weibull model resulted in marginally less bias than the RPSFTM, whereas the IPEexp resulted in marginally higher bias. In this scenario, only the Weib2m method produced less bias than the ITT analysis, resulting in very low percentage bias of −0.02%.

The only substantive difference between Scenario 25 and Scenario 26 was that the treatment effect was larger in Scenario 26, with mean survival time increased to 579 days, associated with an average HR of 0.52. Owing to this, the percentage bias associated with each of the adjustment methods increased, typically by approximately 2–3% points. However, the pattern of comparative percentage bias between the methods remained similar – the Weib2m produced least bias (percentage bias 0.01%) and did not produce more bias than in Scenario 25; the SNM produced next least bias (with SNMn marginally outperforming SNM, percentage bias 6.96% and 7.41%, respectively), followed by the RPSFTM/IPE methods (percentage bias −8.08% to −8.15%). In this scenario, the PPexc approach produced slightly less bias than the IPCW (percentage bias 8.40% compared to −9.13%), but the PPcens method led to much higher levels of bias (percentage bias 69.13%). Due to the higher treatment effect, the ITT analysis gave higher percentage bias (12.65%) than in Scenario 25 and produced higher bias than all adjustment methods except PPcens.

Table 2 presents detailed results of Scenario 27 and Scenario 28. Scenario 27 is approximately equivalent to Scenario 26, and Scenario 28 is approximately equivalent to Scenario 25, except the ‘common treatment effect’ assumption holds – i.e. the treatment effect is not time-dependent. This has an important impact upon the results of the adjustment methods. While the Weib2m method continued to produce very low levels of bias (percentage bias 0.00% to 2.12%), the RPSFTM/IPE methods also performed very well, with percentage bias between −0.09% and 0.09%. The IPCW and IPCWn methods produced similar levels of bias to those found in Scenarios 25 and 26 (percentage bias −5.94% to −9.36%), but the SNM and SNMn methods produced higher levels of bias (percentage bias 6.96% to 11.25%). The PPcens method again produced very high levels of bias (percentage bias 77.04% to 82.57%), while the PPexc method did comparatively well (percentage bias 1.11% to 2.56%).

Tables 1 and 2 show a substantial difference in the levels of coverage and MSE achieved by each of the adjustment methods. However, comparing levels of coverage and MSE is not particularly meaningful when levels of bias are high. Despite this, it is important to note the low levels of coverage achieved by the Weib2m, the SNM, the IPE and the IPEexp – particularly in cases where the Weib2m, IPE and IPEexp produced low levels of bias. In contrast, the RPSFTM method produced higher levels of coverage (98.60–98.70%) in Scenarios 26 and 27, in which the ‘common treatment effect’ assumption held. This is likely to be due to the RPSFTM retaining the ITT analysis *p* value in the estimation of the adjusted treatment effect, and the width of the associated CIs which are used to estimate CIs for the mean restricted survival time. It is notable that the Weib2m produced a substantially lower MSE than the other methods in Scenarios 25–28, even when levels of bias were similar. Whilst this suggests that the Weib2m is associated with relatively low variability, the poor coverage associated with the method means that this should be interpreted with care. Table 1 demonstrates that in scenarios where similar levels of bias were produced by IPCW, SNM and RPSFTM/IPE methods, the MSE associated with the RPSFTM/IPE methods was comparatively low, suggesting that the randomisation-based methods exhibited lower variability than the observational-based methods.

The coverage associated with the Weib2m and SNM methods was poor because CIs for mean counterfactual survival times were estimated by using the 95% CIs for ψ_B in equation (6). This only takes into account the uncertainty in the treatment effect itself – it does not take into account the uncertainty in the underlying survival distribution. In reality, if a two-stage approach were to be taken, uncertainty around mean survival estimates would need to be taken into account using bootstrapping. For the IPE approaches, the CIs around the parameter estimates supplied by the final iteration of the IPE algorithm were used to generate restricted mean CIs. These represent an underestimate of the true CI.[29] Morden et al.[5] showed that if a bootstrapping approach to estimate CIs for the treatment effect is taken, coverage is satisfactory with the IPE method (in scenarios where low bias is achieved).

The adjustment methods converged to give an estimate in the vast majority of cases across Scenarios 25–28, with only the SNM method failing to converge in 0.1% of simulations in Scenario 28.

4.2 Scenarios with high switching proportions

Tables 3 and 4 present detailed results from Scenarios 37–40. These are illustrative of the results of scenarios in which the switching proportion simulated was approximately 85–95%. Scenarios 37–40 are similar to Scenarios 25–28, respectively, with the only substantive difference the switching proportion.

The increased switching proportion has an important impact on the bias associated with the adjustment methods across all scenarios – particularly for the observational-based adjustment methods. The bias associated with the IPCW and IPCWn increased by factors of approximately 4–7 compared to the scenarios with moderate switching proportions (percentage bias 31.69% to 46.61%) and similar was true for the SNM and SNMn (percentage bias 16.56% to 38.17%). The randomisation-based methods were less affected by the very high switching proportions – in Scenarios 37 and 38, where the ‘common treatment effect’ assumption did not hold, the bias associated with the RPSFTM, IPE and IPEexp increased, but did not double (percentage bias −9.33 % to −11.84%). These methods continued to produce very low bias in Scenarios 39 and 40, in which the ‘common treatment effect’ assumption held (percentage bias 0.74% to 1.05%). While the bias associated with the Weib2m method increased approximately fivefold in the scenarios with very high switching proportions, the level of bias remained relatively low (percentage bias 5.34% to 8.55%).

In Scenario 37, the ITT analysis resulted in lower bias than all the adjustment methods. This is due to the low level of bias associated with the ITT in this scenario (percentage bias 3.58%) – owing to the relatively small treatment effect (average HR of 0.75) combined with the 15% treatment effect decrement received by switchers – and because the adjustment methods do not perform well in this scenario. In the corresponding scenario with a moderate switching proportion (Scenario 25), only the Weib2m method produced lower bias than the ITT analysis, but the Weib2m produces a higher level of bias in Scenario 37 due to the higher switching proportion. In Scenario 38, the bias associated with the ITT analysis increased to 17.30%, and the Weib2m, RPSFTM, IPE and IPEexp produced lower bias, with the Weib2m leading to less than half the bias of any other method (percentage bias 5.34%). In the corresponding scenario with a moderate switching proportion (Scenario 26), the IPCW, IPCWn, SNM and SNMn also produced less bias than the ITT analysis, but this was no longer the case in scenarios with very high switching proportions. Similar was true for Scenarios 39 and 40 – only the RPSFTM, IPE, IPEexp and Weib2m produced lower bias than the ITT analysis.

Table 3. Estimates of the true mean survival (days) in the control group – Scenarios 37 and 38, 85–95% switch proportion, reduced effect in switchers – Results.

Scenario details	Method	Mean estimate	SE of mean[a]	95% Confidence interval		Bias	Percentage bias	MSE	Coverage (%)	Convergence (%)
				Lower	Upper					
Scenario number: 37	ITT	385.39	19.63	347.75	423.04	13.33	3.58	563.25	89.70	100.00
True mean survival:	PPexc	218.64	86.59	57.46	379.83	−153.42	−41.23	31033.71	51.90	100.00
Control: 372.06	PPcens	942.85	65.21	860.64	1025.05	570.79	153.41	330052.47	100.00	100.00
Experimental: 462.27	IPCW	545.48	176.64	338.30	763.39	173.42	46.61	61276.86	59.90	100.00
	IPCWn	504.18	182.80	303.51	729.98	132.12	35.51	50869.71	66.00	100.00
Mean switch %: 93.31%	Weib2m	394.19	67.63	308.45	600.50	22.13	5.95	5062.98	96.68	99.30
True average HR: 0.75	SNM	440.70	145.09	349.70	570.50	68.64	18.45	25761.46	55.35	98.10
Mean censored: 7.23%	SNMn	433.65	143.96	348.45	556.19	61.60	16.56	24517.46	53.21	98.10
	RPSFTM	337.00	33.77	264.06	431.21	−35.06	−9.42	2369.74	91.30	100.00
Treatment effect:	IPE	337.35	33.71	291.81	387.54	−34.71	−9.33	2341.27	64.90	100.00
15% decrement	IPEexp	336.61	33.34	282.77	397.14	−35.45	−9.53	2368.33	74.10	100.00
Scenario number: 38	ITT	436.41	21.08	395.53	477.30	64.35	17.30	4585.94	11.90	100.00
True mean survival:	PPexc	208.61	93.76	48.14	369.09	−163.45	−43.93	35504.79	47.40	100.00
Control: 372.06	PPcens	949.29	69.48	872.15	1026.43	577.23	155.15	338022.50	100.00	100.00
Experimental: 579.21	IPCW	544.68	182.22	331.85	770.77	172.62	46.40	63001.42	60.10	100.00
	IPCWn	505.55	188.35	298.65	737.24	133.49	35.88	53295.38	65.30	100.00
Mean switch %: 93.41%	Weib2m	391.91	71.54	310.44	610.02	19.85	5.34	5512.37	94.05	99.20
True average HR: 0.52	SNM	462.58	171.31	364.21	628.94	90.52	24.33	37539.95	57.86	99.90
Mean censored: 13.78%	SNMn	450.28	156.04	362.47	616.19	78.22	21.02	30468.38	57.86	99.90
	RPSFTM	328.03	34.39	253.58	423.67	−44.02	−11.83	3120.78	89.30	100.00
Treatment effect:	IPE	328.29	34.35	281.62	380.25	−43.77	−11.76	3096.07	56.30	100.00
15% decrement	IPEexp	328.00	34.25	272.29	391.46	−44.06	−11.84	3114.18	68.70	100.00

MSE: mean squared error; HR: hazard ratio; ITT: intention-to-treat; PPexc: exclude switchers; PPcens: censor at switch; IPCW: inverse probability of censoring weights; IPCWn: IPCW excluding the time-dependent prognostic covariate; Weib2m: two-stage Weibull estimation; SNM: structural nested model; SNMn: SNM excluding the baseline prognostic covariate; RPSFTM: rank preserving structural failure time model; IPE: iterative parameter estimation with a Weibull model; IPEexp: iterative parameter estimation with an exponential model.

[a] The standard error presented and included within the MSE calculation is that associated with the mean restricted mean estimated by each adjustment method, reflecting the variation in estimated mean survival across the 1000 replications for each scenario.

Table 4. Estimates of the true mean survival (days) in the control group – Scenarios 39 and 40, 85–95% switch proportion, common treatment effect – Results.

				95% Confidence interval						
Scenario details	Method	Mean estimate	SE of mean[a]	Lower	Upper	Bias	Percentage bias	MSE	Coverage (%)	Convergence (%)
Scenario number: 39	ITT	460.61	24.60	413.06	508.17	116.14	33.72	14094.54	0.20	100.00
True mean survival:	PPexc	179.50	65.69	57.25	301.74	−164.97	−47.89	31531.17	31.90	100.00
Control: 344.47	PPcens	935.87	37.73	870.37	1001.36	591.40	171.69	351176.50	100.00	100.00
Experimental: 568.12	IPCW	493.36	150.82	293.18	712.24	148.89	43.22	44915.68	62.80	100.00
	IPCWn	458.61	161.84	263.66	687.09	114.14	33.14	39221.53	68.30	100.00
Mean switch %: 87.07%	Weib2m	373.93	54.48	315.92	553.73	29.47	8.55	3836.63	83.43	99.60
True average HR: 0.52	SNM	475.94	185.95	368.91	697.85	131.47	38.17	51861.19	44.34	99.90
Mean censored: 20.65%	SNMn	461.14	176.29	365.49	681.39	116.67	33.87	44689.89	45.65	99.00
	RPSFTM	347.78	46.10	248.49	483.68	3.32	0.96	2136.03	99.30	100.00
Treatment effect:	IPE	347.99	46.08	288.77	412.16	3.53	1.02	2135.89	82.10	100.00
0% decrement	IPEexp	348.09	46.27	295.32	404.78	3.62	1.05	2153.98	76.70	100.00
Scenario number: 40	ITT	387.89	21.93	344.43	431.35	43.42	12.61	2366.16	51.40	100.00
True mean survival:	PPexc	181.57	62.91	61.46	301.69	−162.89	−47.29	30491.11	30.90	100.00
Control: 344.47	PPcens	925.86	39.46	857.05	994.67	581.39	168.78	339573.87	100.00	100.00
Experimental: 437.88	IPCW	494.56	142.96	300.18	703.91	150.10	43.57	42967.97	63.10	100.00
	IPCWn	453.64	153.78	264.12	672.54	109.18	31.69	35568.88	69.90	100.00
Mean switch %: 86.83%	Weib2m	363.36	68.34	295.97	553.00	18.89	5.48	5026.65	91.69	99.90
True average HR: 0.75	SNM	454.94	176.76	353.62	610.69	110.47	32.07	43449.37	55.12	99.60
Mean censored: 11.72%	SNMn	447.05	174.97	348.22	597.46	102.58	29.78	41137.69	56.73	100.00
	RPSFTM	347.05	41.81	253.75	476.85	2.58	0.75	1754.36	99.30	100.00
Treatment effect:	IPE	347.14	41.53	291.70	406.87	2.67	0.78	1731.75	84.40	100.00
0% decrement	IPEexp	347.02	41.27	298.00	399.38	2.55	0.74	1709.34	78.90	100.00

MSE: mean squared error; HR: hazard ratio; ITT: intention-to-treat; PPexc: exclude switchers; PPcens: censor at switch; IPCW: inverse probability of censoring weights; IPCWn: IPCW excluding the time-dependent prognostic covariate; Weib2m: two-stage Weibull estimation; SNM: structural nested model; SNMn: SNM excluding the baseline prognostic covariate; RPSFTM: rank preserving structural failure time model; IPE: iterative parameter estimation with a Weibull model; IPEexp: iterative parameter estimation with an exponential model.

[a] The standard error presented and included within the MSE calculation is that associated with the mean restricted mean estimated by each adjustment method, reflecting the variation in estimated mean survival across the 1000 replications for each scenario.

Tables 3 and 4 also show a substantial difference in the levels of coverage and MSE achieved by each of the adjustment methods. As was the case previously, this is not particularly meaningful when levels of bias are high. However, it is relevant to note that the Weib2m produced improved levels of coverage in Scenarios 37–40. This may appear counterintuitive, given that the method produces higher bias in these scenarios. However, the increased coverage is due to the higher uncertainty in the estimation of the treatment effect ψ_B, owing to the increased switching proportion. The RPSFTM method again led to very high levels of coverage in scenarios in which it produced low bias, while the IPE and IPEexp produced lower coverage because CIs around the parameter estimates supplied by the final iteration of the estimation algorithm were used. In contrast to results for scenarios with moderate switching proportions, the Weib2m did not produce a lower MSE than all other methods in scenarios with very high switching proportions. In Scenarios 37–40, the Weib2m produced a higher MSE than RPSFTM/IPE methods, despite producing lower bias in Scenarios 37 and 38. This demonstrates that the variability associated with the Weib2m method increases when the treatment switching proportion is very high, and with respect to variability the Weib2m method is proportionately more affected by high switching proportions than are RPSFTM/IPE methods.

The adjustment methods again converged to give an estimate in the vast majority of cases across Scenarios 37–40, with only a slight increase in convergence failure associated with the higher switching proportions. The SNM method failed to converge in up to 1.9% of simulations included in Scenarios 37–40. The Weib2m failed to produce an upper CI for ψ_B in 0.1–0.8% of simulations.

4.3 Other scenarios

We have presented detailed results for eight scenarios which we believe provide a clear illustration of the key findings of our study. In the other scenarios, the pattern of the results remained similar – PPcens and PPexc approaches generated very high levels of bias; RPSFTM, IPE and IPEexp methods produced low bias when the 'common treatment effect' assumption held; the Weib2m method produced generally low levels of bias across all scenarios although levels were slightly higher when the switching proportion increased; SNM, SNMn, IPCW and IPCWn methods generally produced high bias when the switching proportion was very high. However, there were some exceptions that violated this pattern – though we believe that when this occurred it was explainable. Figures D1 to D6 show the percentage bias for the key methods across all scenarios, presented in Online Appendix D.

Although the bias associated with the IPCW method was highly related to the switching proportion, in some scenarios (for example, Scenarios 13–23, and to a lesser extent Scenarios 61–72) relatively low bias (percentage bias −0.07% to −7.13% and −0.51% to −14.83%, respectively) resulted from the IPCW and IPCWn methods even when the switching proportion was very high. In these scenarios, patients with relatively good prognosis were more likely to switch treatments. However, the poorest prognosis patients who died before 21 simulated days were not able to switch and these made up 3–13% of control group patients, depending upon the severity of the simulated disease. In scenarios where patients with better prognosis are more likely to switch, the pseudo population formulated by the IPCW will not include those with better prognosis or those with the very worst prognosis – the remaining patients may be reasonably reflective of average prognosis patients. In these situations, the IPCW could produce reasonably low bias, even in the presence of very high switching proportions. In contrast, when poor prognosis patients are more likely to switch, and when the vast majority of patients do switch, the remaining basis for the IPCW pseudo population is likely to consist of those patients with the very longest survival times – a proportion of which are likely to be censored due to end of follow-up – leading to high levels of error in the IPCW analysis.

It was clear that the proportion of control group patients who switched had the most important impact on the bias associated with the majority of the adjustment methods. The exception to this was that for the RPSFTM, IPE and IPEexp methods, where the 'common treatment effect' assumption was the most important factor. When the 'common treatment effect' assumption held the RPSFTM, IPE and IPEexp approaches generally produced least bias compared to the range of alternative methods, with only the Weib2m producing less bias on occasions when the switching proportion was low. When the treatment effect received by switching patients was approximately 15% less than that received by patients in the experimental group, the RPSFTM, IPE, IPEexp and IPCW approaches produced similar levels of bias when the switching proportion was moderate. In these scenarios, the Weib2m produced least bias of the adjustment methods, but the ITT analysis produced less bias when the treatment effect was low (average HR approximately 0.75). In scenarios where the treatment effect received by switching patients was approximately 25% less than that received by patients in the experimental group (Scenarios 9–12, 21–24, 33–36, 45–48, 57–60, 69–72), the IPCW and SNM methods produced less bias than the RPSFTM, IPE and IPEexp, provided the switching proportion was moderate. Again, the Weib2m produced least bias of the adjustment methods in these scenarios, but the ITT analysis often produced less bias.

The IPCW method was not affected by the complexity of the switching mechanism – similar bias was observed when switching depended only upon biomarker level (Scenarios 25–72) compared to when switching depended on both biomarker level and time-to-progression (Scenarios 1–24). Although the IPCWn produced marginally lower bias than the IPCW in each of the scenarios presented in Tables 3 and 4, overall the IPCW produced lower bias than the IPCWn in 65.3% of scenarios. Conversely, the SNM only produced lower bias than the SNMn in 2.8% of scenarios. As expected, the bias associated with the IPCW, IPCWn, SNM, SNMn and Weib2m methods was not affected by the size of treatment effect decrement assigned to switchers, since these methods are not reliant upon the 'common treatment effect' assumption. As illustrated by the results presented in Section 4.2, an increased treatment effect led to an increase in the bias associated with the adjustment methods, but this was of less importance than the switching proportion and the 'common treatment effect' assumption. The disease severity modelled did not importantly affect the adjustment methods other than the SNM method, which failed to converge in a significant number of simulations when disease severity was high combined with a low treatment effect received by switchers. In the worst case (Scenario 71), the SNM converged in only 9.6% of simulations.

In the scenarios in which the treatment effect was time-dependent, the RPSFTM, IPE and IPEexp methods always led to negative bias – that is, they overadjusted for the treatment switching effect. This is likely to be due to the re-censoring involved in the treatment effect estimation procedure. Re-censoring involves basing the treatment effect estimation upon shorter term data, and where the experimental group treatment effect decreases over time this may lead to an overestimate of the true treatment effect. This appears to have been the case across all scenarios.

5 Discussion

Our simulation study demonstrates that randomisation-based methods for adjusting for treatment switching, such as the RPSFTM and IPE algorithm, produce low bias in a wide range of scenarios, provided the relative treatment effect received by switching patients is equal to that received by experimental group patients (that is, the ‘common treatment effect’ assumption holds). Importantly, the applicability of this assumption is difficult to test based upon trial data. Our results confirm the findings of previous research.[5] However, when the treatment effect is strongly time-dependent, and the ‘common treatment effect’ assumption does not hold, these methods produce high levels of bias and in some circumstances (for instance, when the treatment effect is low) may be less preferable than an ITT analysis.

In the presence of time-dependent treatment effects, treatment switching adjustment methods are limited and are all prone to important bias. Observational-based methods such as the IPCW and SNM – which do not require the ‘common treatment effect’ assumption – require large amounts of data and are particularly sensitive to bias when the switching proportion is very high. Our simulations suggest that the relatively small size of RCT datasets may cause these methods to work sub-optimally, even when the ‘no unmeasured confounders’ assumption holds. This is unsurprising given the reliance of these methods on observational modelling. In our simulations, in which the sample size was 500, a 90% switching proportion in control group patients who became at risk of switching (those who lived longer than 21 days) led to approximately 20 control group patients not switching in each simulation. The IPCW method uses these patients as a pseudo population upon which to base adjusted control group survival estimates, and such low numbers are clearly problematic. This confirms the findings of other researchers – Howe et al.[40] found that IPCW was prone to bias in small samples, if selection bias was very strong, and if there were unmeasured confounders.

We found that observational-based methods worked similarly well even when a prognostic covariate was excluded from their models. In retrospect, this was not very surprising. The baseline biomarker group covariate is directly related to the baseline prognosis group covariate (see equation (7)), and the biomarker level at the time of disease progression is essentially a time-dependent measure of prognosis which is likely to be highly correlated with the time to disease progression covariate. Hence, excluding biomarker-related covariates from the observational models may be expected to have a relatively minor impact providing baseline prognosis and time-to-progression covariates are included, or vice-versa. Importantly, across all scenarios the IPCW method outperformed the naïve censoring analysis – which the IPCW would reduce to if all important covariates were unmeasured.

As expected, naïve methods (such as simple censoring and exclusion approaches) produced high levels of bias consistently across all scenarios and thus should be avoided. When the ‘common treatment effect’ assumption is expected to hold, the RPSFTM and IPE algorithm represent the optimal switching adjustment methods. However, when a time-dependent treatment effect is suspected identifying an optimal method is much more difficult. If the treatment switching mechanism is similar to that simulated in our simulation study (that is, switching can only occur after disease progression and must happen very soon after disease progression) and data on key prognostic variables are collected upon disease progression, a simple two-stage Weibull method may represent the most appropriate adjustment method.

In reality, it is unlikely that any one switching adjustment method will be optimal across all situations. Instead, the problem of identifying an appropriate method should be tackled on a case-by-case basis, following a step-by-step process as suggested by Latimer et al.[6] The ‘no unmeasured confounders’ assumption of the IPCW and SNM methods is untestable, and assessment of the ‘common treatment effect’ assumption relied upon by the RPSFTM and IPE algorithm is likely to be prone to bias. However, analyses can be undertaken and information can be gathered to shed light on the plausibility of these assumptions. To consider the ‘no unmeasured confounders’ assumption, previous studies may be reviewed in order to determine whether any important covariates were identified that were excluded from the study affected by treatment switching. Clinical expert opinion could also be sought in order to determine whether any indicators of switching were not collected in the study. An important consideration is that patient preference for switching is not routinely collected in clinical trials, yet this may influence the switching decision. This may cast doubt on the use of observational methods for adjusting for treatment switching in RCTs. Thus, if treatment switching is anticipated steps should be taken to ensure the collection of all relevant covariate data, and the potentially appropriate adjustment methods should be pre-specified.

Any data-based assessment of the 'common treatment effect' assumption is prone to bias due to time-dependent confounding. If there is a disease progression-related time-point after which switching becomes possible, the treatment effect in switching patients could be estimated by considering this period as an observational dataset – using either an IPCW or SNM approach, or using the first step of the simpler two-stage Weibull method. This could then be compared to the estimated treatment effect in the experimental group, adjusted for switching. However, as our simulations show, these methods are prone to bias and so using them to assess the 'common treatment effect' assumption is problematic. They may, however, provide some information on whether the treatment effects experienced in the different groups were broadly similar. Alternatively, if patients at different stages of disease were randomised into the trial in question, the effects on these groups could be compared. However, this comparison will be confounded by switching which may impact upon the groups differently. It is therefore important that the clinical and biological plausibility of a common treatment effect is ascertained through eliciting clinical opinion and by considering the mechanism of action of the novel therapy.

Various authors have attempted to apply multi-parameter versions of the RPSFTM, in order to allow a relaxation of the 'common treatment effect' assumption.[19,28,41] However, relying solely on the randomisation assumption to allow two different treatment effects to be estimated for different groups has proven unsuccessful, with meaningful point estimates difficult to determine. This is because there is too little information available to allow us to estimate more than one treatment effect based on comparability of treatment groups at randomisation. Hence, this represents an outstanding problem with randomisation-based methods. Specifying a one-parameter randomisation-based model that covers the control group and the experimental group in a biologically plausible way is theoretically problematic. Treatment with experimental drugs is usually stopped at the point of disease progression in clinical trials suggesting that clinicians expect that the treatment will have no further beneficial impact once progression has occurred in a subject already taking the treatment. The fact that switching occurs suggests that there is an expectation that patients whose disease has progressed without the experimental treatment having been received may benefit from post-progression treatment – presumably because resistance to the beneficial effects of the drug have not been developed. However, these two groups of patients are at different stages of the disease pathway and assuming that the treatment has an identical effect on them appears biologically implausible. It is important to analyse trial data and to obtain expert opinion on the biological mechanism of action of the experimental treatment in an attempt to gather more information as to whether (and to what extent) the 'common treatment effect' assumption is likely to be untrue.

SNMs and the two-stage method as applied in this study are less problematic with regard to biological plausibility, because they are applied only to the control group, before their resulting treatment effects are used to produce a counterfactual dataset in which the adjusted control group survival times can be compared to the experimental group survival times. Hence, these approaches do not attempt to specify a model that covers both the treatment effect in switchers and the in experimental group simultaneously. Similar is true of the IPCW, which estimates a treatment effect specific to the experimental group. However, despite these theoretical advantages, our simulations show that the SNM and IPCW methods are prone to substantial bias when switching proportions are very high. Bias also increased with the simplified two-stage method when the switching proportion was very high, but to a lesser extent compared to the SNM and IPCW methods. The simplified two-stage method essentially represents a simplification of the SNM, and owing to its strong assumption that there is no time-dependent confounding between the secondary baseline (disease progression) and the time of treatment switch is theoretically inferior. However, this method generally produced low levels of bias across the scenarios we investigated, suggesting that the bias associated with its theoretical limitations was less than the bias produced by the theoretically superior methods, given the specific treatment switching mechanism and trial characteristics simulated in this study.

6 Limitations

We attempted to include all the most important and most relevant scenarios given results of the Morden et al. study,[5] realistic cancer trial characteristics and the characteristics of the methods that were being assessed. However, there remain potentially interesting and relevant scenarios that were not included. In particular, the proportion of patients who switched was an extremely important factor in the results. A range of relatively high switching proportion scenarios were considered, under the assumption that the adjustment methods would struggle with these most. However, given the high levels of bias associated with the assessed methods in the scenarios that incorporate a time-dependent treatment effect, it would be valuable to identify whether levels of bias fall with lower switching levels. This is particularly important for the IPCW and SNM approaches, which were particularly sensitive to the switching proportion. Switching proportions vary widely in reality; for instance in a trial assessing cetuximab for use in head and neck cancer 6% of patients in the control group received cetuximab after disease progression,[42] whereas a NICE appraisal of sunitinib for the treatment of gastrointestinal stromal tumours reported a switching proportion of 84% in a pivotal trial.[43]

In addition, we showed that the IPCW approach often produced lower bias than RPSFTM or IPE methods when there was a time-dependent treatment effect (thus, when the 'common treatment effect' assumption did not hold). However, only two levels of treatment effect decrements applied to switching patients were investigated (15% and 25%). It would be interesting to consider a larger range of treatment effect decrements to gain a better understanding of what level of treatment decrement is required for the IPCW method to outperform the RPSFTM/IPE methods.

Also, in all the scenarios tested, the assumed sample size was 500 patients. This broadly reflects the size of metastatic oncology trials and matches the assumption made by Morden et al.[5] However, clearly different trials will have different sizes in response to sample size calculations that forecast different effect sizes. Given the data requirements of methods such as IPCW and SNMs, this could be an important factor. In fact, if very large proportions of patients switched, but the trial was extremely large, there may still be a substantial number of patients upon which to base the 'pseudo' population and so bias might be avoided. Equally, in smaller trials, these methods may work less well. Hence, investigating the performance of the adjustment methods according to sample size would be valuable.

A technical limitation of our study surrounds the use of the *stgest* Stata command to implement the SNM method. The method often failed to converge and this appeared to be due to the 'round' option included in the command. This might be argued to be a limitation associated with the Stata command rather than the SNM method itself.

A general limitation of simulation studies is that the results are likely to always be linked in some way to the chosen data generating process. Attempts were made to limit this by testing different distributions for parametric switching adjustment methods. Given that a Weibull model was used to generate the underlying survival times, the data generating mechanism may have favoured Weibull-based approaches such as the IPE algorithm, and the two-stage Weibull method. However, the results showed that the IPE algorithm method performed similarly well in estimating the treatment effect when it was applied using an exponential model. Also, due to the inclusion of a time-dependent covariate in the data generating model, the resulting survival times no longer followed a true Weibull distribution. Despite this, it may be of value to conduct similar studies using different data generating models. In addition, whilst the two-stage estimation method was applied using a Weibull model in this study, it could be applied using any AFT distribution, including more flexible mixture models.[44] Further, it would be of value to run scenarios in which switching could occur at time-points more distant from the point of disease progression, in order to assess the sensitivity of the simple two-stage approach to this.

It may also be of value to re-run the simulations using different methods to estimate the treatment effect received by switching patients and patients in the experimental group. It is difficult to simulate survival data and varying treatment effects in a biologically plausible way. We simulated a treatment effect that decreased over time in the experimental group, and the resulting Kaplan–Meier survivor functions (such as those presented in Figure 1) appeared plausible. For switchers, the treatment effect was not linked to time; instead the baseline treatment effect was multiplied by a factor to ensure that these patients received a plausible effect. An alternative would be to link this to time and other covariates using equation (12). However, this would not be expected to alter the performance of the key complex adjustment methods. The randomisation-based methods do not attempt to model the treatment effect process and so the important factor that determines their bias is the extent to which the average treatment effect differs between switching and experimental group patients – not how this treatment effect difference is estimated. The IPCW approach censors switching patients and so the treatment effect received (and how it is estimated) by these patients does not influence the bias with which the method estimates the treatment effect received in the experimental group.

We used a 'survivor function' approach to derive estimates of mean survival, described in Section 3.6. For some of the adjustment methods, alternative approaches could have been taken – for example, an 'extrapolation' approach could be taken for the RPSFTM and IPE methods, whereby a parametric model is fitted to the counterfactual survival times generated. However, this approach would be more prone to bias created by extrapolating from re-censored datasets. Given that we were interested in studying the bias associated with the switching adjustment methods – rather than extrapolation methods – the 'survivor function' approach was less problematic. In fact, we included versions of the RPSFTM and IPE that took this 'extrapolation' approach in our study, and results were very similar to those obtained using the 'survivor function' approach but were sensitive to the parametric model used to extrapolate. However, given the key role that extrapolation often plays in HTA, due to a life-time horizon often being required, further work exploring the process of extrapolation in the presence of treatment switching is required.

7 Conclusions

We conclude that the characteristics of trials and novel therapies should be considered on a case-by-case basis when an analyst is attempting to identify appropriate methods for adjusting for treatment switching in clinical trials but that there are some circumstances in which we should expect certain methods to perform badly. In particular, when the proportion of control group patients that switch treatments is very high (above approximately 85% in a trial with 250 patients randomised to the control group), the IPCW method is prone to substantial error. RPSFTM and IPE approaches will produce substantial bias when the 'common treatment effect' assumption is false, with the resulting bias proportionate to the extent to which the assumption is false. Provided the treatment switching mechanism is amenable, the simple two-stage Weibull method represents an appropriate adjustment method that is less sensitive to the switching proportion and is unaffected by the 'common treatment effect' assumption.

Acknowledgements

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

International Society for Pharmacoeconomics and Outcomes Research, 15th Annual European Conference, Berlin, Germany, November 2012.

Health Technology Assessment International, 9th Annual Meeting, Bilbao, Spain, June 2012.

Survival Analysis for Junior Researchers conference. 'Methods for adjusting survival estimates to adjust for treatment crossover: a simulation study'. Leicester, UK, 2012.

Part of this work was carried out whilst PCL was on study leave from the University of Leicester. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Financial support for this study was provided in part by grants from the National Institute for Health Research and the Pharmaceutical Oncology Initiative. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. JP Morden works for the ICR-CTSU, which receives core funding from Cancer Research UK.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Nicholas R Latimer was funded by a National Institute for Health Research Doctoral Research Fellowship [DRF 2009-02-82]. This article/paper/report presents independent research funded by the National Institute for Health Research (NIHR).

KR Abrams is partially supported as a Senior Investigator by the National Institute for Health Research (NIHR) in the UK [NI-SI-0508-10061].

MJ Crowther is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship [DRF-2012-05-409].

This work was supported by the Pharmaceutical Oncology Initiative, a group of pharmaceutical companies who are part of the Association of the British Pharmaceutical Industry (ABPI); and the National Institute for Health Research (NIHR) [grant numbers DRF 2009-02-82, NI-SI-0508-10061, DRF-2012-05-409].

References

1. National Institute for Health and Clinical Excellence. *Guide to the methods of technology appraisal*. London: NICE, 2008, http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (accessed 5 March 2012).
2. Briggs A, Claxton K and Sculpher M. *Decision modelling for health economic evaluation*. New York: Oxford University Press Inc, 2006.
3. Gold MR, Siegel JE, Russell LB, et al. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, Inc, 1996.
4. Canadian Agency for Drugs and Technologies in Health. *Guidelines for the economic evaluation of health technologies*, 3rd ed. Canada: Canadian Agency for Drugs and Technologies in Health, 2006.
5. Morden JP, Lambert PC, Latimer NR, et al. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Med Res Methodol* 2011; **11**: 4. DOI: 10.1186/1471-2288-11-4.
6. 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*. Epub ahead of print 21 January 2014. DOI: 10.1177/0272989X13520192.

7. Tappenden P, Chilcott J, Ward S, et al. Methodological issues in the economic analysis of cancer treatments. *Eur J Cancer* 2006; **42**: 2867–2875.
8. Watkins C, Huang X, Latimer N, et al. Adjusting overall survival for treatment switches: commonly used methods and practical application. *Pharm Stat* 2013; **12**: 6.
9. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. *Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics*, http://www.fda.gov/downloads/Drugs/.../Guidances/ucm071590.pdf (2007, accessed 8 November 2014).
10. Committee for Medicinal Products for Human Use (CHMP). Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man (CHMP/EWP/205/95 REV.3). Methodological considerations for using progression-free survival (PFS) as primary endpoint in confirmatory trials for registration. European Medicines Agency, http://www.ema.europa.eu/docs/en_GB/document_library/Other/2009/12/WC500017749.pdf (2008, accessed 8 November 2014).
11. Robins JM and Tsiatis AA. Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Commun Stat Theory Methods* 1991; **20**: 2609–2631.
12. White IR. Uses and limitations of randomization-based efficacy estimators. *Stat Methods Med Res* 2005; **14**: 327–347.
13. Lee Y, Ellenberg J, Hirtz D, et al. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* 1991; **10**: 1595–1605.
14. Horwitz R and Horwitz S. Adherence to treatment and health outcomes. *Arch Intern Med* 1993; **153**: 1863–1868.
15. White IR, Walker S, Babiker AG, et al. Impact of treatment changes on the interpretation of the Concorde trial. *Aids* 1997; **11**: 999–1006.
16. Hernan MA, Brumback B and Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Statist Assoc* 2001; **96**: 440–448.
17. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Ser B* 1972; **34**: 187–220.
18. Robins JM and Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**: 779–788.
19. Robins JM and Greenland S. Adjusting for differential rates of prophylaxis therapy for Pcp in high-dose versus low-dose Azt treatment arms in an aids randomized trial. *J Am Stat Assoc* 1994; **89**: 737–749.
20. Yamaguchi T and Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: structural nested models and marginal structural models to test and estimate treatment arm effects. *Stat Med* 2004; **23**: 1991–2003.
21. Howe CJ, Cole SR, Chmiel JS, et al. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am J Epidemiol* 2011; **173**: 569–577.
22. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME and Berry D (eds) *Statistical models in epidemiology: the environment and clinical trials*. New York: Springer-Verlag, 1999, pp.95–134.
23. Robins JM. Structural nested failure time models. In: Andersen PK and Keiding N (eds) *Survival analysis*. Chichester, UK: John Wiley and Sons; Armitage P and Colton T (eds) *The encyclopedia of biostatistics*, 1998, pp. 4372–4389.
24. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H and Mulley A (eds) *Health service research methodology: a focus on AIDS*. Washington, DC: U.S. Public Health Service, National Center for Health Services Research, 1989, pp.113–159.
25. Sterne JAC and Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *Stata J* 2002; **2**: 164–182.
26. Mark SD and Robins JM. A method for the analysis of randomized trials with compliance information - an application to the multiple risk factor intervention trial. *Controlled Clin Trials* 1993; **14**: 79–97.
27. Robins JM. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow DG and Kessler R (eds) *Methodological issues of AIDS mental health research*. New York: Plenum Publishing, 1993, pp.213–290.
28. White IR, Babiker AG, Walker S, et al. Randomization-based methods for correcting for treatment changes: examples from the Concorde trial. *Stat Med* 1999; **18**: 2617–2634.
29. Branson M and Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Stat Med* 2002; **21**: 2449–2463.
30. Stata statistical software intercooled, Version 11.0, Texas, USA, 2009.
31. Crowther MJ, Abrams KR and Lambert PC. Joint modeling of longitudinal and survival data. *Stata J* 2013; **13**: 165–184.
32. Crowther MJ and Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013; **32**: 4118–4134.
33. Crowther MJ, Abrams KR and Lambert PC. Flexible parametric joint modelling of longitudinal and survival data. *Stat Med* 2012; **31**: 4456–4471.
34. Bender R, Augustin T and Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005; **24**: 1713–1723.
35. GlaxoSmithKline. *Manufacturer submission to the national institute for health and clinical excellence. Submission to address the question of whether and how lapatinib falls within the Supplementary Advice to Appraisal Committees on appraising treatments that extend life at the end of life*. UK: GlaxoSmithKline, http://www.nice.org.uk/guidance/gid-tag387/documents/final-submission2 (2009, accessed 8 November 2014).
36. Merck Serono LTD. *Single technology appraisal submission: Erbitux (cetuximab) for the first-line treatment of recurrent and/or metastatic squamous cell carcinoma of the head and neck*. UK: Merck Serono LTD, 2009.
37. Collett D. *Modelling survival data in medical research*, 2nd ed. Boca Raton: Chapman & Hall/CRC CRC Press LLC, 2003.
38. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Stat Med* 2006; **25**: 4279–4292.
39. White IR, Walker S and Babiker AG. strbee: Randomization-based efficacy estimator. *Stata J* 2002; **2**: 140–150.
40. Howe CJ, Cole SR, Chmiel JS, et al. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am J Epidemiol* 2011; **173**: 569–577.

41. Yamaguchi T and Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: an application in a clinical trial of unresectable non-small-cell lung cancer. *Stat Med* 2004; **23**: 2005–2022.
42. Vermorken JB, Mesia R, Rivera F, et al. Platinum-based chemotherapy plus Cetuximab in head and neck cancer. *N Engl J Med* 2008; **359**: 1116–1127.
43. Bond M, Hoyle M, Moxham T, et al. The clinical and cost-effectiveness of Sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, http://www.nice.org.uk/nicemedia/live/12040/43430/43430.pdf (2009, accessed 30 November 2014).
44. Demeris N, Lunn D and Sharples LD. Survival extrapolation using the poly-Weibull model. *Stat Methods Med Res* 2011. DOI: 10.1177/0962280211419645 [Epub ahead of print].