The University Of Sheffield.

School Of Health And Related Research.

Health Economics and Decision Science (HEDS)

Discussion Paper

Assessing methods for dealing with treatment crossover in clinical trials: A follow-up simulation study

NR Latimer, KR Abrams, PC Lambert, MJ Crowther, JP Morden

DP 14.01

HEDS Discussion Paper

No. 14.01

Assessing methods for dealing with treatment crossover in clinical trials: A follow-up simulation study

NR Latimer[1], KR Abrams[2], PC Lambert[2,3], MJ Crowther[2], JP Morden[4].

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

[2] Department of Health Sciences, University of Leicester

[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.

Assessing methods for dealing with treatment crossover in clinical trials: A follow-up simulation study

Latimer NR[1], Abrams KR[2], Lambert PC[2,3], Crowther MJ[2], Morden JP[4].

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

[2] Department of Health Sciences, University of Leicester

[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.

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

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

Michael 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).

Conflict of interests statement

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.

James Morden works for the ICR-CTSU, which receives core funding from Cancer Research UK.

Part of this work was carried out whilst PCL was on study leave from the University of Leicester.

Abstract

Background Treatment switching commonly occurs in clinical trials of novel interventions, particularly in the advanced or metastatic cancer setting, which causes important problems for health technology assessment. Previous research has demonstrated which adjustment methods are suitable in specific scenarios, but scenarios considered have been limited.

Objectives We aimed to assess statistical approaches for adjusting survival estimates in the presence of treatment switching in order to determine which methods are most appropriate in a new range of realistic scenarios, building upon previous research. In particular we consider smaller sample sizes, reduced switching proportions, increased levels of censoring, and alternative data generating models.

Methods We conducted a simulation study to assess the bias, mean squared error and coverage associated with alternative switching adjustment methods across a wide range of realistic scenarios.

Results Our results generally supported those found in previous research, but the novel scenarios considered meant that we could make conclusions based upon a more robust evidence base. Simple methods such as censoring or excluding patients that switch again resulted in high levels of bias. More complex randomisation-based methods (e.g. Rank Preserving Structural Failure Time Models (RPSFTM)) were unbiased when the "common treatment effect" held. Observational-based methods (e.g. inverse probability of censoring weights (IPCW)) coped better with time-dependent treatment effects but are heavily data reliant, and generally led to higher levels of bias in our simulations. Novel "two stage" methods produced relatively low bias across all simulated scenarios. All methods generally produced higher bias when the simulated sample size was smaller and when the censoring proportion was higher. All methods generally produced lower bias when switching proportions were lower. We find that the size of the treatment effect in terms of an acceleration factor has an important bearing on the levels of bias associated with the adjustment methods.

Conclusions Randomisation-based methods can accurately adjust for treatment switching when the treatment effect received by patients that switch is the same as that received by patients randomised to the experimental group. When this is not the case observational-based methods or simple two-stage methods should be considered, although the IPCW is prone to substantial bias when the proportion of patients that switch is greater than approximately 90%. Simple methods such as censoring or excluding patients that switch should not be used.

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 (RCT) 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 estimates of overall survival (OS) are key.[1,2,3,4] 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.[5,6,7,8] 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 (PFS) as a primary endpoint, rather than overall survival (OS), because drug regulatory agencies such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) 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 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] In previous research we completed a simulation study that included a full comparison of all relevant adjustment methods across a range of realistic scenarios – including scenarios where key methodological assumptions are not satisfied.[12] Such a study had not previously been undertaken. The aim of this paper is to describe a second simulation study that complements the previous study, providing further

information on the performance of switching adjustment methods in realistic scenarios. In Section 2 we summarise the findings of our previous study and present the aims of the current study. In Section 3 we briefly describe the switching adjustment methods. Section 4 presents the methods we have used to conduct the current study and Section 5 presents results of the study, and discusses these. Section 6 considers the implications of our results, offering conclusions and recommendations, and also considers the limitations of the study.

2. Findings from previous research

In our previous study we simulated 72 scenarios in order to analyse the bias and coverage associated with a range of switching adjustment methods in a wide range of different situations.[12] Simple adjustment methods, such as a standard intention to treat (ITT) analysis, censoring switchers at the point of switch (PPcens), excluding switching patients from the analysis (PPexc), and including the treatment received as a time-dependent covariate were compared to more complex methods. The more complex methods were categorised as observational-based or randomisation-based. Observational-based methods included inverse probability of censoring weights (IPCW) and structural nested models (SNM). Randomisation-based methods included the rank preserving structural failure time model (RPSFTM) and the iterative parameter estimation algorithm (IPE). In addition, we considered a novel two-stage Weibull method.

Our simulation study demonstrated that naïve methods (such as simple censoring and exclusion approaches) produced high levels of bias consistently across all scenarios and thus should be avoided. We found 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). 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 may not be preferable to an ITT analysis.

We found that 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 suggested that the relatively small size of RCT datasets may cause these methods to work sub-optimally – these methods produced important levels of bias (approximately 5-10%) even when the “no unmeasured confounders” assumption held and the switching proportion was moderate (approximately 60%). The bias associated with the observational-based methods increased substantially when switching proportions increased to around 90%.

We found that the novel two-stage Weibull method performed well across the majority of scenarios, often producing less bias than any of the other adjustment methods. Although the method was sensitive to the switching proportion, it was much less sensitive to this than the IPCW and SNM methods.

Although the findings of our previous study were valuable, we did not obtain information on all scenarios that may be of interest. In particular, we did not test any scenarios with a switching proportion of lower than 52%. We focussed upon high switching proportions because we hypothesised that adjustment methods would struggle most when a high proportion of control group patients switched treatments – and thus, if they performed well in these scenarios they could be expected to perform well at lower switching proportions. However, given that several of the adjustment methods did not perform well at high switching proportions it would be valuable to investigate whether they perform better at lower switching proportions. Also, our previous study only considered a sample size of 500, with 1:1 randomisation to the control and experimental groups. Feedback suggested that metastatic oncology RCTs often have sample sizes of less than 500, and that they are often randomised using a 2:1 ratio in favour of the experimental group. Given the reliance of observational-based methods on patient and event numbers, analysing the performance of the alternative methods with a lower sample size is important. Additionally, in our previous study we only considered administrative censoring proportions of 1-21% across the 72 scenarios simulated, whereas in fact censoring proportions for overall survival may be higher than this in oncology trials. Hence, further investigation of this parameter is warranted. Similarly, further investigation of the impact of missing data on potentially important confounders is important. Feedback from clinicians suggested that usually patients have a choice of whether they wish to accept a clinician's offer of switching treatments – and usually data on this choice would not be collected in a clinical trial. This represents an "unmeasured confounder" for the probability of switching, and could impact upon the performance of the IPCW method. Finally, as with any simulation study, there may be suspicion that results were driven by the methods used to generate the simulated data. Hence, we wished to investigate the use of alternative data generating models.

Therefore, the aims of the current study were to provide further information on the performance of switching adjustment methods, focussing on five main areas:

- Lower switching proportions
- Lower sample sizes
- Higher censoring proportions
- Missing predictors of switch
- Alternative data generating models.

3. Adjustment methods

In this section we briefly introduce the switching adjustment methods. The different switching adjustment methods are grouped into simple methods (those which are currently widely used),[6] and more complex methods. Further, the more complex methods are classified as "observational-based" methods and "randomisation-based" methods.

3.1 Simple methods

3.1.1 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 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.[20]

3.1.2 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 that 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.[21,22]

3.1.3 Treatment as a time-dependent covariate

Under this approach data are analysed according to treatment received, using a Cox proportional hazards model [23] in which a binary time-dependent covariate indicates time-periods in which treatment was received. The model takes the form:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta X_i(t)\right) \quad (1)$$

where $\lambda_0(t)$ is the baseline hazard function and $X_i(t)$ takes the value of zero while a patient is receiving the control and 1 while they are receiving the experimental treatment. Again, this approach may break the randomisation balance and is therefore prone to selection bias.[24]

3.2 Complex methods

3.2.1 Observational-based complex methods

3.2.1.1 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. In the context of treatment switching, patients are artificially censored at the time of switch, and remaining observations are weighted based upon covariate values in an attempt to remove selection bias.

Stabilised weights ($\widehat{W}(t)$) applied to each individual for time interval (t), as specified by Hernan *et al.* (2001) are:[25]

$$\widehat{W}(t) = \prod_{k=0}^{t} \frac{Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),V,T>k]}{Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),\bar{L}(k),T>k]} \quad (2)$$

where $C(k)$ is an indicator function demonstrating whether or not 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 (2) represents the probability of an individual remaining uncensored (not 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.

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. 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 IPCW versions.[26]

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. This assumption cannot be tested using the observed data.[27,28] Models for switching and survival must be correctly specified,[29] and the method fails if there are any covariates which ensure (that is, the probability equals 1) that treatment switching will occur.[25,28,30]

3.2.1.2 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 datasets.[31] However, these models can also be used to address treatment switching in an RCT. 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 (1998):[31]

$$U = \int_0^T \exp[\psi A_i(t)]dt \qquad (3)$$

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. The SNM is used to estimate counterfactual survival times for a range of possible treatment effects and g-estimation is used to determine a value ψ_0 for which treatment exposure at each time-point is independent of counterfactual survival. The model used for the g-test, as specified by Robins (1998),[31] is a time-dependent Cox proportional hazards model for the hazard of treatment change:

$$\lambda_0(t)\exp[\alpha' W(t)] \qquad (4)$$

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 *θ = 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. The confidence interval for ψ_0 is given by the values of ψ that result in the g-test not being rejected at the 0.05 level.[31]

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.[32].

3.2.2 Randomisation-based efficacy estimators

3.2.2.1 Rank Preserving Structural Failure Time Model

The RPSFTM method represents a SNM approach designed specifically for an RCT context.[11] The RPSFTM uses a counterfactual framework to estimate the causal effect of the treatment in question, but relies only upon the randomisation of the trial, treatment history and observed survival times to identify the treatment effect. 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 compliance is full 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 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:

$$U_i = T_{A_i} + e^{\psi 0} T_{B_i} \quad (5)$$

$e^{-\psi 0}$ represents the acceleration factor 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:

$$U_i = \int_0^{T_i} \exp[\psi X_i(t)] dt \quad (6)$$

which is identical to the SNM introduced in equation (3). The value of ψ is estimated using a grid search. For each value of ψ equation (5) 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.[33] The point estimate of ψ is that for which the test (z) statistic equals zero.

White *et al.* (1999) demonstrate that censoring is problematic for the RPSFTM due to an association between treatment received, counterfactual censoring time, and prognosis.[34] The authors suggest that possible bias be avoided by breaking the dependence between censoring time and treatment received by recensoring $U_i(\psi)$ 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 RPSFTM is rank preserving, and therefore assumes that if two patients have the same observed event time and neither have received treatment, those two patients would also have the same event time if they both received treatment. Further, it is assumed 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]

3.2.2.2 Iterative Parameter Estimation algorithm

Branson and Whitehead (2002) extended the RPSFTM method using parametric methods, developing a novel iterative parameter estimation (IPE) procedure.[35] 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 (6), 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 is continued 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.[35] Bootstrapping is recommended to obtain standard errors and confidence intervals for the treatment effect.[35]

The IPE procedure makes similar assumptions to the RPSFTM method – for example the randomisation assumption is made, as is the "common treatment effect" assumption. An additional assumption is that survival times follow a parametric failure time distribution.

3.2.2.3 Two-stage estimation – a novel method

In our previous simulation study we considered a novel method for adjusting for treatment switching, designed in accordance with the type of switching often observed in metastatic oncology RCTs.[12] Usually switching is only permitted after disease progression, but is likely to happen soon after this time-point. In this case, disease progression can be used as a secondary baseline for patients in the control group and data on these patients can be treated as an observational dataset. Fitting an accelerated failure time (AFT) model (such as a Weibull model) to this data including covariates measured at the secondary baseline and including a time-varying covariate indicating treatment switch would be expected to produce a reasonable estimate of the treatment effect received by patients who switched, 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 can then be obtained using:

$$U_i = T_{A_i} + \frac{T_{B_i}}{\mu_B} \qquad (7)$$

Where T_{A_i} represents the time spent on control treatment, T_{B_i} represents the time spent on the new intervention, and $\mu_B$is the treatment effect (acceleration factor) in switching patients.

Robins and Greenland (1994) and Yamaguchi and Ohashi (2004) have previously used a similar approach to adjust for treatment switches,[27,28] but have utilised an SNM to estimate the treatment effect in the control group, rather than a less complex AFT model as suggested here. The simplified approach suggested in our previous report and re-stated here makes no attempt to adjust for time-dependent confounding beyond disease progression, but requires less data (the "no unmeasured confounders" assumption is only required at the secondary baseline timepoint) and does not require

modelling of the treatment switching process. If switching occurs soon after the secondary baseline bias caused by time-dependent confounding may be minimal. Unlike the RPSFTM and IPE methods, the simple two-stage method suggested here does not require the "common treatment effect" assumption, and our previous study showed that this method performed very well across a wide range of scenarios. Hence we investigate this method further in the current study. The method is not restricted to a Weibull model – any AFT model may be used, and hence in this study we consider both a Weibull model and a Generalised Gamma model. In addition, in our previous study we did not incorporate recensoring within this method. However, because the method involves "shrinking" both censoring times and survival times for switching patients there is scope for informative censoring, as demonstrated by White *et al.* (1999) for the RPSFTM approach.[34] Hence, in the current study we incorporate full recensoring within the two-stage estimation, whereby $U_i(\mu_B)$ (where μ_B is the acceleration factor associated with treatment estimated by the AFT model fitted to the control group post-progression data) is recensored at the minimum of the administrative censoring time C_i and $C_i \exp \mu_B$. $U_i(\mu_B)$ is then replaced by the censoring time of the counterfactual event time $D_i^*(\mu_B)$ if $D_i^*(\mu_B) < U_i(\mu_B)$.

4. Novel simulation study

We simulated independent datasets in which the true survival differences between treatment options were known. We then applied each of the switching adjustment methods and compared their bias, mean squared error 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. The simulation study was conducted using Stata software, version 11.2.[39]

4.1 Underlying survival times

We used a joint survival and longitudinal model to simultaneously generate a continuous time-dependent covariate (referred to as "antigen") and survival times.[40] 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 antigen value for the i^{th} patient at time t was:

$$antigen_i(t) = \beta_{0_i} + \beta_1 t + \beta_2 t \times trt_i + \beta_4 badprog_i \qquad (8)$$

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.

In our previous study we simulated survival times using a Weibull baseline distribution and the antigen model changed linearly with log time (rather than time shown in Equation (8)). This allowed use of the inversion simulation method described by Bender *et al.* (2005),[41] which is a computationally simple method to implement, with all required formulae having closed form solutions.

In order to simulate survival times from a more complex underlying distribution, to reflect those seen in real datasets, we used the general survival simulation framework described by Crowther and Lambert (2013),[40] which uses a combination of numerical integration and root finding to simulate survival dependent on a time-varying biomarker. In particular, we assume a 2-component mixture Weibull baseline hazard function to incorporate the desired flexibility. This can be written as

$$h_0(t) = \frac{\lambda_1\gamma_1 p t^{\gamma_1-1}\exp(-\lambda_1 t^{\gamma_1})+\lambda_2\gamma_2(1-p)t^{\gamma_2-1}\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1})+(1-p)\exp(-\lambda_2 t^{\gamma_2})} \quad (9)$$

where $\lambda_1, \lambda_2 > 0$ and $\gamma_1, \gamma_2 > 0$ are scale and shape parameters, respectively. We have the mixture parameter, *p*, with $0 \leq p \leq 1$, therefore p represents the contribution of the first Weibull to the overall survival model, and $1 - p$ represents the contribution of the second Weibull. The linear predictor of the survival model is then incorporated as follows:

$$h_i(t) = h_0(t)\exp[X_i(t)\beta(t)] \quad (10)$$

where,

$$X_i(t)\beta(t) = \delta_1(trt_i) + \big(\eta(t)\big)trt_i + \delta_2 badprog_i + \alpha\big(antigen_i(t)\big) \quad (11)$$

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 antigen level.

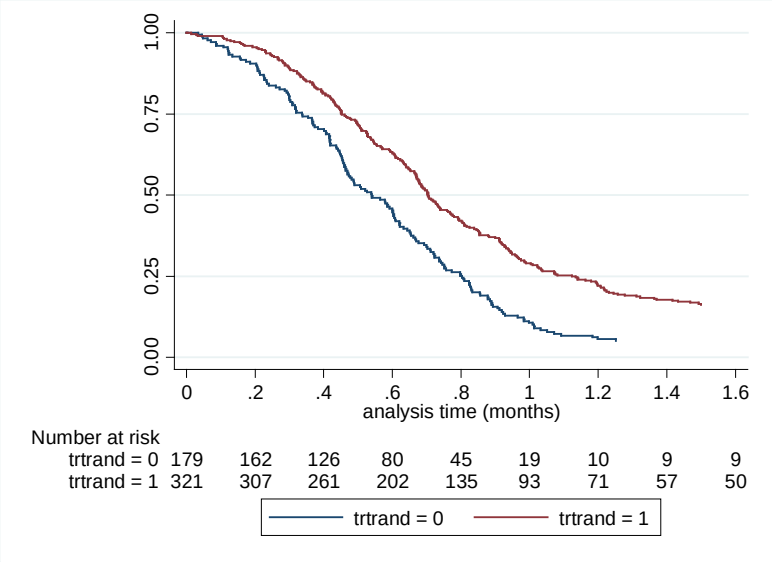
Simulating using a mixture model allows us to simulate complex hazard functions that could not be produced using one Weibull model. The result is a hazard function that does not represent that associated with any standard parametric distribution. This is important because there is no reason to expect that real-world survival data will follow standard parametric distributions, and also because simulating complex hazard functions means that none of the switching adjustment methods should be advantaged due to underlying assumptions.

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

$\beta_0 = 20$, $\sigma_0^2 = 1$, $\beta_1 = 15$, $\beta_2 = -8$, $\beta_4 = 10$, $\delta_1 = -0.75$, $\delta_2 = 0.5$, $\alpha = 0.02$, $\lambda_1 = 1.8$, $\gamma_1 = 2.1$, $\lambda_2 = 0.1$, $\gamma_2 = 0.5$, $pmix = 0.7$, $\eta = 0.3$

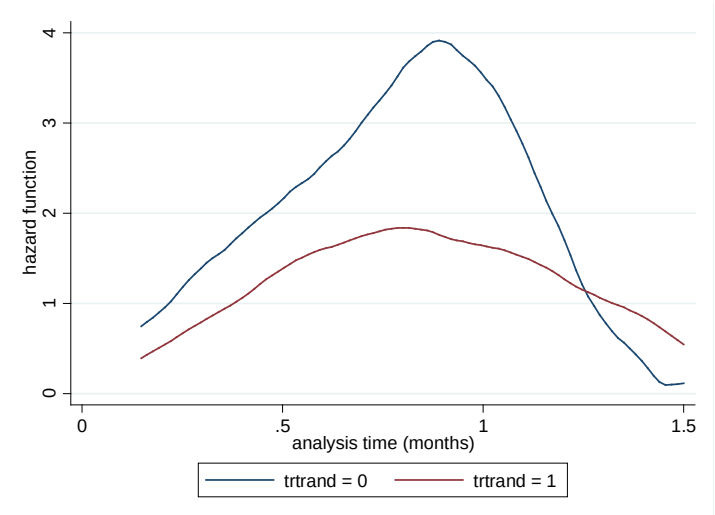
One example of the Kaplan-Meier curves produced by the simulation model (in the absence of treatment switching) using these parameter values are presented in Figure 1. Note that trtrand=0 represents the control group, and trtrand=1 represents the experimental group.

Figure 1: Overall Survival Kaplan-Meier from one simulated dataset Scenario 1: No switching



The hazard function associated with Figure 1 is illustrated in Figure 2. This demonstrates that we simulated a hazard function that was initially low, which then steadily increased before decreasing towards the end of the trial follow-up. We believe that this is representative of the types of hazards that would be expected within an metastatic oncology RCT setting – the initial hazard is likely to be low, reflecting the inclusion criteria usually used in RCTs which means that patients with the worst prognosis are usually excluded. The hazard is then likely to rise, reflecting the seriousness of the disease, before falling in the longer-term as those who remain alive are of relatively better prognosis.

Figure 2: Hazard function from one simulated dataset Scenario 1: No switching



4.2 Treatment effect in the experimental group

We cannot write down the treatment effect experienced in the experimental group over time because our hazard function includes "*t*" terms. However, as demonstrated by Figure 2, the treatment effect (as observed by the difference between the hazard functions) initially increases during the period of greatest hazard, before falling in the longer-term. We believe that this is representative of a realistic treatment effect, which falls in the longer-term when the initial treatment effect may have worn off, and when only better prognosis patients remain alive.

4.3 Treatment effect in switchers

The treatment effect applied to patients who switched from the control group to the experimental treatment was calculated in the same way as for our previous study. 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% and 20%. This allowed us to test scenarios in which the "common treatment effect" assumption did not hold.

4.4 The switching mechanism

We allowed the probability of treatment switching to depend upon the antigen value at the time of disease progression and the time of progression itself, in a similar way to that modelled in our previous simulation study. 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 consultations immediately following disease progression (including the consultation at which progression was first observed), and the probability of switching declined in each of these consultations. Consultations 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 consultation at which disease progression was observed. In addition, switching was only permitted in patients who were randomly assigned a value of '1' for a "choice" variable. 80% of patients were assigned a value of '1' and 20% were assigned a value of '0'. Hence patients were only "at risk" of switching if they had a "choice" covariate value of '1', and if they had had 3 or less consultations since their disease progression was observed (including the initial consultation at which disease progression was observed). Patients never became "at risk" of switching if they had a "choice" covariate value of '0', and if they died before 21 days since randomisation.

The probability of switching was calculated for each control group patient who had a "choice" covariate value of '1' using a logistic function. In the base case the probability of switching increased if the antigen value was high at the time of disease progression, and if time-to-progression was high.

Both these factors indicate that patients with a relatively long progression free survival period are more likely to switch treatments than patients with shorter progression free survival times – hence, patients with better prognosis were more likely to switch. The probability of switching in the different progression and antigen groups at the three consultations following disease progression for the base case scenario are presented in Appendix 1.

4.5 Scenarios investigated

The simulated data generating mechanism had several variables for which values had to be assumed. These are listed in Appendix 2. The variables altered within the simulations related to:

- Sample size: moderate (n=500); small (n=300); large (n=1,000)
- Data generating model: 2-component mixture Weibull baseline hazard function; 2-component mixture Gompertz baseline hazard function
- Treatment effect decrement received by switchers: 0% (zero time-dependency); 20%
- Switch proportion: moderate (approximately 50% of control group); low (approximately 20% of control group); very low (approximately 7% of control group); very high (approximately 94% of "at-risk" control group, which is equivalent to approximately 70% of all control group patients)
- Treatment effect: moderate (average HR approximately 0.75); high (average HR approximately 0.50)
- Disease severity: moderate (restricted mean survival in control group approximately 285 days, censoring approximately 15%); high (restricted mean survival in control group approximately 365 days, censoring approximately 55%)

Varying the switch proportion, the treatment effect size, the commonality of the treatment effect and disease severity resulted in 16 scenarios. In addition, we tested the impact of different sample sizes and data generating models. All 16 of the base scenarios were tested again in simulations in which the sample size was reduced from 500 to 300. In all scenarios the randomisation was 2:1 in favour of the experimental group. Then, each of the 32 scenarios were replicated using a 2-component mixture Gompertz baseline hazard function instead of the Weibull function used for the base scenarios.

As an addition to our main 64-scenario analysis, we re-ran selected scenarios to investigate the impact of specific alterations to parameters. We selected 8 base scenarios to run with a sample size of 1,000 instead of 500, in order to investigate the impact of this on the performance of the adjustment methods. Given the observational-analysis nature of the IPCW method, we hypothesised that a larger sample size may improve its performance. Four base scenarios were re-run incorporating a very low switching proportion because we wished to investigate whether the IPCW method performed poorly if very few patients switched (which could lead to problems with accurately modelling the switching mechanism). Finally, 4 scenarios were re-run incorporating a very high switching proportion in order to assess the consistency of our results with our previous study. In total 80 scenarios were run. One-thousand simulations were run for each scenario.

4.6 Performance measures

Because of the inclusion of a time-varying confounder in our simulated dataset, the treatment effect is a function both of the initially assigned treatment effect, and the effect that occurs through the antigen (CEA), as the experimental treatment reduces the antigen value over time and the antigen value carries a risk of death. Therefore, the treatment effect is not constant over time, and it is not possible to produce a single true HR or acceleration factor that the results of the adjustment methods can be compared to. Instead, as in our previous study, we used restricted mean survival time as our true value upon which to base our performance measures. Because this study seeks to provide information on the performance of switching adjustment methods it is of most relevance to consider mean survival times restricted to the trial follow-up period, so as not to confuse bias associated with extrapolation methods with bias associated with switching adjustment methods.

In our previous study we were able to integrate the survivor functions associated with our simulated survival data. However, this is no longer possible because our hazard function includes "*t*" terms. Instead, for each scenario we simulated data for 1,000,000 patients without incorporating treatment switching, and estimated the mean survival at 18 months (the administrative censoring time in the simulated dataset). We used this as our "truth" upon which to base our performance measures. Because this value is the product of a simulation rather than a calculation it is prone to error, but this is likely to be extremely minimal given the large number of patients simulated.

We evaluated the performance of the switching adjustment methods according to the bias in their estimate of the true area under the curve (restricted mean at 18 months) for the control group. Bias (δ) was measured by the difference between the true restricted mean (β) and the estimated restricted mean ($\bar{\hat{\beta}}$). Relative bias was calculated as $\frac{\delta}{\beta}$. The mean squared error (MSE) was also calculated, where the standard error was that associated with the mean restricted mean estimated by each adjustment method over the 1,000 simulations run for each scenario.

The coverage of each method was also calculated, defined as the proportion of simulations where the 95% confidence intervals of the restricted mean estimated by each method contained the true restricted mean. We also calculated the proportion of times that each method resulted in an estimate of the treatment effect (i.e. the proportion of times they converged), which helps illustrate whether any of the methods are potentially unreliable and unsuitable for use in the real world. Where methods do not converge the bias and coverage performance measures were calculated based upon simulations in which convergence did occur.

4.7 Adjustment methods to be included

The methods tested in our simulation study were those described in Section 3, with some exceptions. It is clear that the "simple" switching adjustment methods described in Section 3.1 are highly prone to bias – this is demonstrated by our previous simulation study.[12] In our current study we retained the

ITT and PP approaches since they are commonly used, and because the ITT analysis represents the standard statistical analysis of a clinical trial. We excluded the treatment as a time-dependent covariate approach in the current study, since this is seldom used and because it produced extremely high bias across the range of scenarios included in our previous study – we deemed it unnecessary to investigate this method any further. We also excluded the SNM method, because this typically performed more poorly than the IPCW method across the scenarios included in our previous study, and often failed to converge.[12]

For the RPSFTM method we used a Cox test within the g-estimation procedure, and 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. For the RPSFTM and IPE methods we included baseline covariates in the estimation procedure. For the IPCW method we used stabilised weights and included two versions – one in which all covariates were included in the relevant models, and one in which the "choice" covariate was excluded – in order to test the sensitivity of the method to the availability of this covariate.

We applied the two-stage method using both a Weibull model and a Generalised Gamma model, so that the performance of different AFT models could be compared. We fitted these models to control group patients using disease progression as the secondary baseline time-point, and included covariates for switching, baseline prognosis group, baseline antigen value, time-to-disease progression, antigen value at disease progression, and the "choice" covariate.

The specific approach we used to apply each of the adjustment methods are explained in detail in Appendix 3.

5. Results

The performance of each adjustment method differed importantly depending upon the scenario investigated. Due to the large number of methods and scenarios assessed it is not helpful to present detailed results for every method and every scenario. Instead, we present detailed results from 8 scenarios that clearly illustrate the key findings. In Section 5.1 we report key results in scenarios that involved moderate (approximately 50%) switching proportions, and in Section 5.2 we report key results in scenarios that involved low (approximately 20%) switching proportions. In Section 5.3 we summarise the extent to which the 8 scenarios focussed upon reflect the results of the 24 other base scenarios completed. We then summarise the results of the additional scenarios run – those that tested the sensitivity of the results to the data generating model, those that tested a larger sample size, and those that tested extreme high and low switching proportions. In this section, method names are abbreviated as follows: Intention-to-Treat (ITT), Exclude switchers (PPexc), Censor at switch (PPcens), Inverse Probability of Censoring Weights (IPCW), IPCW excluding the "choice" covariate (IPCWn), Rank Preserving Structural Failure Time Model (RPSFTM), Iterative Parameter

Estimation with a Weibull model (IPE), Iterative Parameter Estimation with an exponential model (IPEexp), two-stage Weibull estimation (Weib2m), two-stage Generalised Gamma model (Gam2m).

In Appendix 4 we present an overview of each scenario run, with regard to average treatment effects, true restricted mean survival (area under the curve), switching proportions and censoring proportions.

5.1 Scenarios with moderate switching proportions

Tables 1 and 2 present detailed results from Scenarios 1, 3, 9 and 11. These are illustrative of the results of scenarios in which the switching proportion simulated was approximately 58 – 61% of at-risk patients. In Scenario 1 mean true survival in the control group (in the absence of treatment switching) was 0.56 years, and in the experimental group was 0.79 years, which was associated with an average HR of 0.51. The mean switching proportion was 44% of all control group patients, which equated to 58% of those who became at risk of switching. 14% of patients were administratively censored at 1.5 years and the treatment effect applied to switching patients was 20% 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 antigen covariate value at the time of disease progression, and the time to disease progression itself. Patients with higher antigen values at longer progression free survival times were more likely to switch. These patients tended to be of relatively good prognosis.

As expected, the ITT analysis overestimated the true, unconfounded control group mean survival time in this scenario. The absolute bias was 0.05 years, equivalent to 8.22% relative to the true mean survival time. Simple adjustment methods (PPexc, PPcens) produced substantially higher relative bias than the ITT analysis, ranging from -9.94% to 17.66%. The IPCW and the IPCWn both underestimated mean survival in the control group (hence over-estimated the treatment effect), but produced lower bias than the ITT analysis, with the version that included the "choice" covariate resulting in marginally less bias than the version that excluded this covariate (relative bias of -2.30% compared to -2.46%). In this scenario the RPSFTM, IPE and IPEexp all produced very similar levels of bias (relative bias -1.40%, -1.70% and -1.90% respectively), underestimating mean survival in the control group but producing lower bias than the ITT and IPCW analyses. The Weib2m and Gam2m methods produced less bias than all other methods, resulting in very low relative bias of 0.48% and 0.41% respectively.

The only substantive difference between Scenario 1 and Scenario 3 was that the treatment effect was lower in Scenario 3, with mean survival time 0.64 years in the control group and 0.74 years in the experimental group, associated with an average HR of 0.76. Owing to this, the relative bias associated with each of the adjustment methods generally marginally decreased. However, the best performing adjustment methods remained the same – the Weib2m and Gam2m produced least bias (relative bias -0.28% and -0.36% respectively). However, in this scenario the IPCW and IPCWn produced lower bias than the RPSFTM, IPE and IPEexp (relative bias 1.19% and 0.51% compared to

-1.40%, -1.69% and -1.79% respectively). It is interesting that the IPCWn produced lower bias than the IPCW – a finding that we will discuss further in Section 5.7.2. In this scenario the simple adjustment methods again produced substantially higher levels of bias (PPcens relative bias: 22.13%; PPexc relative bias: -5.37%). Due to the lower treatment effect the ITT analysis gave lower relative bias (2.94%) than in Scenario 1, but still produced higher bias than all adjustment methods except PPcens and PPexc.

Table 1: Scenarios 1 and 3 - Results

Scenario details	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Relative bias	MSE	Coverage (%)	Successful estimation (%)
				Lower	Upper					
Scenario number: 1	ITT	0.60	0.03	0.55	0.66	0.05	8.22	0.0029	65.40	100.00
True mean survival:	PPExc	0.50	0.04	0.43	0.58	-0.06	-9.94	0.0044	66.40	100.00
Control: 0.56	PPcens	0.65	0.04	0.57	0.74	0.10	17.66	0.0114	36.10	100.00
Experimental: 0.79	IPCW	0.54	0.03	0.47	0.63	-0.01	-2.30	0.0013	97.60	100.00
	IPCWn	0.54	0.03	0.47	0.63	-0.01	-2.46	0.0012	97.40	100.00
Mean switch %: 58.26%	Weib2m	0.56	0.03	0.54	0.58	0.00	0.48	0.0008	52.10	100.00
True Average HR: 0.51	Gam2m	0.56	0.03	0.54	0.58	0.00	0.41	0.0008	52.10	100.00
Mean censored: 13.59%	RPSFTM	0.55	0.03	0.49	0.62	-0.01	-1.40	0.0011	94.70	100.00
Treatment effect:	IPE	0.55	0.03	0.48	0.62	-0.01	-1.70	0.0012	94.20	100.00
20% decrement	IPEexp	0.55	0.03	0.48	0.62	-0.01	-1.90	0.0012	93.90	100.00
Scenario number: 3	ITT	0.66	0.03	0.60	0.72	0.02	2.94	0.0014	91.90	100.00
True mean survival:	PPExc	0.61	0.05	0.51	0.70	-0.03	-5.37	0.0034	86.50	100.00
Control: 0.64	PPcens	0.78	0.05	0.69	0.88	0.14	22.13	0.0223	14.70	100.00
Experimental: 0.74	IPCW	0.65	0.04	0.55	0.76	0.01	1.19	0.0018	98.40	100.00
	IPCWn	0.64	0.04	0.55	0.74	0.00	0.51	0.0015	98.60	100.00
Mean switch %: 61.20%	Weib2m	0.64	0.03	0.62	0.67	0.00	-0.28	0.0012	50.60	100.00
True Average HR: 0.76	Gam2m	0.64	0.03	0.62	0.67	0.00	-0.36	0.0012	51.80	99.90
Mean censored: 15.03%	RPSFTM	0.63	0.04	0.56	0.71	-0.01	-1.40	0.0018	93.30	100.00
Treatment effect:	IPE	0.63	0.04	0.56	0.71	-0.01	-1.69	0.0019	92.80	100.00
20% decrement	IPEexp	0.63	0.04	0.56	0.71	-0.01	-1.79	0.0019	92.50	100.00

Table 2 presents detailed results of Scenario 9 and Scenario 11. Scenario 9 is approximately equivalent to Scenario 1 and Scenario 11 is approximately equivalent to Scenario 3, except the "common treatment effect" assumption holds. This has an important impact upon the results of the adjustment methods. While the Weib2m and Gam2m methods continued to produce very low levels of bias (relative bias -0.36% to 0.44%), the RPSFTM/IPE methods also performed very well, with relative bias between -0.01% and -0.55% in Scenario 9, and between -0.77% and -1.17% in Scenario 11. The RPSFTM again produces lower bias than the IPE method, which in turn produces lower bias than the IPEexp method. In Scenarios 9 and 11 the IPCW and IPCWn methods produced similar levels of bias to those found in Scenarios 1 and 3 (relative bias -2.27% and -2.49% in Scenario 9, and 1.29% and 0.60% in Scenario 11), with lower bias again produced in the scenario exhibiting the lower treatment effect. The PPcens and PPexc methods again produced substantially higher levels of bias

(relative bias -5.46% to 21.74%) than all methods, including the ITT analysis (relative bias associated with the ITT analysis was 9.83% in Scenario 9 and 3.48% in Scenario 11).

Table 2: Scenarios 9 and 11 - Results

Scenario details	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Relative bias	MSE	Coverage (%)	Successful estimation (%)
				Lower	Upper					
Scenario number: 9	ITT	0.61	0.03	0.55	0.67	0.05	9.83	0.0039	56.20	100.00
True mean survival:	PPExc	0.50	0.04	0.42	0.57	-0.06	-10.54	0.0049	63.40	100.00
Control: 0.56	PPcens	0.65	0.04	0.57	0.74	0.10	17.16	0.0109	40.10	100.00
Experimental: 0.79	IPCW	0.54	0.03	0.47	0.63	-0.01	-2.27	0.0013	98.00	100.00
	IPCWn	0.54	0.03	0.47	0.63	-0.01	-2.49	0.0012	97.70	100.00
Mean switch %: 58.27%	Weib2m	0.56	0.03	0.54	0.58	0.00	0.44	0.0008	54.90	100.00
True Average HR: 0.51	Gam2m	0.56	0.03	0.54	0.58	0.00	0.39	0.0008	55.40	100.00
Mean censored: 13.74%	RPSFTM	0.56	0.04	0.49	0.63	0.00	-0.01	0.0012	95.10	100.00
Treatment effect:	IPE	0.55	0.03	0.49	0.63	0.00	-0.31	0.0012	95.30	100.00
0% decrement	IPEexp	0.55	0.03	0.49	0.62	0.00	-0.55	0.0012	94.90	100.00
Scenario number: 11	ITT	0.66	0.03	0.60	0.73	0.02	3.48	0.0016	89.10	100.00
True mean survival:	PPExc	0.61	0.05	0.51	0.70	-0.04	-5.46	0.0036	84.60	100.00
Control: 0.64	PPcens	0.78	0.05	0.69	0.88	0.14	21.74	0.0218	16.20	100.00
Experimental: 0.74	IPCW	0.65	0.04	0.55	0.76	0.01	1.29	0.0020	98.00	100.00
	IPCWn	0.65	0.04	0.55	0.74	0.00	0.60	0.0017	98.30	100.00
Mean switch %: 60.90%	Weib2m	0.64	0.04	0.62	0.67	0.00	-0.36	0.0013	47.70	100.00
True Average HR: 0.76	Gam2m	0.64	0.04	0.62	0.67	0.00	-0.44	0.0013	48.60	100.00
Mean censored: 15.09%	RPSFTM	0.64	0.04	0.56	0.72	0.00	-0.77	0.0020	92.10	100.00
Treatment effect:	IPE	0.63	0.04	0.56	0.71	-0.01	-1.06	0.0020	91.20	100.00
0% decrement	IPEexp	0.63	0.04	0.56	0.71	-0.01	-1.17	0.0021	90.70	100.00

Tables 1 and 2 show a substantial difference in the levels of coverage and MSE achieved by each of the adjustment methods. It is important to note the low levels of coverage achieved by the Weib2m and Gam2m – particularly because these produce low levels of bias. These methods exhibit poor coverage because confidence intervals for mean counterfactual survival times were estimated by using the 95% confidence intervals for ψ_B in equation (7). 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. The MSE results provide information on the variability of the estimates obtained using the different adjustment methods, combined with the bias results. The MSE results suggest that the levels of variability associated with the different adjustment methods are generally similar relative to the bias levels – i.e. higher levels of bias are generally associated with higher MSEs. Variability is useful to consider, because if methods produce similar levels of bias, but one produces much more variable estimates, then the method that produces less variability may be preferred. However, our results suggest that different levels of

variability across methods do not seem to be of key importance – bias is the most important indicator of the performance of these methods.

In contrast, the IPCW method produced higher levels of coverage (97.40% - 98.60%) in Scenarios 1, 3, 9 and 11, indicating the relatively wide confidence intervals associated with the treatment effect estimated using the IPCW approach.

The RPSFTM, IPE and IPEexp approaches led to similar levels of coverage – a result that may not be expected as the RPSFTM generally retains the ITT analysis p-value in its estimating of the adjusted treatment effect, whereas the IPE analysis does not – the confidence intervals around the parameter estimates supplied by the final iteration of the IPE algorithm were used to generate restricted mean confidence intervals. However, because we have used baseline covariates to increase the power associated with these methods, the RPSFTM does not retain the simple ITT analysis p-value, and as a result coverage is reduced (compared, for example, to scenarios presented in our previous study for variations of the RPSFTM and IPE methods where baseline covariates were not included), though remains over 90% in these scenarios.

Successful estimation was achieved with all of the adjustment methods across Scenarios 1, 3, 9 and 11, with only the Gam2m method failing to converge in 0.1% of simulations in Scenario 3. It should be noted that the IPCW and IPCWn methods failed to converge in one of the weighting regressions in several simulations, but that estimations were still obtained from these. Restricting the results of these methods only to simulations in which full convergence was achieved in each of the regression models had only a minor impact on their performance – this will be discussed further in Section 5.8.1.

5.2 Scenarios with low switching proportions

Tables 3 and 4 present detailed results from Scenarios 5, 7, 13 and 15. These are illustrative of the results of scenarios in which the switching proportion simulated was approximately 23 – 26% of at-risk patients. Scenarios 5, 7, 13 and 15 are similar to Scenarios 1, 3, 9 and 11 respectively, with the only substantive difference the switching proportion.

The reduced switching proportion has a limited impact on the bias associated with the adjustment methods. In Scenarios 5 and 13 the bias associated with the IPCW and IPCWn methods increased (relative bias ranged from -3.52% to -4.05%) compared to the bias produced in Scenarios 1 and 9 (relative bias ranged from -2.27% to -2.49%), whereas the opposite was true in Scenarios 7 and 15 (relative bias ranged from -0.15% to -0.83%) compared to Scenarios 3 and 11 (relative bias ranged from 0.51% to 1.29%). This may suggest the IPCW method is susceptible to more fluctuation than the other adjustment methods. The bias associated with the two-stage Weib2m and Gam2m methods generally marginally decreased when the switching proportion reduced to the levels simulated in Scenarios 5, 7, 13 and 15, with relative bias remaining substantially less than 1% in each

scenario. Similarly, the bias associated with the RPSFTM, IPE and IPEexp methods marginally reduced in Scenarios 5, 7, 13 and 15 compared to Scenarios 1, 3, 9 and 11.

As expected, the bias associated with the ITT analyses and the simple adjustment methods (PPcens and PPexc) reduced in Scenarios 5, 7, 13 and 15 compared to Scenarios 1, 3, 9 and 11, owing to the lower switching proportion. However, in each Scenario PPcens and PPexc continued to produce higher bias than the ITT analysis and all of the complex adjustment methods (with the exception of Scenario 13, in which the PPcens method produced slightly less bias than the IPCW, IPCWn and ITT analyses (relative bias 3.73% compared to -4.02%, -4.05% and 3.99% respectively). The complex adjustment methods produced lower bias than the ITT analysis in Scenarios 5, 7, 13 and 15, with the exception of the IPCW and IPCWn methods, which produced marginally more bias than the ITT analyses in Scenarios 5 and 13 (relative bias of IPCW and IPCWn -3.52% and -3.53% compared to 3.47% in Scenario 5; relative bias -4.02% and -4.05% compared to 3.99% in Scenario 13). The RPSFTM produced least bias in Scenario 5, the IPCW produced least bias in Scenario 7, and the Gam2m produced least bias in Scenarios 13 and 15.

Table 3: Scenarios 5 and 7 - Results

Scenario details	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Relative bias	MSE	Coverage (%)	Successful estimation (%)
				Lower	Upper					
Scenario number: 5	ITT	0.58	0.03	0.52	0.63	0.02	3.47	0.0011	90.00	100.00
True mean survival:	PPExc	0.53	0.03	0.47	0.59	-0.03	-4.76	0.0016	83.60	100.00
Control: 0.56	PPcens	0.58	0.03	0.52	0.64	0.02	4.18	0.0015	89.90	100.00
Experimental: 0.79	IPCW	0.54	0.03	0.47	0.61	-0.02	-3.52	0.0012	95.80	100.00
	IPCWn	0.54	0.03	0.47	0.61	-0.02	-3.53	0.0012	95.40	100.00
Mean switch %: 23.78%	Weib2m	0.56	0.03	0.55	0.57	0.00	0.46	0.0007	33.80	100.00
True Average HR: 0.51	Gam2m	0.56	0.03	0.55	0.57	0.00	0.45	0.0007	32.90	100.00
Mean censored: 13.42%	RPSFTM	0.56	0.03	0.49	0.63	0.00	-0.17	0.0008	97.90	100.00
Treatment effect:	IPE	0.55	0.03	0.49	0.63	0.00	-0.31	0.0008	97.90	100.00
20% decrement	IPEexp	0.55	0.03	0.49	0.63	0.00	-0.33	0.0008	98.10	100.00
Scenario number: 7	ITT	0.65	0.03	0.59	0.72	0.01	1.69	0.0012	94.50	100.00
True mean survival:	PPExc	0.62	0.04	0.55	0.69	-0.02	-3.10	0.0018	88.70	100.00
Control: 0.64	PPcens	0.68	0.04	0.61	0.75	0.04	6.19	0.0029	82.00	100.00
Experimental: 0.74	IPCW	0.64	0.04	0.56	0.73	0.00	-0.15	0.0014	98.20	100.00
	IPCWn	0.64	0.04	0.56	0.73	0.00	-0.21	0.0014	98.10	100.00
Mean switch %: 25.87%	Weib2m	0.65	0.03	0.63	0.67	0.00	0.63	0.0011	32.30	100.00
True Average HR: 0.76	Gam2m	0.65	0.03	0.63	0.67	0.00	0.59	0.0011	33.40	100.00
Mean censored: 15.15%	RPSFTM	0.64	0.04	0.57	0.72	0.00	-0.17	0.0014	96.80	100.00
Treatment effect:	IPE	0.64	0.04	0.56	0.72	0.00	-0.29	0.0014	96.40	100.00
20% decrement	IPEexp	0.64	0.04	0.56	0.72	0.00	-0.30	0.0014	96.50	100.00

Tables 6 and 7 again show the low levels of coverage associated with the Weib2m and Gam2m methods, demonstrating the inadequacy of using the 95% confidence intervals for ψ_B to estimate

confidence intervals for mean counterfactual survival times. It is notable that the coverage of the RPSFTM, IPE and IPEexp is improved in Scenarios 5, 7, 13 and 15 compared to Scenarios 1, 3, 9 and 11, suggesting that the performance of these methods is improved in scenarios with lower switching proportions.

Successful estimation was achieved with all of the adjustment methods across Scenarios 5, 7, 13 and 15, with only the Gam2m method failing to converge in 0.3% of simulations in Scenarios 13 and 15. Again, it should be noted that the IPCW and IPCWn methods failed to converge in one of the weighting regressions in several simulations, but that estimations were still obtained from these. Restricting the results of these methods only to simulations in which full convergence was achieved in each of the regression models had only a minor impact on their performance.

Table 4: Scenarios 13 and 15 - Results

Scenario details	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Relative bias	MSE	Coverage (%)	Successful estimation (%)
				Lower	Upper					
Scenario number: 13	ITT	0.58	0.03	0.52	0.63	0.02	3.99	0.0013	88.40	100.00
True mean survival:	PPExc	0.53	0.03	0.47	0.59	-0.03	-5.19	0.0017	80.80	100.00
Control: 0.56	PPcens	0.58	0.03	0.52	0.64	0.02	3.73	0.0014	90.90	100.00
Experimental: 0.79	IPCW	0.53	0.03	0.47	0.61	-0.02	-4.02	0.0014	94.90	100.00
	IPCWn	0.53	0.03	0.47	0.61	-0.02	-4.05	0.0014	95.00	100.00
Mean switch %: 23.76%	Weib2m	0.56	0.03	0.55	0.57	0.00	0.12	0.0007	32.70	100.00
True Average HR: 0.51	Gam2m	0.56	0.03	0.55	0.57	0.00	0.09	0.0007	32.80	99.70
Mean censored: 13.52%	RPSFTM	0.56	0.03	0.49	0.63	0.00	0.11	0.0009	97.90	100.00
Treatment effect:	IPE	0.56	0.03	0.49	0.63	0.00	0.01	0.0009	97.90	100.00
0% decrement	IPEexp	0.56	0.03	0.49	0.63	0.00	-0.06	0.0009	98.20	100.00
Scenario number: 15	ITT	0.65	0.03	0.59	0.71	0.01	1.43	0.0011	94.90	100.00
True mean survival:	PPExc	0.62	0.04	0.55	0.69	-0.02	-3.68	0.0019	89.10	100.00
Control: 0.64	PPcens	0.68	0.04	0.60	0.75	0.04	5.51	0.0025	86.20	100.00
Experimental: 0.74	IPCW	0.64	0.04	0.56	0.72	0.00	-0.76	0.0013	97.70	100.00
	IPCWn	0.64	0.04	0.56	0.72	-0.01	-0.83	0.0013	97.80	100.00
Mean switch %: 25.61%	Weib2m	0.64	0.03	0.63	0.66	0.00	0.06	0.0011	34.10	100.00
True Average HR: 0.76	Gam2m	0.64	0.03	0.63	0.66	0.00	0.03	0.0011	34.90	99.70
Mean censored: 15.04%	RPSFTM	0.64	0.04	0.56	0.72	0.00	-0.45	0.0013	96.50	100.00
Treatment effect:	IPE	0.64	0.04	0.56	0.72	0.00	-0.55	0.0013	96.40	100.00
0% decrement	IPEexp	0.64	0.04	0.56	0.72	0.00	-0.58	0.0013	96.40	100.00

5.3 Other “base” scenarios

We have presented detailed results for 8 scenarios which we believe provide a clear illustration of the key findings of our study. It is notable that across all of these scenarios the complex adjustment methods generally performed well, with relative bias usually below 1.0%. Some caution should be taken with this conclusion, since in this study we worked in yearly units – hence in the context of

survival estimates and cost-effectiveness analysis a 1.0% error in the estimate of mean survival could still be important. However, it appears that in scenarios with moderate to low switching proportions the complex adjustment methods perform well – with the possible exception of the IPCW and IPCWn methods in scenarios where the treatment effect is high (in which relative bias increased to 2-4%).

In the other "base" scenarios run (that is, the first 16 scenarios, as described in Appendix 4) 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 – in particular when the "common treatment effect" assumption held and when the switching proportion was low; the Weib2m and Gam2m methods produced generally low levels of bias across all scenarios; IPCW and IPCWn methods generally produced lower levels of bias than the ITT analysis and sometimes produced similar levels of bias to the RPSFTM, IPE and IPEexp methods in scenarios where the "common treatment effect" assumption did not hold, but relative bias fluctuated more between scenarios. Relative bias graphs for the key methods across the first 16 scenarios are presented in Appendix 5.

It is notable that in all scenarios the RPSFTM, IPE and IPEexp methods led to negative bias – that is, they over-adjusted for the treatment switching effect. This is likely to be due to the recensoring involved in the treatment effect estimation procedure. Recensoring 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 over-estimate of the true treatment effect. This appears to have been the case across all scenarios. This pattern was not as clear for the IPCW and IPCWn methods – negative bias was produced in Scenarios 1, 5, 7, 9, 13 and 15, suggesting a tendency to over-estimate the treatment effect in circumstances where the treatment effect was high and the proportion of administrative censoring was relatively low. There was a tendency for the IPCW and IPCWn to produce positive bias when the treatment effect was lower and when censoring proportions were higher. The IPCW and IPCWn were not affected by the "common treatment effect" assumption. In general the Weib2m and Gam2m methods were prone to positive bias (underestimating the treatment effect), although Scenarios 3, 6 and 11 were exceptions to this.

5.4 Impact of sample size

Scenarios 17-32 replicated Scenarios 1-16, but with the sample size simulated in each scenario reduced to 300. We anticipated that this would cause a worsening in the performance of the adjustment methods, particularly for the IPCW and IPCWn due to their observational basis. In fact, there was a marginal increase in bias for all methods, and the increase was not greater for the IPCW approaches when compared to the RPSFTM, IPE and IPEexp. Relative bias across Scenarios 17-32 increased by 0.15 percentage points for the IPCW, 0.21 percentage points for IPCWn, and 0.15, 0.14 and 0.14 percentage points for the RPSFTM, IPE and IPEexp methods respectively. Relative bias associated with the Weib2m increased by 0.11 percentage points, and by 0.03 percentage points for the Gam2m. The reduced sample size had a significant impact upon the convergence of the Gam2m

method. Across Scenarios 1-16 the Gam2m method converged in 98.8% of simulations, but this fell to 88.1% in Scenarios 17-32. Convergence problems were heightened in scenarios that included a high proportion of administrative censoring – in these scenarios the Gam2m method converged in 76-84% of simulations. Hence, the results for the Gam2m method should be treated with caution in these scenarios. The relative bias associated with the key adjustment methods across Scenarios 17-32 is illustrated in Appendix 6.

5.5 Impact of censoring

Scenarios with an odd number incorporated administrative censoring proportions of 13-15%, whereas scenarios with an even number incorporated administrative censoring proportions of 47-56% (for more details, see Appendix 4). The mean relative bias associated with “odd” and “even” scenarios within the first 32 scenarios of our study are presented in Table 5.

Table 5: Comparison of relative bias by censoring proportion, Scenarios 1-32

Method	Relative bias, "odd" scenarios (low censoring)	Relative bias, "even" scenarios (high censoring)	Increase in relative bias in "even" (high censoring) scenarios
ITT	4.37	2.65	-1.71
PPexc	6.05	2.83	-3.22
PPcens	12.24	9.01	-3.23
IPCW	2.06	2.12	0.06
IPCWn	1.94	1.35	-0.59
Weib2m	0.38	0.59	0.21
Gam2m	0.40	0.42	0.02
RPSFTM	0.63	0.94	0.31
IPE	0.81	0.96	0.16
IPEexp	0.90	0.97	0.07

Table 5 demonstrates that for the simple adjustment methods (PPexc, PPcens) and the ITT analysis relative bias substantially fell in scenarios that had higher administrative censoring proportions. This is because in these scenarios the proportion of patients in the control group as a whole who switched treatment was lower (as demonstrated by Table A4, in Appendix 4) than in the corresponding “odd” scenarios. The average proportion of all control group patients that switched in the “even” scenarios was 21.7%, compared to 30.6% in the “odd” scenarios. This makes it difficult to identify the specific effect that increases in the censoring proportion has on the performance of the adjustment methods.

For the IPCW method the proportion of patients who switch of those who become at-risk of switching is important, and this was similar between the “odd” and “even” scenarios. An increase in the administrative censoring proportion may cause problems for IPCW estimation if relatively few events are observed in patients in the control group “at-risk” population who do not switch treatment. For the IPCW method this led to a marginal increase in relative bias in the “even” scenarios, despite the lower overall switching proportion in these scenarios. For the IPCWn method this was not the case – this is

likely to be due to the larger “at-risk” population used by this version of the method (because it uses all control group patients, rather than only those with a “choice” covariate value of 1), an issue that will be discussed further in Section 5.8.1.

Relative bias marginally increased in the “even” scenarios for the Weib2m, Gam2m, RPSFTM, IPE and IPEexp methods, as demonstrated by Table 5. This is despite the lower proportion of switching in the “even” scenarios, and suggests that the reduction in observed events caused by the increase in the administrative censoring proportion more than counteracts the reduction in bias that would be expected due to the lower switching proportion. Hence, it appears that if all else remained equal it would be reasonable to expect an increase in bias associated with the adjustment methods in the presence of increasing levels of administrative censoring.

5.6 Impact of the data generation model

In Scenarios 33-64 we replicated Scenarios 1-32, replacing the 2-component mixture Weibull baseline hazard function data generation model with a 2-component mixture Gompertz baseline hazard function data generation model. We chose parameter values for the Gompertz models such that the simulated scenarios were similar to the mixture-Weibull based scenarios. Although the probabilistic nature of the simulations and the different characteristics of the Weibull and Gompertz models meant that we could not make the corresponding scenarios identical, the scenario details summarised in Table A4, Appendix 4, demonstrate that these scenarios were very similar. It was important to test the sensitivity of our results to different data generation models, in order to investigate whether the results of the adjustment methods were sensitive to the parametric distributions used to generate the data. However, we found that the performance of the adjustment methods in Scenarios 33-64 was very similar to that observed in Scenarios 1-32.

5.7 Additional analyses

5.7.1 Larger sample size

Because the IPCW method is reliant upon observational modelling of the switching mechanism and survival, we anticipated that the method may perform better in larger trials, and may perform more poorly in smaller trials. Section 5.4 demonstrates that the impact of reducing the sample size to 300 had only a marginal impact on the relative bias associated with the IPCW methods, and that in the scenarios investigated the reduction in sample size was no more important for the IPCW method than it was for the RPSFTM, IPE and IPEexp methods. However, to provide further information on this in Scenarios 65—72 we re-ran Scenarios 1, 2, 3, 4, 9, 10, 11 and 12 with a sample size of 1,000 patients, rather than 500 (retaining the 2:1 randomisation in favour of the experimental group).

Increasing the sample size to 1,000 had very little impact upon the bias associated with the adjustment methods. Further detail on this is provided in Table 6, which shows that the relative bias associated with all of the adjustment methods was very similar irrespective of whether the sample

size was 500 or 1,000. The methods that benefited most from the increase in the sample size were the two-stage estimation methods (Weib2m and Gam2m), for which relative bias reduced by 0.06-0.10 percentage points. It is logical that these methods benefit from a larger sample size, as the "observational" part of the approach – that is, the estimation of the treatment effect in control group patients who switch – is based upon an increased number of patients. It is also notable that in the scenarios with n=1,000, the Gam2m method converged in all simulations, which was not the case in all of the n=500 scenarios.

Table 6: Comparison of relative bias by sample size

Method	Relative bias, n=500 Scenarios	Relative bias, n=1000 Scenarios	Reduction in relative bias in n=1000 Scenarios
ITT	4.88	4.87	-0.01
PPEXC	6.38	6.34	-0.04
PPCENS	16.62	16.60	-0.02
IPCW	2.41	2.36	-0.05
IPCWn	1.65	1.66	0.01
Weib2m	0.44	0.34	-0.10
Gam2m	0.40	0.34	-0.06
RPSFTM	1.11	1.13	0.02
IPE	1.28	1.24	-0.04
IPEexp	1.37	1.33	-0.05

5.7.2 Extreme switching proportions

In our previous simulation study we found that observational-based methods such as the IPCW were prone to very substantial levels of bias when the switching proportion was extremely high (approximately 90% of "at-risk" patients).[12] In the current simulation study our results show that the IPCW performs much better – producing much lower bias – at moderate levels of switching (approximately 60% of "at-risk" patients) and lower levels of switching (approximately 25% of "at-risk" patients). In these scenarios all complex adjustment methods perform well, generally producing low levels of bias, with relatively rare exceptions for the IPCW method when the treatment effect was high (equivalent to an average HR of approximately 0.51). The RPSFTM, IPE and IPEexp methods produced higher levels of bias when the "common treatment effect" assumption did not hold, but even in such circumstances the resulting bias was relatively low when the switching proportion was low. In order to provide further information on the impact of the switching proportion in Scenarios 73-80 we re-ran Scenarios 1-8 with extreme switching proportions. In Scenarios 73-76 a very low switching proportion (5%-8% of all control group patients, equivalent to 10-11% of "at-risk" patients) was simulated. In Scenarios 77-80 a very high switching proportion (47%-70% of all control group patients, equivalent to 94-95% of "at-risk" patients) was simulated.

Reducing the switching proportion had a significant impact upon the results of the ITT analysis – as expected the bias associated with the ITT analysis is lower when the switching proportion is lower. In contrast, the relative bias associated with the IPCW and IPCWn methods generally increased. This is

logical, because when very few patients switch the IPCW may struggle to accurately model the probability of switching in "at-risk" patients. The impact of simulating a very low switching proportion is much less clear for the RPSFTM, IPE, IPEexp, Weib2m and Gam2m methods, with relative bias marginally increasing in some scenarios and marginally decreasing in others. However, it is important to note that in these scenarios the ITT analysis is more likely to produce least bias.

The scenarios in which an extremely high switching proportion was simulated resulted in much more noticeable changes in the bias associated with the adjustment methods. The extreme increase in switching proportion increased the bias associated with all adjustment methods, but this increase was much more significant for the IPCW method than all other methods. This reflects the results of our previous simulation study. However, of particular note is that the bias associated with the IPCWn method did not increase substantially, and that in these scenarios the IPCWn produced substantially less bias than the IPCW method. Initially, this may seem counterintuitive, because the IPCWn excludes information on the "choice" covariate, which influences the probability of switching, and thus it would be expected that this version of the IPCW would perform sub-optimally. In fact, across all scenarios we found that the IPCW method produced very similar levels of bias as the IPCWn method, and the IPCW method only produced less bias than the IPCWn in 11 of the 32 base scenarios.

We believe that the reason that the IPCWn method produced similar and often less bias than the IPCW method relates to the data required by the weighting regressions, and also to the characteristics of the "choice" covariate. Firstly, because the "choice" covariate value was randomly assigned to simulated patients, and was not related to prognosis, including data on this within the IPCW weighting regressions may not be expected to have a substantial impact – the impact would be expected to be greater if the "choice" covariate was associated with prognosis in some way.

Secondly, because the "choice" covariate represents a perfect predictor of switching – that is, it is impossible for any patients to switch if they have a "choice" covariate value of '0', it cannot be included as a covariate in the weighting regressions. Instead, we incorporated this information by only applying the time-dependent weighting regression (the denominator of the stabilised weight) to patients who had experienced disease progression and who had a "choice" covariate value of '1'. Because only 80% of patients had a "choice" covariate value of '1', this cut down the size of the sample size informing this regression by 20%. This means that the remaining patients are more likely to be assigned high weights, which leads to the IPCW adjusted treatment effect becoming prone to substantial error. The results of Scenarios 77-80 suggest that this becomes very important at extremely high switching proportions. In Scenarios 77-80 the number of control group patients who were at-risk of switching but did not switch ranged between 4 and 8 for the IPCW method, whereas for the IPCWn method this number ranged between 25 and 40. In our previous study, we concluded that the IPCW method produced significant bias when this number fell below 19.[12] The results of our current study support this finding – the IPCW method can perform relatively well when the number of at-risk patients who do not switch is 25 or above, even when data on a non-prognostic variable that

influences the probability of switching is not incorporated. However, when the number of at-risk patients who do not switch is lower than 10, the method is prone to very high levels of bias.

5.8 Comparisons of variations of methods

Thus far, we have presented results across the range of scenarios that we simulated, but have largely grouped similar methods together in our discussion. However, it is important to consider the relative performance of the similar methods.

5.8.1 IPCW

In Section 5.7.2 we discussed the relative performance of the IPCW and IPCWn methods. However, in our simulation study we also examined the relative performance of these methods according to the convergence of the logistic weighting regressions (for the numerator and denominator of the stabilised weight). Stata provides information on three relevant indicators of the performance of the regression:

a. Whether or not the regression converged
b. The number of completely determined successes
c. The number of completely determined failures

Convergence is clearly an issue, and if any successes or failures are completely determined this is a sign of potential hidden collinearity. Across our 32 base scenarios, convergence of the IPCW method occurred in 59.9% of simulations, and convergence of the IPCWn method occurred in 63.8% of simulations. Convergence combined with zero completely determined successes or failures occurred in just 21.0% of simulations for the IPCW method, and 25.1% of simulations for the IPCWn method. Convergence was lower in simulations in which relatively lower proportions of patients switched treatments, and was particularly low (sometimes as low as 4-7% for convergence combined with zero completely determined successes or failures) in scenarios with lower switching proportions combined with a simulated sample size of 300 patients. The lowest level of convergence (irrespective of whether any successes or failures were completely determined) was 33.2%, in Scenario 21.

In the results previously presented in this Section we have included all simulations for the IPCW method, since Stata provides coefficient estimates even if regressions fail to converge. However, convergence and possible collinearity are clear problems associated with the IPCW method, and therefore in practice the application of the IPCW method will need to be considered carefully on a case-by-case basis, and models may need to be adapted in order to achieve convergence. Given the high proportions of simulations in which convergence was not achieved, comparisons of the results of the IPCW method according to the extent to which convergence was achieved is problematic. However, we found that both IPCW and IPCWn methods only produced marginally lower levels of bias in instances where full convergence was achieved, compared to instances where full convergence was not achieved. Table 7 demonstrates that in instances where the IPCW converged, the mean relative bias across Scenarios 1-32 was 0.24 percentage points lower than when all IPCW

results were analysed irrespective of convergence. When convergence was achieved and zero successes or failures were completely determined, mean relative bias reduced by a further 0.15 percentage points. Similar reductions in relative bias were observed for the IPCWn method.

Stata has strict convergence criteria, and this may explain why the IPCW methods appear to have produced reasonable results even when one or more of the logistic regressions did not converge. In addition we anticipate that the convergence problems may have resulted from the use of splines within the logistic weighting regressions. To create these splines we used the spbase Stata program, as recommended by Fewell *et al,*[43] with 5 knots placed according to percentiles of the survival time distribution. However, we believe that using an alternative spline basis, using the Stata rcsgen program and generating knots based upon the event time distribution, may allow convergence issues to be avoided. As a lack of convergence does not seem to be of key importance in our simulations we do not anticipate that this has had an important impact upon our results, but from a practical perspective models may need to be adapted to achieve convergence.

Table 7: Relative bias by IPCW convergence status – Scenarios 1-32

Method	Relative bias (Scenarios 1-32)	Proportion of simulations (Scenarios 1-32)
IPCW all	2.09	100%
IPCW converged	1.85	59.9%
IPCW converged and zero completely determined	1.70	21.0%
IPCWn all	1.65	100%
IPCWn converged	1.40	63.8%
IPCWn converged and zero completely determined	1.32	25.1%

5.8.2 RPSFTM, IPE and IPEexp

The RPSFTM, IPE and IPEexp methods all use the same underlying counterfactual survival model to estimate a treatment effect adjusted for treatment switching. They only differ in their estimation procedure – with the RPSFTM using non-parametric g-estimation and the IPE and IPEexp using an iterative procedure based upon parametric distributions. In our study the IPE uses a Weibull parametric distribution and the IPEexp uses an exponential distribution. Across our 32 base scenarios we found that the RPSFTM resulted in least bias, with relative bias across all 32 scenarios averaging at 0.79%, compared to 0.89% for the IPE method and 0.94% for the IPEexp. Given the complex hazard functions used to generate our survival data it is not surprising that the non-parametric RPSFTM method produces marginally lower bias than the IPE method, and similarly that the IPE method produces marginally lower bias than the IPEexp method. The similarity in the results of these methods reflects the findings of our previous simulation study.[12]

5.8.3 Two-stage methods

In our previous simulation study we tested a novel two-stage Weibull method. It performed well, and so in this study we examined the two-stage method more closely, considering both a two-stage Weibull method (Weib2m) and a two-stage Generalised Gamma method (Gam2m). Both methods performed well in the current study. The Gam2m method produced marginally lower bias across the 32 base scenarios (relative bias 0.41% compared to 0.48%), but the Gam2m experienced some convergence issues in scenarios, with convergence dropping to 76.4% in Scenario 26. Convergence was lower in scenarios in which the sample size was 300, and where the proportion of administrative censoring was high. However, in general the convergence issues were not very serious, with the Gam2m converging in 98.8% of simulations when the sample size was 500, and 88.1% when the sample size was 300. In practice, the accelerated failure time model used within the two-stage estimation method should be determined through a consideration of which model best fits the data.

5.9 Results – Summary

Our results show that in the scenarios simulated each of the switching adjustment methods generally performed well, although they were sensitive to key scenario parameters. For the RPSFTM, IPE and IPEexp the most influential parameters were the "common treatment effect" and the switching proportion. Across Scenarios 1-32 the mean relative bias for the RPSFTM was 0.46%, for the IPE was 0.55% and for the IPEexp was 0.60% in scenarios where the "common treatment effect" assumption held. This relative bias more than doubled in scenarios where the "common treatment effect" assumption did not hold (in which the treatment effect received by switchers was approximately 20% lower than the treatment effect received by patients initially randomised to the experimental group. In these scenarios the mean relative bias was 1.11%, 1.22% and 1.27% for the RPSFTM, IPE and IPEexp respectively. The switching proportion had a similarly important impact on the performance of these methods. Across the 32 base scenarios relative bias for the RPSFTM, IPE and IPEexp was 1.25%, 1.43% and 1.51% respectively in scenarios where the switching proportion ranged between 30% and 44% of all control group patients. This was reduced significantly in scenarios where the switching proportion ranged between 13% and 18%. In these scenarios the relative bias associated with the RPSFTM, IPE and IPEexp was 0.32%, 0.34% and 0.36% respectively. These findings were supported by the results of our analyses that considered more extreme switching proportions. Our results also suggest that the randomisation-based methods perform better in the presence of lower administrative censoring proportions, in larger sample sizes, and when the treatment effect is relatively low – but the impact of these parameters is relatively minor. These methods are likely to produce very low bias when the "common treatment effect" holds and when switching proportions are low.

For the IPCW and IPCWn the switching proportion and the size of the treatment effect had the most important influences on performance. For the IPCW. mean relative bias increased from 1.52% in scenarios between 1 and 32 that had switching proportions of 13-18% (equivalent to approximately 25% of "at-risk" patients) to 2.67% in those scenarios with switching proportions of 30-44% (equivalent to approximately 60% of "at-risk" patients). For the IPCWn the mean relative bias across

these scenarios increased from 1.48% to 1.81%. In scenarios where the average treatment effect was equivalent to a HR of approximately 0.51 the mean relative bias for the IPCW was 2.49% and for the IPCWn was 2.17%. These fell to 1.70% and 1.13% in scenarios where the average treatment effect was equivalent to a HR of approximately 0.76. Our additional analyses demonstrated that these methods were prone to extremely high bias in scenarios where approximately 95% of "at-risk" patients switched treatments – reflecting the findings of our previous study. Bias also increased at very low levels of switching for these methods, but this increase was only marginal. As expected, the commonality of the treatment effect had little impact on the performance of the IPCW methods. More surprising, increasing the proportion of administrative censoring from around 15% to around 50% also had a relatively minor impact, although as for the randomisation-based methods our results indicated that the IPCW and IPCWn performed better with lower administrative censoring proportions. Similarly, we found that reducing the sample size from 500 to 300 had a relatively minor impact on the IPCW methods – although again, as for the randomisation-based methods, a larger sample size is likely to lead to reduced bias. However, it is notable that the "critical number" of control group patients who do not switch (that is, the number of control group non-switchers that is required in order for the IPCW method to perform acceptably, which appears to be between 10 and 20) will be reached with lower switching proportions in smaller trials. The IPCW and IPCWn methods usually produced higher levels of bias than the randomisation-based methods, although this was often only marginal (for example, across the 16 scenarios within Scenarios 1-32 in which the "common treatment effect" assumption did not hold the mean relative bias associated with the IPCW and IPCWn was 1.99% and 1.55% respectively, compared to 1.11% for the RPSFTM). The bias associated with the IPCW and IPCWn was often in the opposite direction to that produced by the randomisation-based methods, and fluctuated more between scenarios. The randomisation-based methods usually had a greater advantage over the IPCW methods when the "common treatment effect" assumption held.

Importantly, the IPCW method and the IPCWn method performed similarly, except when the switching proportion of "at-risk" patients was extremely high – in these circumstances the IPCWn produced much lower bias than the IPCW. This is because in these scenarios the IPCWn retained substantially more patients in its time-dependent weighting regression, and it appeared that excluding information on patient choice was not important, probably because the covariate was not prognostic for survival.

The two-stage accelerated failure time model methods (Weib2m and Gam2m) were much less sensitive to the scenario parameters than the other adjustment methods. Their relative bias differed only marginally between different sets of scenarios, with the biggest impacts arising from increasing the censoring proportion to approximately 50% from approximately 15% (mean relative bias increased from 0.38% to 0.59% for the Weib2m, and the Gam2m failed to converge in a significant proportion of simulations when the censoring proportion was high). Relative bias increased with these methods when the sample size decreased, and when the switching proportion decreased. These methods often produced least bias compared to all other adjustment methods, even when the "common treatment effect" assumption held. Where this was less likely was where the "common treatment

effect" assumption held and there was a lower switching proportion – in these scenarios it was more common for the randomisation-based methods to produce least bias.

The simple adjustment methods (PPexc and PPcens) produced higher bias than all of the other adjustment methods and the ITT analysis across all scenarios. In general, the more complex adjustment methods produced less bias than the ITT analysis. However, the ITT analysis produced less bias than the IPCW, IPCWn, RPSFTM, IPE and IPEexp in some circumstances in which the administrative censoring proportion was high (causing the adjustment methods to work slightly less well) combined with a relatively low treatment effect (meaning the bias associated with the ITT was relatively low), particularly when the "common treatment effect" assumption did not hold (causing the randomisation-based methods to produce higher bias). This was the case in Scenarios 4 and 20. Similar results – where the ITT analysis produced less bias than some (but not all) of the complex adjustment methods) – were found in Scenarios 8, 12, 16, 24, 28 and 32. In all of these scenarios the administrative censoring proportion was relatively high and the treatment effect was relatively low. The ITT analysis produced very low bias when the switching proportion was extremely low (Scenarios 73-76), and produced less bias than the IPCW, IPCWn, RPSFTM, IPE and IPEexp in two of these scenarios, but in each scenario at least one of the complex adjustment methods produced lower bias.

6. Conclusions, limitations, recommendations and research priorities

In this Section we compare the results of the current study to those found in our previous simulation study and make conclusions and recommendations on the use of switching adjustment methods. We also address the limitations of our study.

6.1 Comparison to previous study

In general, the results of our current study support our previous findings, and offer further insights. However, there are some key differences in the results of the two studies. Firstly, the methods generally produce less bias in the current study than in our previous study. We often found levels of bias of between 5% and 10% in our previous study, or even higher in scenarios with high switching proportions and where the "common treatment effect" assumption did not hold (for the randomisation-based methods). In our present study, levels of relative bias rarely exceeded 2-3% across all scenarios except those that incorporated extreme switching proportions. There are two critical reasons for this difference in results. Firstly, in our previous study half of the 72 scenarios run included extremely high switching proportions equivalent to 90-95% of the control group at-risk of switching. We re-tested these circumstances in our additional analyses in the current study, and again found that bias increased substantially – particularly for the IPCW methods.

Secondly, in our previous study we used a less complex data generating mechanism, compared to the mixture models used in the current study, which allowed hazard function turning points to be simulated. Although we simulated similar average treatment effects in terms of hazard ratios in our

two studies, the survival functions simulated were different and the corresponding average acceleration factors (AF) in each scenario were very different: because hazard ratios work on the hazard scale, and acceleration factors work on the time scale, a given hazard ratio can be compatible with a range of different acceleration factors. For example, in our previous study the average AF across the 72 scenarios varied between 1.44 and 3.58, and was over 2.0 in 60 of the 72 scenarios. In the current study the AF across the 32 base scenarios ranged between 1.22 and 1.78, despite the fact that we simulated data that produced similar average hazard ratios in the two studies. Given that we have found that the performance of each of the switching adjustment methods is affected by the size of the treatment effect – particularly the IPCW methods, which produce more bias when the treatment effect is higher – this is likely to be a key reason behind the lower biases found in the current study. In addition, in the scenarios that violated the "common treatment effect" assumption the decrement in the treatment effect was calculated as a proportion of the average true AF in the experimental group – when this true AF is higher the absolute decrement in the treatment effect applied to switching patients will be higher given the same proportional decrement. This is likely to explain why the randomisation-based methods were more sensitive to departures from the "common treatment effect" assumption in our previous study – it is the absolute difference in the AF between switchers and patients randomised to the experimental group that determines the subsequent bias of these methods. In our new study we found that the randomisation-based methods generally produced low (in the region of 1-2%) bias, and slightly less bias than the IPCW methods even when the treatment effect received by switchers was 20% lower than that received by patients randomised to the experimental group. This is likely to be because the true AF was relatively low compared to our previous study, where we found that decrements in the treatment effect of 20% or more were associated with very significant increases in bias associated with the randomisation-based methods. This in itself is a useful finding of our new study, and illustrates that it is important to assess the size of the treatment effect in terms of an AF, rather than only a HR.

In addition to this, the current simulation study has provided more information on the performance of two-stage AFT methods, as well as the impact of sample size, administrative censoring, and more moderate switching proportions. Hence, we are able to able to add to our previous recommendations drawing upon our new findings.

6.2 Recommendations

Below, we make a series of recommendations on the use of switching adjustment methods, based upon the findings of our two simulation studies. In addition, in Figure 3 we provide a step-by-step analytical guide that – alongside our more detailed recommendations – could be used by analysts on a case-by-case basis to help determine which adjustment methods may be appropriate.

1. If a detailed analysis of the trial data suggests that the treatment effect received by switchers is unlikely to be different to that received by experimental group patients an RPSFTM or IPE

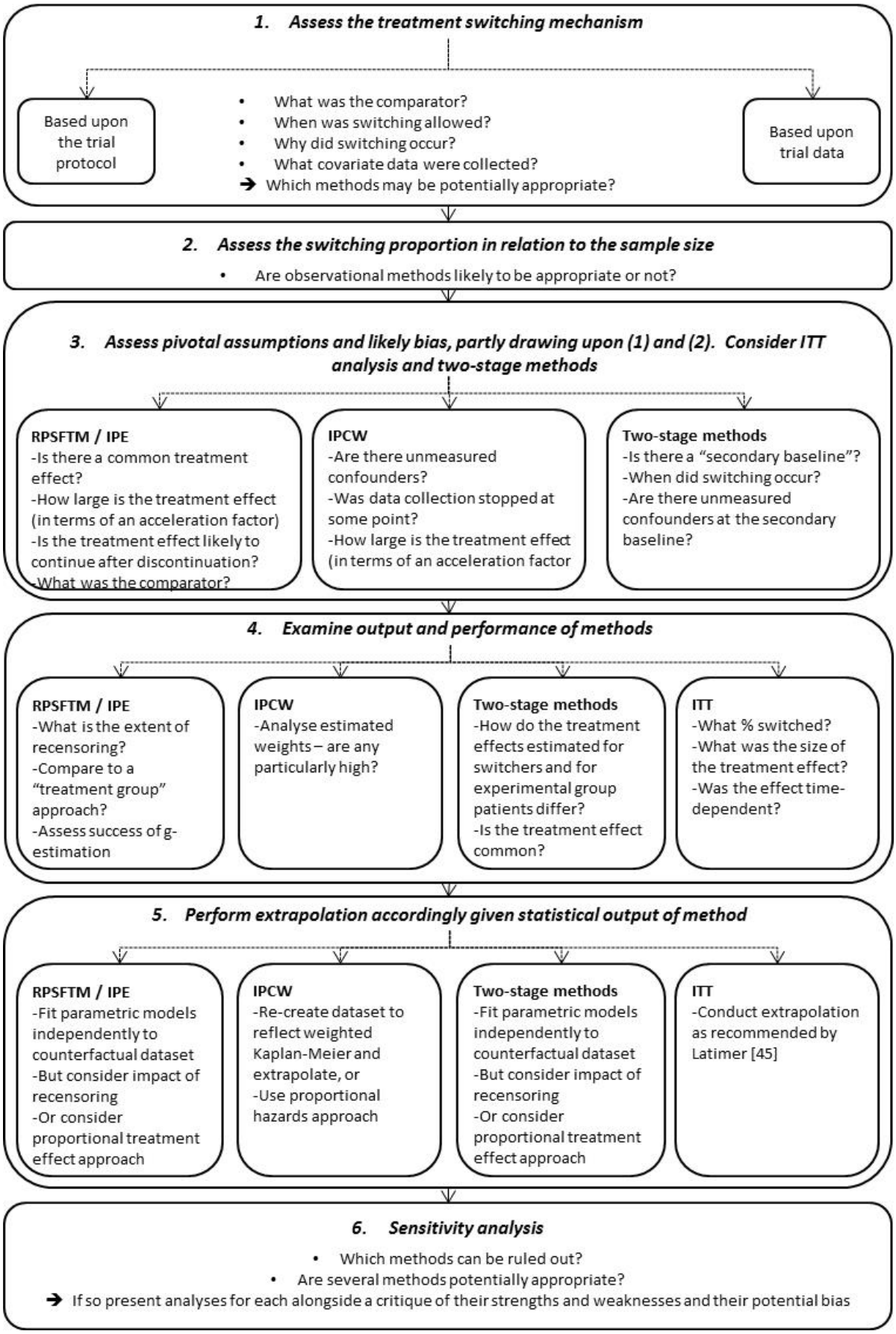
approach should be used to adjust for treatment switching. If it is feasible to apply a two-stage accelerated failure time model approach, this is also likely to produce low bias. An IPCW approach may also produce low bias, but this is less certain.

2. If the proportion of patients that switch is 90% or greater of those that became at risk of switching the IPCW method is highly prone to bias (assuming a trial sample size of approximately 300-500, with 1:1 or 2:1 randomisation in favour of the experimental group). Randomisation-based methods and two-stage accelerated failure time models are relatively less affected by high levels of switching and therefore should be given precedence (unless there is evidence of a strong time-dependent treatment effect).
3. If there is evidence of a time-dependent treatment effect the strength of that effect should be assessed if possible. If it is feasible to apply a two-stage method the suitability of this approach should be considered based upon the treatment switching mechanism. The importance of departures from the "common treatment effect" assumption depends upon the absolute size of the acceleration factor and the absolute difference between the acceleration factor received by switchers compared to patients randomised to the experimental group. If the AF in the experimental group is less than approximately 1.8 randomisation-based methods may produce only relatively minor levels of bias (1-2%) and are likely to produce lower bias than the IPCW method even if the AF decrement in switchers is up to 20%. However, if the AF in the experimental group is higher – for example, 2.0-4.0 – the randomisation-based methods can be expected to produce substantial bias (in the region of 10%) if the AF decrement in switchers is 15% or greater. In this case the IPCW method is likely to be preferable if a two-stage technique cannot be applied – although it may still be prone to bias of around 10%, depending upon the switching proportion and the exact size of the treatment effect.
4. In circumstances where the treatment effect is very small (with an AF of less than approximately 1.4) and where the switching proportion is low, the ITT analysis will produce low levels of bias, but may still produce more bias than the complex adjustment methods, depending upon the other characteristics of the trial. On the other hand, in circumstances where the complex adjustment methods do not work well – for instance, where the switching proportion is extremely high, or where the AF is high (2.0-4.0) in combination with a treatment effect decrement in switchers of 20% or more – the ITT analysis may produce least bias (but will still contain bias).
5. If the decrement in the treatment effect received by switchers is likely to be above 30% RPSFTM and IPE methods are likely to be unsuitable, but this remains dependent upon the size of the AF – if the AF is low (less than 1.8) these methods may still produce relatively low levels of bias.
6. When there is a time-dependent treatment effect and the RPSFTM/IPE methods are applied it is likely that these will over-estimate the treatment effect – this should be taken into account if such methods are to be used to produce 'least bias' estimate of the treatment effect. It is less

certain whether the IPCW method will lead to an over- or under-estimate of the treatment effect.

7. Reducing the sample size of a trial causes all methods to produce marginally more bias, whereas increasing the sample size has the opposite effect. For non-extreme switching proportions all adjustment methods are affected by this to a similar extent. However, it is notable that with smaller sample sizes the critical number of "at-risk" patients who do not switch treatments will be reached at lower switching proportions, and the opposite is true for larger sample sizes.
8. The IPCW method is more adversely affected than the other adjustment methods when very low (less than 10%) proportions of patients switch treatments.
9. Under the guidance above, the switching adjustment methods can still produce low levels of bias when administrative censoring is up to approximately 50%. However, higher censoring levels generally lead to increased bias associated with the adjustment methods, and cause some problems with the convergence of two-stage methods. Censoring is increasingly important in small trials (with sample size of around 300).
10. If extrapolation is required to estimate mean survival the impact of recensoring should be considered when RPSFTM or IPE methods are used. An analysis should be undertaken to identify whether recensoring is likely to lead to inappropriate extrapolations. A survivor function approach whereby the treatment effect is applied to an extrapolation of un-recensored experimental group survival times is likely to be preferable.
11. When preliminary analysis of trial data suggests that the choice of preferable switching adjustment method is unclear, sensitivity analysis should be undertaken to demonstrate the uncertainty (and the value of this uncertainty) associated with the methodology used.

Figure 12: Treatment switching analysis framework



6.3 Limitations

A limitation associated with any simulation study is that not all scenarios that we observe in practice can be investigated. We attempted to include all the most important and most relevant scenarios given results of our previous study, realistic cancer trial characteristics and the characteristics of the methods we were assessing. We also ran additional analyses where we selected specific scenarios to re-run with altered parameter values in order to test specific hypotheses. However, there remain further questions that would be useful to investigate.

For instance, we found that the randomisation-based methods were much more robust to departures from the "common treatment effect" assumption in this study, compared to our previous study. This is highly likely to be due to the difference in the size of the acceleration factors simulated in our current study, compared to in our previous study. It would be of particular value to analyse real-world datasets in order to determine realistic AF sizes, and, if necessary, to re-run our simulations using these in order that we can better understand the potential problems caused by departures from the "common treatment effect" assumption in the real world.

Secondly, between our previous study and our current study we have gathered useful information regarding what number of "at-risk" patients in the control group who do not switch treatments is required in order for the IPCW to produce low levels of bias. It seems that this number is likely to be in the region of 10-20, but it would be very useful to run further scenarios with different sample sizes and switching proportions aimed specifically at providing more information on this. Linked to this, we could further analyse our simulations in order to attempt to ascertain what level IPCW stabilised weights can rise to before substantial bias is observed in the results. From a practical perspective, this would provide useful guidance to analysts.

As it was in our previous study, a technical limitation of our simulation study was that we could not estimate the weighted Kaplan Meier successfully and without bias in our simulations. However, we instead used the IPCW survivor function approach which is likely to closely resemble results that would have been obtained for the WKM.

Another general limitation of simulation studies is that the results are likely to always be linked in some way to the chosen data generating process. However, we have gone some way to demonstrating that this is not the case in this study, since we tested each scenario using two different data-generating models. In addition, our results generally support those found in our previous study, which used a different data generating model.

Finally, our previous study and our current study give us confidence that the two-stage AFT method represents a potentially valuable method for adjusting for treatment switching. However, so far we have only tested this method in scenarios where switching happens soon after the "secondary

baseline" of disease progression. It would be important to identify how sensitive this method is to switching that occurs further from the secondary baseline time-point, and also to test how sensitive this method is to violations of the "no unmeasured confounders" assumption.

Appendix 1: Treatment switching probabilities

Table A1 presents the probability of switching for different patient groups at different time-points in our base scenarios. Higher group numbers represent higher values for that group (that is, "time to progression group" 0 are the control group patients that had time-to-progression times in the lowest 33.3% of the control group). Note however that these groups only refer to patients who became "at-risk" of switching – that is, those control group patients that survived for longer than 21 days. Hence the lowest 33% represent the lowest third of the at-risk group, not the control group as a whole.

Table A1: Probability of treatment switch by prognostic groups and consultation

Consultation 1		Antigen group at progression		
		0	1	2
Time to progression group	0	0.10	0.18	0.28
	1	0.25	0.40	0.54
	2	0.40	0.57	0.70
Consultation 2		Antigen group at progression		
		0	1	2
Time to progression group	0	0.08	0.15	0.24
	1	0.21	0.35	0.48
	2	0.35	0.52	0.65
Consultation 3		Antigen group at progression		
		0	1	2
Time to progression group	0	0.05	0.10	0.16
	1	0.14	0.25	0.37
	2	0.25	0.40	0.54

In the base case scenario the mean switching proportion in the control group across the 1,000 simulations was 43.60%, which was equivalent to 58.26% of control group patients who became at-risk of switching – i.e. those that experienced disease progression and had a "choice" covariate value of '1'. This proportion of switching led to an increase in the average HR based on an ITT analysis from 0.51 to 0.60, reflecting the beneficial effect on survival of switching from the control group onto the experimental treatment.

Appendix 2: Scenario parameter values

In Table A2, values for each variable in Scenario 1 are quoted, as are alternative values for different scenarios.

Table A2: Simulated scenarios – Parameter values and alternatives tested

Variable	Value (Scenario 1)	Alternative Values
Sample size	500 (2:1 randomisation)	300; 1,000 (2:1 randomisation)
Number of prognosis groups (prog)	2	-
Probability of good prognosis	0.5	-
Probability of poor prognosis	0.5	-
Maximum follow-up time	3 years (1095 days)	-
Choice covariate (probability of value of '1')	0.8	-
Multiplication of OS survival time due to bad prognosis group	Log hazard ratio = 0.5	-
Survival time distribution	Weibull parameters: Mix 1: Shape parameter 2.1 Scale parameter 1.8 Mix 2: Shape parameter 0.5 Scale parameter 0.1 p = 0.7 (mix parameter)	Weibull parameters to represent a less severe disease with more censoring: Mix 1: Shape parameter 2.1 Scale parameter 1.5 Mix 2: Shape parameter 0.5 Scale parameter 0.05 p = 0.25 (mix parameter) Gompertz parameters: Mix 1: Shape parameter -1.6 Scale parameter 0.15 Mix 2: Shape parameter 2.2 Scale parameter 0.5 p = 0.3 (mix parameter) Gompertz parameters to represent a less severe disease with more censoring: Mix 1: Shape parameter -1.6 Scale parameter 0.1 Mix 2: Shape parameter 2.2 Scale parameter 0.4 p = 0.75 (mix parameter)
Progression free survival	Overall survival time multiplied by a value from a beta distribution with shape parameters (10,10) – this implies the assumption that time to progression is approximately half of OS. This is not an important assumption – time to progression is only included because we model a situation where switching cannot occur before disease progression	-
Baseline treatment effect (note this is not the true treatment effect as this does not take into account the effect of the treatment that occurs through the time-dependent confounder, antigen level, or the time-dependent part of the treatment effect, η)	Baseline log hazard ratio in scenarios that include an additional time-dependent effect = -0.75	Alter log hazard ratio to -0.35 to represent a smaller treatment effect
Antigen intercept	Calculated using a normal distribution with mean of 20 and standard deviation of 1. Increased by 10 in patients who are in the poor prognosis group.	-
Antigen value progression over time	As demonstrated by formula (8). $\beta_2 = -8$ to represent that the antigen value increases more slowly in the experimental group, and $\beta_1 = 15$ to indicate that the antigen value increases over time	-

Impact of antigen value on overall survival	As demonstrated by formulas (10) and (11). Increased antigen value increases the risk of death. The strength of this relationship depends on the variable α, which equals 0.02 in Scenario 1	-
Impact of antigen value on treatment effect	Because treatment reduces the progression of the antigen value and increased antigen values increase the risk of death, the treatment has an additional effect through the antigen. The strength of this relationship depends on the variable α, which equals 0.02 in Scenario 1	All scenarios include a time-dependent treatment effect in the experimental group. However, in selected scenarios the treatment effect received by switchers equals the average treatment effect in the experimental group, satisfying the "common treatment effect" assumption
Time-dependent portion of treatment effect, η	η =0.3 to generate a reduction in the treatment effect over time	All scenarios include a time-dependent treatment effect in the experimental group. However, in selected scenarios the treatment effect received by switchers equals the average treatment effect in the experimental group, satisfying the "common treatment effect" assumption
Assumed frequency of consultations	One every 3 weeks (21 days)	-
Probability of switching treatment over time	As shown in Table 1. This results in a switching proportion of approximately 44% in Scenario 1	Test a low switching scenario where all probabilities are decreased – to an extent where approximately 20% of control group patients switch. Test a very low switching scenario where all probabilities are decreased – to an extent where approximately 7% of control group patients switch. Test a very high switching scenario where all probabilities are increased – to an extent where approximately 94% of "at-risk" control group patients switch
Prognosis of switching patients	As shown in Table 1. This makes switching more likely in good prognosis patients, via a mechanism that takes into account both time to progression and antigen value at progression	-
Treatment effect in switching patients	Equal to baseline treatment effect multiplied by ω. Set ω such that treatment effect received by switching patients is 80% of the average effect received by experimental group patients in base scenarios.	Alter ω such that the "common treatment effect" assumption holds – the treatment effect received by switching patients equals 100% of the average effect received by experimental group patients.

Appendix 3: Applying the methods

- Intention to Treat (ITT)

The area under the curve associated with an ITT analysis was calculated simply by applying the Stata stci command to the unadjusted dataset which is subject to confounding by treatment switching.

- Per Protocol – exclude switchers (PPexc)

The area under the curve associated with a per-protocol analysis where switching patients are excluded will be calculated by simply excluding all switchers and using the Stata stci command.

- Per Protocol – censor switchers (PPcens)

The area under the curve associated with a per-protocol analysis where switching patients are censored was calculated by censoring all switchers at the time at which they switch treatments, and using the Stata stci command on this adjusted dataset.

- Inverse Probability of Censoring Weights

IPCW was applied in line with the example given by Fewell *et al.* (2004), although we applied the IPCW method rather than a full marginal structural model.[43] In addition, we only applied weights to patients in the control group, as our context is an RCT rather than an observational study.

To apply the IPCW method using stabilised weights first the data is split into time intervals and time-dependent covariate values are recorded for each of these intervals. Data is excluded for switching patients beyond the point of switch, and OS is censored for these patients. IPCWs are then estimated for each patient and for each time interval. The numerator of each stabilised weight is the cumulative probability of remaining uncensored (i.e. not switching) from the beginning of follow-up to the end of the interval given only baseline covariates. This is estimated for all control group patients for all time periods. The denominator of the stabilised weight is the cumulative probability of remaining uncensored (i.e. not switching) to the end of the interval given both baseline and time-dependent covariates. These weights are only different from 1 for time periods during which patients are at risk of switching – that is after disease progression has been observed, and before 3 consultations after disease progression have taken place. The probabilities of remaining uncensored are obtained by fitting pooled logistic models with informative censoring due to treatment switching as the dependent variable. A Cox proportional hazards model can then be run, weighted by the stabilised weights, in order to estimate an adjusted HR that estimates the average treatment effect which theoretically avoids confounding due to switching.

To test the sensitivity of the IPCW method to the no unmeasured confounders assumption we included two versions of the method. In the first we included all baseline and time-dependent covariates and considered that patients were only at risk of switching if their "choice" covariate equalled 1:

- Baseline prognosis group
- Baseline antigen value
- Time-to-disease progression
- Antigen value at disease progression
- Antigen value (which differs at each observation)

In the second we simulated a situation whereby the "choice" covariate was unknown, and hence weights were calculated irrespective of the value of the "choice" covariate.

The IPCW approach allows a weighted Kaplan-Meier (WKM) curve to be obtained, which would provide the optimal area under the curve measure for the method. However, the Stata code provided by Fewell *et al.* does not provide us with this curve, and there are problems with calculating this in the context of a simulation study. For the WKM to be estimated we must calculate the sum of the weights for all patients at-risk, and all patients who experienced the event, for each time point. Because it is possible in our study that control group patients with the longest follow-up times may switch and be censored at an earlier date a new administrative censoring time would need to be generated for each simulation in order to avoid biased overestimates of mean survival being produced. This means that generating the WKM accurately in a simulation study would be a very computationally-intensive process and therefore we took a different approach to calculating mean survival for the control group associated with the IPCW approach. First we fitted a flexible parametric model (FPM) to the experimental group survival data (as an FPM will provide a better fit to the complex hazards simulated than standard parametric models). We then predicted the survivor function and hazard function for the experimental group, and multiplied the experimental group hazard function by the inverse of the IPCW HR to obtain the control group hazard function. From this we calculated the control group survivor function and calculated the area under the survival curve up to 18 months. We termed this the "IPCW survivor function" approach. This should represent a close approximation of the IPCW WKM, were this to be extrapolated to 1095 days. CIs for the mean survival estimate were calculated by applying the 95% CIs of the estimated treatment effect in the "survivor function" process.

- Rank Preserving Structural Failure Time Model

The RPSFTM included in the simulation study was applied using the strbee program developed by White *et al.* (2002).[44] The strbee program incorporates recensoring and allows baseline covariates to be included in the estimation procedure.

The RPSFTM method provides an estimate of the treatment effect adjusted for treatment switching in the form of an acceleration factor. It also provides counterfactual survival times – i.e. survival times that would have been observed if nobody had received treatment. In our previous study, we tested three approaches to calculating the control group area under the curve:[12]

- Shrinkage approach. The inverse of the AF is used to shrink survival times in patients who switched and the stci Stata command is used to estimate the area under the adjusted survival curve. This approach does not involve full recensoring as although the AF is estimated using recensoring and survival times of switching patients are recensored, survival times of all other control group patients are not. This creates the potential for bias.
- Extrapolation approach. Under this approach the recensored counterfactual survival times produced by the strbee command are extrapolated out to 1095 days and the area under the extrapolated survival curve is estimated.
- Survivor function approach. This approach is similar to the "survivor function" approach described above for the "IPCW survivor function" method, except the control group survival curve is derived in a slightly different way because the RPSFTM provides an acceleration factor rather than a hazard ratio. An FPM is fitted to the experimental group data, and the survivor function derived. The time associated with each survivor function probability is divided by the RPSFTM AF in order to obtain the survival times associated with the survival probabilities for the control group, and the area under the resulting curve is estimated up to 18 months.

Our previous simulation study demonstrated that the "extrapolation" and "survivor function" approaches produced similar results.[12] We concluded that the "shrinkage" approach should not be relied upon due to its inherent bias.[12] Because the "survivor function" approach is more consistent with the approach used for the IPCW method, and because it is less prone to potential bias associated with extrapolating from recensored counterfactual survival times, we only included the "survivor function" in the current study. CIs for the area under the curve estimate were calculated by applying the 95% CIs of the estimated treatment effect in the "survivor function" process.

- Iterative Parameter Estimation Algorithm

The IPE algorithm approach can also be applied using the strbee Stata program. We applied the method using full recensoring (rather than the partial recensoring initially recommended by Branson and Whitehead (2002)[35]), and included baseline covariates. We applied the method using both a Weibull distribution and an exponential distribution in order to examine the sensitivity of the method to the parametric form chosen in the treatment effect estimation process.

In addition to an AF adjusted for treatment switching, the IPE method provides us with the parameter values of the final parametric model used to estimate the adjusted treatment effect and these could

have been used to estimate mean survival at 18 months. We tested this "extrapolation" approach alongside "survivor function" and "shrinkage" approaches in our previous study, and found that the "extrapolation" and "survivor function" approaches produced similar results. To aid consistency with the IPCW and RPSFTM analyses included in this study, we used a "survivor function" approach, as described for the RPSFTM method above. As for the other adjustment methods, CIs for the area under the curve estimate were calculated by applying the 95% CIs of the estimated treatment effect in the area under the curve estimation process. As noted by Morden *et al.* (2011) this is likely to provide relatively poor coverage as the confidence intervals associated with the treatment effect from the final IPE iteration are underestimates.[5]

- Two-stage accelerated failure time model

In our previous simulation study we tested a novel "two-stage Weibull" method.[12] The method performed well, generally producing low levels of bias and often producing less bias than any other adjustment method. Hence, we investigated this method further in the current study. Rather than relying only upon a Weibull model, we generalised the method to a two-stage AFT method – because it is applicable using any AFT model. We chose to apply the method using a Weibull model, as before, and also using a Generalised Gamma model. The Generalised Gamma distribution is a more flexible distribution than the Weibull distribution, with an additional parameter included in the model. Hence it may be hypothesised that this model could produce more accurate results than the Weibull. In reality, this will depend upon the fit of each model to the observed data, and including both methods will demonstrate how different results may be if different AFT models are chosen.

In addition, in the current study we implement the two-stage AFT method incorporating full recensoring – an approach that was not taken in our initial study, but which avoids a potential bias associated with estimating counterfactual survival and censoring times, as discussed by White *et al.* (1999).[34]

The following covariates were included in the AFT models used to estimate the treatment effect associated with switching patients:

- Baseline prognosis group
- Baseline antigen value
- Time-to-disease progression
- Antigen value at time of disease progression
- Choice covariate

These are the "baseline" covariates in the secondary dataset that only covers the post-progression period for the control group. The resulting treatment effect was then used to shrink survival times in switching patients using formula (7) presented in Section 3.2.2.3. The area under the curve of the

adjusted dataset produced by this method was then calculated using the Stata stci command, and confidence intervals were calculated using the confidence intervals of the treatment effect.

Appendix 4: Overview of simulation scenarios

Table A4 presents key details associated with each of the scenarios simulated. Scenarios 1-32 are the base scenarios using a 2-component mixture Weibull baseline hazard function. Scenarios 33-64 replicate these scenarios using a 2-component mixture Gompertz baseline hazard function. Scenarios 65-72 are additional scenarios investigating the impact of a larger sample size, and scenarios 73-80 are additional scenarios investigating the impact of extreme switching proportions.

The true area under the curve (restricted mean survival) unconfounded by treatment switching is presented, along with the average treatment effect in terms of a hazard ratio (calculated using a Cox model) and an acceleration factor (calculated using a Weibull model). These reflect the treatment effect calculated before switching is applied averaged across each of the 1000 simulations run for each scenario. This represents only an approximation of the true treatment effect as the proportional hazards assumption does not hold. In terms of a hazard ratio, the treatment effect varied between 0.51 and 0.77.

The proportion of control group patients that switch, averaged across the 1000 simulations that made up each scenario, is also presented. The switching proportion varied between 5% and 70% of all control group patients. Scenarios 5-8, 13-16, 21-24, 29-32 and corresponding Gompertz-based scenarios (37-40, 45-48, 53-56 and 61-64) were designed to result in moderately low levels of switching, although these levels are probabilistic and are reliant on other characteristics. Scenarios 73-76 investigated very low switching proportions, and Scenarios 77-80 investigated very high switching proportions. Table A4 also presents the switching proportion as a percentage of the control group patients that became “at-risk” of switching. In our simulations control group patients could only switch treatments if they were alive at their first ‘consultation’ at 21 days, if their disease progressed before the end of the simulated follow-up, and if they had a “choice” covariate value of ‘1’. The switching proportion as a percentage of patients that became at-risk of switching is higher than when it is measured as a percentage of all control group patients – it ranged from 10% to 95%. This is particularly important to consider for observational-based approaches such as IPCW as these methods are reliant upon differentiating between the patient characteristics of switchers and non-switchers and applying inverse probability weightings based upon these characteristics. This can only be achieved by comparing the patients who were at risk of switching treatments and this will become increasingly difficult at the extremes – either when almost all patients switch, or when very few patients switch. The IPCW formulates a ‘pseudo population’ whose survival times are based upon on those of uncensored patients, and thus if there are very few of these patients high weightings will be applied which could lead to bias. We estimated the proportion of patients who become at risk of switching in each scenario by collecting data on the number of patients for whom disease progression

was observed in each simulation. We then calculated the mean for this value across the scenario, and multiplied this by 0.8, representing the proportion of patients who had a “choice” covariate value of ‘1’. This is therefore approximate, but appropriately indicative for our purposes.

Table A4 also presents details on whether the treatment effect was assumed to be “common” – that is, whether the treatment effect received by switchers was the same as the average treatment effect received by patients initially randomised to the experimental group. In scenarios 9-16, 25-32, 41-48, 57-64 and 69-72 the “common treatment effect” assumption held. To provide further information on the strength of the time-dependent effect in each scenario we also include details on the treatment effect size received by switchers.

Table A4 also presents details on the mean proportion of patients that were censored in each scenario – that is, the proportion for whom death was not observed. This varied between 13% and 56%.

Table A4: Overview of Scenarios

Scenario	Truth (years)		Average treatment effects		Mean switcher % of total	Mean switcher % of at risk	Mean censoring proportion (%)	Sample size	Data generating model	Common treatment effect?	Treatment effect in switchers (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
1	0.56	0.79	0.51	1.54	43.60%	58.26%	13.59%	500	Weibull	No	1.43	80%
2	0.99	1.20	0.52	1.78	30.03%	60.33%	55.80%	500	Weibull	No	1.63	80%
3	0.64	0.74	0.76	1.22	43.09%	61.20%	15.03%	500	Weibull	No	1.17	80%
4	0.99	1.08	0.77	1.25	30.52%	61.13%	46.60%	500	Weibull	No	1.20	80%
5	0.56	0.79	0.51	1.54	17.77%	23.78%	13.42%	500	Weibull	No	1.43	80%
6	0.99	1.20	0.52	1.78	12.88%	25.86%	55.21%	500	Weibull	No	1.63	80%
7	0.64	0.74	0.76	1.22	18.18%	25.87%	15.15%	500	Weibull	No	1.17	80%
8	0.99	1.08	0.77	1.25	13.24%	26.55%	46.58%	500	Weibull	No	1.20	80%
9	0.56	0.79	0.51	1.54	43.63%	58.27%	13.74%	500	Weibull	Yes	1.54	100%
10	0.99	1.20	0.52	1.78	30.04%	60.53%	56.35%	500	Weibull	Yes	1.78	100%
11	0.64	0.74	0.76	1.22	42.86%	60.90%	15.09%	500	Weibull	Yes	1.22	100%
12	0.99	1.08	0.77	1.25	30.66%	61.63%	46.82%	500	Weibull	Yes	1.25	100%
13	0.56	0.79	0.51	1.54	17.78%	23.76%	13.52%	500	Weibull	Yes	1.54	100%
14	0.99	1.20	0.52	1.78	12.86%	25.86%	55.53%	500	Weibull	Yes	1.78	100%
15	0.64	0.74	0.76	1.22	18.00%	25.61%	15.04%	500	Weibull	Yes	1.22	100%
16	0.99	1.08	0.77	1.25	13.14%	26.44%	46.70%	500	Weibull	Yes	1.25	100%
17	0.56	0.79	0.51	1.54	43.65%	58.34%	13.48%	300	Weibull	No	1.43	80%
18	0.99	1.20	0.52	1.78	29.90%	60.27%	55.81%	300	Weibull	No	1.63	80%
19	0.64	0.74	0.76	1.22	43.08%	61.18%	15.00%	300	Weibull	No	1.17	80%
20	0.99	1.08	0.77	1.25	30.35%	60.91%	46.69%	300	Weibull	No	1.20	80%
21	0.56	0.79	0.51	1.54	17.82%	23.81%	13.45%	300	Weibull	No	1.43	80%
22	0.99	1.20	0.52	1.78	12.95%	26.03%	55.34%	300	Weibull	No	1.63	80%
23	0.64	0.74	0.76	1.22	18.22%	25.90%	15.05%	300	Weibull	No	1.17	80%
24	0.99	1.08	0.77	1.25	13.07%	26.43%	46.74%	300	Weibull	No	1.20	80%
25	0.56	0.79	0.51	1.54	43.58%	58.25%	13.64%	300	Weibull	Yes	1.54	100%

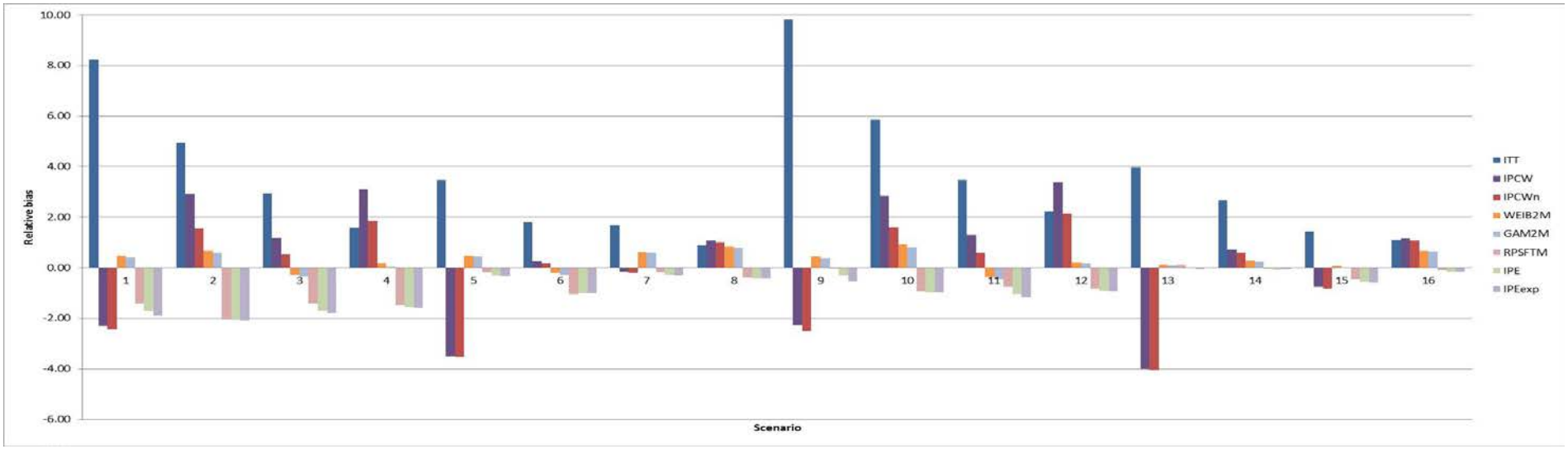
Scenario	Truth (years)		Average treatment effects		Mean switcher % of total	Mean switcher % of at risk	Mean censoring proportion (%)	Sample size	Data generating model	Common treatment effect?	Treatment effect in switchers (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
26	0.99	1.20	0.52	1.78	30.34%	60.85%	56.14%	300	Weibull	Yes	1.78	100%
27	0.64	0.74	0.76	1.22	42.97%	61.12%	15.11%	300	Weibull	Yes	1.22	100%
28	0.99	1.08	0.77	1.25	30.54%	61.54%	46.85%	300	Weibull	Yes	1.25	100%
29	0.56	0.79	0.51	1.54	17.97%	23.96%	13.46%	300	Weibull	Yes	1.54	100%
30	0.99	1.20	0.52	1.78	13.02%	26.12%	55.58%	300	Weibull	Yes	1.78	100%
31	0.64	0.74	0.76	1.22	18.07%	25.71%	15.17%	300	Weibull	Yes	1.22	100%
32	0.99	1.08	0.77	1.25	13.05%	26.36%	46.86%	300	Weibull	Yes	1.25	100%
33	0.54	0.78	0.51	1.60	40.89%	55.89%	13.82%	500	Gompertz	No	1.48	80%
34	0.99	1.19	0.52	1.77	33.77%	54.62%	55.23%	500	Gompertz	No	1.62	80%
35	0.63	0.74	0.76	1.24	42.71%	59.91%	15.66%	500	Gompertz	No	1.19	80%
36	0.99	1.08	0.77	1.25	36.23%	58.54%	46.17%	500	Gompertz	No	1.20	80%
37	0.54	0.78	0.51	1.60	16.78%	22.91%	13.59%	500	Gompertz	No	1.48	80%
38	0.99	1.19	0.52	1.77	13.64%	22.04%	54.59%	500	Gompertz	No	1.62	80%
39	0.63	0.74	0.76	1.24	18.00%	25.26%	15.77%	500	Gompertz	No	1.19	80%
40	0.99	1.08	0.77	1.25	15.28%	24.70%	46.11%	500	Gompertz	No	1.20	80%
41	0.54	0.78	0.51	1.60	40.82%	55.77%	13.89%	500	Gompertz	Yes	1.60	100%
42	0.99	1.19	0.52	1.77	33.77%	54.59%	55.61%	500	Gompertz	Yes	1.77	100%
43	0.63	0.74	0.76	1.24	42.52%	59.78%	15.78%	500	Gompertz	Yes	1.24	100%
44	0.99	1.08	0.77	1.25	36.24%	58.57%	46.26%	500	Gompertz	Yes	1.25	100%
45	0.54	0.78	0.51	1.60	16.64%	22.74%	13.65%	500	Gompertz	Yes	1.60	100%
46	0.99	1.19	0.52	1.77	13.59%	21.94%	54.84%	500	Gompertz	Yes	1.77	100%
47	0.63	0.74	0.76	1.24	17.98%	25.21%	15.64%	500	Gompertz	Yes	1.24	100%
48	0.99	1.08	0.77	1.25	15.24%	24.66%	46.13%	500	Gompertz	Yes	1.25	100%
49	0.54	0.78	0.51	1.60	40.30%	55.16%	13.67%	300	Gompertz	No	1.48	80%
50	0.99	1.19	0.52	1.77	33.99%	54.89%	55.21%	300	Gompertz	No	1.62	80%
51	0.63	0.74	0.76	1.24	42.60%	59.79%	15.85%	300	Gompertz	No	1.19	80%

Scenario	Truth (years)		Average treatment effects		Mean switcher % of total	Mean switcher % of at risk	Mean censoring proportion (%)	Sample size	Data generating model	Common treatment effect?	Treatment effect in switchers (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
52	0.99	1.08	0.77	1.25	36.45%	58.84%	46.15%	300	Gompertz	No	1.20	80%
53	0.54	0.78	0.51	1.60	16.42%	22.46%	13.64%	300	Gompertz	No	1.48	80%
54	0.99	1.19	0.52	1.77	13.71%	22.13%	54.58%	300	Gompertz	No	1.62	80%
55	0.63	0.74	0.76	1.24	18.01%	25.27%	15.54%	300	Gompertz	No	1.19	80%
56	0.99	1.08	0.77	1.25	15.24%	24.65%	46.10%	300	Gompertz	No	1.20	80%
57	0.54	0.78	0.51	1.60	40.79%	55.73%	13.94%	300	Gompertz	Yes	1.60	100%
58	0.99	1.19	0.52	1.77	33.86%	54.61%	55.51%	300	Gompertz	Yes	1.77	100%
59	0.63	0.74	0.76	1.24	42.54%	59.71%	15.70%	300	Gompertz	Yes	1.24	100%
60	0.99	1.08	0.77	1.25	36.23%	58.52%	46.31%	300	Gompertz	Yes	1.25	100%
61	0.54	0.78	0.51	1.60	16.63%	22.70%	13.59%	300	Gompertz	Yes	1.60	100%
62	0.99	1.19	0.52	1.77	13.72%	22.19%	54.81%	300	Gompertz	Yes	1.77	100%
63	0.63	0.74	0.76	1.24	17.81%	25.05%	15.65%	300	Gompertz	Yes	1.24	100%
64	0.99	1.08	0.77	1.25	15.27%	24.68%	46.06%	300	Gompertz	Yes	1.25	100%
65	0.56	0.79	0.51	1.54	43.44%	58.07%	13.57%	1000	Weibull	No	1.43	80%
66	0.99	1.20	0.52	1.78	30.04%	60.39%	55.90%	1000	Weibull	No	1.63	80%
67	0.64	0.74	0.76	1.22	43.01%	61.13%	15.05%	1000	Weibull	No	1.17	80%
68	0.99	1.08	0.77	1.25	30.35%	61.10%	46.78%	1000	Weibull	No	1.20	80%
69	0.56	0.79	0.51	1.54	43.57%	58.24%	13.77%	1000	Weibull	Yes	1.54	100%
70	0.99	1.20	0.52	1.78	30.16%	60.53%	56.21%	1000	Weibull	Yes	1.78	100%
71	0.64	0.74	0.76	1.22	42.87%	60.90%	15.14%	1000	Weibull	Yes	1.22	100%
72	0.99	1.08	0.77	1.25	30.56%	61.27%	46.79%	1000	Weibull	Yes	1.25	100%
73	0.56	0.79	0.51	1.54	7.34%	9.81%	13.27%	500	Weibull	No	1.43	80%
74	0.99	1.20	0.52	1.78	5.41%	10.85%	54.85%	500	Weibull	No	1.63	80%
75	0.64	0.74	0.76	1.22	7.68%	10.92%	14.95%	500	Weibull	No	1.17	80%
76	0.99	1.08	0.77	1.25	5.58%	11.18%	46.30%	500	Weibull	No	1.20	80%
77	0.56	0.79	0.51	1.54	70.20%	93.80%	13.67%	500	Weibull	No	1.43	80%

Scenario	Truth (years)		Average treatment effects		Mean switcher % of total	Mean switcher % of at risk	Mean censoring proportion (%)	Sample size	Data generating model	Common treatment effect?	Treatment effect in switchers (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
78	0.99	1.20	0.52	1.78	47.22%	94.45%	56.06%	500	Weibull	No	1.63	80%
79	0.64	0.74	0.76	1.22	66.39%	94.32%	15.15%	500	Weibull	No	1.17	80%
80	0.99	1.08	0.77	1.25	47.11%	94.91%	46.73%	500	Weibull	No	1.20	80%

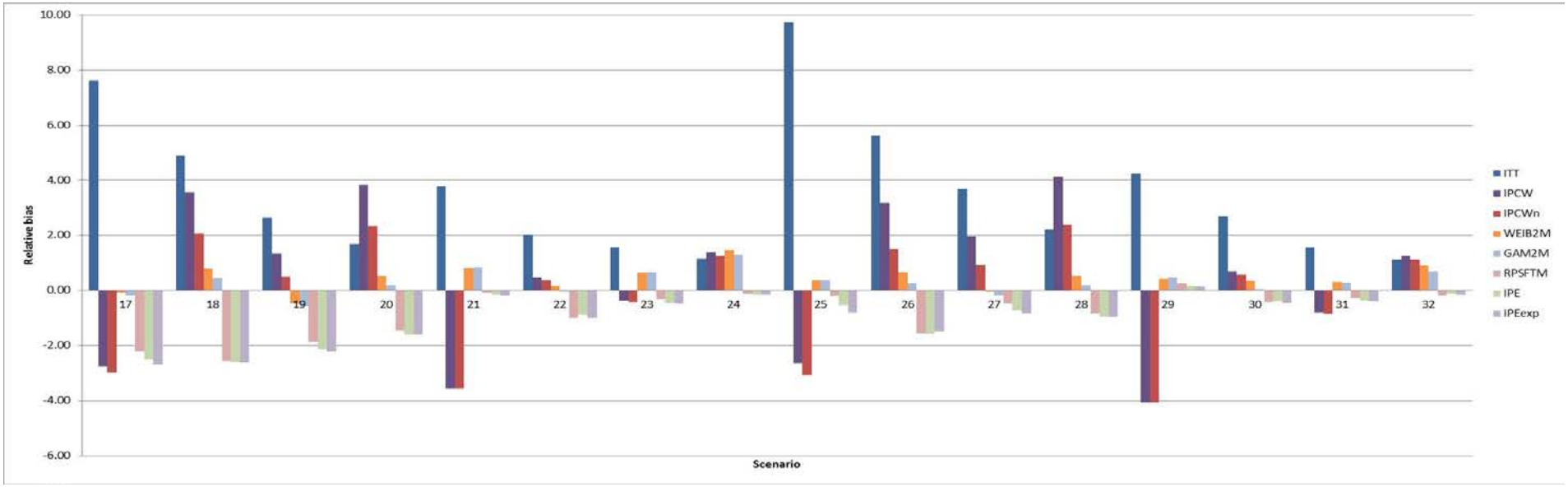
Appendix 5: Relative Bias Scenarios 1-16

Figure A5: Relative bias Scenarios 1-16



Appendix 6: Relative Bias Scenarios 17-32

Figure A6: Relative bias in Scenarios 17-32



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, Sculpher M. Decision modelling for health economic evaluation. Oxford University Press Inc., New York, 2006
3. Gold M.R, Siegel J.E, Russell L.B, Weinstein M.C. Cost-effectiveness in health and medicine. Oxford University Press, Inc., New York, 1996
4. Canadian Agency for Drugs and Technologies in Health, Guidelines for the economic evaluation of health technologies: Canada, 3rd Edition, 2006
5. Morden JP, Lambert PC, Latimer NR, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. BMC Med Res Methodol. 2011; 11.
6. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – an economic evaluation context: Methods, limitations and recommendations. Medical Decision Making, Published online before print: January 21, 2014, doi:10.1177/0272989X13520192
7. Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. European Journal of Cancer 2006;42(17):2867-75
8. Watkins C, Huang X, Latimer N, Tang Y, Wright EJ. Adjusting overall survival for treatment switches: commonly used methods and practical application. Pharmaceutical Statistics. 2013; 12; 6.
9. U.S.Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Clinical trial endpoints for the approval of cancer drugs and biologics. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research, editors. 2007.
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. 201. European Medicines Agency.

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

12. Latimer N, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – a simulation study. Health Economics and Decision Science Discussion Paper DP 13/06, March 2013, University of Sheffield

13. 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.

14. Roche Products Ltd. Achieving clinical excellence in the treatment of relapsed non-small cell lung cancer, Tarceva (erlotinib) NICE STA Submission. 2006.

15. Bond M, Hoyle M, Moxham T, Napier M, Anderson R. The clinical and cost-effectiveness of sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer. 2009. 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.

16. Lewis R, Bagnall AM, Forbes C, Shirran E, Duffy S, Kleijnen J et al. A rapid and systematic review of the clinical effectiveness and cost effectiveness of trastuzumab for breast cancer. 2001. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.

17. National Institute for Health and Clinical Excellence. Final Appraisal Determination: Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours, TA86. 2004. London, NICE.

18. Jannsen-Cilag Ltd. STA submission to NICE: Velcade (Bortezomib) for the treatment of multiple myeloma patients at first relapse. 2006.

19. Hoyle M, Rogers G, Garside R, Moxham T, Stein K. The clinical and cost effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene. 2008. 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.

20. White IR. Uses and limitations of randomization-based efficacy estimators. Stat Methods Med Res. 2005;14(4):327-47.

21. Lee Y, Ellenberg J, Hirtz D, Nelson K. Analysis of clinical trials by treatment actually received: is it really an option? Stat Med. 1991;10:1595–1605.

22. Horwitz R, Horwitz S. Adherence to treatment and health outcomes. Arch Intern Med. 1993;153:1863–1868.

23. Cox DR. Regression Models and Life Tables (with Discussion). Journal of the Royal Statistical Society, Series B 1972; 34:187-220

24. White IR, Walker S, Babiker AG, Darbyshire JH. Impact of treatment changes on the interpretation of the Concorde trial. Aids 1997,11(8):999-1006

25. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. J Am Statist Assoc. 2001; 96(454):440-448.

26. Robins JM, 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(3):779-788.

27. Robins JM, 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. Journal of the American Statistical Association 1994; 89(427):737-749.

28. Yamaguchi T, 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. Statistics in Medicine 2004; 23(13):1991-2003.

29. Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias. American Journal of Epidemiology 2011; 173(5):569-577.

30. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag; 1999. 95-134.

31. Robins JM. Structural Nested Failure Time Models. Andersen PK, Keiding N, editors. Survival Analysis. 4372-4389. 1998. Chichester, UK, John Wiley and Sons. The Encyclopedia of Biostatistics. Armitage, P. and Colton, T.

32. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. The Stata Journal 2[2], 164-182. 2002.

33. Mark SD, Robins JM. A Method for the Analysis of Randomized Trials with Compliance Information - An Application to the Multiple Risk Factor Intervention Trial. Controlled Clinical Trials 1993; 14(2):79-97.

34. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. Statistics in Medicine 1999; 18(19):2617-2634.

35. Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. Stat Med. 2002; 21(17):2449-2463.

36. Law MG, Kaldor JM. Survival analyses of randomized clinical trials adjusted for patients who switch treatments. Statistics in Medicine 1996;15(19):2069-76.

37. Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. Statistics in Medicine 2001;20(24):3753-67.

38. Walker AS, White IR, Babiker AG. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. Statistics in Medicine 2004;23(4):571-90.

39. Stata statistical software intercooled, Version 11.2, Texas, USA: 2011.

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

41. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; 24:1713-1723.

42. Collett D. Modelling Survival Data in Medical Research, 2nd ed. Boca Raton: Chapman & Hall/CRC CRC Press LLC; 2003.

43. Fewell Z, Hernan MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. The Stata Journal 4[4], 402-420. 2004.

44. White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. The Stata Journal 2[2], 140-150. 2002.

45. Latimer NR. Survival analysis for economic evaluations alongside clinical trials – extrapolation with patient-level data: Inconsistencies, limitations and a practical guide. Med Decis Making. published online 22 January 2013.