The University Of Sheffield.

School Of Health And Related Research.

Health Economics and Decision Science (HEDS)

Discussion Paper

Adjusting survival time estimates to account for treatment switching in randomised controlled trials – a simulation study

Nicholas R Latimer , Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ , Morden JP , Akehurst RL , Campbell MJ

DP 13/06

March 2013

HEDS Discussion Paper

No. 13.06

Adjusting survival time estimates to account for treatment switching in randomised controlled trials – a simulation study

Nicholas R Latimer, 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

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

[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.

Nicholas Latimer was funded by a National Institute for Health Research Doctoral Research Fellowship (DRF 2009-02-82). This article/paper/report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Michael Crowther is funded by a NIHR Doctoral Research Fellowship (DRF-2012-05-409).

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

The Pharmaceutical Oncology Initiative, a group of pharmaceutical companies who are part of the Association of the British Pharmaceutical Industry (ABPI) provided financial support for the simulation study referred to in this manuscript.

Financial support for this study was provided in part by grants from the NIHR 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.

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

Abstract

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. It is unclear which methods to adjust for switching are most appropriate in realistic scenarios.

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 range of realistic scenarios.

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 Simple methods such as censoring or excluding patients that switch always resulted in high levels of bias. More complex randomisation-based methods (e.g. Rank Preserving Structural Failure Time Models (RPSFTM)) were unbiased only when the treatment effect was not time-dependent. Observational-based methods (e.g. inverse probability of censoring weights (IPCW)) coped better with time-dependent treatment effects but are heavily data reliant, are sensitive to model misspecification and often produced high levels of bias in our simulations, particularly when the proportion of patients that switched treatments was very high (approximately 90%). We introduce a novel "two stage" method, whereby a specific disease-related time-point is used to define a secondary baseline after which switching is permitted, allowing treatment effects to be estimated separately for patients that switch and patients randomised to the experimental group. We find that this method can perform well provided the treatment switching mechanism is amenable.

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. This involves an economic evaluation and for interventions that impact upon survival a life-time horizon is typically modelled.[1,2,3,4] Typically, the evidence to support the effectiveness of the drug comes from randomised controlled trials (RCT). However, treatment switching – where patients allocated to the control group of an RCT are permitted to cross over onto the experimental treatment at some point during follow-up – is common in trials of oncology treatments, and causes problems for HTAs.[5,6] 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. Switching the other way, i.e. from active to no treatment does not cause a problem, since in this case the ITT represents what is likely to happen in practice.

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, 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.[7,8] Hence, there is less motivation for pharmaceutical companies to ensure that randomised groups are maintained beyond disease progression.

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),[9] perform very well when their key methodological assumptions are satisfied.[5] However, a full comparison of

all relevant methods across a range of realistic scenarios – including scenarios where key methodological assumptions are *not* satisfied – has not previously been undertaken. The aim of this paper is to evaluate the performance (in terms of bias, coverage and mean squared error) of alternative switching adjustment methods in a range of realistic scenarios, related to their ability to adjust for treatment switching in order to estimate the "true" mean survival time in control groups affected by treatment switching. We introduce a novel "two-stage" method for adjusting for treatment switching, and run a broad range of novel scenarios when testing the various methods.

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

2. Switching adjustment methods

The different switching adjustment methods considered in this paper can be grouped into simple methods (those which are currently widely used),[10] and more complex methods. Further, the more complex methods can be classified as 'observational-based' methods and 'randomisation-based' methods. Here, relevant methods are briefly described. More details on the complex switching adjustment methods are provided in Appendix A.

2.1 Simple methods

2.1.1 Intention to treat

An ITT analysis does not attempt to adjust for treatment switching, but represents the standard analysis undertaken on an RCT. Groups are compared as 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.[11]

2.1.2 Per protocol – excluding and censoring switchers

Where treatment received in an RCT differs from what was planned, a common approach to analysing the resulting data is to conduct a per protocol (PP) analysis. In the case of treatment switching, data from patients 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.[12,13]

2.1.3 Treatment as a time-dependent covariate

Under this approach data are analysed according to treatment received, using a Cox proportional hazards model in which a binary time-dependent covariate indicates time-periods in which treatment was received.[14] Again, this approach may break the randomisation balance and is therefore prone to selection bias.[15]

2.2 Complex methods

2.2.1 Observational-based complex methods

2.2.1.1 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.[16] 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, and it is assumed that exposure to treatment accelerates the time to event (such as death) by a factor $\exp(-\psi)$. It is assumed that exposure to treatment is independent of counterfactual survival times, conditional on a "no unmeasured confounders" assumption. This requires that all variables that contribute to the process that determines whether a patient switches treatment are measured.[17]. A 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 key limitation of the SNM method is the “no unmeasured confounders” assumption. This assumption cannot be tested using the observed data.[18,19] In an RCT context, this assumption becomes more problematic than in the observational setting that the method was developed for, because RCT datasets are typically much smaller. When fewer data are available SNMs may become less stable and confidence intervals may become wide. For an SNM to be applied, a substantial amount of data need to be collected in the clinical trial, and also it must be possible to define for each patient when they became “at risk” of treatment crossover so that the SNM is only applied to the relevant “observational” dataset.

2.2.1.2 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. The method is reliant on the “no unmeasured confounders” assumption – if there are data on all time-dependent prognostic factors for mortality that independently predict informative censoring (switching), then the dependence between the censoring and failure can be corrected for by replacing the Kaplan-Meier estimator, log-rank test, and Cox partial likelihood estimator of the hazard ratio by their IPCW versions.[20]

The IPCW method has a number of limitations in common with the SNM approach. In particular, the untestable “no unmeasured confounders” assumption is critical, and therefore data requirements may restrict the practicality of the method. Also, as for the SNM method, models must be correctly specified.[20] In addition, the IPCW approach fails if there are any covariates which ensure (that is, the probability equals 1) that treatment switching will or will not occur.[19,21,22]

2.2.2. Randomisation-based complex methods

2.2.2.1 Rank Preserving Structural Failure Time Model

The RPSFTM method represents a SNM approach designed specifically for an RCT context.[9] Like the SNM approach, the RPSFTM uses a counterfactual framework to estimate the causal effect of the

treatment in question. However rather than modelling the treatment process using treatment and prognostic covariate history to identify the causal treatment effect, the RPSFTM identifies the treatment effect using only the randomisation of the trial, observed survival and observed treatment history. It is assumed 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. This assumption is linked to the associated assumptions that the 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.[9]

The RPSFTMs primary limitations involve the "common treatment effect" assumption and the randomisation assumption (that randomised groups are well balanced at baseline). The latter should be reasonable in the context of an RCT, but when sample sizes are particularly small differences in baseline characteristics between randomised groups could occur. The "common treatment effect" assumption is more problematic. If patients who switch on to the experimental treatment part way through the trial receive a different treatment effect (relative to the time for which the treatment is taken) compared to patients randomised to the experimental group, the RPSFTM estimate of the treatment effect received by patients in the experimental group will be biased. Given that treatment switching is often only permitted after disease progression – at which time the capacity for a patient to benefit may be different compared to pre-progression – the "common treatment effect" assumption may not be clinically plausible.

2.2.2.2 Iterative Parameter Estimation algorithm

Branson and Whitehead (2002) extended the RPSFTM method using parametric methods, developing a novel iterative parameter estimation (IPE) procedure.[23] The same type of accelerated failure time model is used, but a parametric failure time model is fitted to the original, unadjusted ITT data to obtain an initial estimate of the treatment effect. The failure times of switching patients are then re-estimated using this, and this iterative procedure continues until the new estimate is very close to the previous estimate (the authors suggest within 10^{-5} of the previous estimate), at which point the process is said to have converged.[23]

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 distribution, and thus the authors state that it is important to identify suitable parametric models using tools such as log-cumulative hazard plots.

2.3 Two-stage estimation – A novel method

We tested a method in our simulation study that we have not seen used before in the literature, but that appears to provide a good fit to the treatment switching mechanism often observed in oncology RCTs. 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 a Weibull model (or any other accelerated failure time model) to this data including covariates measured at the secondary baseline (including a 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. The resulting acceleration factor associated with switching could then be used to "shrink" survival times in switching patients to derive a counterfactual survival dataset upon which standard survival analysis could be undertaken. This "two-stage Weibull" approach implicitly acknowledges that the RCT is appropriately randomised up until the point of disease progression, but after that time-point the trial essentially becomes an observational study.

2.4 Methods summary

It is clear that the "simple" switching adjustment methods described in Section 2.1 are highly prone to bias. The complex methods described in Section 2.2 may improve upon these, but they too are associated with important limitations. While SNM, IPCW, RPSFTM and IPE methods will be unbiased in scenarios in which their methodological assumptions hold, their assumptions are limiting and may not be plausible in an RCT context. For instance, observational-based methods are reliant upon the "no unmeasured confounders" assumption and require sufficient data availability to allow the treatment and survival processes to be accurately modelled. Randomisation-based methods are less

data-reliant, but depend upon the "common treatment effect" assumption, which may not be clinically plausible. The "two-stage Weibull" approach described in Section 2.3 offers a simple novel approach to adjust for the type of treatment switching often seen in oncology trials. It relies upon the "no unmeasured confounders" assumption, but only at the point of disease progression. It is limited in that it can only be applied in certain circumstances (such that there is an appropriate secondary baseline), but in practice this may often be the case. The method is further limited by the fact that unless all switching occurs immediately at the secondary baseline time-point it will be prone to time-dependent confounding – however this may often be the case (or may be close to being the case) and so the resultant bias may not be great.

In order to inform the practical use of these methods, we will compare the relative bias associated with them in realistic scenarios – particularly in situations in which key methodological assumptions do not hold.

3. Simulation study design

We simulated independent datasets in which the true survival differences between treatment options were known. 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, based upon input from pharmaceutical companies and our own knowledge. The simulation study was conducted using Stata, version 11.0.[24]

3.1 Underlying survival times

We used a joint longitudinal and survival model to simultaneously generate a time-dependent covariate (referred to as "antigen") and survival times.[25] 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:

$$\text{antigen}_i(t) = \beta_{0i} + \beta_1 \log(t) + \beta_2 \log(t)\, \text{trt}_i + \beta_4 \text{badprog}_i \qquad (1)$$

where,

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

β_{0i} is the random intercept, β_1 the slope against log(time) for a patient in the control group, $\beta_1 + \beta_2$ the slope for a patient in the experimental treatment group (all assuming the same "badprog" status). 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. Antigen follows a linear relationship with log(t) to avoid numerical integration when simulation survival times.

Based on methods described in detail by Bender *et al*,[26] the antigen level was incorporated into the survival simulating process based upon the Weibull model hazard function:

$$h(t) = \lambda\gamma t^{\gamma-1}\exp(X\beta) \qquad (2)$$

where,

$$X\beta = \delta_1 * trt_i + (\eta * \log(t)) * trt_i + \delta_2 badprog_i + \alpha(antigen(t)) \qquad (3)$$

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

$$S(t) = \exp\left(\frac{-\lambda\gamma}{\gamma+\alpha(\beta_1+\beta_2)+\eta}\exp\left(\delta_1 + \delta_2 * badprog_i + \alpha(\beta_{0i} + \beta_4 * badprog_i)\right)\left(t^{\gamma+\alpha(\beta_1+\beta_2)+\eta}\right)\right) \qquad (4)$$

and for the control group:

$$S(t) = \exp\left(\frac{-\lambda\gamma}{\gamma+\alpha(\beta_1)}\exp\left(\delta_2 * badprog_i + \alpha(\beta_{0i} + \beta_4 * badprog_i)\right)\left(t^{\gamma+\alpha(\beta_1)}\right)\right) \qquad (5)$$

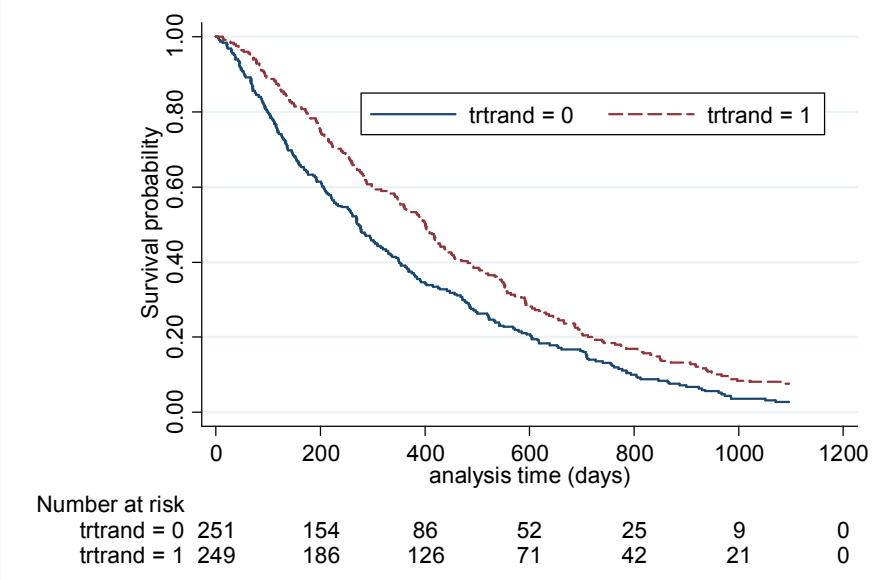
Through this model, treatment and baseline prognosis were allowed to affect the antigen level over time. Treatment group, baseline prognosis and antigen level all also impact upon survival probabilities over time, and antigen level and time influence the treatment effect over time. We believe that the model allows the kind of association between variables that would be expected in reality.

Values for the survival model parameters were selected in order to ensure that the simulated data suitably represented the type of dataset that the study was designed to replicate. To ensure this, a variety of parameter values were tested. In the "base case" (Scenario 1) simulation the parameter values for the Weibull survival model and the longitudinal antigen model were:

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

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

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



These Kaplan-Meier curves provide a close match to Kaplan-Meier curves for OS presented in two recent NICE technology appraisals of treatments for metastatic cancer in which treatment switching was an issue (see Figure B1 and Figure B2 in Appendix B).[27,28]

3.2 Treatment effect in the experimental group

The treatment effect received by experimental group patients was related to antigen level and time. Collett (2003) showed the Acceleration Factor φ (AF) relates the survivor function in the experimental S_E and control S_C group by:[29]

$$S_E = S_C\left(\frac{t}{\varphi}\right) \quad (6)$$

Given the survivor functions presented in equations (4) and (5) the AF over time can be estimated as:

$$\varphi = \exp\left[\frac{-\log\left(\frac{(\gamma+\alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2+\eta}}{(\gamma+\alpha\beta_1+\alpha\beta_2+\eta)}\right)}{(\gamma+\alpha\beta_1)}\right] \quad (7)$$

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

3.3 Treatment effect in switchers

The treatment effect applied to patients who switched from the control group to the experimental treatment was not calculated using the time-dependent AF equation, since this equation estimates the treatment effect at a certain time-point assuming the patient had been receiving the treatment since baseline. Instead, the baseline treatment effect was applied to switchers, but was multiplied by a factor (ω) such that the effect received was lower than the average effect received by experimental group patients. The magnitude of ω was varied to test scenarios in which the treatment effect received by switchers was different to the average effect received by experimental group patients to a greater and lesser extent. Values of ω were chosen to test scenarios in which the treatment effect

received by patients who switched was 0%, 15% and 25% lower than the average effect received by patients originally randomised to the experimental group. This allowed us to test scenarios in which the "common treatment effect" assumption did not hold to varying degrees.

3.4 The switching mechanism

In the base case scenario the probability of treatment switching was allowed to depend upon the antigen value at the time of disease progression (split into three categories – referred to as "antigen group at progression") and the time of progression itself (also split into three categories – referred to as "time to progression group"). 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. This reflects that in a clinical trial setting if a patient has not switched onto the experimental treatment soon after experiencing disease progression, they are very unlikely to switch.

The probability of switching was calculated for each control group patient using a logistic function. In the base case the probability of switching increased if the antigen value was low at the time of disease progression, and if time-to-progression was high. Details on the probability of switching in the different progression and antigen groups at the three consultations following disease progression are presented in Appendix C, Table C1.

3.5 Scenarios investigated

The simulated data generating mechanism had several variables for which values had to be assumed. These are listed in Appendix D, Table D1. The variables altered within the simulations related to:

- Treatment effect decrement received by switchers: 0% (zero time-dependency); 15%; 25%
- Switch proportion: moderate (approximately 50-60% of control group); high (approximately 90-95% of control group)
- Treatment effect: moderate (average HR approximately 0.75); high (average HR approximately 0.50)
- Disease severity: moderate (mean survival in control group approximately 200-220 days); high (mean survival in control group approximately 340-375 days)

This resulted in 24 scenarios. In addition, we tested the impact of alternative switching mechanisms. All 24 of the base scenarios were tested again in simulations in which the treatment switching mechanism was based only upon "antigen group at progression". In Scenarios 25-48 it was assumed that treatment switching was more likely if the antigen value was low at progression (hence patients who progressed quickly and were therefore more likely to have poor prognosis were more likely to switch). In Scenarios 49-72, Scenarios 25-48 were repeated, but under the assumption that switching was more likely if the antigen value was high at progression (hence patients who progressed slowly and were therefore more likely to have good prognosis were more likely to switch).

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

3.6 Performance measures

The time-dependency of the simulated treatment effect meant that it was not possible to produce a single "true" HR or AF that the results of the switching adjustment methods could be compared to. Instead, the survivor functions given in equations 8 and 9 were integrated in order that the true mean survival time in the control group could be calculated, and restricted mean survival at 1095 days (the administrative censoring time in the simulated dataset) was used as the "truth" upon which performance measures were based. Hence methods were assessed based upon how well they could estimate counterfactual survival times in the control group, in the presence of treatment switching from the control group to the experimental treatment. The restricted mean was convenient to use as it avoided potential additional biases associated with having to extrapolate survival times and placed the focus of the simulation study on switching adjustment methods, rather than extrapolation

methods. The methods used to obtain estimates of mean survival at 1095 days for each of the switching adjustment methods are described in Appendix E.

We used bias, mean squared error (MSE) and coverage to assess the performance of the methods, as recommended by Burton *et al*.[30] Absolute bias was measured by the difference between the true restricted mean and the restricted mean estimated by the crossover methods, and relative bias was estimated to represent the absolute bias as a percentage of the true mean survival time. Where methods did not converge the bias and coverage performance measures were calculated based upon simulations in which convergence did occur.

3.7 Methods included

The methods tested within the simulation study were those described in Section 2. For the RPSFTM method we used a log-rank test within the g-estimation procedure, as this has proven most reliable in previous studies.[5] For the IPE algorithm we tested alternative methods using exponential and Weibull models within the estimation procedure, in order to assess whether the performance of the method is sensitive to this. For the IPCW method we used stabilised weights. For both the IPCW and SNM methods we included two versions – one in which all covariates were included in the relevant models, and one in which key covariates were excluded – in order to test their sensitivity to the "no unmeasured confounders" assumption.

For the two-stage Weibull method we fitted a Weibull model to control group patients using disease progression as the secondary baseline time-point. We included covariates for switching, baseline prognosis group, baseline antigen group, time-to-disease progression group and antigen at disease progression group. Given that our simulations ensured that switching occurred soon after disease progression, and thus the potential impact of time-dependent confounding between disease progression and switch was low, we anticipated that this method would perform well.

The recensoring involved in the RPSFTM method caused some problems with estimating mean survival at 1095 days, and therefore a degree of extrapolation was required. We tested two approaches to deal with this – an "extrapolation" approach and a "survivor function" approach, which are described in more detail in Appendix E. Briefly, the "extrapolation" approach involved fitting a

Weibull parametric model to the counterfactual control group dataset in order to estimate mean survival at 1095 days. The "survivor function" approach involved fitting a Weibull model to the experimental group survival data and then applying the inverse of the RPSFTM acceleration factor to derive a survivor function for the counterfactual control group, allowing mean survival at 1095 days to be calculated.

4. Results

The performance of each method differed importantly depending upon the scenario investigated. Due to the large number of methods and scenarios assessed we present only key findings. In Section 4.1 we report key bias results, before focussing on the coverage and MSE of specific methods in Sections 4.2 and 4.3, respectively. A summary table describing the characteristics of each scenario is presented in Appendix F. We refer to figures illustrating our results that are presented in Appendices G, H and I.

4.1 Bias

4.1.1 Simple methods

The relative bias associated with simple adjustment methods (censoring, exclusion, treatment as a time-dependent covariate, ITT) is shown in Figures G1, G2 and G3 in Appendix G and was high across all scenarios. The simple censoring, exclusion and treatment as a time-dependent covariate approaches led to higher bias than a standard ITT analysis across 100%, 91.3% and 93.1% of scenarios respectively. It is important to note that in 12 scenarios the ITT analysis produced very low levels of bias (less than 0.33%). This occurred in scenarios where the treatment effect decrement applied to switching patients was large (approximately 25%) and the true treatment effect received by patients randomised to the experimental group was relatively small (HR of approximately 0.75). In these circumstances patients who switched received very little benefit from switching.

4.1.2 Inverse Probability of Censoring Weights

As shown in Figure G4 in Appendix G the IPCW approach outperformed the ITT analysis in 60% of scenarios (see Table 1). The approach was marginally more successful when antigen covariates

were included in the model, but the difference was not substantial. In scenarios where the IPCW method worked relatively well it generally led to underestimates of survival in the control group (over-estimates of the treatment effect) (relative bias was generally around 5-10%). In scenarios where the switching proportion was very high the method produced much higher relative bias (up to 30-50%) and generally led to substantial underestimates of the treatment effect. This was particularly evident in Scenarios 37 to 48, in which the switching proportion was very high (approximately 95% of patients who survived more than 21 days).

Table 1: Proportion of scenarios with lower bias than ITT analysis

Method	All scenarios	"Zero TDC" scenarios	"TDC" scenarios	"Additional TDC" scenarios
IPCW	59.7%	83.3%	54.2%	41.7%
RPSFTM extrapolation	54.2%	100.0%	50.0%	12.5%
RPSFTM survivor function	54.2%	100.0%	50.0%	12.5%
IPE Weibull extrapolation	54.2%	100.0%	50.0%	12.5%
IPE Weibull survivor function	54.2%	100.0%	50.0%	12.5%
IPE exponential extrapolation	55.6%	100.0%	50.0%	16.7%
IPE exponential survivor function	54.2%	100.0%	50.0%	12.5%
SNM	37.5%	50.0%	33.3%	29.2%
Two-stage Weibull	81.9%	100.0%	95.8%	50.0%

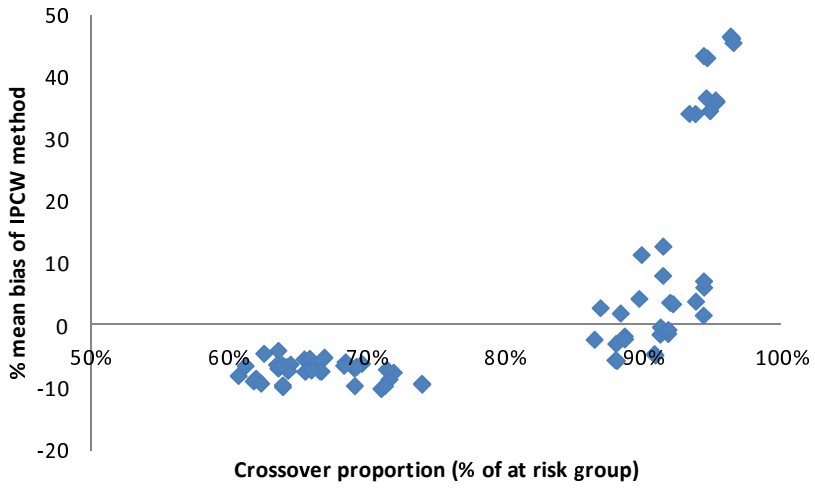
The IPCW method did not seem to be affected by the complexity of the switching mechanism – similar bias was observed when switching depended only upon antigen level (Scenarios 25-72) compared to when switching depended on both antigen level and time to progression (Scenarios 1-24).

The bias associated with the IPCW method was highly related to the switching proportion. Very high levels of switching caused the method to produce more bias. The relationship between the mean percentage bias associated with the IPCW method and the mean proportion of the at-risk population that switched is illustrated in Figure 3. It is clear that once the switching proportion increased to approximately 90% and beyond the method became susceptible to very high levels of relative bias (up to 30-50%).

The bias of the IPCW method was not affected by the additional treatment effect decrement applied in Scenarios 9-12, 21-24, 33-36, 45-48, 57-60, 69-72. This is logical because switching patients are censored at the time of switch, and thus the treatment effect received by these patients is unimportant.

The IPCW method generally produced less bias when the treatment effect was relatively small. In scenarios in which the method worked reasonably well, bias was lower when the treatment effect was lower (bias was lower in Scenarios 1, 4, 5, 8, 9, 25, 28, 29, 32, 33, 49, 52, 56, 57, 64, 65, 68 and 71 than in Scenarios 2, 3, 6, 7, 10, 26, 27, 30, 31, 34, 50, 51, 55, 58, 63, 66, 67 and 72 respectively).

Figure 3: Mean relative bias of the IPCW method compared to crossover proportion



4.1.3 Rank Preserving Structural Failure Time Model and Iterative Parameter Estimation

Illustrations of relative bias for these methods are presented in Figures G5 and G6 in Appendix G. In the scenarios with no time-dependent treatment effect ("zero TDC" scenarios) the "common treatment effect" assumption held. All variations of the RPSFTM and IPE methods produced less bias than the ITT analysis in these scenarios (see Table 1). This was not reliant on the switching proportion, the size of the treatment effect, or the prognosis of switching patients. Differences between the alternative versions of the RPSFTM and IPE methods were minor. Table 2 presents the relative bias associated with each variation of the RPSFTM and IPE methods across the zero TDC scenarios.

Table 2: Relative bias of RPSFTM and IPE approaches in zero TDC scenarios

Method	Extrapolation	Survivor Function
	Relative bias (%)	Relative bias (%)
RPSFTM	0.531	0.551
IPE Weib	0.547	0.573
IPE Exp	1.177	0.594

Note: For the "extrapolation" and "survivor function" approaches the bias was sometimes positive and sometimes negative and thus relative bias is presented. "Zero TDC scenarios" are scenarios 3, 4, 7, 8, 15, 16, 19, 20, 27, 28, 31, 32, 39, 40, 43, 44, 51, 52, 55, 56, 63, 64, 67, 68 (see Table F1 in Appendix F)

Table 2 demonstrates that the "extrapolation" and "survivor function" approaches led to very similar levels of bias, with the "extrapolation" approach producing marginally less bias. The IPE exponential "survivor function" approach led to much less bias than the IPE exponential "extrapolation" approach. This was expected, because under the "survivor function" approach the exponential model was used within the IPE estimation procedure, but the resulting acceleration factor was then applied to a Weibull survival model fitted to the experimental group. If the acceleration factor had been applied to an exponential survival model fitted to the experimental group the results would probably have been much worse, as observed for the "extrapolation" approach. The IPE exponential methods were outperformed by the equivalent RPSFTM and IPE Weibull methods as expected (owing to the use of a Weibull distribution in the data-generating mechanism), but this was only marginal. This demonstrates that using the exponential model within the IPE estimation procedure did not significantly reduce the success of the method, even though the underlying data did not resemble an exponential distribution. However, where extrapolation was undertaken it was important to select the most appropriate parametric model.

In scenarios where the treatment effect was time-dependent ("TDC" scenarios) the RPSFTM and IPE methods produced substantially more bias. In these scenarios the "common treatment effect" assumption did not hold. The impact of this is illustrated in Figure G10 in Appendix G. All scenarios except 3-4, 7-8, 15-16, 19-20, 27-28, 31-32, 39-40, 43-44, 51-52, 55-56, 63-64, and 67-68 included a time-dependent treatment effect, and the bias associated with the RPSFTM and IPE methods was much higher in these scenarios. The results also show, as expected, that when a larger treatment effect decrement was applied to switching patients the RPSFTM and IPE methods produced even more bias (bias was increased in Scenarios 9-12, 21-24, 33-36, 45-48, 57-60 and 69-72 compared to the corresponding Scenarios 1-2, 5-6, 13-14, 17-18, 25-26, 29-30, 37-38, 41-42, 49-50, 53-54, 61-62 and 65-66 respectively) (see Figures G10 and G11 in Appendix G).

In the scenarios in which switching patients received a 15% treatment effect decrement, the RPSFTM and IPE methods led to lower bias than the ITT analysis in 50% of scenarios (see Table 1). The scenarios where these approaches did not improve upon the ITT analysis were those in which the treatment effect was relatively low (equivalent to an average HR of approximately 0.75). In the scenarios in which switching patients received a 25% treatment effect decrement ("additional TDC" scenarios), the RPSFTM and IPE methods led to lower bias than the ITT analysis in only 12.5% of scenarios (see Table 1).

In the TDC scenarios the RPSFTM and IPE methods always 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.

4.1.4 Structural Nested Models

The SNM method performed better than the ITT analysis in 38% of scenarios (31% when there was a time-dependent treatment effect and 50% when there was not) (see Table 1), and performed particularly poorly when the switching proportion was very high (see Figure G7 in Appendix G). The method worked very similarly whether or not the bad prognosis covariate was included in the model. In several scenarios the method failed to converge in a high proportion of simulations. This occurred particularly when the disease severity was high and this was exacerbated when combined with a high switching proportion and a low treatment effect (in such scenarios convergence ranged between 10% and 28%).

The fact that the SNM method seems reliant on disease severity may be due to the stgest Stata command.[17] The command has a "round" option which allows the treatment effect to be estimated allowing counterfactual survival times to be rounded to the nearest unit specified by "round". In the simulations, a value of "1" was used, which is reasonable given that a day time-scale was used. However, often the method failed because estimates of the treatment effect did not allow all

counterfactual survival times to be within one day of the value required in order for the method to converge.

4.1.5 Two-stage Weibull

The two-stage Weibull approach consistently performed well and produced less bias than the ITT analysis in 82% of scenarios (see Table 1). Levels of bias only increased significantly in scenarios where switching proportions were very high (see Figure G8 in Appendix G). The scenarios in which the two-stage Weibull method produced higher bias than the ITT analysis were those in which the treatment benefit received by switching patients was very low. In general, the two-stage Weibull approach led to positive bias when the treatment effect was large, and when poor prognosis patients were more likely to switch. The method was prone to negative bias when good prognosis patients were more likely to switch.

4.1.6 Bias summary

When there was no time-dependent covariate that causes a reduction in the treatment effect over time – meaning that the "common treatment effect" assumption holds – the RPSFTM and IPE approaches produced least bias (see Figure G9 in Appendix G). The results were much more variable when the treatment effect is time-dependent. When the treatment effect received by switching patients was approximately 15% less than that received by patients in the experimental group, the RPSFTM/IPE and IPCW approaches produced similar levels of bias, provided less than 90% of at-risk control group patients switched (see Figure G10 in Appendix G). The SNM method also produced similar levels of bias in these circumstances (providing disease severity is low), but conclusions on this cannot be drawn due to the large proportion of times that this method did not converge.

When the treatment effect received by switching patients was approximately 25% less than that received by patients in the experimental group the IPCW and SNM methods produced less bias than all RPSFTM/IPE variants, provided less than 90% of at-risk control group patients switched (see Figure G11 in Appendix G). However, in these circumstances the ITT analysis often produced least

bias when the treatment effect was relatively small. The simple two-stage Weibull approach produced least bias in several of the TDC scenarios.

4.2 Coverage

Graphs illustrating the coverage of the complex switching adjustment methods are presented in Appendix H. The coverage associated with the RPSFTM "survivor function" approach was very good in the "zero TDC" scenarios, at slightly over 95%. We did not calculate coverage for the RPSFTM "extrapolation" approach in our study because this would require bootstrapping, which would be computationally very expensive in a simulation study of this size. The coverage of the IPE Weibull "survivor function" approach was less good, often between 80% and 90%. This was expected as the confidence intervals around the final iteration of the IPE algorithm were used to generate restricted mean confidence intervals. These represent an underestimate of the true confidence interval.[23] Morden *et al* (2011) showed that if a bootstrapping approach to estimate confidence intervals for the treatment effect is taken, coverage is satisfactory with the IPE method (in zero TDC scenarios).[5]

The coverage of the IPCW method was also good, although slightly lower than the RPSFTM "survivor function" method, which is likely to be due to its decreased accuracy (increased relative bias). Its coverage was also substantially worse in the scenarios in which the method produced relatively high bias.

The coverage associated with the two-stage Weibull method was poor, often being around 50% to 60%. This is despite the low bias associated with the method and is because the method involved shrinking the survival times of switching patients to obtain an adjusted dataset before restricted mean survival was estimated. The higher and lower confidence intervals of the treatment effect estimated in switching patients were used to obtain higher and lower confidence intervals for restricted mean survival. However, 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. This is also true for the SNM approach which was also applied using a two-stage approach (see Appendices A and E). In reality, if a two-stage approach is to be taken, uncertainty around mean survival estimates would need to be taken into account using bootstrapping.

In the “TDC” scenarios the coverage of the RPSFTM and IPE “survivor function” approaches reduced from approximately 95-100% to approximately 90-95%, reflecting the bias associated with these methods in these scenarios. The coverage of the IPCW method was marginally higher than the RPSFTM method in scenarios in which the two methods produced similar levels of bias. In the TDC scenarios in which the IPCW method produced high levels of bias (those in which the switching proportion was very high) the coverage of the IPCW method again decreased substantially.

4.3 Mean Squared Error

Graphs illustrating the MSE of the complex switching adjustment methods are presented in Appendix I. Generally a lower MSE value represents a more accurate measure, but it is important to bear in mind bias and coverage results while considering MSE values, as a low MSE may be misleading if bias is relatively high or coverage is relatively low.

In the “zero TDC” scenarios the RPSFTM and IPE methods produced relatively low MSE. Generally the two-stage Weibull method produced the lowest MSE, but this should be interpreted with care because the method is associated with low coverage. The IPCW method produced a lower MSE than the SNM method across all scenarios which confirms its better performance in these scenarios, especially considering it generally provided lower bias and improved coverage.

As would be expected given the results already presented, in the “TDC” scenarios the MSE associated with the RPSFTM and IPE methods was higher than in the “zero TDC” scenarios, particularly in the “additional TDC” scenarios. In a selection of the “additional TDC” scenarios the IPCW method (and occasionally the SNM method) produced lower MSE than the RPSFTM and IPE methods. This is a similar pattern to that observed in the bias results and demonstrates that relative bias is the key distinguishing factor in the MSE results. This suggests that the standard errors of the mean restricted mean survival estimates associated with the IPCW, RPSFTM and IPE methods were similar (at least in the relatively low switching scenarios).

5. Discussion

Our simulation study demonstrates that randomisation-based methods for adjusting for treatment switching, such as the RPSFTM and IPE algorithm, produce lower bias than all other methods and

provide good coverage 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). This confirms the findings of previous research.[5] However, when the treatment effect is strongly time-dependent, and the "common treatment effect" assumption does not hold, these methods produce high levels of bias and in some circumstances may not be preferable to an ITT analysis.

In the presence of time-dependent treatment effects, treatment switching adjustment methods are limited and are all prone to important bias. Observational-based methods such as the IPCW and SNM – which do not require the "common treatment effect" assumption – require large amounts of data and are particularly sensitive to bias when the switching proportion is very high. Our simulations suggest that the relatively small size of RCT datasets may cause these methods to work sub-optimally, even when the "no unmeasured confounders" assumption holds. This is unsurprising given the reliance of these methods on observational modelling. In our simulations, in which the sample size was 500, a 90% switching proportion in control group patients who became at risk of switching (those who lived longer than 21 days) led to approximately 20 control group patients not switching. The IPCW method uses these patients as a pseudo population upon which to base adjusted control group survival estimates, and such low numbers are clearly problematic. This confirms the findings of other researchers – Howe *et al* found that IPCW was prone to bias in small samples, if selection bias was very strong, and if there were unmeasured confounders.[31]

We found that observational-based methods worked similarly even when a prognostic covariate was excluded from their models. In retrospect, this was not very surprising. The baseline antigen group covariate is directly related to the baseline prognosis group covariate (see equation (1)), and the "antigen group at disease progression" covariate is essentially a time-dependent measure of prognosis which is likely to be highly correlated with the "time to progression group" covariate. Hence, excluding antigen-related covariates from the observational models may be expected to have a relatively minor impact providing baseline prognosis and "time to progression group" covariates are included, or vice-versa. Importantly, the IPCW method outperformed the naïve censoring analysis – which the IPCW would reduce to if all important covariates were unmeasured – across all scenarios.

As expected, naïve methods (such as simple censoring and exclusion approaches) produced high levels of bias consistently across all scenarios and thus should be avoided. When the "common treatment effect" assumption is expected to hold, the RPSFTM and IPE algorithm represent the optimal switching adjustment methods. However when a time-dependent treatment effect is suspected identifying an optimal method is much more difficult. If the treatment switching mechanism is similar to that simulated in our simulation study (that is, switching can only occur after disease progression and must happen very soon after disease progression) and data on key prognostic variables are collected upon disease progression, a simple two-stage Weibull method may represent an appropriate adjustment method and is worthy of consideration. Our findings associated with this method should be interpreted with a degree of caution, because our simulated switching mechanism guaranteed that switching only occurred after disease progression, and happened soon after progression – fitting in with the requirements of the two-stage method. However, this mechanism represents what is observed in reality and so should not devalue the potential usefulness of the method. It may also be observed that we used a Weibull model to generate initial survival times, thus suggesting the two-stage Weibull method would be likely to perform well. However, because we used a joint longitudinal and survival model the resulting survival distributions did not follow a Weibull distribution.

In reality, there is not one method that is likely to be optimal across all scenarios, and the problem of identifying an appropriate method should be tackled on a case-by-case basis. The "no unmeasured confounders" assumption of the IPCW and SNM methods is untestable, and assessment of the "common treatment effect" assumption relied upon by the RPSFTM and IPE algorithm is likely to be prone to bias. However, analyses can be undertaken and information can be gathered to shed light on the plausibility of these assumptions. To consider the "no unmeasured confounders" assumption, previous studies may be reviewed in order to determine whether any important covariates were identified that were excluded from the study affected by treatment switching. Clinical expert opinion could also be sought in order to determine whether any indicators of switching were not collected in the study. An important consideration is that patient preference for switching is not routinely collected in clinical trials, yet this may influence the switching decision. This may cast doubt on the use of observational methods for adjusting for treatment switching in RCTs and also has important

implications for clinical trialists: if switching is to be allowed, suitable data should be collected within the trial to allow for the application of adjustment methods.

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

Various authors have attempted to apply multi-parameter versions of the RPSFTM, in order to allow a relaxation of the "common treatment effect" assumption.[18,32,33] However, relying solely on the randomisation assumption to allow two different treatment effects to be estimated for different groups has proven unsuccessful, with meaningful point estimates difficult to determine. Hence this represents an outstanding problem with randomisation-based methods.

6. Limitations

We attempted to include all the most important and most relevant scenarios given results of the Morden *et al* (2011) study,[5] realistic cancer trial characteristics and the characteristics of the methods that were being assessed. However, there remain potentially interesting scenarios that were not included. In particular, the proportion of patients that switched was an extremely important factor in the results. A range of relatively high switching proportion scenarios were considered, under the assumption that the adjustment methods would struggle with these most. However, given the high levels of bias associated with the assessed methods in the scenarios that incorporate a time-

dependent treatment effect it would be interesting to identify whether levels of bias fall with lower switching levels. This is particularly important for the IPCW and SNM approaches, which were particularly sensitive to the switching proportion.

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

Also, in all the scenarios tested the assumed sample size was 500 patients. This approximately reflects the common size of metastatic oncology trials, and matches the assumption made by Morden *et al* (2011).[5] However, given the data requirements of methods such as IPCW and SNMs, this could be an important factor. In fact, if very large proportions of patients switched, but the trial was extremely large, there may still be a substantial number of patients upon which to base the "pseudo" population and so bias might be avoided.

A technical limitation of our study surrounds the use of the stgest Stata program to implement the SNM method.[17] The method often failed to converge and this appeared to be due to the "round" option included in the program. This might be argued to be a limitation associated with the Stata program rather than the SNM method itself.

A general limitation of simulation studies is that the results are likely to always be linked in some way to the chosen data generating process. Attempts were made to limit this by testing different distributions for parametric switching adjustment methods. Given that a Weibull model was used to generate the underlying survival times, the data generating mechanism may have favoured Weibull-based approaches such as the IPE algorithm and the two-stage Weibull method. However, the results showed that IPE method performed similarly well in estimating the treatment effect when it was applied using an exponential model. Also, due to the inclusion of a time-dependent covariate in the data generating model, the resulting survival times no longer followed a true Weibull distribution. Despite this, it may be of value to conduct similar studies using different data generating models.

It may also be of value to re-run the simulations using different methods to estimate the treatment effect received by switching patients. This was not linked to time, instead the baseline treatment effect was multiplied by a factor to ensure that these patients received a plausible effect. An alternative would be to link this to time and other covariates using formula (7). However this would not be expected to alter the performance of the key complex adjustment methods. The randomisation-based methods do not attempt to model the treatment effect process and so the important factor that determines their bias is the extent to which the average treatment effect differs between switching and experimental group patients – not how this treatment effect difference is estimated. The IPCW approach censors switching patients and so the treatment effect received (and how it is estimated) by these patients does not influence the bias with which the method estimates the treatment effect received in the experimental group.

Finally, we attempted to generate the survival data as realistically as possible, and in such a way that did not satisfy the requirements of any of the adjustment methods. In itself this could be regarded as a limitation, as data were not generated in such a way that satisfies the requirements of methods such as SNMs or the IPCW. Hence these methods could not be expected to produce unbiased results. However, the aim of the simulation study was to demonstrate the performance of these methods in realistic situations – and it is highly likely that in the real world data will not be generated in a way that satisfies the requirements of these models.

7. Conclusions

We conclude that the characteristics of trials and novel therapies should be considered on a case-by-case basis when an analyst is attempting to identify appropriate methods for adjusting for treatment switching in clinical trials but that there are some circumstances in which we should expect certain methods to perform badly. In particular, if the proportion of control group patients that switch is very high (approximately 90% in trials with a sample size of approximately 500) observational-based methods such as IPCW and SNM are unlikely to perform well. Randomisation-based methods are less reliant upon the switching proportion, but themselves will perform poorly if the treatment effect received by patients who switch is likely to be substantially different (approximately 25% different) to that received by patients originally randomised to the experimental group. Simple two-stage methods may represent useful analyses in situations in which other methods are prone to substantial bias.

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. 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
7. 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.
8. 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.
9. 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.
10. Latimer N, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival estimates to account for treatment switching in randomised controlled trials – an economic evaluation context: Methods, limitations and recommendations. Submitted to Medical Decision Making 2013.

11. White IR. Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research 2005; 14(4):327-347.
12. 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.
13. Horwitz R, Horwitz S. Adherence to treatment and health outcomes. Arch Intern Med. 1993;153:1863–1868.
14. Cox DR, Oakes D. Analysis of Survival Data. Boca Raton: Chapman & Hall/CRC; 1984.
15. 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.
16. 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.
17. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. The Stata Journal 2[2], 164-182. 2002.
18. 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.
19. 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.
20. 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.
21. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association 2001; 96(454):440-448.
22. 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.

23. Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. Stat Med 2002; 21(17):2449-2463.
24. Stata statistical software intercooled, Version 11.0, Texas, USA: 2009.
25. Crowther MJ, Abrams KR and Lambert PC. Flexible parametric joint modelling of longitudinal and survival data. Statistics in Medicine 2012;31(30):4456-4471.
26. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine 2005; 24:1713-1723.
27. GlaxoSmithKline UK. Manufacturer Submission to the National Institute for Health and Clinical Excellence. Submission to address the question of whether and how lapatinib falls within the Supplementary Advice to Appraisal Committees on appraising treatments that extend life at the end of life. 25-8-2009.
28. Merck Serono LTD. Single technology appraisal submission: Erbitux (cetuximab) for the first-line treatment of recurrent and/or metastatic squamous cell carcinoma of the head and neck. 2009.
29. Collett D. Modelling Survival Data in Medical Research, 2nd ed. Boca Raton: Chapman & Hall/CRC CRC Press LLC; 2003.
30. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine 2006; 25:4279-4292.
31. 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. Am J Epidemiol. 2011; 173(5):569-577.
32. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. Stat Med. 1999; 18(19):2617-2634.
33. Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. Stat Med. 2004; 23(13):2005-2022.
34. Witteman JCM, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MAJ et al. G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. American Journal of Epidemiology 1998; 148(4):390-401.

35. 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.
36. Fewell Z, Hernan MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. The Stata Journal 2004; 4(4):402-420.
37. White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. The Stata Journal 2002; 2(2):140-150.

Appendix A: Complex switching adjustment methods

Structural Nested Models

Counterfactual survival times for a range of different values of ψ can be estimated using a SNM such as that presented by Robins (1998):[16]

$$U = \int_0^T \exp[\psi A_i(t)]dt \quad \text{(A1)}$$

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. A G-estimation procedure is then used to search for the 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), is a time-dependent Cox proportional hazards model for the hazard of treatment change:[16]

$$\lambda_0(t)\exp[\alpha' W(t)] \quad \text{(A2)}$$

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.[16]

Censoring presents a problem for SNMs because it means that counterfactual survival times can only be estimated for a subset of patients (those that were not censored). However, researchers have shown how both uninformative and informative censoring can be dealt with within an SNM, either by replacing counterfactual survival times in the g-estimation procedure with a function of counterfactual

survival time and censoring time which is observed for all patients, or using inverse probability of censoring weights.[16,34]

Because SNMs were designed for estimating causal treatment effects in observational studies they do not make use of the randomisation element of an RCT. The method could not be applied simply to an RCT dataset, because it would be inappropriate to attempt to model the treatment process (that is, the treatment received by patients over time) when patients are randomised to treatment groups – patients randomised to the control group would not receive the intervention (until switching was allowed) irrespective of their covariates, and similarly patients in the intervention group would remain in that group (though they may discontinue treatment) irrespective of their covariates.

However, this does not mean that the SNM method cannot be useful in an RCT context. The control group after the point at which treatment switching becomes possible could be treated as an observational dataset. The SNM method could then be applied to this dataset to estimate the treatment effect specific to control group (switching) patients. Given the resulting treatment effect estimate (in terms of an acceleration factor – working on the time scale) the survival times of control group switching patients could be adjusted to estimate counterfactual survival times had switching not occurred. This is similar to the approach taken by Robins and Greenland (1994) and Yamaguchi and Ohashi (2004).[18,19]

Inverse Probability of Censoring Weights

Robins and Finkelstein (2000) recommend using "stabilised" inverse probability of censoring weights, as these are shown to be more efficient.[20] Unstabilised weights are simply the inverse of the conditional probability of having remained uncensored until time t conditional on baseline and time-dependent covariates, whereas stabilised weights are the conditional probability of having remained uncensored until time t given baseline covariates, divided by the conditional probability of having remained uncensored until time t given baseline and time-dependent covariates. The stabilised weight will be equal to 1 for all t if the history of the included prognostic factors for failure do not impact upon the hazard of censoring at t – thus there would be no informative censoring and treatment switching would be random.[20]

Formally, the stabilised weights applied to each individual for time interval (t), as specified by Hernan *et al* (2001) are:[21]

$$\widehat{W}(t) = \prod_{k=0}^{t} \frac{\Pr[C(k)=0 | \bar{C}(k-1)=0, \bar{A}(k-1), V, T>k]}{\Pr[C(k)=0 | \bar{C}(k-1)=0, \bar{A}(k-1), \bar{L}(k), T>k]} \quad \text{(A3)}$$

Here $C(k)$ 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 (A3) 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. Robust variance estimators or bootstrapping should be used to estimate confidence intervals.[21,22]

Rank Preserving Structural Failure Time Model

An accelerated failure time counterfactual survival model similar to that presented in equation (A1) is used. Counterfactual survival time is a product of observed time spent on treatment and observed time spent off treatment, where time spent on treatment is multiplied by the factor exp(ψ). The value of the treatment effect (ψ_0) is estimated as the value of ψ for which counterfactual survival 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.[35] The point estimate of

ψ is that for which the test (z) statistic equals zero. Because the RPSFTM is a randomisation-based efficacy estimator the p-value from the ITT analysis is maintained.[32]

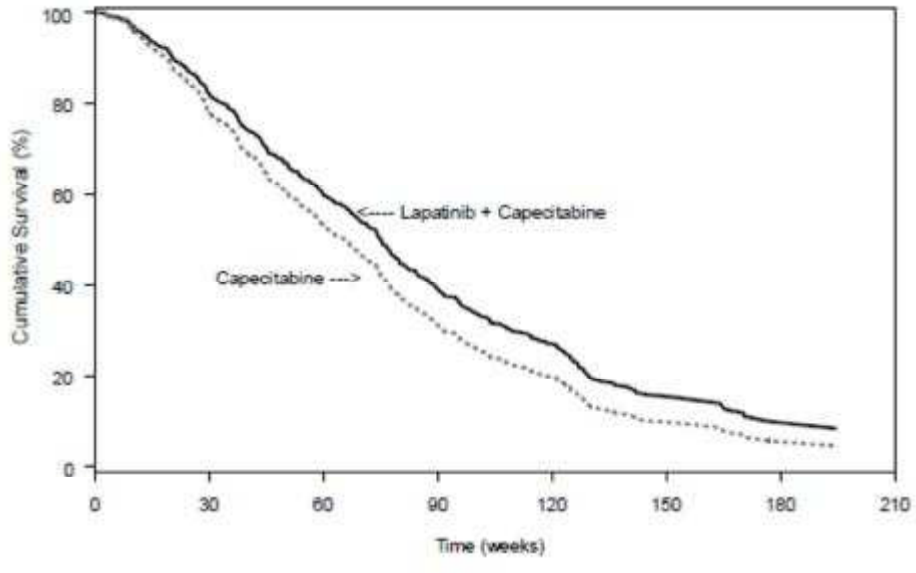
White *et al* (1999) demonstrate that censoring is problematic for the RPSFTM.[32] A positive or negative treatment effect may increase or decrease the probability that the survival time of an individual is censored, and, where treatment switching occurs, treatment received is likely to be associated with prognosis. In turn, this means that the censoring of counterfactual survival times may depend on prognostic factors and therefore be informative.[32] Bias associated with this can be avoided by recensoring counterfactual survival times at the earliest possible censoring time given the treatment effect ψ.[32] Thus for each patient in treatment groups at risk of switching the recensored censoring time is the minimum of the observed administrative censoring time (C_i) and the product $\exp(\psi)C_i$. If the a patient experienced an event, but the recensoring time is less than the event time, that patient has their survival time recensored and their event is no longer observed.

Iterative Parameter Estimation algorithm

This method uses the same type of accelerated failure time model as the RPSFTM, but 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 exp(ψ) and the counterfactual survival time model presented in equation (A1), 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(ψ) 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.[ref] Bootstrapping is recommended to obtain standard errors and confidence intervals for the treatment effect.[23]

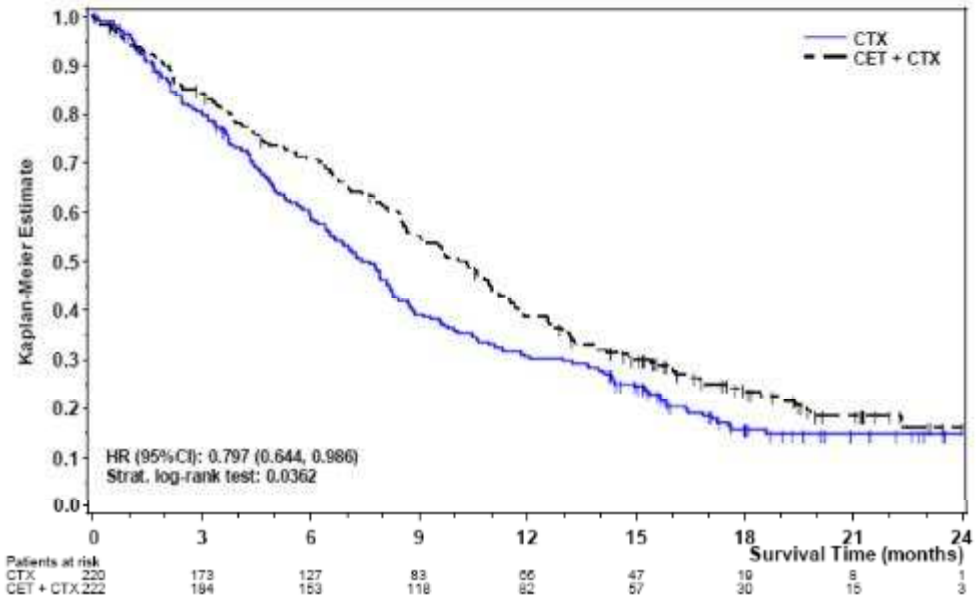
Appendix B

Figure B1: Overall Survival Kaplan-Meier - lapatinib+capecitabine for metastatic breast cancer



Note: This is Figure 2 in GSK's submission to NICE, August 2009.[27]

Figure B2: Overall Survival Kaplan-Meier - cetuximab for metastatic squamous cell carcinoma of the head and neck



Note: This is Figure B2 from: Merck Serono's submission to NICE, 2009.[28]

Appendix C

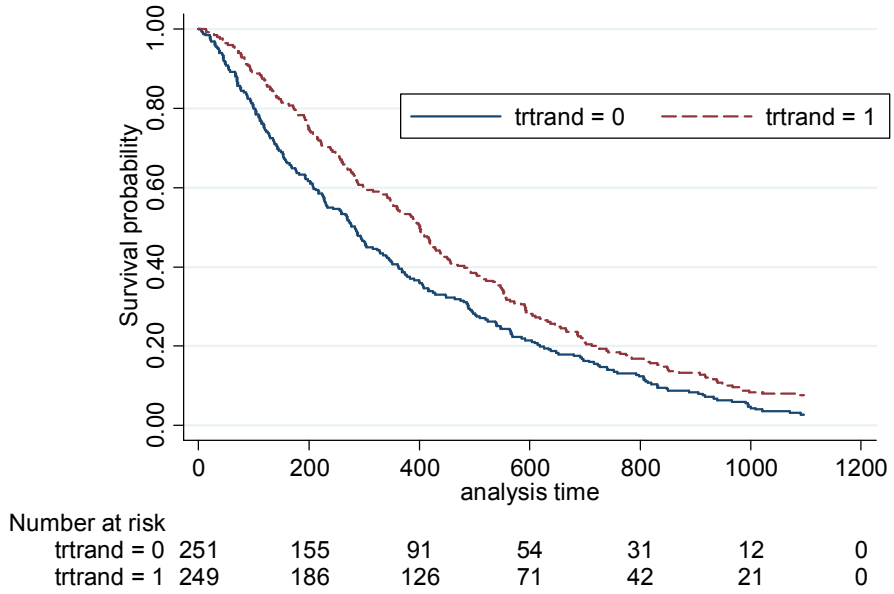
The probability of switching in the different progression and antigen groups at the three consultations following disease progression for the base case scenario are presented in Table C1. Higher group numbers represent higher values for that group (that is, "time to progression group" 0 are the control 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 C1: Probability of treatment crossover by prognostic groups and consultation

Consultation 1		Antigen group at progression		
		0	1	2
Time to progression group	0	0.30	0.20	0.10
	1	0.51	0.41	0.30
	2	0.88	0.75	0.60
Consultation 2		Antigen group at progression		
		0	1	2
Time to progression group	0	0.23	0.15	0.07
	1	0.42	0.32	0.23
	2	0.84	0.68	0.51
Consultation 3		Antigen group at progression		
		0	1	2
Time to progression group	0	0.14	0.09	0.04
	1	0.29	0.22	0.15
	2	0.75	0.55	0.38

When switching is incorporated into Scenario 1 using the switching probabilities presented in Table C1, 156 (62%) control group patients switch treatments. The resulting Kaplan-Meier curve is as shown in Figure C1. As desired, the control group Kaplan-Meier curve moves towards the experimental group Kaplan-Meier curve (as can be seen by comparing Figure 1 with Figure C1). The ITT HR in this instance increases to 0.743 from 0.719, demonstrating the effect of the switching.

Figure C1: Overall Survival Kaplan-Meier from simulated dataset Scenario 1: With switching



Number at risk

trtrand = 0	251	155	91	54	31	12	0
trtrand = 1	249	186	126	71	42	21	0

Appendix D

Values for each variable in Scenario 1 are quoted, as are alternative values for different scenarios.

Table D1: Simulated scenarios – Parameter values and alternatives tested

Variable	Value (Scenario 1)	Alternative Values
Sample size	500	-
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)	-
Baseline effect of being in the bad prognosis group	Log hazard ratio = 0.5	-
Survival time distribution	Weibull parameters when time-dependent effect included: Shape parameter 0.9 (mortality decreasing over time) Scale parameter 0.0005 Weibull parameters when time-dependent effect not included: Shape parameter 0.9 (mortality decreasing over time) Scale parameter 0.004	Alter scale parameter to 0.001 to represent a more severe disease (and hence less censoring) in scenarios with time-dependent effect Alter scale parameter to 0.007 to represent a more severe disease (and hence less censoring) in scenarios without time-dependent effect
Progression free survival	Overall survival time multiplied by a value from a beta distribution with shape parameters (5,5) – this implies the assumption that PFS is approximately half of OS. This is not an important assumption – PFS is only included because I 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 = -0.7 Log hazard ratio in scenarios that do not include an additional time-dependent effect = -0.3	Alter log hazard ratio to -1.1 to represent a larger treatment effect in scenarios with time-dependent effect Alter log hazard ratio to -0.7 to represent a larger treatment effect in scenarios without time-dependent effect
Antigen intercept	Calculated using a normal distribution with mean of 20 and standard deviation of 1	-
Antigen value progression over time	As demonstrated by formula [11]. $\beta_2 = -4$ to represent that the antigen value increases	-

	more slowly in the experimental group, and $\beta_4 = 5$ so that bad prognosis patients start with higher levels of the antigen	
Impact of antigen value on overall survival	As demonstrated by formulas [12-21]. 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	Remove impact of the antigen value by setting α=0
Impact of antigen value on treatment effect	As demonstrated by formulas [22-28]. 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	Remove impact of antigen value by setting α=0 Model larger time-dependent effect by applying additional decrement multiplier to switching patients
Time-dependent portion of treatment effect, η	η =0.15 to generate a reduction in the treatment effect over time, in scenarios where a time-dependent treatment effect is assumed	Remove impact of η by setting η =0 in scenarios where the treatment effect is not time-dependent
Assumed frequency of consultations	One every 3 weeks (21 days)	-
Probability of switching treatment over time	As shown in Table 6.1. This results in a switching proportion of approximately 63% in Scenario 1	Test a high switching scenario where all probabilities are increased – to an extent where approximately 90% of patients that survive longer than 21 days switching
Prognosis of switching patients	As shown in Table 6.1. 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	Make switching more likely in poor prognosis patients. Test scenarios where switching is based on a simpler mechanism (only based on the antigen value)
Treatment effect in switching patients	Equal to baseline treatment effect multiplied by ω. Set ω such that treatment effect received by switching patients is 85% of the average effect received by experimental group patients in base scenarios.	Alter ω such that treatment effect received by switching patients equals 75% - 78% of the average effect received by experimental group patients. In scenarios where the treatment effect is not time-dependent, set ω to 1 – such that the treatment effect received by switching patients is the baseline effect received by the experimental group.

Based on the variables and alternative values presented in Table D1, 72 scenarios were run.

Appendix E: Applying the methods

Here we describe precisely how each switching adjustment method was applied and how restricted mean survival was estimated.

- *Intention to treat*

The restricted mean associated with an ITT analysis was calculated by applying Stata's "stci, rmean" command which computes the mean survival time restricted to the longest follow-up time (that is, 1095 days) for the specified patient group. For the ITT analysis this was applied to the control group data confounded by treatment crossover.

- *Per protocol – exclude switchers*

The restricted mean associated with a per-protocol analysis where switching patients were excluded was calculated by excluding all switching patients and using Stata's stci command.

- *Per protocol – censor switchers*

The restricted mean associated with a per-protocol analysis where switching patients were censored was calculated by censoring all switching patients at the time at which they switched, and using Stata's stci command on this adjusted dataset.

- *Treatment as a time-dependent covariate*

Treatment was included as a time-dependent covariate within a Cox model including covariates. This method was expected to be biased, because including covariates adjusts for factors through which the treatment has an effect. However excluding covariates would also be expected to lead to bias, because the prognosis of switching patients would not be accounted for. The restricted mean was estimated using a "survivor function" approach. A Weibull model was fitted to the experimental group survival data and the associated survivor function and hazard function were derived. The experimental group hazard function was multiplied by the inverse of the HR to obtain the control group hazard function, and the control group survivor function was then derived up to 1095 days.

Confidence intervals (CIs) for the mean survival estimate were calculated by applying the 95% CIs of the estimated treatment effect in the "survivor function" process.

- *Structural Nested Models - Observational SNM with g-estimation*

The observational SNM included in the simulation study was applied using the stgest command in Stata, in line with the example given by Sterne and Tilling (2002).[17] Generally, the SNM method is applied to observational datasets and in the context of an RCT applying the method is more complex. The SNM was applied to the control group after disease progression had occurred, to estimate the treatment effect in switching patients compared to control group patients that did not switch. The resulting AF was used to "shrink" survival times in switching patients in order to arrive at a counterfactual dataset adjusted for switching.

The restricted mean associated with the adjusted dataset was estimated using Stata's stci command, and CIs for the restricted mean estimate were calculated by applying the 95% CIs of the estimated treatment effect in the "shrinkage" process. This is likely to represent an underestimate of the true confidence interval as it does not take into account the uncertainty in the underlying survival distribution.

In order to test the sensitivity of the observational SNM method to not including all prognostic confounders, two versions of the method were included. In the first version, all covariates (baseline prognosis group, time-to-disease progression group, antigen at baseline group, antigen at progression group, and antigen over time group) were included, and in the second version the baseline prognostic group covariate was excluded.

- *Inverse Probability of Censoring Weights*

IPCW was applied in line with the example given by Fewell *et al* (2004),[36] although unlike Fewell *et al*'s example the IPCW method rather than a full MSM was applied. In addition, weights were only applied to patients in the control group, as the context is an RCT rather than an observational study.

To apply the IPCW method using stabilised weights first the simulated data were split into time intervals and time-dependent covariate values were recorded for each of these. Data were excluded

for switching patients beyond the point of switch, and OS was censored for these patients. IPCWs were then estimated for each patient and for each time interval. The numerator of each stabilised weight was the cumulative probability of remaining uncensored (that is, not switching) from the beginning of follow-up to the end of the interval given only baseline covariates and the number of consultations since randomisation. This was estimated for all control group patients for all time periods. The denominator of the stabilised weight was the cumulative probability of remaining uncensored (that is, not switching) to the end of the interval given baseline and time-dependent covariates, and the number of consultations since randomisation. These weights were only different from 1 for time periods during which patients were at risk of switching – that is after disease progression had been observed, and before 3 consultations after disease progression had taken place. The probabilities of remaining uncensored (that is, not switched) were obtained by fitting pooled logistic models with informative censoring due to treatment switching as the dependent variable. A Cox proportional hazards model that incorporated baseline (but not time-dependent) covariates was then run, weighted by the stabilised weights, in order to estimate an adjusted IPCW HR.

To test the sensitivity of the IPCW method to the "no unmeasured confounders" assumption two versions of the method were applied. The first included all baseline and time-dependent covariates:

- Baseline prognosis group
- Baseline antigen group (antigen level at baseline split into 3 groups)
- Time-to-disease progression group
- Antigen at disease progression group
- Antigen group (which splits patients into 6 groups based upon their antigen level as this changes over time)
- An interaction term of time-to-disease progression group multiplied by antigen at time of progression group.

The second version excluded all antigen-related covariates.

The IPCW approach allows a weighted Kaplan-Meier (WKM) curve to be obtained, which would provide the optimal restricted mean measure for the method. However, Fewell *et al*'s Stata

methodology does not provide this curve,[36] and there are problems with calculating this in the context of a simulation study. For the WKM to be estimated the sum of the weights for all patients at risk, and all patients who experienced the event, for each time point, must be calculated. In the simulation study it was possible that control group patients with the longest follow-up times may switch and be censored at an earlier date and therefore a new administrative censoring time (that is, the time to which the restricted mean is calculated) would need to be generated for each simulation in order to avoid biased overestimates of mean survival being produced. Therefore, generating the WKM accurately in a simulation study would be very computationally-intensive and therefore a "survivor function" approach – as described for the "Treatment as a time-dependent covariate" method – using the IPCW HR was taken to estimate restricted mean survival for the control group at 1095 days. This should represent a close approximation of the IPCW WKM. CIs for the mean survival estimate were calculated by applying the 95% CIs of the IPCW HR in the "survivor function" process.

- *RPSFTM with log-rank test*

The RPSFTM included in the simulation study was applied using the strbee Stata program developed by White *et al* (2002).[37] 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. Given this, we used two approaches to calculate the control group restricted mean:

- "Extrapolation" approach. Under this approach the recensored counterfactual survival times produced by the strbee command were extrapolated out to 1095 days and the area under the extrapolated survival curve (that is, the restricted mean) was estimated. A Weibull model was used to extrapolate. The extrapolated portion was relatively small as mean survival was restricted to 1095 days, but a poor extrapolation could still impact upon the bias associated with this approach.
- "Survivor function" approach. This approach is similar to the "survivor function" approach previously described for the "Treatment as a time-dependent covariate" method, adapted for use with an acceleration factor. A Weibull model was fitted to the experimental group data

and the time associated with each survivor function probability was divided by the RPSFTM AF in order to obtain the survival times associated with the survival probabilities for the control group. Mean survival was estimated up to 1095 days by calculating the area under the survival curve. Confidence intervals (CIs) for the mean survival estimate were calculated by applying the 95% CIs of the estimated treatment effect in the "survivor function" process.

- *IPE algorithm*

The IPE algorithm approach was also applied using the strbee Stata program.[37] The method was applied using full recensoring, and included versions with and without covariates. It was also applied 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 the parameter values of the final parametric model used to estimate the adjusted treatment effect. These were used to estimate the restricted mean survival at 1095 days associated with the final model. This is similar to the "extrapolation" approach described above for the RPSFTM method. The "survivor function" approach was also applied in order to obtain alternative restricted mean estimates. Importantly when the IPE method was applied using an exponential model substantial differences between the "extrapolation" approach and the "survivor function" approach were expected. Under the "extrapolation" approach the extrapolation was undertaken using the final exponential model estimated by the IPE approach, whereas under the "survivor function" approach the treatment effect obtained using the exponential version of the IPE method was applied to the experimental group survival times estimated using a Weibull model.

As for the other adjustment methods, CIs for the restricted mean estimate were calculated by applying the 95% CIs of the estimated treatment effect in the restricted mean estimation process. As noted by Morden *et al* (2011) this was likely to provide relatively poor coverage as the confidence intervals associated with the treatment effect from the final IPE iteration are underestimates.[5]

- *Other methods: Two-stage Weibull*

This method involved first estimating the switching treatment effect in the control group for the period following disease progression. A Weibull model was used to estimate this treatment effect, including the following covariates:

- Baseline prognosis group
- Baseline antigen group
- Time-to-disease progression group
- Antigen at disease progression group

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 the following equation:

$$\text{counterfactual survival time} = \text{crossover time} + \left(\frac{1}{\text{AF}} * (\text{observed survival} - \text{crossover time})\right)$$

The restricted mean survival of the adjusted dataset produced was then calculated using Stata's stci command, and confidence intervals were calculated using the confidence intervals of the treatment effect.

Appendix F: Overview of simulation scenarios

Table F1 provides information on each of the scenarios simulated. The true restricted mean 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). Where there is a time-dependent treatment effect this reflects only an approximation of the true treatment effect as proportional hazards/constant acceleration factor assumptions do not hold. In terms of a hazard ratio, the average treatment effect varied between 0.50 and 0.75.

Table F1 shows that the switching proportion varied between 52% and 94% of all control group patients. Scenarios 13-24, 37-48 and 61-72 were designed to result in higher levels of switching, although these levels are probabilistic and are reliant on other characteristics. This led the level of switching to vary between otherwise equivalent scenarios with switching proportions highest in Scenarios 25-48, followed by Scenarios 49-72 and 1-24. Table F1 also presents the switching proportion as a percentage of the control group patients that became "at-risk" of switching. Control group patients could only switch if they were alive at their first "consultation" at 21 days. The proportion of patients that died before this point and never became at-risk depended upon disease severity. The proportion of switching patients as a percentage of patients that became at-risk ranged from 61% to 96%.

Table F1 shows that Scenarios 1-24 incorporated a complex switching probability mechanism in which better prognosis patients were generally more likely to switch. Scenarios 25-48 and 49-72 incorporated a simpler switching probability mechanism based only upon antigen level at the time of disease progression. In Scenarios 25-48 patients with a relatively poor prognosis were more likely to switch. The opposite was true in Scenarios 49-72.

Table F1 shows that in Scenarios 1, 2, 5, 6, 13, 14, 17 and 18 the treatment effect received by experimental group patients was dependent upon the antigen level and time and in these scenarios switching patients received a reduced treatment effect. In Scenarios 9, 10, 11, 12, 21, 22, 23 and 24 the treatment effect received by experimental group patients was time-dependent and related to the antigen level, and an additional decrement (compared to scenarios 1, 2, 5, 6, 13, 14, 17 and 18) was applied to the effect received by switching patients (in Table F1 these are labelled as "Yes+" in the

"Time-dependent treatment effect" column). In scenarios 3, 4, 7, 8, 15, 16, 19 and 20 the treatment effect was not time-dependent or related to the antigen – in these scenarios the "common treatment effect" assumption held. This pattern across scenarios was repeated in scenarios 25-48 and 49-72 (i.e. scenarios 25 and 49 are equivalent to scenario 1, except with altered switching mechanisms, and so on). Further details on the effect size decrement applied to switching patients in each scenario are included ("Treatment effect in switching patients (AF)"), as well as on the effect size as a proportion of that received by the experimental group. This varied between 75% and 100%.

Table F1: Overview of simulated scenarios

Scenario	Truth		Average treatment effects		Mean switching % of total	Mean switching % of at risk	Mean censoring proportion (%)	Disease severity	Prognosis of switching patients	Time-dependent treatment effect	Treatment effect in switching patients (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
1	372.06	462.27	0.75	1.28	63.37%	65.33%	7.19%	Low	Complex - good	Yes	1.09	85%
2	372.06	579.28	0.52	1.75	61.54%	63.44%	13.42%	Low	Complex - good	Yes	1.48	85%
3	344.47	568.12	0.51	2.15	56.26%	61.07%	19.97%	Low	Complex - good	No	2.15	100%
4	344.47	437.88	0.75	1.39	58.46%	63.45%	11.65%	Low	Complex - good	No	1.39	100%
5	216.96	285.64	0.73	1.32	60.25%	64.04%	0.84%	High	Complex - good	Yes	1.12	85%
6	216.96	381.51	0.50	1.80	58.17%	61.82%	2.74%	High	Complex - good	Yes	1.53	85%
7	201.45	387.21	0.51	2.17	52.48%	60.56%	7.02%	High	Complex - good	No	2.17	100%
8	201.45	271.95	0.75	1.40	54.09%	62.41%	2.80%	High	Complex - good	No	1.40	100%
9	372.06	462.27	0.75	1.28	63.74%	65.71%	6.86%	Low	Complex - good	Yes +	1.00	78%
10	372.06	579.21	0.52	1.75	61.45%	63.36%	12.86%	Low	Complex - good	Yes +	1.31	75%
11	216.96	285.64	0.73	1.32	60.52%	64.32%	0.76%	High	Complex - good	Yes +	1.00	76%
12	216.96	381.51	0.50	1.80	58.02%	61.66%	2.58%	High	Complex - good	Yes +	1.36	75%
13	372.06	462.27	0.75	1.28	88.37%	91.10%	7.25%	Low	Complex - good	Yes	1.09	85%
14	372.06	579.21	0.52	1.75	87.96%	90.68%	13.80%	Low	Complex - good	Yes	1.48	85%
15	344.47	568.12	0.51	2.15	80.99%	87.90%	20.60%	Low	Complex - good	No	2.15	100%
16	344.47	437.88	0.75	1.39	81.29%	88.23%	11.80%	Low	Complex - good	No	1.39	100%
17	216.96	285.64	0.73	1.32	83.30%	88.53%	0.83%	High	Complex - good	Yes	1.12	85%

18	216.96	381.51	0.50	1.80	82.66%	87.85%	2.81%	High	Complex - good	Yes	1.53	85%
19	201.45	387.21	0.51	2.17	74.83%	86.34%	7.24%	High	Complex - good	No	2.17	100%
20	201.45	387.21	0.75	1.40	75.20%	86.77%	2.84%	High	Complex - good	No	1.40	100%
21	372.06	462.27	0.75	1.28	88.39%	91.12%	6.87%	Low	Complex - good	Yes +	1.00	78%
22	372.06	579.21	0.52	1.75	88.02%	90.75%	13.09%	Low	Complex - good	Yes +	1.31	75%
23	216.96	285.64	0.73	1.32	83.30%	88.53%	0.74%	High	Complex - good	Yes +	1.00	76%
24	216.96	381.51	0.50	1.80	82.77%	87.97%	2.65%	High	Complex - good	Yes +	1.36	75%
25	372.06	462.27	0.75	1.28	69.65%	71.80%	7.02%	Low	Simple - poor	Yes	1.09	85%
26	372.06	579.21	0.52	1.75	71.62%	73.84%	13.10%	Low	Simple - poor	Yes	1.48	85%
27	344.47	568.12	0.51	2.15	65.86%	71.48%	19.08%	Low	Simple - poor	No	2.15	100%
28	344.47	437.88	0.75	1.39	64.03%	69.50%	11.15%	Low	Simple - poor	No	1.39	100%
29	216.96	285.64	0.73	1.32	64.89%	68.96%	0.81%	High	Simple - poor	Yes	1.12	85%
30	216.96	381.51	0.50	1.80	66.71%	70.90%	2.60%	High	Simple - poor	Yes	1.53	85%
31	201.45	387.21	0.51	2.17	59.78%	68.98%	6.47%	High	Simple - poor	No	2.17	100%
32	201.45	387.21	0.75	1.40	57.88%	66.79%	2.61%	High	Simple - poor	No	1.40	100%
33	372.06	462.27	0.75	1.28	69.16%	71.30%	6.89%	Low	Simple - poor	Yes +	1.00	78%
34	372.06	579.21	0.52	1.75	71.64%	73.86%	12.52%	Low	Simple - poor	Yes +	1.31	75%
35	216.96	285.64	0.73	1.32	65.02%	69.10%	0.75%	High	Simple - poor	Yes +	1.00	76%
36	216.96	381.51	0.50	1.80	66.95%	71.15%	2.52%	High	Simple - poor	Yes +	1.36	75%
37	372.06	462.27	0.75	1.28	93.31%	96.20%	7.23%	Low	Simple - poor	Yes	1.09	85%
38	372.06	579.21	0.52	1.75	93.41%	96.30%	13.78%	Low	Simple - poor	Yes	1.48	85%
39	344.47	568.12	0.51	2.15	87.06%	94.50%	20.65%	Low	Simple - poor	No	2.15	100%
40	344.47	437.88	0.75	1.39	86.83%	94.25%	11.72%	Low	Simple - poor	No	1.39	100%
41	216.96	285.64	0.73	1.32	89.12%	94.71%	0.84%	High	Simple - poor	Yes	1.12	85%
42	216.96	381.51	0.50	1.80	89.54%	95.16%	2.80%	High	Simple - poor	Yes	1.53	85%
43	201.45	387.21	0.51	2.17	81.17%	93.65%	7.17%	High	Simple - poor	No	2.17	100%
44	201.45	387.21	0.75	1.40	80.77%	93.20%	2.81%	High	Simple - poor	No	1.40	100%
45	372.06	462.27	0.75	1.28	93.31%	96.20%	6.86%	Low	Simple - poor	Yes +	1.00	78%

46	372.06	579.21	0.52	1.75	93.51%	96.40%	12.96%	Low	Simple - poor	Yes +	1.31	75%
47	216.96	285.64	0.73	1.32	88.87%	94.44%	0.75%	High	Simple - poor	Yes +	1.00	76%
48	216.96	381.51	0.50	1.80	89.49%	95.11%	2.62%	High	Simple - poor	Yes +	1.36	75%
49	372.06	462.27	0.75	1.28	66.26%	68.31%	7.15%	Low	Simple -good	Yes	1.09	85%
50	372.06	579.21	0.52	1.75	64.57%	66.57%	13.45%	Low	Simple -good	Yes	1.48	85%
51	344.47	568.12	0.51	2.15	59.10%	64.15%	19.99%	Low	Simple -good	No	2.15	100%
52	344.47	437.88	0.75	1.39	60.78%	65.97%	11.46%	Low	Simple -good	No	1.39	100%
53	216.96	285.64	0.73	1.32	61.53%	65.39%	0.82%	High	Simple -good	Yes	1.12	85%
54	216.96	381.51	0.50	1.80	60.01%	63.78%	2.76%	High	Simple -good	Yes	1.53	85%
55	201.45	387.21	0.51	2.17	53.90%	62.19%	6.96%	High	Simple -good	No	2.17	100%
56	201.45	387.21	0.75	1.40	55.26%	63.76%	2.77%	High	Simple -good	No	1.40	100%
57	372.06	462.27	0.75	1.28	66.15%	68.20%	6.91%	Low	Simple -good	Yes +	1.00	78%
58	372.06	579.21	0.52	1.75	64.45%	66.45%	12.86%	Low	Simple -good	Yes +	1.31	75%
59	216.96	285.64	0.73	1.32	61.95%	65.84%	0.78%	High	Simple -good	Yes +	1.00	76%
60	216.96	381.51	0.50	1.80	60.00%	63.76%	2.58%	High	Simple -good	Yes +	1.36	75%
61	372.06	462.27	0.75	1.28	91.44%	94.27%	7.26%	Low	Simple -good	Yes	1.09	85%
62	372.06	579.21	0.52	1.75	90.87%	93.68%	13.88%	Low	Simple -good	Yes	1.48	85%
63	344.47	568.12	0.51	2.15	84.11%	91.30%	20.81%	Low	Simple -good	No	2.15	100%
64	344.47	437.88	0.75	1.39	84.13%	91.31%	11.77%	Low	Simple -good	No	1.39	100%
65	216.96	285.64	0.73	1.32	86.39%	91.81%	0.85%	High	Simple -good	Yes	1.12	85%
66	216.96	381.51	0.50	1.80	86.25%	91.67%	2.79%	High	Simple -good	Yes	1.53	85%
67	201.45	387.21	0.51	2.17	77.63%	89.57%	7.30%	High	Simple -good	No	2.17	100%
68	201.45	387.21	0.75	1.40	77.78%	89.75%	2.81%	High	Simple -good	No	1.40	100%
69	372.06	462.27	0.75	1.28	91.42%	94.25%	6.87%	Low	Simple -good	Yes +	1.00	78%
70	372.06	579.21	0.52	1.75	91.40%	94.23%	12.98%	Low	Simple -good	Yes +	1.31	75%
71	216.96	285.64	0.73	1.32	86.59%	92.03%	0.75%	High	Simple -good	Yes +	1.00	76%
72	216.96	381.51	0.50	1.80	86.27%	91.69%	2.59%	High	Simple -good	Yes +	1.36	75%

Appendix G: Relative bias graphs

Graphs showing bias across scenarios are presented throughout this Appendix – care should be taken when comparing these because the y-axes use different scales.

Figure G1: Mean bias (%) across scenarios – ITT analysis

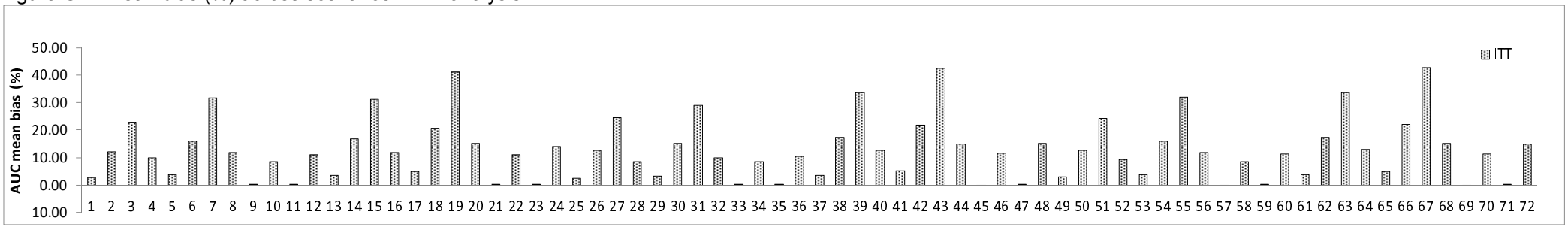


Figure G2: Mean bias (%) across scenarios – Exclusion and censoring approaches

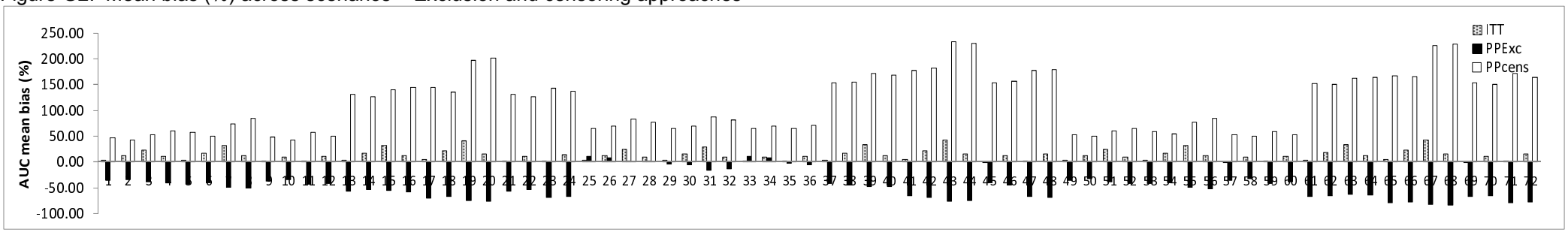


Figure G3: Mean bias (%) across scenarios – TDCM

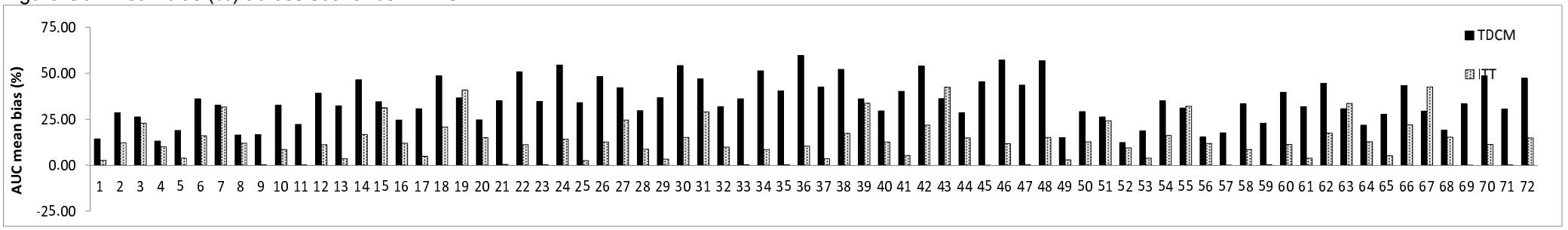


Figure G4: Mean bias (%) across scenarios – IPCW

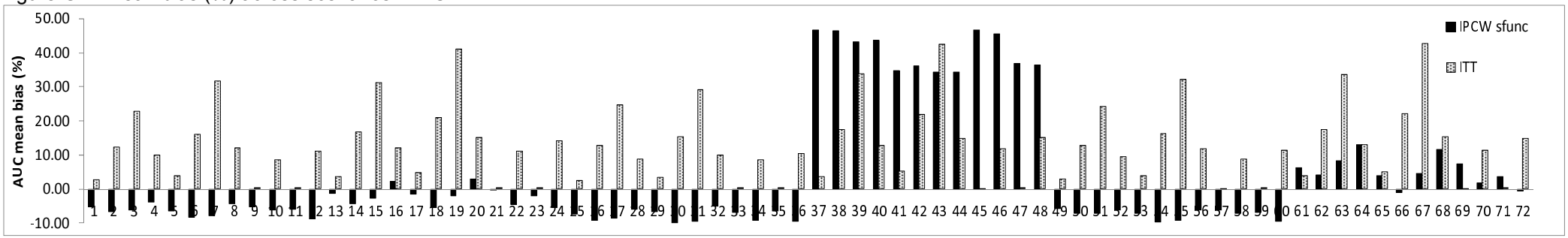


Figure G5: Mean bias (%) across scenarios – RPSFTM and IPE Weibull "survivor function" approaches

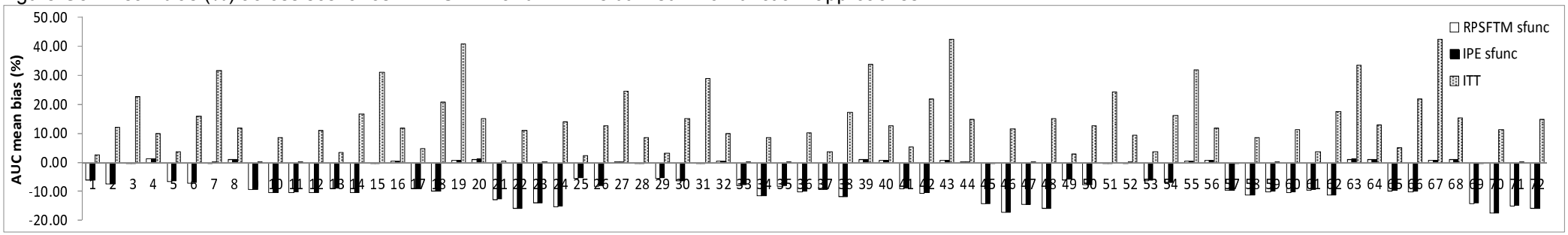


Figure G6: Mean bias (%) across scenarios – RPSFTM and IPE Weibull "extrapolation" approaches

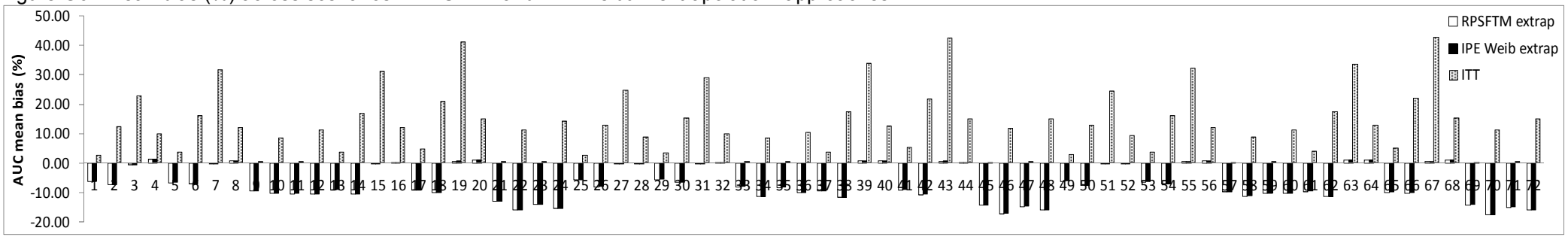


Figure G7: Mean bias (%) across scenarios – SNM with g-estimation

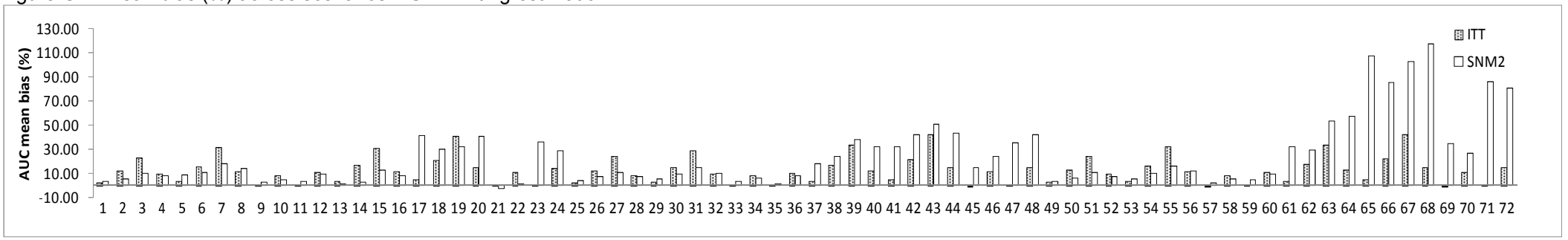


Figure G8: Mean bias (%) across scenarios – Two-stage Weibull

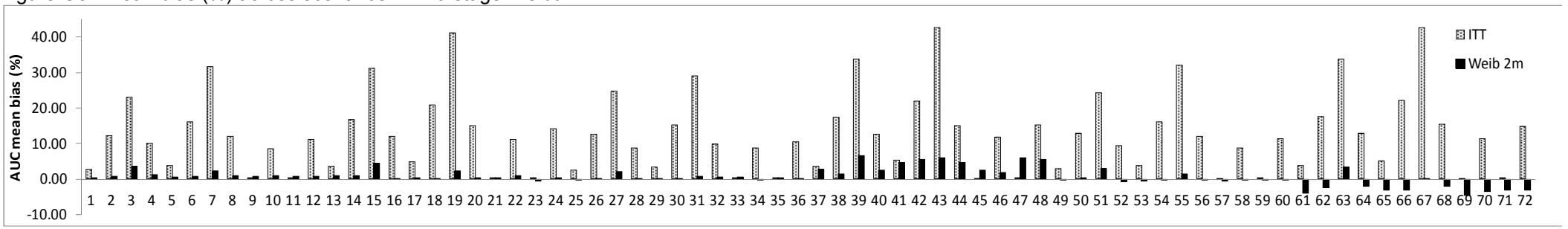


Figure G9: Bias across zero TDC scenarios – selected methods

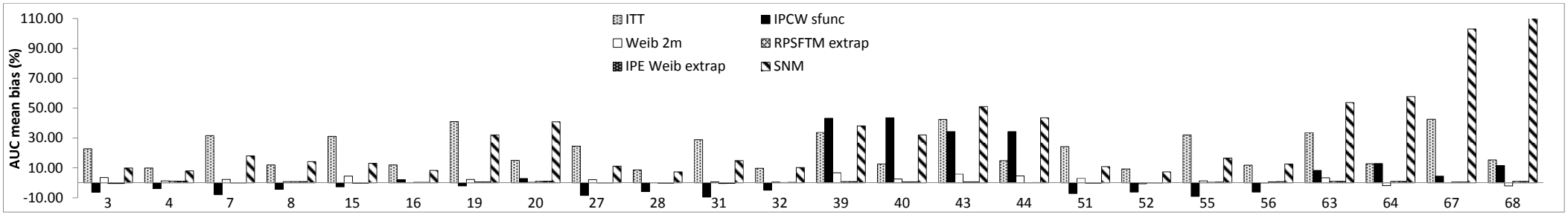


Figure G10: Bias across TDC scenarios – selected methods

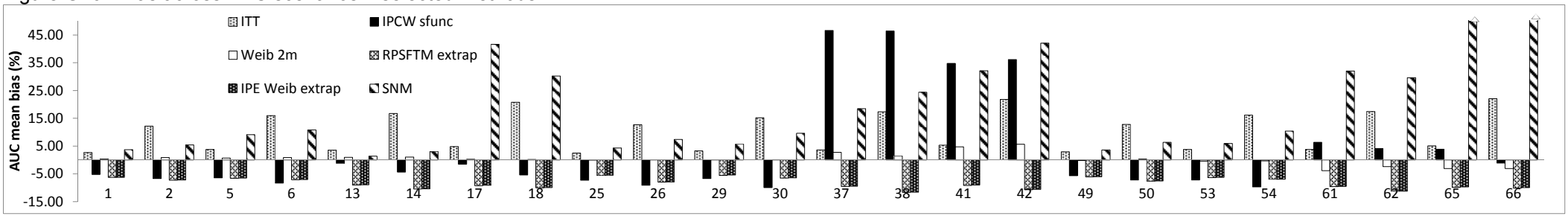
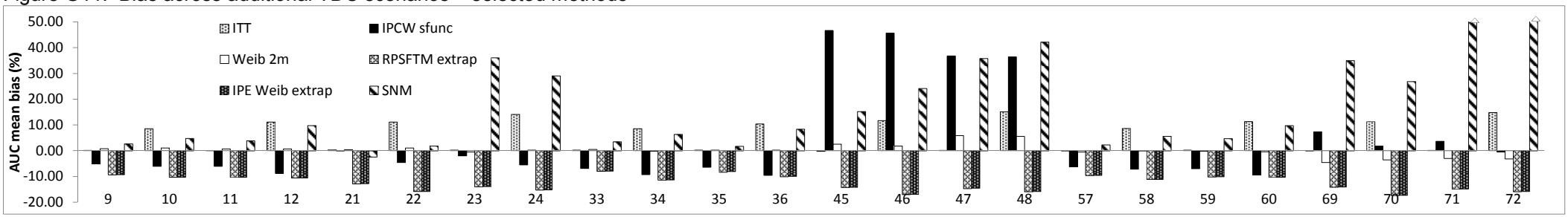


Figure G11: Bias across additional TDC scenarios – selected methods



Appendix H: Coverage

Figure H1: Coverage across zero TDC scenarios – selected methods

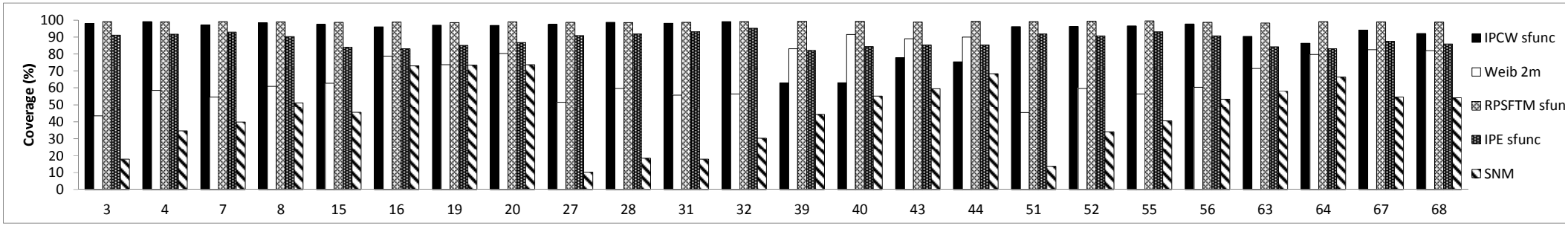


Figure H2: Coverage across TDC scenarios – selected methods

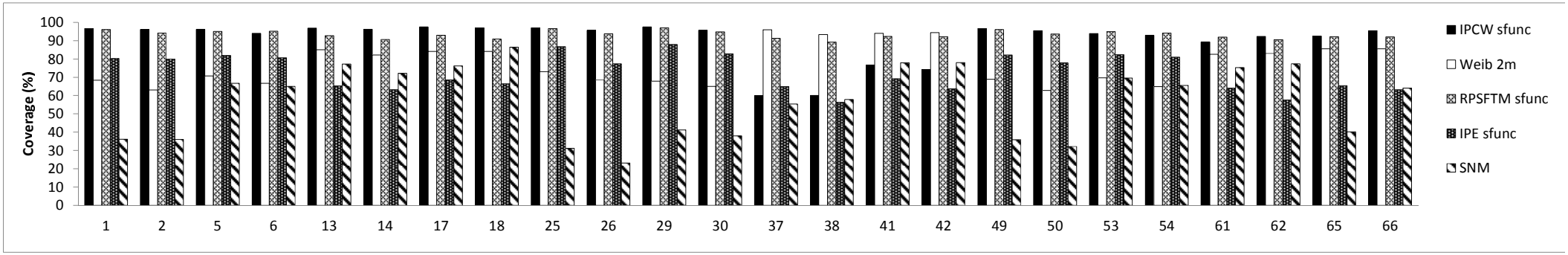
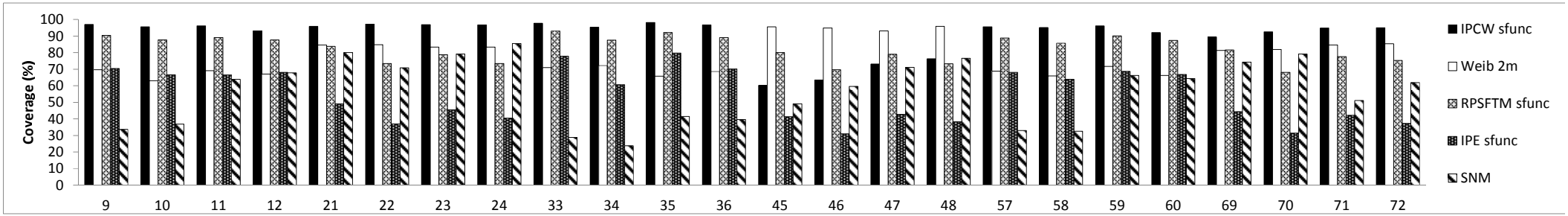


Figure H3: Coverage across additional TDC scenarios – selected methods



Appendix I: MSE

Note some graphs in this section have been truncated in order to remain useful. Truncated bars are signified by an arrow at their peak.

Figure I1: MSE across zero TDC scenarios – selected methods

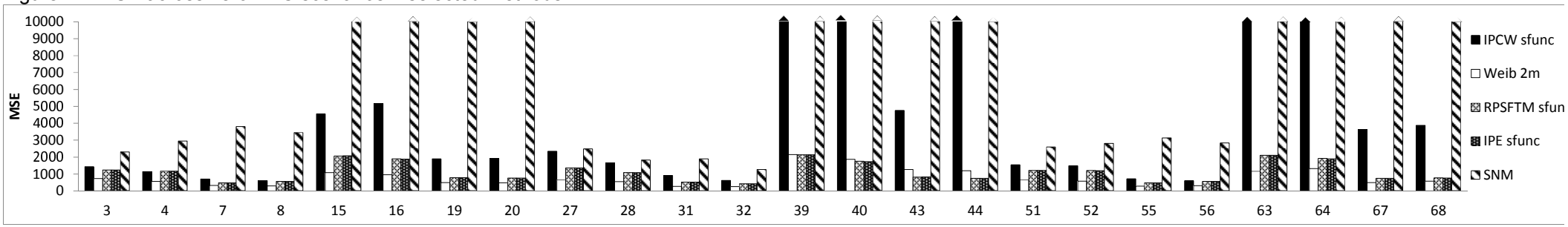


Figure I2: MSE across TDC scenarios – selected methods

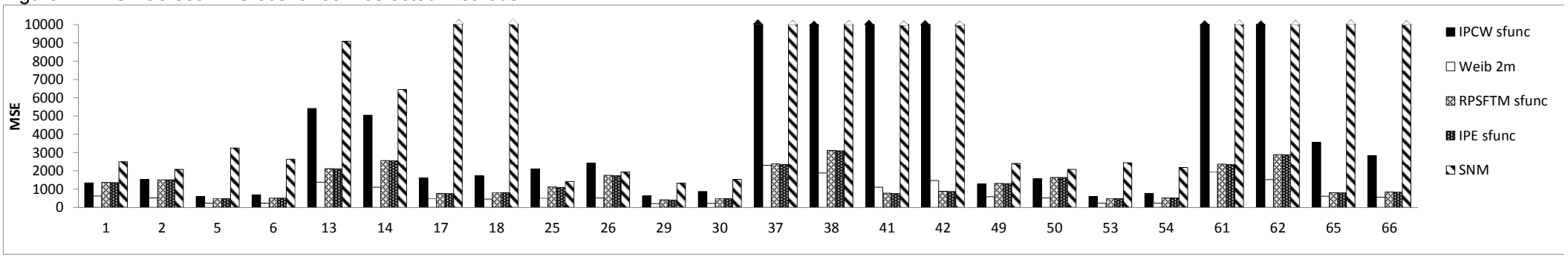


Figure I3: MSE across additional TDC scenarios – selected methods