Assessing methods for dealing with treatment switching in randomised clinical trials

James P Morden*[1,2], Paul C Lambert[2] , Nicholas Latimer[3], Keith R Abrams[2], Allan J Wailoo[3]

[1]ICR-CTSU, The Institute of Cancer Research, Sir Richard Doll Building,Cotswold Road, Sutton, Surrey, UK, SM2 5NG
[2]Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences University of Leicester, 2nd Floor Adrian Building, University Road, Leicester, LE1 7RH
[3]Health Economics and Decision Science, ScHARR, University of Sheffield, Regent Court,30 Regent Street, Sheffield, S1 4DA

Email: James P Morden*- James.Morden@icr.ac.uk; Paul C Lambert - paul.lambert@le.ac.uk; Nicholas Latimer - N.Latimer@sheffield.ac.uk; Keith R Abrams - kra1@le.ac.uk; Allan J Wailoo - a.j.wailoo@sheffield.ac.uk;

*Corresponding author

Abstract

Background: We investigate methods used to analyse the results of clinical trials with survival outcomes in which some patients switch from their allocated treatment to another trial treatment. These included simple methods which are commonly used in medical literature and may be subject to selection bias if patients switching are not typical of the population as a whole. Methods which attempt to adjust the estimated treatment effect, either through adjustment to the hazard ratio or accelerated failure time models, are also considered. A simulation study was conducted to assess the performance of each method in a number of different scenarios.

Results: 16 different scenarios were identified which differed by the proportion of patients switching, underlying prognosis of switchers and size of true treatment effect. 1000 datasets were simulated for each of these and all methods applied. Selection bias was observed in naive methods when the different in survival between switchers and non-switchers were large. A number of methods, particularly the AFT method of Branson and Whitehead were found to give less biased estimates of the true treatment effect in these situations.

Conclusions: Simple methods are often not appropriate to deal with treatment switching. Alternative methods such as the Branson & Whitehead method to adjust for switching should be considered.

Background

Randomized clinical trials (RCTs) are widely used to assess the merits of a new treatment or procedure compared to a control treatment. The outcome for each of the randomised treatment groups are then compared. Survival outcomes are commonly used, with the time to an event such as death or disease progression analysed.

However, in reality it is common for patient to switch from the treatment to which they are randomised. They may switch to the other trial treatment, a non-trial treatment or stop receiving treatment altogether. Switches may occur for a number of reasons, many of which are related to an individual's prognosis or characteristics. Most commonly patients will switch from the control arm to the intervention arm, often as a last resort. A clinician may decided that a patient is responding poorly to their allocated treatment and it is therefore unethical to let them continue on this regime. Switching may also occur in cancer trials when a patient's disease may progress and they therefore switch to an alternative treatment.

An intention-to-treat (ITT) approach is often used where patients are analysed dependent on the treatment they are randomised to, regardless of whether they actually went on to receive this treatment for the entire follow-up period. This pragmatic approach is said to reflect the overall effectiveness of a treatment policy if it were introduced on a wider scale [1]. However, it is often of more interest to measure the true efficacy of a treatment. Efficacy is especially important when assessing the cost-effectiveness of a treatment [2]. There is also no guarantee that departures from allocated treatment observed within a trial setting are representative of what may actually be observed if the treatment was introduced to the general public.

Efficacy is often quantified using a per-protocol (PP) approach which measures how well a patient fares dependent of the treatment they actually receive, regardless of which treatment arm they were randomised to. Patients who switch from their randomised treatment are therefore excluded from the analysis or censored at the time of their switch. This approach can lead to severe selection bias if those excluded differ in prognosis from those retained in the analysis, which is likely in this setting as patients often switch treatments because their condition has deteriorated [3]

An example of the potential dangers of using a per-protocol analysis can be seen in a NICE appraisal of Trastuzumab, a drug for the treatment of metastatic breast cancer [4]. 75% of patients randomised to control treatment eventually switched to the experimental arm. These patients were excluded completely from the analysis and a median survival gain of 17.9 months was found. However, if all control patients had been included, this median survival gain was greatly reduced to just 7 months. The true median survival gain from the treatment is likely to be somewhere between these two values. Because the quality adjusted life year (QALY) gain associated with a new cancer treatment is often driven by the extended period of life, basing a cost effectiveness analysis on a scenario whereby overall survival gains are very likely to have been either under-estimated or over-estimated will cause inaccurate and misleading cost effectiveness results, which could potentially lead to incorrect reimbursement decisions being made.

This issue also arose during the multiple technology appraisal of renal cell carcinoma treatments. In analysis run for NICE by the Decision Support Unit (DSU) the impact of treatment crossover on the estimated incremental cost effectiveness ratio (ICER) is highlighted [5]. For sunitinib an interim analysis of overall survival before treatment crossover was permitted led to an ICER of £59,819 compared to interferon, based on a hazard ratio of 0.65. After treatment crossover was permitted the overall survival hazard ratio increased to 0.82 and other things remaining equal the ICER increased to £118,005. In reality, the ICER is likely to lie somewhere inbetween these values.

Various methods have been proposed to evaluate the true efficacy of a treatment taking into account deviations from the randomised treatment group. The aim of this project is to investigate these methods along with so-called "simple" methods which have been used in existing trial literature. Methods presented include those which adjust the hazard ratio to take switching into account and those which make use of accelerated failure time models to predict how well a patient would have survived if they had remained on the treatment they were allocated to. Section 2 contains an overview of all of these methods. Methods were formally assessed using a simulation study, the design of which can be found in Section 3 and results in Section 4. Section 5 describes the extension of one of the accelerated failure time methods to provide adjusted hazard ratios for ease of interpretation. Section 6 contains a discussion of the findings of this work and future extensions to it.

Overview of Methods

The different methods can be grouped loosely into simple methods (those which are currently widely used), adjusted hazard ratio methods and accelerated failure time model methods.

Simple methods

Various methods have been used in existing literature in situations where patients depart from their randomised treatment. This section will focus on four of these, intention-to-treat, excluding or censoring patients if they switch treatments and modelling treatment as a time-varying covariate.

Intention-to-treat

Many authors take the pragmatic approach and use an intention-to-treat (ITT) analysis. Patients are analysed depending on which treatment arm they were randomised to. The results from an ITT should always be given regardless of whether the effectiveness of the treatment is of interest as it reflects the design of the study. While analysis of this type is perfectly valid, it does tend to underestimate the true efficacy of a treatment [6]. For example, if the experimental treatment truly is superior to the control treatment, and some patients have switched from control to experimental and are therefore receiving the benefits of this, using an ITT analysis will make the treatments appear more similar than they really are. The benefit of this type of analysis is that randomisation balance between groups is maintained, reducing the possibility of bias affecting the results [7–9].

Per-protocol (excluding switchers or censoring at switch)

A per-protocol (PP) or as-treated approach involves analysing patients according to the treatment they actually received rather than that to which they were randomised. This is commonly used to supplement an intention-to-treat analysis [6].

An analysis of this type may involve censoring patients at the point at which they switch, or completely excluded any switching patients from the analysis. Whereas ITT uses randomisation to ensure treatment arms are balanced in all aspects other than treatment, PP analysis may be subject to selection bias as groups may no longer be balanced after a patient is censored or excluded [7]. This type of bias is particularly likely if a patient's probability of switching treatments is strongly related to their underlying prognosis [**?**].

Treatment as a time-varying covariate

Walker et al [10] present a method using a Cox proportional hazards model with treatment as a time-varying covariate to assess the effect of treatment actually recieved by a patient. The model takes the form:

$$\lambda_i(t) = \lambda_0(t)\exp(\beta X_i(t)) \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function and $X_i(t)$ takes a value of zero while a patient is receiving the control and 1 while they are receiving the experimental treatment. However, like the PP methods presented above, this method can break the randomisation balance and is therefore subject to selection bias if switching is related to prognosis [11].

Adjusted hazard ratio methods

Adjusted Cox model (Law and Kaldor, 1996)

Law and Kaldor [12] propose a method of adjusting the standard Cox model to take into account patients who depart from their randomised treatment. The method can be used in situations where patients switch both from control to experimental treatment or in the opposite direction.

The method works on the principle that patients can be divided into four groups depending on their switching pattern. So given an RCT comparing two treatments (A and B), patients are classified as being in Group AA or BB if they were allocated to A or B and did not switch treatments, or to Group AB or BA if they switched from their allocated treatment to the other treatment. The hazard rates in each group are assumed proportional.

A Cox model is then fitted with a time-varying covariate for switching time. Full details of the method can be found in the paper itself [12].

Problems with this method have been raised by White [13]. Patients are grouped as described above according to future events, i.e) a treatment switch which has not yet occurred. So for example, subjects in group AB are said to have a certain hazard function before they switch. However in reality they have a hazard of zero up to the point at which they switch treatment, as they cannot die before this point or they would be in group AA. White states that this is likely to bias the hazard ratio seen towards the null.

Causal proportional hazards estimator (Loeys and Goetghebeur, 2003)

Loeys and Goethebeur [14] present a method for calculating the true treatment efficacy in situations where all patients take their allocated treatment in one arm and compliance is "all-or-nothing" in the other arm. This means that if a patient in this arm switches, the switch is assumed to have happened at time zero, and the patient is assumed to have only received the treatment they switched onto and none of their allocated treatment. The method and its implementation in the Stata package are described further by Kim and White [15].

The authors consider a clinical trial in which patients are randomised to receive either a control treatment or an experimental treatment. The method work on the assumption that all patients in the control arm comply fully, and patients in the experimental arm may either comply fully (complier) or not at all (non-complier). Patients in the control arm are also classed as either being a complier or non-complier depending on how they would have behaved *if* they had been randomised to the experimental arm. The proportion of non-compliers is assumed to be the same in both arms due to randomisation.

The method then makes use of Kaplan-Meier survival estimates and the assumed relationship between control and experimental compliers to find an estimate of the hazard ratio. [14] gives full details of the methodology used.

The all-or nothing compliance assumption may only be appropriate in certain situations, such as a trial to investigate a new screening program where patients may be allocated to attend a screening but may not attend. The method was applied using the Stata program *stcomply* as described by Kim and White [15,16].

Accelerated failure time model methods

The methods in this section make use of accelerated failure time (AFT) models, an alternative form of survival model to the commonly used proportional hazards model. A proportional hazards model assumes that covariates multiply the hazard by some constant, whereas an AFT model assumes that a covariate multiplies the predicted event time by some constant [17].

These methods have been refered to as randomisation-based efficacy estimators (RBEEs) [6] as they compare groups as randomised and therefore are intended to reduce biases which may be introduced by comparing groups as-treated. Note that the adjusted hazard ratio method of Loeys & Goetghebeur [14] described previously (section 2.2.2) is also a RBEEs as it preserves the randomisation balance and the significance level from an ITT analysis.

Rank preserving structural failure time models (Robins and Tsiatis, 1991)

Robins and Tsiatis [18] describe the use of AFT models to estimate the true efficacy of a treatment. A patient's observed event time is related to their counterfactual event time, that which would have been observed for that patient if they had not received any treatment.

Consider a randomised trial with two arms, a control arm (A) and an experimental arm (B). Each patient i has an observed time to event or censoring T_i. $R_i = A$ or B is the patient's randomised treatment arm. Each patient also has a counterfactual event time U_i which is the event time which would have been observed if no treatment had been received. Patients in the control arm who do not switch treatment will have $T_i = U_i$, so their counterfactual event time will be observed. U_i is unobserved for all other patients. The assumption is made that U_i is independent of R_i due to randomisation balance.

Consider the observed event time T_i as being made up of a patient's time on the control treatment T_{Ai} and their time on the experimental treatment T_{Bi}, so $T_i = T_{Ai} + T_{Bi}$. For patients who did not switch treatments, either T_{Ai} or T_{Bi} will be equal to zero. T_i is related to the counterfactual event time U_i by the following causal model:

$$U_i = T_{Ai} + e^{-\psi_0} T_{Bi} \tag{2}$$

e^{ψ_0} is often called the *acceleration factor*, the amount by which a patient's expected time to event is increased by treatment. A value of $e^{\psi_0} > 1$ indicates a beneficial treatment effect whereas $e^{\psi_0} < 1$ suggests treatment has a detrimental effect, increasing the speed at which a patient moves towards their event. e^{ψ_0} is perhaps easier to interpret than $e^{-\psi_0}$ so results will be presented in this form.

By defining a binary process $X_i(t)$ which equals 1 when a patient is on experimental treatment and 0 otherwise, equation (2) can be rewritten as:

$$U_i = \int_0^{T_i} \exp[\psi X_i(t)]dt \tag{3}$$

For a given value of ψ, the hypothesis $\psi_0 = \psi$ can be tested by first calculating $U_i(\psi)$ using equation (2). $Z(\psi)$ is then calculated as the test statistic for the hypothesis $U(\psi) \perp\!\!\!\perp R$.

A number of different tests could be used to calculate $Z(\psi)$, either Wald tests from parametric models such as Weibull and exponential, or non-parametric tests such as the logrank test. The value of ψ for which $Z(\psi)$=0 is taken as the point estimate. This is the value for which U is balanced between treatment arms.

The method has been extended and implemented in Stata (through the *strbee* program) by White et al [19,20]. Define C_i as the administrative censoring time which corresponds to the end of follow-up. Using equation (13), the censoring time for $U_i(\psi)$ is given by:

$$D_i(\psi) = \int_0^{C_i} \exp[\psi X_i(t)]dt \tag{4}$$

Given X_i and therefore D_i may depend on prognosis, censoring of $U_i(\psi)$ is informative. White et al [20] suggest possible bias from this be avoided by recensoring $U_i(\psi)$ as:

$$D_i^*(\psi) = min(C_i, C_i \exp \psi) \tag{5}$$

So if $D_i^*(\psi) < U_i(\psi)$, $U_i(\psi)$ is replaced by $D_i^*(\psi)$.

An interval bisection process can be used to find the point estimate and confidence interval for ψ. Further details of this can be found in the discussion of the *strbee* program [19].

Iterative parameter estimation algorithm (Branson and Whitehead, 2002)

Branson and Whitehead [21] build on the method developed in the previous section by replacing the test-based estimation of ψ with a likelihood-based analysis. Survival times are assumed to have a parametric form, but the relationship between a patient's prognosis and their switching pattern is not modelled.

An iterative parameter estimation (IPE) algorithm is used. This retains all patients to the treatment group to which they were initially randomised. Using the same notation as used in the previous section, consider the model relating counterfactual and observed event times seen previously (equation (2))

An initial estimate for e^{ψ} is obtained by comparing the treatment arms as randomised using an parametric failure time model (equivalent to an intention-to-treat approach). A number of parametric distributions could be chosen for this such as log-logistic, log-normal or gamma. We use a Weibull distribution as it has the advantage of having both AFT model and proportional hazards model parameterisations [17].

Given this initial estimate, the observed survival times of patients who switched from control to experimental treatment are transformed using the current estimate for e^{ψ} and equation (2). Groups are compared again, giving an updated estimate for e^{ψ}. The process is then repeated until the latest value of e^{ψ} becomes sufficiently close to its value from the previous iteration, at which point the process is said to have converged. Further explanation of the algorithm can be found in the original paper [21].

If the algorithm projects a patient's survival time beyond the administrative censoring time C_i, the patient is considered censored and their projected survival time is replaced by C_i. This recensoring is restricted only to patients in the control arm who switch treatments, unlike the recensoring implemented to the Robins and Tsiatis method by White et al [20].

Standard errors can be calculated by either taking the standard error from the final regression in the algorithm or by using bootstrapping [**?**]. The authors discuss how using the standard error from the final regression may give standard errors which are too small. This is because the covariance matrix from the final iteration of the IPE is not does not take into account the fact that control arm patients have had their survival time adjusted by the algorithm.

Parametric randomisation-based methods (Walker et al, 2004)

In the previous two methods ψ is chosen to balance the counterfactual event time U between treatment arms. However as discussed previously and by Robins and Tsiatis [18], these methods can be associated with a loss of information through recensoring and arbitrary differences from the results of ITT analysis. Walker et al [10] present an extension to these semi-parametric methods which involve full parametric modelling of the relationship between U and the treatment a patient actually receives Z. Again we consider a trial with control (A) and experimental (B) arms where some patients who are randomised to control actually switch to receive the experimental treatment at some point during follow-up. Consider U_i as a patient's counterfactual event time and Z_i as the time at which they start receiving experimental treatment. The authors propose specifying a joint parametric model for U_i and Z_i which is made up of three parts:

1. **A causal model relating U_i to a patient's observed failure time T_i.** This is the AFT model seen in previous sections, (equation (2)).
2. **Model for the association between U and Z.** This is a bivariate frailty model. Either a positive stable [22] or gamma [23] frailty are suggested. These models include a parameter ϕ which describes the level of association between U and Z.
3. **Models for the marginal cumulative hazards.** $H_u(u)$ and $H_z(z)$ are the marginal cumulative hazards of U and Z respectively.

Fitting this model using maximum-likelihood techniques would only ensure the original randomisation balance is preserved if all models are correctly specified. Parameter estimates will therefore be very sensitive to inaccuracies in the model specification. To deal with this, the authors suggest an alternative approach to maximum likelihood to estimate parameters. They use an augmented model to maintain the randomisation balance between groups which corresponds to the Cox model test in the semi-parametric approach of Robins & Tsiatis (see section 2.3.1). The model has the form:

$$H_u^*(u) = e^{\rho R} H_u(u) \tag{6}$$

An estimate of ψ can be found so that an estimate of ρ would be equal to zero, indicating there is no relationship between a patient's underlying survival time and the treatment arm they are randomised to so randomisation balance is maintained. Full details of the estimation process are described by Walker et al [10]. The method is implemented in Stata as the *gparm* and *gparmee* programs [16].

Simulation study design

To formally assess the various methods, a simulation study was conducted. Independent datasets were simulated with the true difference between treatments known and each method applied to the data to see how close they came to the truth. Simulated data was designed to reflect data which is obtained from real clinical trials. This section contains details of the design of the simulation study.

Underlying survival times

The starting point for simulating data was to generate a number of patients with an underlying survival time. A sample size of 500 was chosen, with 250 patients allocated each to receive control or experimental treatment. This sample size reflects what is often seen in large cancer trials [4, 24, 25]. Survival times for these patients were then generated from a Weibull distribution as described by Bender et al [26]. The shape parameter γ was set at 0.5 which assumes mortality rate is decreasing over time, a situation often observed in cancer trials [27, 28]. The scale parameter λ was chosen so that approximately 90% of patients who receive no treatment had died after three years of follow-up

Entry and exit times

Patients were assumed to have entered the study at some point during a one-year period, with their entry time generated from a uniform distribution between time zero and 1 year. Patients were then censored at 3 years to represent the end of the follow-up period. Therefore all patients were followed up for between 2 and 3 years, dependent on their entry time, representing what is often seen in a real trial setting.

Patient prognosis

As described previously, bias can often occur when patients with different underlying prognoses have different probabilities of switching between treatment arms. To investigate this, patients were split into two groups, those with good prognosis and those with poor prognosis. The probability of a patient being in the good prognosis group was set at either 30% or 75%. Patients allocated to the good prognosis group were assumed to have their previously generated survival time multiplied by a certain factor. Values of 1.2 and 3 were chosen to represent a small and large difference between prognosis groups. Randomisation should ensure prognosis proportions were balanced between treatment arms.

Switching probability

The probability of a patient switching was then set, dependent on their prognosis group. Only switching from the control to the experimental treatment was considered. The assumption was made that patients in the poor prognosis group were more likely to crossover, as is often the case with the experimental treatment considered as a ”rescue” measure. Two sets of probabilities were considered; probabilities of switching 10% and 25% for good and poor prognosis groups respectively to represent a relatively small proportion of patients switching treatments or 50% and 75% for good and poor groups respectively to represent a trial with a large proportion of control patients switching. These probabilities were then used to generate a binary variable indicating whether or not a patient switches treatments.

Switching time

For patients who switched treatments, a switching time was generated which occurred sometime between their entry into the study and their exit (through either death or censoring). Switching times were generated using a uniform distribution. This assumes that a patient is equally likely to switch at any point between their entry into the study and death or censoring.

Adjusting survival times for treatment received

Given we now have generated a number of patients with an underlying survival time based on their prognosis group, a treatment arm indicator, a switching indicator and switching time for those who switch, the next step is to adjust survival times based on the amount of treatment a patient actually receives. For each patient, survival time is made up of time on control T_{Ai} and time on experimental treatment T_{Bi}. Patients randomised to control who do not switch treatments will have $T_{Bi} = 0$. All patients randomised to experimental treatment will have $T_{Ai} = 0$ as no patients from this arm are allowed to switch treatments. Adjusted patient survival time T_i^* is then calculated using the formula for the causal accelerated failure time model as described by Walker et al [10]:

$$T_i^* = T_{Ai} + e^{\psi} T_{Bi} \tag{7}$$

where e^{ψ} is the true effect of treatment. Patient times are therefore extended using this formula. If a patient's survival time is extended beyond three years they are censored at three years.

Treatment effect

Initial treatment effect hazard ratios of 0.9 and 0.7 were chosen to represent situations with a smaller and larger true difference between treatments, with the experiment treatment considered beneficial.

As the values of λ and γ used to simulate the underlying survival times are known, the hazard ratios β described above can be converted into e^{ψ} form required by equation (7) by using the formula described by Collett [17]:

$$\psi = \frac{-\ln \beta}{\gamma} \tag{8}$$

For example, a hazard ratio of 0.7, with γ=0.5 equates to $\psi = 0.7133$ and therefore $e^{\psi} = 2.04$.

Table 1 gives a summary of all variables considered when simulating patient data and the values chosen for these.

Table 1: Summary of simulation variables

Variable	**Scenarios**	**Details**
Sample size	1	500 patients, 250 in each treatment arm
Weibull shape parameter γ	1	0.5, to represent mortality rate decreasing over time
Weibull scale parameter λ	1	1.33, chosen such that 90% of patients have died after 3 years of follow-up
Probability of patient having good prognosis	2	30% or 75%
Difference in survival between good and poor prognosis groups	2	Survival times of good prognosis group multiplied by a factor of either 1.2 or 3
Probability of switching treatment dependent on prognosis group	2	10% (good prognosis) and 25% (poor prognosis) or 50% (good prognosis) and 75% (poor prognosis)
Switching time	1	Generated from a Uniform distribution
Initial treatment effect	2	Hazard ratio of 0.9 or 0.7

Applying the methods

By considering all possible combinations of the variables described in Table 1, 16 scenarios were identified. For each of these, data was generated as described above, and the various methods applied to this dataset. This process was repeated 1000 times for each scenario. For each method the mean treatment effect $\bar{\hat{\beta}}$ and its standard error $SE(\hat{\beta})$ were calculated. The means of the standard error and 95% confidence limits from each method were also calculated. For the Branson and Whitehead method, standard errors were taken from the final regression of the algorithm rather than bootstrapping as used previously due to the large computing time bootstrapping for each simulated dataset would require.

Performance measures

Measures which can be used to assess the methods presented were calculated as described by Burton et al [29]. The bias of each method δ was calculated as:

$$\delta = \bar{\hat{\beta}} - \beta \tag{9}$$

where β is the true initial treatment effect for that particular scenario.

The mean square error (MSE) is a useful measure of the overall accuracy of a method as it includes both measures of bias and of the variability of estimates given by a method [29]. The MSE is calculated as:

$$MSE = (\bar{\hat{\beta}} - \beta)^2 + (SE(\hat{\beta}))^2 \tag{10}$$

Also calculated was the coverage of each method. This is the defined as the proportion of times the 95% confidence interval for a particular method contains the true initial treatment effect β. Coverage should be approximately equal to 95%, indicating that around 95% of the confidence intervals include the true value.

As some methods may not successfully converge in certain situations, the proportion of times each method successfully gave a parameter estimate was also calculated. Methods which are unsuccessful for a large number of simulated datasets may be of little practical use.

Simulation Results

Table 2 shows details of the parameter values used in each of the 16 scenarios and the table in which the results for this scenario can be found. A selection of results are presented in this section.

For figures in this section, method names were abbreviated as follows: Intention-to-Treat (ITT), Exclude switchers (PP), Censor at switch (CENS), Treatment as time-varying covariate (TVC), Law & Kaldor (LK), Loeys & Goetghebeur (LG), Robins & Tsiatis with logrank test (RT-LR), with Cox test (RT-COX), with exponential test (RT-EXP), with Weibull test (RT-WB), Branson & Whitehead (BW) and Walker et al parametric method (WALK).

Table 2: List of scenarios

Scenario Number	**Treatment effect (HR)**		**% with good prognosis**		**Good prog life**		**Crossover probabilities (good and poor prognosis)**		**Table Number**
	0.9	**0.7**	**30%**	**75%**	×1.2	×3	**10% and 25%**	**50% and 75%**	
1	√		√		√		√		3
2		√	√		√		√		3
3	√			√	√		√		**??**
4		√		√	√		√		**??**
5	√		√		√			√	4
6		√	√		√			√	4
7	√			√	√			√	**??**
8		√		√	√			√	**??**
9	√		√			√	√		5
10		√	√			√	√		5
11	√			√		√	√		**??**
12		√		√		√	√		**??**
13	√		√			√		√	6
14		√	√			√		√	6
15	√			√		√		√	**??**
16		√		√		√		√	**??**

Prognosis and bias

We will first focus on four particular scenarios, 2, 6, 10 and 14. Each of these has 30% of patients with good prognosis, a true treatment difference of $\beta = 0.7$ on the hazard ratio scale or e^{ψ}=2.04 on the AFT scale. The scenarios vary in the difference in lifetime between good and poor prognosis groups, with good prognosis patient's survival multiplied by 1.2 in scenarios 2 and 6 and by 3 in scenarios 10 and 14. The scenarios also differ in the probabilities of switching in good and poor prognosis groups, with probabilities of 10% and 25% respectively in scenarios 2 and 10 and of 50% and 75% respectively in scenarios 6 and 14. Full results from these scenarios can be found in Tables 3, 4, 5 and 6.

Table 3: Scenarios 1 & 2: 30% good prognosis, ×1.2 good prognosis lifetime, 10% and 25% switching probabilities, 100% treatment effect for switchers, uniform distribution for switching times

True HR and e^{ψ}	Method	Mean estimate	Mean SE	SE of mean	95% Confidence interval		Bias	MSE	Coverage (%)	Successful estimation (%)
					Lower	Upper				
0.9 & 1.23	**Hazard ratio methods**									
	ITT	0.9120	0.0884	0.0882	0.7542	1.1027	0.0120	0.0079	94.6	100.0
	Exclude switchers	0.9053	0.0932	0.0929	0.7399	1.1076	0.0053	0.0087	94.3	100.0
	Censor at switch	1.0589	0.1091	0.1086	0.8653	1.2957	0.1589	0.0370	67.5	100.0
	Time-dependent covariate	1.1998	0.1189	0.1211	0.9880	1.4570	0.2998	0.1046	19.1	100.0
	Law and Kaldor	0.9168	0.1136	0.1132	0.7192	1.1688	0.0168	0.0131	94.7	100.0
	Loeys and Goethebeur	0.8900	-	0.1086	0.6999	1.1354	-0.0100	0.0119	94.8	100.0
	AFT methods									
	ITT	1.2357	0.2392	0.2423	0.8457	1.8060	0.0011	0.0587	94.6	100.0
	Exclude switchers	1.2586	0.0932	0.2616	0.8415	1.8830	0.0240	0.0690	94.3	100.0
	Censor at switch	0.9214	0.1864	0.1897	0.6199	1.3700	-0.3132	0.1341	66.3	100.0
	Robins and Tsiatis - Logrank	1.2715	-	0.2852	0.8244	1.9604	0.0370	0.0827	94.7	100.0
	Robins and Tsiatis - Cox	1.2703	-	0.2806	0.8199	1.9888	0.0357	0.0800	95.2	97.2
	Robins and Tsiatis - Exponential	1.2781	-	0.2820	0.9712	1.7776	0.0436	0.0814	83.0	99.7
	Robins and Tsiatis - Weibull	1.2714	-	0.2845	0.8278	1.9933	0.0369	0.0823	95.0	99.7
	Branson and Whitehead	1.2681	0.2455	0.2745	0.8678	1.8536	0.0335	0.0765	92.7	100.0
	Walker et al	2.2108	1.1659	1.1869	0.8323	1218.7190	0.9763	2.3617	76.6	99.4
0.7 & 2.04	**Hazard ratio methods**									
	ITT	0.7315	0.0731	0.0743	0.6014	0.8897	0.0315	0.0065	93.6	100.0
	Exclude switchers	0.7050	0.0744	0.0776	0.5733	0.8669	0.0050	0.0060	94.3	100.0
	Censor at switch	0.8215	0.0868	0.0886	0.6678	1.0106	0.1215	0.0226	69.2	100.0
	Time-dependent covariate	0.9364	0.0950	0.0982	0.7675	1.1424	0.2364	0.0655	18.8	100.0
	Law and Kaldor	0.7376	0.0957	0.0968	0.5720	0.9511	0.0376	0.0108	93.6	100.0
	Loeys and Goethebeur	0.6733	-	0.0876	0.5220	0.8672	-0.0267	0.0084	93.2	100.0
	AFT methods									
	ITT	1.9259	0.3858	0.3992	1.3007	2.8524	-0.1150	0.1726	93.5	100.0
	Exclude switchers	2.0825	0.0744	0.4676	1.3766	3.1516	0.0417	0.2204	94.3	100.0
	Censor at switch	1.5194	0.3140	0.3241	1.0136	2.2786	-0.5214	0.3769	66.5	100.0
	Robins and Tsiatis - Logrank	2.1014	-	0.5024	1.3483	3.3204	0.0606	0.2561	94.5	100.0
	Robins and Tsiatis - Cox	2.0969	-	0.4977	1.3303	73.7821	0.0561	0.2508	94.9	93.5
	Robins and Tsiatis - Exponential	2.1041	-	0.5090	1.4957	3.1128	0.0633	0.2631	87.1	100.0
	Robins and Tsiatis - Weibull	2.1017	-	0.5024	1.3497	3.4566	0.0609	0.2561	94.9	100.0
	Branson and Whitehead	2.0889	0.4188	0.4770	1.4104	3.0949	0.0481	0.2299	92.2	100.0
	Walker et al	3.6507	1.8935	1.5878	1.3852	92.4403	1.6099	5.1127	79.5	93.1

Table 4: Scenarios 5 & 6: 30% good prognosis, ×1.2 good prognosis lifetime, 50% and 75% switching probabilities, 100% treatment effect for switchers, uniform distribution for switching times

True HR and e^{ψ}	Method	Mean estimate	Mean SE	SE of mean	95% Confidence interval Lower	Upper	Bias	MSE	Coverage (%)	Successful estimation (%)
0.9 & 1.23	**Hazard ratio methods**									
	ITT	0.9364	0.0909	0.0909	0.7741	1.1328	0.0364	0.0096	93.6	100.0
	Exclude switchers	0.9204	0.1275	0.1278	0.7016	1.2074	0.0204	0.0167	96.0	100.0
	Censor at switch	2.1663	0.3018	0.2975	1.6488	2.8466	1.2663	1.6921	0.0	100.0
	Time-dependent covariate	3.0639	0.4065	0.4054	2.3625	3.9740	2.1639	4.8466	0.0	100.0
	Law and Kaldor	0.9396	0.1168	0.1144	0.7364	1.1987	0.0396	0.0147	94.2	100.0
	Loeys and Goethebeur	0.8435	-	0.2725	0.4578	1.8486	-0.0565	0.0774	92.8	99.9
	AFT methods									
	ITT	1.1732	0.2280	0.2311	0.8018	1.7173	-0.0613	0.0572	93.2	100.0
	Exclude switchers	1.2492	0.1275	0.3479	0.7276	2.1475	0.0146	0.1212	95.9	100.0
	Censor at switch	0.2373	0.0665	0.0663	0.1371	0.4115	-0.9973	0.9989	0.0	100.0
	Robins and Tsiatis - Logrank	1.2882	-	0.3862	0.6906	2.2552	0.0536	0.1521	93.8	100.0
	Robins and Tsiatis - Cox	1.2855	-	0.3874	0.6868	2.2531	0.0509	0.1527	93.9	96.3
	Robins and Tsiatis - Exponential	1.3036	-	0.3693	0.9172	1.9617	0.0691	0.1412	83.6	100.0
	Robins and Tsiatis - Weibull	1.2879	-	0.3857	0.6988	2.2550	0.0534	0.1516	94.5	100.0
	Branson and Whitehead	1.2841	0.2497	0.3619	0.8772	1.8802	0.0495	0.1334	83.7	100.0
	Walker et al	2.1037	0.7802	0.7791	1.0276	4.4146	0.8692	1.3624	70.0	99.8
0.7 & 2.04	**Hazard ratio methods**									
	ITT	0.8073	0.0812	0.0814	0.6629	0.9833	0.1073	0.0181	71.5	100.0
	Exclude switchers	0.7179	0.1009	0.1019	0.5451	0.9457	0.0179	0.0107	94.8	100.0
	Censor at switch	1.6825	0.2388	0.2454	1.2740	2.2223	0.9825	1.0256	0.0	100.0
	Time-dependent covariate	2.4211	0.3257	0.3396	1.8602	3.1516	1.7211	3.0776	0.0	100.0
	Law and Kaldor	0.8110	0.1064	0.1046	0.6271	1.0488	0.1110	0.0233	81.2	100.0
	Loeys and Goethebeur	0.5248	-	0.1728	0.2544	1.0233	-0.1752	0.0606	83.8	99.0
	AFT methods									
	ITT	1.5845	0.3208	0.3294	1.0657	2.3569	-0.4563	0.3167	72.2	100.0
	Exclude switchers	2.0610	0.1009	0.6062	1.1886	3.5793	0.0202	0.3679	94.4	100.0
	Censor at switch	0.3894	0.1088	0.1126	0.2254	0.6742	-1.6514	2.7398	0.0	100.0
	Robins and Tsiatis - Logrank	2.1120	-	0.6624	1.1204	3.8159	0.0712	0.4439	94.8	100.0
	Robins and Tsiatis - Cox	2.1062	-	0.6612	1.1095	3.8198	0.0654	0.4415	95.1	92.6
	Robins and Tsiatis - Exponential	2.1121	-	0.6636	1.3323	3.3923	0.0713	0.4455	87.4	100.0
	Robins and Tsiatis - Weibull	2.1128	-	0.6631	1.1249	3.8301	0.0719	0.4449	94.9	100.0
	Branson and Whitehead	2.0536	0.4177	0.6096	1.3787	3.0601	0.0127	0.3718	83.0	100.0
	Walker et al	3.6581	1.3293	1.2498	1.8098	7.5537	1.6172	4.1774	63.9	98.0

Table 5: Scenarios 9 & 10: 30% good prognosis, ×3 good prognosis lifetime, 10% and 25% switching probabilities, 100% treatment effect for switchers, uniform distribution for switching times

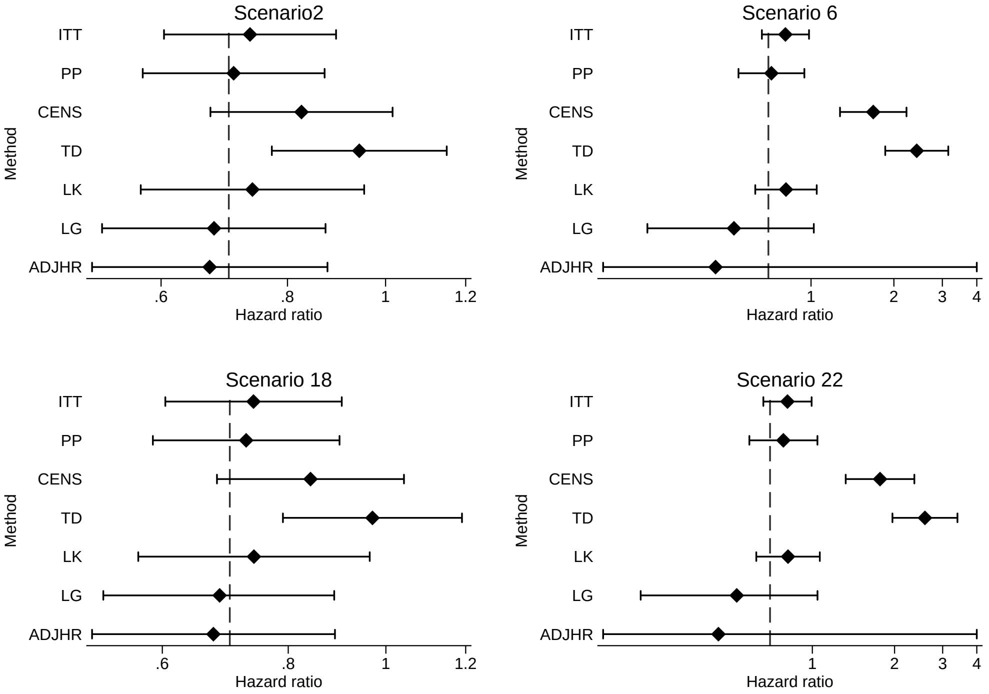
True HR and e^{ψ}	Method	Mean estimate	Mean SE	SE of mean	95% Confidence interval		Bias	MSE	Coverage (%)	Successful estimation (%)
					Lower	Upper				
0.9 & 1.23	**Hazard ratio methods**									
	ITT	0.9136	0.0911	0.0926	0.7515	1.1107	0.0136	0.0088	94.8	100.0
	Exclude switchers	0.9235	0.0979	0.0968	0.7504	1.1367	0.0235	0.0099	94.6	100.0
	Censor at switch	1.0721	0.1137	0.1128	0.8709	1.3197	0.1721	0.0424	63.7	100.0
	Time-dependent covariate	1.2253	0.1249	0.1254	1.0035	1.4962	0.3253	0.1215	15.2	100.0
	Law and Kaldor	0.9121	0.1173	0.1223	0.7088	1.1737	0.0121	0.0151	94.1	100.0
	Loeys and Goethebeur	0.8933	-	0.1137	0.6975	1.1444	-0.0067	0.0130	95.3	100.0
	AFT methods									
	ITT	1.2399	0.2523	0.2519	0.8323	1.8479	0.0054	0.0635	94.7	100.0
	Exclude switchers	1.2159	0.0979	0.2580	0.7959	1.8583	-0.0187	0.0669	94.9	100.0
	Censor at switch	0.8983	0.1911	0.1890	0.5922	1.3632	-0.3363	0.1488	64.4	100.0
	Robins and Tsiatis - Logrank	1.2838	-	0.3018	0.8085	2.0389	0.0492	0.0935	94.6	100.0
	Robins and Tsiatis - Cox	1.2854	-	0.3020	0.8073	2.0820	0.0508	0.0938	94.7	95.2
	Robins and Tsiatis - Exponential	1.2895	-	0.2934	0.9593	1.8561	0.0549	0.0891	84.8	99.8
	Robins and Tsiatis - Weibull	1.2836	-	0.3003	0.8153	2.0802	0.0490	0.0926	94.5	99.8
	Branson and Whitehead	1.2713	0.2589	0.2823	0.8530	1.8954	0.0368	0.0810	92.9	100.0
	Walker et al	2.9706	1.4458	1.4409	1.1904	20.7004	1.7360	5.0898	57.0	98.6
0.7 & 2.04	**Hazard ratio methods**									
	ITT	0.7390	0.0760	0.0777	0.6041	0.9041	0.0390	0.0076	92.0	100.0
	Exclude switchers	0.7267	0.0791	0.0800	0.5872	0.8995	0.0267	0.0071	93.8	100.0
	Censor at switch	0.8418	0.0918	0.0938	0.6798	1.0422	0.1418	0.0289	60.1	100.0
	Time-dependent covariate	0.9698	0.1012	0.1054	0.7904	1.1899	0.2698	0.0839	14.8	100.0
	Law and Kaldor	0.7397	0.0998	0.1027	0.5679	0.9637	0.0397	0.0121	92.8	100.0
	Loeys and Goethebeur	0.6840	-	0.0908	0.5243	0.8886	-0.0160	0.0085	95.5	100.0
	AFT methods									
	ITT	1.9110	0.4010	0.4163	1.2669	2.8838	-0.1298	0.1901	92.5	100.0
	Exclude switchers	1.9836	0.0791	0.4501	1.2841	3.0655	-0.0573	0.2059	94.7	100.0
	Censor at switch	1.4612	0.3165	0.3284	0.9560	2.2344	-0.5796	0.4438	59.4	100.0
	Robins and Tsiatis - Logrank	2.1150	-	0.5430	1.3172	3.4305	0.0742	0.3004	94.7	100.0
	Robins and Tsiatis - Cox	2.1140	-	0.5395	1.2935	27.2997	0.0732	0.2964	94.9	92.2
	Robins and Tsiatis - Exponential	2.1112	-	0.5439	1.4578	3.1966	0.0704	0.3008	87.4	100.0
	Robins and Tsiatis - Weibull	2.1150	-	0.5434	1.3189	3.5350	0.0742	0.3008	94.8	100.0
	Branson and Whitehead	2.0643	0.4343	0.4923	1.3670	3.1186	0.0235	0.2430	92.2	100.0
	Walker et al	4.1987	2.1153	1.6132	1.6252	17.0861	2.1579	7.2589	68.2	75.4

Table 6: Scenarios 13 & 14: 30% good prognosis, ×3 good prognosis lifetime, 50% and 75% switching probabilities, 100% treatment effect for switchers, uniform distribution for switching times

True HR and e^{ψ}	**Method**	**Mean estimate**	**Mean SE**	**SE of mean**	**95% Confidence interval Lower**	**Upper**	**Bias**	**MSE**	**Coverage (%)**	**Successful estimation (%)**
0.9 & 1.23	**Hazard ratio methods**									
	ITT	0.9406	0.0940	0.0965	0.7733	1.1442	0.0406	0.0110	92.2	100.0
	Exclude switchers	1.0018	0.1448	0.1472	0.7547	1.3300	0.1018	0.0320	89.6	100.0
	Censor at switch	2.2865	0.3323	0.3282	1.7198	3.0403	1.3865	2.0301	0.0	100.0
	Time-dependent covariate	3.2815	0.4544	0.4520	2.5017	4.3050	2.3815	5.8757	0.0	100.0
	Law and Kaldor	0.9508	0.1230	0.1263	0.7379	1.2251	0.0508	0.0185	93.7	100.0
	Loeys and Goethebeur	0.8548	-	0.2874	0.4414	1.7657	-0.0452	0.0847	93.6	99.9
	AFT methods									
	ITT	1.1696	0.2386	0.2446	0.7843	1.7447	-0.0650	0.0640	92.6	100.0
	Exclude switchers	1.0631	0.1448	0.3143	0.5983	1.8922	-0.1715	0.1282	90.0	100.0
	Censor at switch	0.2100	0.0626	0.0640	0.1172	0.3769	-1.0246	1.0539	0.0	100.0
	Robins and Tsiatis - Logrank	1.2876	-	0.4161	0.6611	2.3440	0.0530	0.1759	94.4	100.0
	Robins and Tsiatis - Cox	1.2835	-	0.4156	0.6581	2.3451	0.0489	0.1751	94.4	95.9
	Robins and Tsiatis - Exponential	1.3120	-	0.3923	0.9033	2.0554	0.0774	0.1599	84.0	99.7
	Robins and Tsiatis - Weibull	1.2898	-	0.4125	0.6781	2.3512	0.0553	0.1732	94.5	99.8
	Branson and Whitehead	1.2789	0.2613	0.3755	0.8571	1.9090	0.0443	0.1429	82.6	100.0
	Walker et al	2.6653	0.9576	0.9545	1.3297	5.5524	1.4307	2.9579	46.7	99.9
0.7 & 2.04	**Hazard ratio methods**									
	ITT	0.8109	0.0842	0.0843	0.6616	0.9939	0.1109	0.0194	71.4	100.0
	Exclude switchers	0.7834	0.1147	0.1158	0.5880	1.0437	0.0834	0.0204	89.7	100.0
	Censor at switch	1.7695	0.2612	0.2559	1.3250	2.3634	1.0695	1.2092	0.0	100.0
	Time-dependent covariate	2.5841	0.3613	0.3598	1.9647	3.3991	1.8841	3.6791	0.0	100.0
	Law and Kaldor	0.8147	0.1113	0.1125	0.6233	1.0649	0.1147	0.0258	81.4	100.0
	Loeys and Goethebeur	0.5287	-	0.1897	0.2352	1.0444	-0.1713	0.0653	86.9	96.9
	AFT methods									
	ITT	1.5826	0.3351	0.3379	1.0452	2.3973	-0.4582	0.3242	72.8	100.0
	Exclude switchers	1.7560	0.1147	0.5427	0.9817	3.1462	-0.2849	0.3757	90.5	100.0
	Censor at switch	0.3494	0.1032	0.1020	0.1961	0.6240	-1.6914	2.8712	0.0	100.0
	Robins and Tsiatis - Logrank	2.1469	-	0.7376	1.0853	4.0598	0.1061	0.5552	95.7	100.0
	Robins and Tsiatis - Cox	2.1363	-	0.7256	1.0640	4.0757	0.0955	0.5356	95.9	92.4
	Robins and Tsiatis - Exponential	2.1549	-	0.7419	1.3002	3.6453	0.1140	0.5635	88.4	100.0
	Robins and Tsiatis - Weibull	2.1472	-	0.7359	1.0961	4.0901	0.1064	0.5528	95.6	100.0
	Branson and Whitehead	2.0328	0.4340	0.6210	1.3380	3.0899	-0.0080	0.3857	82.6	100.0
	Walker et al	4.2491	1.5202	1.2616	2.1208	8.6357	2.2083	6.4681	45.3	88.3

Figures 1 and 2 show mean estimates and confidence intervals for adjusted hazard ratio and AFT methods from scenarios 2, 6, 10 and 14.

Figure 1: Mean estimates and confidence limits for adjusted hazard ratio methods from Scenarios 2, 6, 10 and 14

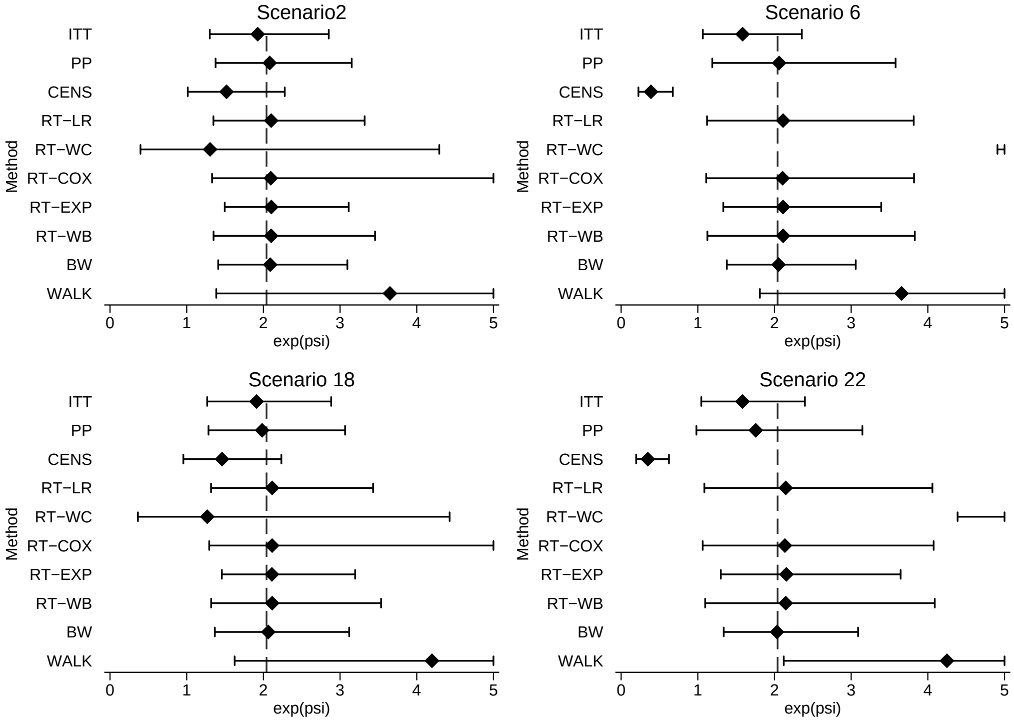


Note: Mean upper confidence limits truncated at $\beta = 4$. Vertical lines show true treatment effect ($\beta = 0.7$)

As expected, the ITT approach underestimated the true treatment effect in each of these four scenarios. This underestimation was relatively small in the scenarios with a small proportion of switchers (2 and 10), around 0.03 - 0.04 on the hazard ratio scale in both cases. This increased to around 0.11 in scenarios 6 and 14 with a large proportion of control patients switching.

Excluding switchers from the analysis gave relatively small bias scenarios 2, 6 and 10. However in scenario 14 where the difference between good and poor prognosis groups and the proportion of switchers were both large, significant bias was seen (0.08 on the hazard ratio scale). The results from this approach are perhaps better than expected with many estimates very close to the true treatment effect, particularly in scenarios where only a small proportion of patients switch treatments. This is possibly explained by the fact that patients who switch treatments have a number of mechanisms acting on them which might cancel each other out. This will be investigated further by looking at differences in scenarios with a smaller and larger

Figure 2: Mean estimates and confidence limits for AFT methods from Scenarios 2, 6, 10 and 14



Note: Mean upper confidence limits truncated at $e^{\psi} = 5$ Vertical lines show true treatment effect ($e^{\psi} = 2.04$)

true treatment effect in the next section.

Perhaps the most striking result from these scenarios is the methods with give particularly large biases, suggesting they are very sensitive to the differences in prognosis between switchers and non-switchers. Of the hazard ratio methods, censoring patients at their switch and considering treatment as a time-dependent covariate both produce large biases, particularly when a large proportion of patients switched treatments (Scenarios 6 and 14) with mean hazard ratio estimates of 1.68 and 1.77 for censoring at switch and 2.42 and 2.58 for treatment as a time-varying covariate. These large biases are reflective of what was seen throughout the simulation study for these methods and suggest they may be inappropriate for use due their large sensitivity to even a relatively weak relationship between switching and prognosis. The parametric method of Walker et al overestimated the true treatment effect in all four scenarios presented here. This overestimation was particularly significant in scenarios with a large difference in lifetime between good and poor prognosis groups (10 and 14), with mean treatment effects of 4.20 and 4.25 over double the true treatment effect of 2.04.

The Law & Kaldor and Loeys & Goetghebeur methods both gave biased estimates in these four scenarios. These biases were particularly large in scenarios with a high proportion of switchers (6 and 14). The Law & Kaldor method seems to underestimate the true treatment effect in all scenarios which may also be due to the way in which the method conditions on future events as described by White [13]. Therefore the assumptions made for this method are incorrect and biases given are likely to be less predictable for a real dataset. The Loeys & Goetghebeur method consistently overestimates the true treatment effect which is perhaps surprising given the method makes the assumption of all-or-nothing compliance, and therefore assumes that a switching patient receives more of the experimental treatment than they actually do. This means that any positive treatment effect seen will actually be due to a smaller amount of treatment than accounted for by the method, so an underestimation of the true treatment effect might be expected.

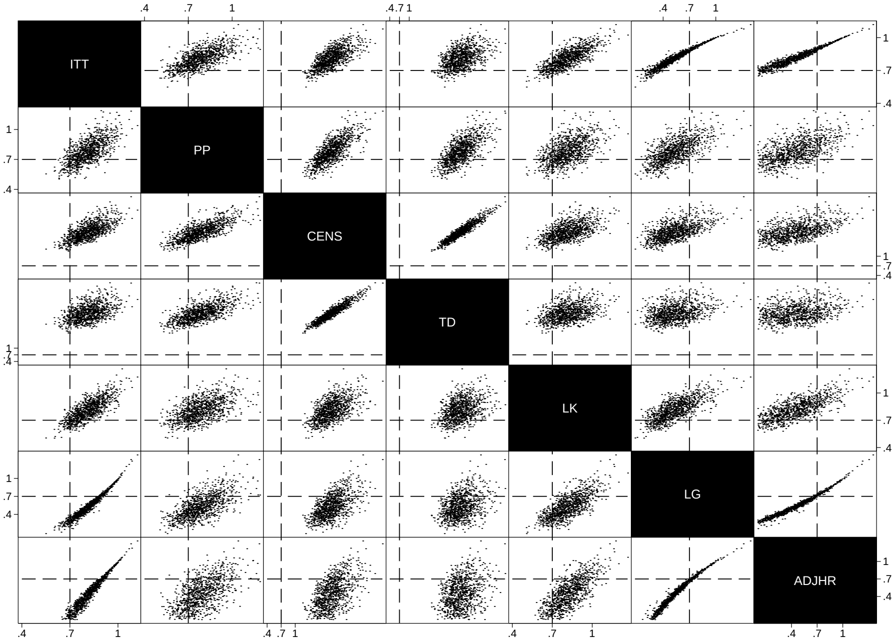
The Robins and Tsiatis method when used with all tests gave very similar mean estimates of e^{ψ}, not differing by any more than 0.02 in these four scenarios. In all cases the mean estimate of e^{ψ} was greater than the true treatment effect of 2.04, suggesting the method is consistently over-adjusting for treatment switching. The mean upper confidence limits given by the Cox test method were very erratic, suggesting they were being unduly influenced by a few large values. There were also some estimation problems with this method, particularly in scenario 14 with 7.6% of simulations unsuccessful when estimating either e^{ψ} or its upper or lower confidence limits.

Very small biases were observed from the Branson and Whitehead method, less than any other AFT method in scenarios 6, 10 and 14. The method also appears to be very robust to more extreme simulated datasets, with 100% successful estimation. Coverage for this method was lower than expected, as low as 82.6% in scenario 14. However as discussed previously, standard errors calculated from the final regression in the algorithm tend to be too small, giving unduly narrow confidence intervals and therefore lower coverage.

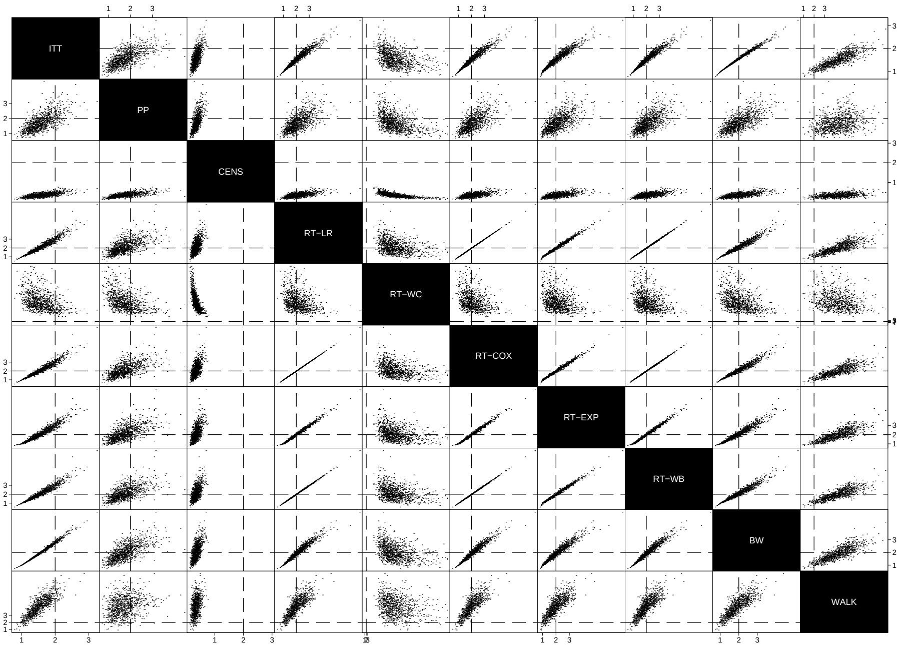
The relationships between point estimates from each method in scenario 14 were further investigated through pairwise scatter plots which can be seen in Figure 3. Vertical and horizontal reference lines show the true treatment effect of $\beta = 0.7$ for adjusted hazard ratio methods or $e^{\psi} = 2.04$ for AFT methods. The relationship between ITT and PP estimates appears to be fairly weak, reflecting the unpredictability of estimates due to biases in this particular scenario. The plots also further illustrate dilution of the true treatment effect when analysing patients as-randomised.

The scatter plot for AFT methods shows the strong relationship between estimates from the Robins & Tsiatis method when using logrank, Cox, exponential or Weibull tests. Realtionships between these

Figure 3: Scatter plot matrix of point estimates from Scenario 14
Adjusted HR methods



AFT methods



estimates and those from the Branson & Whitehead method are also strong, although less so than between the Robins & Tsiatis methods themselves. This is to be expected as the model used by Branson & Whitehead takes the same form as that presented by Robins & Tsiatis, differing only by the way in which the estimate of ψ is found.

Scatter plots for scenarios 2, 6 and 10 showed similar relationships between parameter estimates.

Size of true treatment effect

All scenarios focussed on up to this point have had a large true treatment effect (a hazard ratio β=0.7 or e^{ψ}=2.04). As seen previously, biases seen from excluding all switching patients from the analysis were perhaps not as large as expected. The way in which simulations were set up meant that patients who switch treatments should in general have worse prognosis than those who do not, so excluding these patients from the analysis should make the control group have better survival in general and therefore reduce the difference between and experimental groups observed. However these switching patients also go on to receive a beneficial treatment, perhaps meaning their survival is approximately similar to the control patients who do not switch treatments. If this was the case, excluding these patients would have a relatively small effect on the estimate of the true treatment effect.

To investigate the competing factors acting upon patients who switch treatments in these simulations, we consider scenarios 9 and 13, which are identical to scenarios 10 and 14 respectively except with a smaller true treatment effect of β=0.9 or e^{ψ}=1.23. Scenarios 9 and 10 have probabilities of 10% and 25% of switching treatments in good and poor prognosis groups whereas 13 and 14 have switching probabilities of 50% and 75%. Full details of these scenarios can be found in Table 2. Full results can be found in Tables 5 and 6.

In general, biases observed were greater in scenarios with a larger true treatment effect than a small effect. A notable exception to this can be seen when comparing scenarios 13 and 14 (Table 6). The bias when excluding switchers was greater in scenario 13 with a small treatment effect. This may be because patients in this scenario who switch treatment have worse prognosis but this is ”corrected” to a lesser extent by the treatment they switch onto, making the control arm switchers and non-switchers less similar than in scenario 14 with a larger true treatment effect.

The Branson & Whitehead method also seems to have larger bias in scenarios with a smaller treatment effect. However these biases are still small, with the mean estimate of e^{ψ} closer to the true value than when excluding switchers in both scenarios 13 and 14. There also appears to be a greater difference between estimates given by the various Robins & Tsiatis methods when the true treatment effect is smaller as in scenario 13, although estimates are still strongly related.

Successful estimation

Most of the methods investigated successfully gave an estimate of the treatment effect in all scenarios. However some of the methods experienced problems in certain situations.

The Walker et al parametric method was particulary unsuccessful in scenarios with a large difference in lifetime between good and poor prognosis groups and a large true treatment effect, most notably in scenario 12 where the method was successful for only 43.9% of simulated datasets. This is further evidence that the method may not be suitable for use, especially given the true treatment effect would not be known in real life.

Some estimation problems were also seen with the Robins & Tsiatis methods when used with a Cox, Weibull or exponential test. Given the similarities between estimates when using all tests, the logrank test would seem to be the most appropriate choice for this method as it was 100% successful for all scenarios.

Extension of the Branson & Whitehead method

As seen previously, the method of Branson & Whitehead performed well, giving particularly small biases in scenarios with a large difference in lifetime between good and poor prognosis groups and a large proportion of switchers, scenarios which other methods gave very biased estimates for (see Tables 4, 5 and 6).

One of the limitations of this method and its practical use is that estimates are given in the AFT model form which is less commonly seen in medical literature than hazard ratios from a proportional hazards model [**?**]. However as seen previously, if the shape parameter of the Weibull model γ is known, hazard ratios can be converted to the AFT parameter ψ (see equation (8) in section 3.1)

Rearranging (8) gives the following expression for the hazard ratio β in terms of ψ and γ:

$$\beta = \exp(-\gamma\psi) \tag{11}$$

By taking the value of γ estimated in the final iteration of the IPE algorithm, a hazard ratio β can be estimated from the method using equation (11). The standard error of β can be calculated using the Delta method as described by Collett [17]. However these standard errors are likely to be too small as the standard errors of ψ and γ from which they are calculated are also too small, as described previosuly. Note that this conversion to a hazard ratio would not be possible for the other AFT methods presented here as they do not directly estimate a shape parameter gamma from the data.

To investigate this extension to the Branson and Whitehead method further, simulations for the scenarios focused on previously (2, 6, 10 and 14) were repeated, with γ estimated from the last iteration of the Branson & Whitehead method and used to calculate a hazard ratio and its corresponding standard error as described above. This was compared to hazard ratios from both intention-to-treat and per-protocol approaches for the same simulated data. Table 7 shows mean estimates, bias and the mean standard error for each of the four scenarios.

Table 7: Comparison of mean hazard ratios from the Branson & Whitehead method and ITT and PP approaches

Scenario	**Method**	**Mean HR**	**Bias**	**Mean SE**
2 *(×1.2 Good prognosis lifetime, 10% and 25% switching probabilities)*	ITT	0.7346	0.0346	0.0734
	Exclude switchers	0.7071	0.0071	0.0746
	Branson and Whitehead	0.7077	0.0077	0.0708
6 *(×1.2 Good prognosis lifetime, 50% and 75% switching probabilities)*	ITT	0.8030	0.1030	0.0808
	Exclude switchers	0.7153	0.0153	0.1004
	Branson and Whitehead	0.7172	0.0172	0.0724
10 *(×3 Good prognosis lifetime, 10% and 25% switching probabilities)*	ITT	0.7411	0.0411	0.0763
	Exclude switchers	0.7280	0.0280	0.0793
	Branson and Whitehead	0.7165	0.0165	0.0738
14 *(×3 Good prognosis lifetime, 50% and 75% switching probabilities)*	ITT	0.8121	0.1121	0.0843
	Exclude switchers	0.7810	0.0810	0.1142
	Branson and Whitehead	0.7325	0.0325	0.0762

As seen previously, estimates from the ITT approach are biased towards the null in all four scenarios. This bias is particularly large in scenarios 6 and 14 which have a higher proportion of patients switching from the control arm.

There is very little difference between the mean hazard ratios for the PP and Branson & Whitehead method in scenarios 2 and 6, with the PP approach giving relatively unbiased estimates due to the small difference in lifetime between good and poor prognosis patients. However, when this difference is increased in scenarios 10 and 14, the bias from the PP method increases, most notably in scenario 14 where the difference between prognosis groups is coupled with a large proportion of patients switching.

The Branson & Whitehead method gives estimates close to the true treatment effect for all four scenarios. The method copes particularly well with the large potential biases in scenario 14, giving a mean hazard ratio of 0.73 compared to 0.78 and 0.81 from the PP and ITT approaches respectively.

The Branson & Whitehead method seems to be robust and to correct for treatment switching most successfully of all methods investigated in this report in situations where a patient's switching pattern is strongly related to their prognosis. The fact that the method can give hazard ratios providing γ is estimated from the final iteration of the algorithm is a further advantage if the method were to be more widely used in the analysis of clinical trials.

Conclusions

Summary of results

As expected, adopting an intention-to-treat approach underestimated the true size of the treatment effect, most notably in scenarios where a high proportion of patients switched treatments. Results of the ITT analysis are important as they reflect the overall effectiveness of a treatment policy if it were introduced on a wider scale, but should be presented along with a measure of the true efficacy of a treatment in certain situations.

Commonly adopted approaches of censoring patients at their switching time or considering treatment as a time-dependent covariate were found to be particularly inappropriate, giving biased estimates of the true treatment effect in situations where a patient's switching pattern is strongly related to their underlying prognosis. Excluding switching patients from the analysis altogether gave relatively small biases in situations with a low proportion of switchers, but selection bias increased as switching probabilities were increased. Biases from this approach were fairly predictable in this study, but are likely to far less so if the approach was applied to real life trials where the underlying prognosis of each patient and the true treatment effect are not known.

The Loeys & Goetghebeur method generally gave biased estimates which may be due to the fact that simulations conducted here assumed patients received at least some of their initial treatment, making the ”all-or-nothing” assumption incorrect. This method may have its uses in other situations such as trials looking at the impact of a screening program where patients will either attend their screening or not. Law & Kaldor's method gave fairly small biases in some scenarios, although the direction of these was difficult to predict. However, questions remain about the way in which the method conditions on future events which may bias results towards the null [13, 20].

The method of Branson & Whitehead gave the smallest biases of all methods in situations where the potential for selection bias was high. The method performed particularly well when the difference in lifetime between good and poor prognosis patients was high, which meant patients who switched had worse underlying survival than those who did not. The method was also particulary robust in scenarios with a high proportion of switching patients, and successfully gave a parameter estimate for all simulated datasets in all of the scenarios presented here. The method did not suffer any convergence problems unlike some of the other methods investigated. It was also demonstrated how the estimates of e^{ψ} can be converted to hazard ratios, overcoming one of the main problems with the method being adopted on a wider scale for the analysis of clinical trials with switching patients.

The method of Robins & Tsiatis gave estimates close to the true treatment effect, but biases were larger than those from the Branson & Whitehead method. Its interval bisection method used is also more computationally intensive than the IPE algorithm used in the Branson & Whitehead method. Concerns have been raised previously about how Branson & Whitehead deal with censoring, with the recensoring used as part of the Robin & Tsiatis method said to be more appropriate [30]. Further investigation into situations with a higher proportion of censored observations are needed.

Problems were seen with the Walker parametric method which gave biased estimates and had estimation problems, most notably in scenarios with a high proportion of switchers. However, these problems may have been due to the simplistic way in which the method was implemented in this report.

Limitations

There is a limit to the number of possible scenarios that could be looked at in any simulation study. It may have been of interest to consider scenarios with even greater potential for selection bias and see how well each method performed. An even greater difference in lifetime between good and poor prognosis groups could have been introduced which should ensure that patients who switch and those who do not differ greatly in their underlying survival.

Only two true treatment effects were looked at, hazard ratios of 0.9 and 0.7 to represent a small and large treatment effect. More values could be investigated, possibly an even larger true effect such as a hazard ratio of 0.5, or a scenario where the treatment which patients were switching onto actually had a negative effect, so a hazard ratio of greater than 1 (or $e^{\psi} < 1$). The second of these scenarios would involve a patient's observed survival time being shorter than their underlying event time, a situation in which the recensoring used by the Branson & Whitehead method may not be correct [30].

The method of Branson & Whitehead involves fitting parametric models to the data. In this report a Weibull distribution was used for this, which allowed the conversion to hazard ratios as described in Section 5. Other parametric distributions commonly used with AFT models could also be used to find an estimate of ψ although the same conversion to the hazard ratio scale would not be possible. The log-logistic distribution could be used which can deal with non-monotonic hazard functions unlike the

Weibull [17]. Given that data in this report was simulated from a Weibull distribution, the Weibull approach used was probably sufficient, but further work may be done to investigate the choice of distribution in situations where hazard rate is not constantly increasing or decreasing over time.

As discussed previously, standard errors given from the last iteration of the IPE algorithm in the Branson & Whitehead method may be too small, with bootstrapping required to give standard errors of the correct size. Given the large number of scenarios considered, and the fact that each of these required 1000 simulations, it was not possible to perform bootstrapping for every one of these. An initial investigation into this was made by repeating simulations for scenario 14 (for which the Branson & Whitehead method previously gave a low coverage of 82.6%) with confidence intervals calculated from 100 bootstrapped samples using the normal approximation method. Coverage from bootstrapping improved to 94.1% compared to 81.5% when using standard errors from the last iteration of the IPE algorithm.

The simulation study presented only considered the situation where patients switch from the control arm to receive experimental treatment. In reality patients may switch in both directions. For example some patients may suffer severe side-effects from the experimental treatment and be advised to switch to the control arm. The method of Robins & Tsiatis as implemented through the *strbee* program in Stata does allow switches in both directions to be adjusted for. Branson & Whitehead also state their method can be extended to deal with switching in both directions, although this is yet to be implemented. Further investigation could be done into the way these methods perform in this more complex situation.

No mention was made in this report of adjusting for baseline covariates which may be used to control for imbalances between treatment arms (although this is unlikely in large randomised trials) [31]. Differences in baseline covariates may also account for some of the differences in switching pattern between patients, for example patients of a certain age may be more or less likely to switch treatment groups. Adjusting for these baseline covariates could therefore reduce the biases seen when using some of the simple methods. Branson & Whitehead describe how their method is easily extended to inclusion of baseline covariates by simply including variables in the models fitted as part of the IPE algorithm. Investigations could be performed into this and the extent to which adjusting for baseline covariates can reduce the selection bias observed from the simple methods.

All methods presented in this report give one overall treatment effect and are therefore not necessarily suitable in situations where the treatment effect for patients who switch onto a treatment is not the same as for those who were initially allocated to the experimental treatment arm. This is particuarly important is disease areas such as cancer where treatment switching typically occurs upon disease progression. For example, a NICE appraisal of treatments for colorectal cancer [32] found treatment to be around half as effective for patients who switched onto the treatment as those who received it from the start of the trial. To properly deal with this situation, new methodology may be needed which gives two different estimates of treatment effect, one for patients who are allocated to the treatment from the start of the study and one for patients who switch onto the treatment.

Another approach to the analysis of a trial of this sort would be to make use of any external information there is about a treatment. Patients in the control group who switch treatments could have their survival adjusted using this prior information to estimate the survival time they may have experienced if they had not switched. Comparisons between treatment groups could then be made as usual. This would of course depend on the availability of external information about the treatment and also the way in which switching had been dealt with in the previous trials (if there was any).

Implications

We have illustrated the problem of analysing data from trials in which patients switch treatments and why the ITT approach may not always be sufficient if the efficacy of a treatment is of interest rather than the effectiveness of introducing the treatment policy overall. The susceptibility of simple methods to selection bias was also seen, particularly if patients who switch treatments were not representative of all patients in the trial.

Given a trial in which a significant proportion of patients switch treatments, a method to adjust for this switching could be used to find the true efficacy of the treatment. When reporting a trial with treatment crossover, the authors should report the proportion of switchers, a summary of the distribution of switching times and any evidence of a relationship between switching and relevant prognostic variables. Of the methods investigated here, the Branson & Whitehead method gave the smallest bias and was seen to be robust in a variety of scenarios. Further advantages of this method is the conversion of AFT estimates to hazard ratios and its possible extension to trials in which patients switch in both directions between treatment arms, thus easily enabling inclusion of the results into an economic decision model.

Authors contributions

Text for this section

Acknowledgements

We would like to thanks Ian White for his advice on applying the methods in Stata. James Morden completed part of this work as part of an MSc at the University of Leicester funded by the EPSRC. Part of the work was funded by the NICE decision support unit.

References

1. Peduzzi P, Wittes J, Detre K: **Analysis as-randomized and the problem of nonadherence - An example from the Veterans Affairs randomized trial of coronary-artery bypass-surgery**. *Statistics in Medicine* 1993, **12**(13):1185–1195.
2. Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S: **Methodological issues in the economic analysis of cancer treatments**. *European Journal of Cancer* 2006, **42**(17):2867–2875.
3. White IR, Carpenter J, Pocock SJ, Henderson RA: **Adjusting treatment comparisons to account for non-randomized interventions: an example from an angina trial**. *Statistics in Medicine* 2003, **22**(5):781–793.
4. for Health NI, Excellence C: **The clinical effectiveness and cost effectiveness of trastuzumab for breast cancer**. *TA 34, http://guidance.nice.org.uk/TA34* 2002.
5. Abrams K, Palmer, Wailoo A: **Bevacizumab, sorafenib, sunitinib and temsirolimus for renal cell carcinoma**. *Decision Support Unit Report, http://www.nice.org.uk/nicemedia/pdf/RenalCellCarcinomaExtraWorkPreparedByDSU.pdf* 2008.
6. White IR: **Uses and limitations of randomization-based efficacy estimators**. *Statistical Methods in Medical Research* 2005, **14**(4):327–347.
7. Lee Y, Ellenberg J, DG H, KB N: **Analysis of clinical trials by treatment actually received: is it really an option?** *Stat Med* 1991, **10**:1595 – 1605.
8. Fergusson D, Aaron SD, Guyatt G, P H: **Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis**. *BMJ* 2002, **325**:652–654.
9. Goetghebeur E, Loeys T: **Beyond intention to treat**. *Epidemiologic Reviews* 2002, **24**:85–90. [Times Cited: 14].
10. Walker AS, White IR, Babiker AG: **Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment**. *Statistics in Medicine* 2004, **23**(4):571–590.
11. 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.
12. Law MG, Kaldor JM: **Survival analyses of randomized clinical trials adjusted for patients who switch treatments**. *Statistics in Medicine* 1996, **15**:2069–2076.
13. White IR: **Letters to the editor : Survival analyses of randomized clinical trials adjusted for patients who switch treatments**. *Statistics in Medicine* 1997, **16**:2619–2625.
14. Loeys T, Goetghebeur E: **A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance**. *Biometrics* 2003, **59**:100–105.
15. Kim LG, White IR: **Compliance-adjusted intervention effects in survival data**. *The Stata Journal* 2004, **4**(Number 3):257–264.
16. White IR: **Stata software**. *http://www.mrc-bsu.cam.ac.uk/BSUsite/Software/Stata.shtml* 2009.
17. Collett D: *Modelling survival data in medical research.* Chapman and Hall 2003.
18. Robins JM, Tsiatis AA: **Correcting for non-compliance in randomized trials using rank preserving structural failure time models**. *Communications in Statistics-Theory and Methods* 1991, **20**(8):2609–2631.
19. White IR, Walker S, Babiker A: **strbee: Randomization-based efficacy estimator**. *The Stata Journal* 2002, **2**(Number 2):140–150.

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

21. Branson M, Whitehead J: **Estimating a treatment effect in survival studies in which patients switch treatment**. *Statistics in Medicine* 2002, **21**:2449–2463.

22. Hougaard P: **A class of multivariate failure time distributions**. *Biometrika* 1986, **73**(3):671–678.

23. Oakes D: **Bivariate survival models induced by frailties**. *Journal of the American Statistical Association* 1989, **84**(406):487–493.

24. for Health NI, Excellence C: **Guidance on the use of capecitabine for the treatment of locally advanced or metastatic breast cancer, (TA 62)**. *TA 62, http://guidance.nice.org.uk/TA62* 2003.

25. for Health NI, Excellence C: **Gemcitabine for the treatment of metastatic breast cancer, (TA 116)**. *TA 116, http://guidance.nice.org.uk/TA116* 2005.

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

27. Coleman M, Babb P, Damieckil P, Grosclaude P, Honjo S, Jones J, Knerer G, Pitard A, Quinn M, Sloggett A, De Stavola B: *Cancer survival trends in England and Wales, 1971-1995: deprivation and NHS region*. Office of National Statistics 1999.

28. Lambert P, Smith L, Jones D, Botha J: **Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects**. *Statistics in Medicine* 2005, **24**:3871–3885.

29. Burton A, Altman DG, Royston P, Holder RL: **The design of simulation studies in medical statistics**. *Statistics in Medicine* 2006, **25**:4279–4292.

30. White IR: **Letter to the editor : Estimating treatment effects in randomized trials with treatment switching**. *Statistics in Medicine* 2006, **25**:1619–1622.

31. White IR, Pocock SJ: **Statistical reporting of clinical trials with individual changes from allocated treatment**. *Statistics in Medicine* 1996, **15**(3):249–262.

32. for Health NI, Excellence C: **Irinotecan, Oxaliplatin and Raltitrexed for Advanced Colorectal Cancer (review of TA33)**. *TA 93, http://guidance.nice.org.uk/TA93* 2005.

Figures

Figure 1 - Sample figure title

A short description of the figure content should go here.

Figure 2 - Sample figure title

Figure legend text.