

An approximate randomisation-respecting
adjustment to the hazard ratio for
time-dependent treatment switches in clinical
trials.

Ian R. White¹ A. Sarah Walker²
Abdel G. Babiker²

¹ MRC Biostatistics Unit
Institute of Public Health
Robinson Way
Cambridge CB2 2SR
Tel: 01223 330399
Fax: 01223 330388
Email: ian.white@mrc-bsu.cam.ac.uk

² MRC Clinical Trials Unit
222 Euston Road, London NW1 2DA

May 11, 2004

An approximate randomisation-respecting adjustment to the hazard ratio for time-dependent treatment switches in clinical trials.

Abstract

We consider a clinical trial with a survival outcome, in which subjects may switch in a non-randomised manner from their randomised treatment to control treatment during follow-up. We extend a method for binary outcomes, due to Sommer and Zeger, by viewing a survival outcome as a sequence of binary outcomes. This leads to a simple expression for the hazard ratio among non-switchers at each time in terms of the unadjusted hazard ratio and the proportion of switchers among subjects who experience events. We also estimate an overall adjusted hazard ratio. Although our method is approximate, it has the advantage of preserving the significance level of the log rank test. A simulation study shows that the method works well for the overall adjusted hazard ratio but that interval-specific estimates are under-adjusted as the degree of switching increases and as the association between switching and events becomes stronger. The method is illustrated using data from the Concorde trial of immediate versus deferred treatment in HIV infection.

1 Introduction

Trials with long-term follow-up and survival outcomes frequently have substantial numbers of subjects who change from their randomised treatment. These changes may include switches to the other trial treatment or changes to non-trial treatment. We will consider the analysis of trials with treatment switches, including trials of treatment versus no treatment where subjects may discontinue treatment.

Sometimes the aim of analysis is to compare treatment policies (a pragmatic approach [1]), where the policies involve starting with a particular treatment but then making changes as necessary. The appropriate analysis here is intention-to-treat (ITT) analysis which compares the groups as randomised, ignoring the treatment changes.

Sometimes, however, the aim of analysis is to compare the treatments themselves and to estimate the explanatory treatment effect or efficacy [1, 2]. Efficacy may be defined as the treatment effect that would have been observed if, contrary to fact, no individuals switched from their randomised treatment. Alternatively and more pragmatically, efficacy may be defined as the actual treatment effect in a subgroup of ‘compliers’ who would receive treatment if and only if randomised to it – the complier average causal effect (CACE) [3] or local average treatment effect [4]. Treatment switches typically make the treatment experiences of the randomised groups more similar than they would otherwise have been, so that estimates from the ITT analysis are biased towards the null as estimates of efficacy. The efficacy may be of pharmaceutical interest, or it may help to inform a comparison of different treatment policies. Consider, for example, a treatment that showed an unex-

pected benefit in a clinical trial, despite substantial rates of discontinuation. After the trial results are published, discontinuation rates may well decrease, so that the benefit of the treatment policy will become larger and approach the efficacy.

Explanatory analysis could be based on comparisons of groups defined by treatment actually received. However, this is commonly biased since subjects who stop their randomised treatment may have very different prognosis from those who continue [5–7], and the benefits of randomisation are lost. For this reason ITT analysis is commonly advocated as the only safe approach. There are, however, methods which estimate efficacy while still being based on comparisons of the randomised groups. We previously termed these randomisation-based efficacy estimators (RBEEs) [8]: they have the desirable property that their significance level equals that of a specified ITT analysis.

One class of RBEEs applies to situations of “all-or-nothing compliance” – for example, where some subjects randomised to a surgical procedure do not receive that procedure. In these situations the CACE is estimated in the subgroup of compliers – individuals who would not switch from the allocated treatment, whether it was treatment or control. When interest lies in the mean difference between groups, the CACE may be estimated as the ITT difference in outcome divided by the difference between the two randomised groups in the proportion of subjects receiving the new treatment [4, 9–11]. Related approaches have been used for survival outcomes with all-or-nothing compliance [12, 13].

An alternative class of RBEEs is based on causal models relating observed

outcomes to potential outcomes in the presence or absence of treatment. Estimation exploits the equality of potential outcomes between randomised groups. These methods have been successfully used for survival outcomes with treatment changes occurring throughout follow-up using the causal accelerated life model [5, 14–19]. The practical usefulness of these methods is limited because they do not directly estimate the hazard ratio and because they may be unduly influenced by small treatment-time interactions [8, 19].

At present, ad-hoc methods are commonly used. For example, a recent major trial divided the observed log hazard ratio by the fraction of person-years on treatment [20]. We will show that this method may be inappropriate. There is therefore a clear need for a randomisation-based method for estimating the hazard ratio.

This paper proposes a new approximate hazard-based RBEE for estimating efficacy in clinical trials with survival outcomes. Our method is an extension of the randomisation-based efficacy estimator of Sommer and Zeger [10] which we review in Section 2. In Section 3 we extend the method for a survival outcome with treatment changes occurring throughout follow-up. Standard errors and confidence intervals are discussed in Section 4. In Section 5 we show how to estimate the adjusted hazard ratio under the assumption that it is constant. In Section 6 we extend the methods to work in continuous time. The extension to time-dependent treatment changes and survival outcomes involves approximations which we investigate algebraically and by simulation in Section 7. In Section 8 the methods are applied to the Concorde trial of immediate versus deferred zidovudine treatment in asymptomatic HIV infection, in which, following a protocol amendment in line with clinical practice,

individuals in the deferred group were allowed to start zidovudine early in the face of persistently low CD4 counts, but without clinical symptoms. We conclude in Section 9 with a discussion.

2 Method of Sommer and Zeger

Sommer and Zeger [10] discussed a trial in which villages in rural Indonesia were randomised to treatment (vitamin A supplementation) or no treatment. Some villages randomised to treatment actually received none, while all villages randomised to no treatment were assumed to receive no treatment. The outcome was child mortality, and the cluster randomisation was ignored for this analysis.

Define the following random variables for each child: $D = 1$ if the child died, 0 otherwise; $R = E$ if randomised to experimental treatment, C if randomised to control; *actual treatment* $Z = 1$ if they received the treatment, 0 otherwise; *compliance-type* $W = 1$ if they would receive treatment if randomised to treatment, 0 otherwise. For children in the experimental arm, W is observed and equals Z ; for children in the control arm, W is unobserved. By definition, W and R are independent. Define the non-compliance probability $\alpha = P(W = 0)$. We assume that $\alpha < 1$, so that at least some children would receive treatment if randomised to it. For $r = E, C$, define the event probabilities in non-compliers $\pi_{r0} = P(D = 1 | R = r, W = 0)$, compliers $\pi_{r1} = P(D = 1 | R = r, W = 1)$ and all subjects $\pi_r = P(D = 1 | R = r)$, so that

$$\pi_r = (1 - \alpha)\pi_{r1} + \alpha\pi_{r0}. \quad (1)$$

Define $ITT = \pi_E/\pi_C$ as the overall risk ratio and $CACE = \pi_{E1}/\pi_{C1}$ as the

complier average causal effect on the risk ratio scale. Because W is unobserved in the control arm, π_{C0} , π_{C1} and $CACE$ are not directly estimable, but ITT , α , π_{E0} , π_{E1} and π_C are directly estimable.

In order to estimate $CACE$, Sommer and Zeger made the *exclusion restriction* assumption [4] that outcome is independent of randomised allocation among the treatment-non-compliers:

$$\pi_{C0} = \pi_{E0} \tag{2}$$

Assumption (2) is likely to be true in a blind trial; in the vitamin A trial, it is likely to be true because villages who did not receive treatment were unaware of their randomised allocation. In some unblinded trials, however, the knowledge that a patient was allocated a treatment could affect their outcome even if the treatment was not received. Sensitivity to breach of the exclusion restriction has been discussed, and may be reduced by the use of covariates that predict compliance [21–23].

Estimation of $CACE$ is now possible, because the assumed outcomes of the non-compliers in the control arm may be subtracted from the observed outcomes in the control arm. Using (1) and (2) we may write $CACE$ in terms of directly estimable quantities:

$$CACE = \frac{\pi_{E1}(1 - \alpha)}{\pi_C - \alpha\pi_{E0}}. \tag{3}$$

In later sections we will use an equivalent expression which may be derived from (3):

$$CACE = \frac{ITT(1 - \lambda)}{1 - \lambda ITT} \tag{4}$$

where

$$\lambda = P(Z = 0 | D = 1, R = E) \tag{5}$$

is the non-compliance probability among individuals in the treatment arm who experience events ($\lambda = \alpha\pi_{E0}/\pi_E$). Note that the denominator of (3) and (4) is non-negative under the assumed model.

Suppose we observe d_{E0} , d_{E1} , d_C events among n_{E0} , n_{E1} , n_C subjects, in an obvious notation. Then ITT and λ may be estimated as $\frac{d_E/n_E}{d_C/n_C}$ and d_{E0}/d_{E+} , where $+$ denotes summation over all possible values. Substituting in equation (4) gives an estimate of $CACE$:

$$C\widehat{A}CE = \frac{d_{E1}/n_E}{d_C/n_C - d_{E0}/n_E} \quad (6)$$

If $n_E = n_C$ this expression is $\frac{d_{E1}}{d_C - d_{E0}}$: the numerator is the observed number of events among treatment-compliers in the treatment arm, while the denominator uses the exclusion restriction assumption to estimate the number of events among treatment-compliers in the control arm.

Estimate (6) is the maximum likelihood estimate provided that $\hat{\pi}_{C1} = \frac{d_C/n_C - d_{E0}/n_E}{1 - n_{E0}/n_E}$ lies between 0 and 1. This happens if two boundary conditions hold: (1) $d_C/n_C > d_{E0}/n_E$ and (2) $(n_C - d_C)/n_C > (n_{E0} - d_{E0})/n_E$ [24]. If condition (1) is violated then the maximum likelihood estimate of π_{C1} is 0: if $d_{E1} > 0$ then the profile likelihood increases with $CACE$ and the maximum likelihood estimate is $+\infty$, while if $d_{E1} = 0$ then the data provide no information about $CACE$ [11]. If condition (2) is violated then the maximum likelihood estimate of π_{C1} is 1 and $C\widehat{A}CE = d_{E1}/n_E$. In practice, violations of these conditions casts doubt on the exclusion restriction (2): for example, with $n_E = n_C$, if more deaths occur in non-complying treatment arm subjects than in control arm subjects ($d_{E0} > d_C$), then the former may be harmed by their randomised allocation despite not receiving the treatment.

We note several special cases. Firstly, for risk ratios near 1, a Taylor series

expansion of equation (4) about $\log ITT = 0$ gives $\log CACE \approx \log ITT / (1 - \lambda)$. This shows that $CACE$ lies further from the null than ITT . Secondly, if switching is independent of prognosis, then we may set $\pi_{E0} = \pi_C$ in (3). Using $\pi_{E1}(1 - \alpha) = \pi_E - \alpha\pi_{E0}$ and rearranging gives $ITT = \alpha + (1 - \alpha)CACE$, a simple weighted average of the risk ratios in non-compliers and compliers [25].

We have so far assumed $\alpha < 1$. If $\alpha = 1$, so that there are no compliers and $\lambda = 1$, then $CACE$ is undefined; also, the exclusion restriction assumption (2) implies that $ITT = 1$, and $CACE$ is undefined in (3). One might in practice observe no compliers ($n_{E1} = 0$). In this case one of the boundary conditions is always breached. In practice, one would not wish to estimate $CACE$ in this circumstance.

This method has been extended to deal with switches from control to treatment ('contamination') under a second exclusion restriction assumption [11]. Extending the above notation, it can be shown that

$$CACE = \frac{ITT(1 - \lambda_E) - \lambda_C}{(1 - \lambda_C) - ITT\lambda_E} \quad (7)$$

where λ_E is the probability that individuals who experience events in the treated arm did not receive treatment, and λ_C is the probability that individuals who experience events in the control arm did receive treatment.

3 Methods of Sommer and Zeger extended to a survival outcome

A survival outcome may be regarded as a sequence of binary outcomes, so it is natural to seek to extend the methods of Section 2 for this case. We divide follow-up into intervals $i = 1, \dots, I$. Let random variable T denote

the interval in which an event occurs, or $I + 1$ if no event occurs. Then the ITT risk ratio in interval i is

$$ITT_i = \frac{\pi_{Ei}}{\pi_{Ci}} = \frac{\text{P}(T = i | T \geq i, R = E)}{\text{P}(T = i | T \geq i, R = C)} \quad (8)$$

In Section 6 we will let the intervals become very narrow so that ITT_i becomes a hazard ratio.

We consider the case where individuals in the control arm are never treated. Assume that individuals in the experimental arm may stop treatment but only at the start of an interval, and define the random variable W (compliance-type) as the interval in which this stopping occurs, or $I + 1$ if no stopping occurs. For individuals in the control arm, let W be the interval in which stopping *would have occurred*, had they been allocated to the experimental arm. Let random variable Z_i indicate actual receipt of treatment in interval i . We assume an *extended exclusion restriction* that randomised allocation has no effect on any survivors to interval i who would stop treatment by the start of interval i :

$$\text{P}(T = i | T \geq i, W = w, R = E) = \text{P}(T = i | T \geq i, W = w, R = C) \quad (9)$$

for all i and for all $w \leq i$. For individuals who would stop treatment immediately after randomisation ($w = 1$), this is the same as the exclusion restriction discussed in Section 2. For other individuals, it is plausible if past treatment may be assumed to have no effect on current risk; otherwise, it may be appropriate to lag treatment changes, which we do in Section 8.

We define $CACE$ in interval i as the risk ratio among those who survive to interval i and who would not stop treatment before the end of interval i :

$$CACE_i = \frac{\text{P}(T = i | T \geq i, W > i, R = E)}{\text{P}(T = i | T \geq i, W > i, R = C)} \quad (10)$$

Using the extended exclusion restriction assumption, we show in appendix A that

$$CACE_i \approx CACE_i^* = \frac{ITT_i(1 - \lambda_i)}{1 - \lambda_i ITT_i} \quad (11)$$

where

$$\lambda_i = P(Z_i = 0 | T = i, R = E) \quad (12)$$

is the probability that an individual who has an event in interval i has previously stopped treatment. Approximation (11) is valid either if the total event rate is small for each value of W or if the treatment effect $CACE_i$ is small in each interval (see Appendix A). One of these assumptions is likely to hold in many situations encountered in practice.

We will first consider estimation of $CACE_i$ separately for each interval, deferring estimation of a common value until Section 5. $CACE_i$ is a simple function of ITT_i and λ_i . Let numbers of deaths in interval i and numbers of subjects at risk be d_{E0i} , d_{E1i} , d_{Ci} and n_{E0i} , n_{E1i} , n_{Ci} . Then ITT_i may be estimated as $\frac{d_{E+i}/n_{E+i}}{d_{Ci}/n_{Ci}}$, the observed risk ratio. With large numbers of intervals it would be desirable to model ITT_i over time, for example using Poisson or Cox regression with linear predictor $\alpha_i + (\beta_0 + \beta_1 i)r$ in arm r and interval i . λ_i is the fraction of events in the experimental arm in which treatment was previously stopped and may be estimated as d_{E0i}/d_{E+i} . One could also model λ_i using logistic regression of previous stopping on interval, restricted to individuals in the experimental arm who have events. Boundary conditions apply as in Section 2.

In the presence of censoring, the estimate of $CACE_i$ is valid provided that (1) the estimate of ITT_i is valid, which happens if censoring is independent of event time, and (2) the estimate of λ_i is valid, which happens if censoring

in the experimental arm is independent of actual treatment, conditional on event time. We therefore make different assumptions in the two arms. In the treated arm, we assume that censoring is independent of both events and actual treatment, whereas in the control arm we only assume that censoring is independent of events. By contrast, with all-or-nothing compliance, Frangakis and Rubin produced consistent estimates under the assumption that censoring is independent of events given compliance-type, an assumption that is weaker than ours in the treated arm and arguably more plausible than ours in the control arm [26].

In the special case of all-or-nothing compliance, the extended exclusion restriction (9) follows directly from a simple exclusion restriction. However, our method remains approximate, and exact methods are available [12, 13, 26]. Our method should be most valuable with time-dependent treatment switches.

4 Standard errors and confidence intervals

Approximate standard errors and confidence intervals for $CACE_i^*$ may be computed by ignoring uncertainty in the estimated λ_i . A confidence interval for the adjusted hazard ratio $CACE_i^*$ can then be obtained simply by replacing ITT_i in equation (11) with its confidence limits. Despite ignoring uncertainty in λ_i , this procedure has the desirable property that it exactly preserves the ITT significance level: \widehat{ITT}_i is significantly different from 1 if and only if \widehat{CACE}_i^* is significantly different from 1.

A second approach is to use bootstrap methods [27], sampling whole observations with replacement. ITT_i and λ_i , and hence $CACE_i^*$, are esti-

mated for each bootstrap sample. If the bootstrap estimates of $CACE_i^*$ are Normally distributed then a standard error estimated from 200 bootstrap samples should be adequate to give a Normal-theory confidence interval, but otherwise more bootstrap samples will be required.

Sommer and Zeger estimated standard errors using the delta method [10]. This approximate method has been shown to yield incorrect type I error rates [24]. We therefore suggest that standard errors conditional on $\hat{\lambda}_i$ should be used for preliminary work but that they should be checked by bootstrap methods.

5 Estimating a global adjusted treatment effect

In many applications it will be reasonable to assume that $CACE_i^*$ is constant, say θ . We propose estimating θ conditionally on $\hat{\lambda}_i = d_{E0i}/d_{E+i}$. equation (4) then implies that the ITT risk ratio between the treated and control arms in interval i is

$$ITT_i(\theta) = \frac{\theta}{1 + \lambda_i(\theta - 1)} \quad (13)$$

The model may be fitted by maximising the partial log-likelihood

$$\sum_i d_{Ti} \log ITT_i(\theta) - d_{+i} \log \{n_{Ci} + n_{Ti} ITT_i(\theta)\} \quad (14)$$

with respect to θ . Here d_{ri} and n_{ri} are the number of events and the number of subjects at risk in arm $r = E, C$ in interval i . The standard error for the adjusted log hazard ratio, conditional on $\hat{\lambda}_i$, can be obtained from the observed information matrix. Bootstrap methods yield unconditional inference as in Section 4.

This method yields a test statistic for $\theta = 1$ which is different from the ITT test statistic. It is debatable whether this is desirable. In most situations, the focus of secondary analysis is to estimate θ , and differences in the significance level are either of minor interest or are prone to misinterpretation. We may produce an estimator that is consistent with the ITT significance level by using a weighted partial log-likelihood

$$\sum_i w_i [d_{Ti} \log ITT_i(\theta) - d_{+i} \log \{n_{Ci} + n_{Ti} ITT_i(\theta)\}] \quad (15)$$

because the score statistic for $\theta = 1$ is

$$\sum_i w_i (1 - \hat{\lambda}_i) \left(d_{Ti} - \frac{d_{+i} n_{Ti}}{n_{+i}} \right) \quad (16)$$

and the choice $w_i = 1/(1 - \hat{\lambda}_i)$ yields the log-rank score statistic. This *ITT-weighted* method therefore exactly preserves the log rank significance level if the score test is used with standard errors conditional on $\hat{\lambda}_i$.

Proportional hazards models can not in general hold exactly for both the unadjusted ITT and the adjusted ITT. However, in practice, each may fit well enough to give useful global estimates of unadjusted and adjusted ITT.

6 Extension to continuous time

So far we have grouped follow-up into discrete intervals, which is arbitrary and involves the unrealistic assumption that treatment is constant through intervals. We now adapt our method for continuous time by letting the interval width tend to zero and making modelling assumptions. We modify the notation of Section 3, replacing the subscript i by an argument t (time), and redefining $\pi_{E0}(t)$ etc. as hazard functions. The quantities

$ITT(t) = \pi_E(t)/\pi_C(t)$ and $CACE(t) = \pi_{E1}(t)/\pi_{C1}(t)$ are now hazard ratios, and equation (11) becomes

$$CACE(t) \approx CACE^*(t) = \frac{ITT(t)(1 - \lambda(t))}{1 - \lambda(t)ITT(t)} \quad (17)$$

We propose fitting parametric models for $ITT(t)$ and $\lambda(t)$, yielding a parametric form for $CACE^*(t)$.

To estimate $ITT(t)$, we fit a proportional hazards model with time to event as the outcome. The linear predictor in the control arm is 0 and in the treated arm is a function of t to be specified by the data analyst:

$$h(t) = h_0(t) \exp [\beta_0 R + \beta_1 f(t) R]. \quad (18)$$

For example, the choice $f(t) = t$ specifies that $\log ITT(t)$ is linear in time.

To estimate $\lambda(t)$, we fit a logistic regression model

$$\text{logit } \lambda(t) = \alpha_0 + \alpha_1 g(t) \quad (19)$$

to all treatment arm subjects who experienced events. The outcome is 1 if the subject was a non-complier at the time of the event and 0 if a complier. $g(t)$ is specified by the data analyst: a polynomial or spline in t or $\log t$ could be an appropriate model.

In the simple case where both $\lambda(t)$ and $ITT(t)$ are modelled as constants, $CACE^*(t)$ is also constant with the estimator $C\widehat{ACE}^* = \frac{\widehat{ITT}(1-\hat{\lambda})}{1-\hat{\lambda}\widehat{ITT}}$. Using the approximation for $ITT \approx 1$ developed in Section 2, $\log C\widehat{ACE}^* \approx \log \widehat{ITT}/(1 - \hat{\lambda})$. By contrast, a recent major trial [20] divided the observed log hazard ratio by the fraction of person-years on treatment, which amounts to $\log C\widehat{ACE}^* \approx \log \widehat{ITT}/(1 - \hat{\alpha})$. This divisor is incorrect unless event rates are similar among compliers and non-compliers; even then, the method obscures the dubious assumption that $\lambda(t)$ is constant.

7 Simulation study and numerical evaluation

To evaluate how well our approximate method may perform in practice, we evaluated its bias numerically and by simulation. We considered a model in continuous time in which time to switch, W , is generated for individuals in both arms: in the treated arm it is observable, while in the control arm it is understood as the switch time that would have been observed if the subject had been randomised to treatment. Values of W in the control arm are used to estimate true parameter values but are not used by the proposed estimators. To allow for association between time to event T and time to switch W , we used the following frailty model.

1. Frailty U is distributed with mean 1 and Laplace transform $E[e^{-kU}] = \mathcal{L}(k)$.
2. Switch time W has constant conditional hazard $h_W(t|U) = \rho_W U$.
3. Event time T has constant conditional hazard

$$h_T(t|U, W) = \begin{cases} \rho_T U & \text{if } R = C \text{ or } t > W \\ \rho_T \theta U & \text{if } R = E \text{ and } t \leq W \end{cases}$$

We let U follow a gamma distribution $\mathcal{L}(k) = (1 + kV)^{-1/V}$ with variance $V = 0.1, 0.5, \text{ or } 2$. V controls the association between events and non-compliance. To better understand these values of V , we used Cox regression to compute the hazard ratio between non-compliers and compliers $\pi_{E0}(t)/\pi_{E1}(t)$ on simulated data in the absence of a treatment effect. This hazard ratio was 1.1, 1.6 and 3.5 for $V = 0.1, 0.5$ and 2 respectively. We took $\rho_T = 1$; $\rho_W = 0.5$ and 1 (giving 33% and 50% switching with $V = 0$ and $\theta = 1$); and $\theta = 0.6, 1$ and 1.6.

For each parameter combination, we used the formulae in appendix B to evaluate $ITT(t)$, the true $CACE(t)$, and the proposed $CACE^*(t)$. Marginal treatment effects such as these differ from conditional treatment effects such as θ , whenever $\text{var}(U) > 0$. For example, a conditional proportional hazards model without treatment switches with a positive stable frailty also obeys marginal proportional hazards but with different hazard ratios, while other frailty distributions yield marginal models with non-proportional hazards [28].

For each parameter combination, we also simulated 1000 samples of size 1000. We censored each sample at time 1. Dividing follow-up into 10 intervals of equal width, we estimated ITT_i and ITT by Cox regression; $CACE_i$ and $CACE$ by Cox regression, censoring in both arms at W ; and $CACE_i^*$ and $CACE^*$ by the methods of Sections 3 and 6. Because of the use of Cox regression, we call these quantities hazard ratios. Standard errors and confidence intervals were computed conditional on the estimated $\hat{\lambda}_i$. To improve relative precision in estimating confidence interval coverage, we used 90% confidence intervals instead of the commoner 95%. Boundary condition (1) discussed in Section 2 requires $1 - \hat{\lambda} \widehat{ITT}_i > 0$. When this was violated, $C\widehat{ACE}_i^*$ was taken as $+\infty$. This occurred in 1.7% of cases overall, rising to 5.6% in the last time interval. Boundary condition (2) was not checked since it is very unlikely to be violated when, as here, interval-specific event rates are low. In 0.05% of cases, $\hat{\lambda}_i$ was 1, so that $C\widehat{ACE}_i^*$ was zero. $C\widehat{ACE}_i^*$ was positively skewed and was summarised by its median. Computing the coverage of nominal 90% confidence intervals for $\log CACE^*$ required a true value: this was taken as 0 when $\theta = 1$, and as the simulation mean of $CACE^*$

when $\theta \neq 1$.

Figure 1 shows, for $\theta = 0.6$, exact values of $\log ITT(t)$, $\log CACE(t)$ and $\log CACE^*(t)$, together with the median of $\log \widehat{CACE}_i^*$ from the simulation. As noted above, the presence of frailty makes $\log CACE(t)$ move from $\log 0.6 = -0.51$ towards 0 as time increases. $\log ITT(t)$ is much closer to 0, while $\log CACE^*(t)$ approximates $\log CACE(t)$ well for $V=0.1$ and $V=0.5$. For $V=2$, $\log CACE^*(t)$ approximates $\log CACE(t)$ well for $t < 0.5$ (corresponding to a cumulative incidence in the control group up to 29%) but thereafter is substantially closer to 0. Simulated results for $\log \widehat{CACE}_i^*$ had no appreciable bias; the Monte Carlo error in these results ranges from < 0.01 to $0.02-0.03$ as time increases. For $\theta = 1$, $\log \widehat{CACE}_i^*$ showed no appreciable bias in any case (results not shown).

Figure 1 about here

Using the global estimates, the adjustments remove nearly all of the bias in the unadjusted estimator, performing least well for larger ρ_W and V (Table 1). ITT-weighting slightly increases the standard error compared with the unweighted estimator. The model standard error and coverage are accurate, despite being evaluated conditional on $\hat{\lambda}_i$.

Table 1 about here

There were differences between unweighted and ITT-weighted P-values (results not shown). Cases where the unweighted test was significant at the 10% level, but ITT analysis was not, are of greatest concern, since these could lead analysts to dismiss the ITT significance level. This happened in 434 of all 18000 simulated data sets (2.4%), and in 134 of 1000 simulated data sets with $\theta = 0.6$, $V = 2$ and $\rho_W = 1$.

8 Example: the Concorde trial

The Concorde trial evaluated zidovudine (ZDV) therapy in asymptomatic HIV infection [29]. 1749 participants were randomised to immediate ZDV (the Imm group) or deferred ZDV (the Def group). Initially it was intended that the Def group would not receive ZDV until they progressed to AIDS-related complex (ARC) or AIDS. However, one year after the start of the trial, the protocol was amended in line with changes in clinical practice to allow open zidovudine on the basis of persistently low CD4 cell counts before the onset of ARC or AIDS and to allow primary PCP prophylaxis for participants with CD4 cell count below 200 cells/ μ L. The final results of the trial showed that the Imm group had a small but non-significant reduction in progression to ARC, AIDS or death, no difference in progression to AIDS or death, and a small but non-significant increase in mortality compared with the Def group. We have since investigated how the treatment changes associated with the protocol change could have affected the trial results [5, 8].

Here we analyse progression to ARC, AIDS or death, and attempt to estimate the effect of treatment if no ZDV had been received before ARC, AIDS or death in the deferred treatment arm. The framework described above allows for switches from treatment to control only, so we define the ‘treatment’ to be deferring ZDV: the ‘treated’ arm switch to the ‘control’ treatment if they start ZDV. Of 284 events in Def, 103 occurred in participants who had started ZDV.

In Concorde, the extended exclusion restriction assumption (9) means that the event rate among subjects in the Def group who switched to treatment by time t is the same as in those subjects in the Imm group who would

have switched to treatment by time t had they been randomised to the Def group: this is reasonable provided we believe that any benefits of treatment are independent of the length of time for which it has been received. Approximation (11) holds because event rates are low. Finally, we assume that censoring is independent of events in the Imm group and independent of both events and switches in the Def group.

We divided follow-up into 6 intervals of width 200 days and a 7th open-ended interval. The data are shown in Table 2. For each interval, events in the Def group were classified according to whether open ZDV had been started by the time of the event. Person-years are not used by our method but are shown for illustrative purposes in Table 2, classified by whether open ZDV was started by the end of the interval.

Table 2 about here

We estimated λ_i for each interval using the data in Table 2, and ITT_i using Cox regression. The interval-specific results for ITT_i and $CACE_i^*$ are given in Figure 2, with 95% confidence intervals conditional on $\hat{\lambda}_i$.

Figure 2 about here

We implemented the modelling approach of Section 6 in two ways. Firstly, we modelled $ITT(t)$ and $\lambda(t)$ using fractional polynomials [30] of degree 1: that is, $\log ITT(t)$ and $\text{logit } \lambda(t)$ were modelled as linear in t^n where n is selected from the set $(-2,-1,-0.5,0,0.5,1,2,3)$. $n = 0$, corresponding to the log transformation, gave the highest reduction in deviance for both models, and was chosen as the first model.

Secondly, we expressed both $\log ITT(t)$ and $\text{logit } \lambda(t)$ as restricted cubic splines in t [31]. 5 knots were used, fixed at 100, 400, 700, 1000 and 1300 days: these choices were motivated by the fact that events were fairly evenly spread over most of the interval from 0 to 1500 days but with some tailing off at higher values.

Data analysis was performed in Stata. For computational convenience, time was grouped into 155 small bands within which $ITT(t)$ was taken as constant.

Results for the log model and the cubic spline model are shown graphically in Figure 3. The fitted models for $\text{logit } \lambda(t)$ are reasonably similar, showing an increase in the cumulative proportion of treatment switchers with time. The fitted models for the unadjusted hazard ratio are also reasonably similar, showing a benefit of treatment only in the first 500 days or so, although the spline fit shows more curvature. The adjusted hazard ratio is close to the unadjusted hazard ratio for the first 800 days, first because $\lambda(t)$ is small and then because the hazard ratio is small. After 800 days $\lambda(t)$ is large and the adjusted hazard ratio becomes unstable, especially with the spline models.

Figure 3 about here

Despite the treatment-time interaction suggested by Figure 3, it is also useful to produce overall estimates of treatment effects (Table 3). The unweighted adjusted estimate is almost 3 times as large as the unadjusted estimate, and has a much smaller significance level: this is because it implicitly assigns more weight to earlier times, when treatment is more beneficial. ITT-weighting restores the significance level of the unadjusted analysis and gives a less extreme point estimate. Similar results are achieved whether $\lambda(t)$

is modelled in terms of log time or cubic splines, but wrongly modelling it as a constant somewhat reduces the effect.

Table 3 about here

Some treatment switches could in principle be consequences of disease processes that precede disease events, in which case correction for these switches would be inappropriate. Observing a large number of switches just before events would warn of this danger. This is not the case in the Concorde data: of the 103 individuals who switched and subsequently had an event, only 13 switched in the last 10% of their follow-up time before the event. However, to protect against any possible bias, we lagged treatment changes by 180 days. This reduced to 73 the number of individuals who had an event after switching, and so the adjusted treatment effect was somewhat less extreme than when all switches were included (Table 3).

Table 3 also shows that the two methods of computing standard errors discussed in Section 4 – conditional on $\hat{\lambda}$ and by bootstrap methods – give very similar results. Bootstrap confidence intervals tend to be slightly asymmetrical but remain very similar to the normal-theory conditional confidence intervals.

9 Discussion

We have proposed methods for adjusting for time-dependent treatment changes in a hazard-based framework. In this setting, some assumption about long-term effects of treatment is essential. We chose to assume that past treatment has no effect on current hazard. This motivated the extended exclusion restriction assumption that the hazard at time t is independent of randomised

arm for individuals who, if allocated to treatment, would have stopped it by time t . If in fact subjects who have recently stopped treatment have a risk intermediate between that of compliers and total non-compliers, then our method yields an over-correction for treatment changes; one solution to this difficulty is to lag treatment changes.

Our approximate method yields a simple expression relating the adjusted hazard ratio to the unadjusted hazard ratio and the proportion of non-compliers among individuals experiencing events, where all these quantities may be time-dependent. The approximation is widely plausible since it holds either if the treatment effect is small or if the total event rate is small. The approximate method allows the analyst to model the unadjusted hazard ratio and the proportion of non-compliers over time using standard modelling techniques, and then either to combine them into a model for the adjusted hazard ratio, or to estimate a constant adjusted hazard ratio. It has been implemented as a Stata program `adjhr.ado` which is available from the first author.

Instead of our approximate method, an exact method should be possible by using maximum likelihood methods in the more general framework laid out in Appendix A. It should also be possible to extend this framework to incorporate more general assumptions about long-term effects of treatment, for example that the treatment effect is halved in the interval following treatment cessation, and to allow for more general censoring mechanisms. Further work is ongoing in this area.

Both unweighted and weighted likelihood methods give valid significance levels, but the unweighted method achieves slightly smaller standard errors,

while the weighted method preserves log rank test significance levels. We regard preserving ITT significance levels as an advantage for two reasons. Firstly, in up to 13% of simulated datasets, the unweighted approach yielded significant results when the ITT results were not significant. This could lead potentially misleading claims of significance based on an adjusted analysis. Given the dangers of multiple testing, an analysis that preserves the P-value is desirable. Secondly, the unweighted method gains precision by assigning greater weight to earlier intervals in which compliance is greater. Trials in which the estimated treatment effect tends to be larger in earlier follow-up than in later follow-up, such as Concorde, are common. In such trials, the unweighted method will tend to make both the point estimate and the significance level more extreme. However, adjustment for treatment switching is usually a secondary analysis, and one would want any differences from the primary analysis to reflect treatment switching itself rather than a change in weighting. One might consider using the unweighted adjusted analysis as a primary analysis, but the small potential gain in power would not make up for the loss of credibility in departing from familiar rank tests [25].

Our results indicate potential pitfalls in simple approaches to correcting hazard ratios for non-compliance. In particular, dividing the observed log hazard ratio by the fraction of person-years on treatment may be inappropriate. It is better to divide by the fraction of treatment among individuals who experience events, but if this fraction is not constant then full use of our method is required.

Extension of the hazard-based approach to switches between two treatments, or allowing switching from control to treatment as well as treatment

to control, would be valuable. This requires either the assumption of no defiers (subjects who will always do the opposite of their randomisation) or the assumption that the relative causal effect of treatment is the same for all types of subject [11]. The approximation described in Appendix A should apply in this setting. Allowing for a continuous compliance variable such as the proportion of time so far spent on treatment would also be valuable. Finally, our method applies to placebo-controlled trials in which some treated subjects stop treatment, provided that the placebo effect may be ignored, which is often reasonable [32].

A Full model and approximation for the complier average causal risk ratio

A.1 Notation

Let follow-up be divided into intervals $i = 1, \dots, I$. We first assume there is no censoring except at the end of interval I . We drop individuals from the notation. As before, $r = E, C$ denotes randomised group. Define compliance-type W as a random variable which combines the individual's compliance if randomised to treatment *and* the individual's compliance if randomised to control. In the terminology of Frangakis and Rubin [33], W determines principal strata. Let z_{wi}^r denote the actual treatment of an individual in principal stratum w in interval i if they are randomised to r , and Z_i the corresponding random variable. Actual treatment Z_i is a post-randomisation random variable whereas compliance-type W is independent of randomised group. Let \mathcal{W} be the set of possible values of W . Define the following subsets of \mathcal{W} :

$$\begin{aligned} \text{Compliers: } \mathcal{C}_i &= \{w : z_{wi}^E = 1, z_{wi}^C = 0\} \\ \text{Always-takers: } \mathcal{A}_i &= \{w : z_{wi}^E = 1, z_{wi}^C = 1\} \\ \text{Never-takers: } \mathcal{N}_i &= \{w : z_{wi}^E = 0, z_{wi}^C = 0\} \\ \text{Defiers: } \mathcal{D}_i &= \{w : z_{wi}^E = 0, z_{wi}^C = 1\} \end{aligned}$$

Let T be a random variable equal to the interval in which an individual dies, or $I + 1$ if the individual does not die. Let

$$\begin{aligned} \tau_w &= P(W = w) \\ h_{wi}^r &= P(T = i | T \geq i, W = w, R = r) \end{aligned}$$

$$S_{wi}^r = \text{P}(T \geq i | W = w, R = r) = \prod_{j=1}^{i-1} (1 - h_{wj}^r)$$

$$S_i^r = \text{P}(T \geq i | R = r) = \sum_w \tau_w S_{wi}^r$$

and for any $A \subset \mathcal{W}$, $h_{Ai}^r = \text{P}(T = i | T \geq i, W \in A, R = r)$

$$S_{Ai}^r = \text{P}(T \geq i | W \in A, R = r) = \sum_{w \in A} \tau_w S_{wi}^r / \sum_{w \in A} \tau_w$$

A.2 Model

We make two assumptions:

1. that the set \mathcal{D}_i is empty, that is, that there are no defiers [11].
2. the exclusion restriction $h_{wi}^E = h_{wi}^C$ whenever $w \in \mathcal{A}_i \cup \mathcal{N}_i$.

We want to estimate the risk ratios in the group \mathcal{C}_i of compliers at interval i :

$$CACE_i = h_{\mathcal{C}_i i}^E / h_{\mathcal{C}_i i}^C.$$

One approach would be to write down the likelihood and maximise it by an EM algorithm. We propose a simpler, approximate approach.

A.3 Approximate estimation

We may show that

$$\begin{aligned} CACE_i &= \frac{\text{P}(T = i, W \in \mathcal{C}_i | R = E)}{\text{P}(T = i, W \in \mathcal{C}_i | R = C)} \times \frac{\text{P}(T \geq i, W \in \mathcal{C}_i | R = C)}{\text{P}(T \geq i, W \in \mathcal{C}_i | R = E)} \\ &= \frac{\sum_{w \in \mathcal{A}_i \cup \mathcal{C}_i} \tau_w S_{wi}^E h_{wi}^E - \sum_{w \in \mathcal{A}_i} \tau_w S_{wi}^E h_{wi}^E}{\sum_{w \in \mathcal{N}_i \cup \mathcal{C}_i} \tau_w S_{wi}^C h_{wi}^C - \sum_{w \in \mathcal{N}_i} \tau_w S_{wi}^C h_{wi}^C} \times \frac{S_{\mathcal{C}_i i}^C}{S_{\mathcal{C}_i i}^E} \\ &= \frac{\frac{1}{S_i^E} \sum_{w \in \mathcal{A}_i \cup \mathcal{C}_i} \tau_w S_{wi}^E h_{wi}^E - \frac{1}{S_i^C} \sum_{w \in \mathcal{A}_i} \tau_w S_{wi}^C h_{wi}^C \left[\frac{S_{wi}^E S_i^C}{S_{wi}^C S_i^E} \right]}{\frac{1}{S_i^C} \sum_{w \in \mathcal{N}_i \cup \mathcal{C}_i} \tau_w S_{wi}^C h_{wi}^C - \frac{1}{S_i^E} \sum_{w \in \mathcal{N}_i} \tau_w S_{wi}^E h_{wi}^E \left[\frac{S_{wi}^C S_i^E}{S_{wi}^E S_i^C} \right]} \times \left[\frac{S_i^E S_{\mathcal{C}_i i}^C}{S_i^C S_{\mathcal{C}_i i}^E} \right] \end{aligned}$$

where the third equality follows from the exclusion restriction.

The terms in square brackets are near 1 if the total event rate is small for all w , since then each $S_{wi}^r, S_{C_i}^r, S_i^r \approx 1$. They are also near 1 if the treatment effect is small, since then $S_{wi}^E \approx S_{wi}^C, S_{C_i}^E \approx S_{C_i}^C$ and $S_i^E \approx S_i^C$. Setting these terms equal to 1 yields the approximation

$$\begin{aligned} CACE_i \approx CACE_i^* &= \frac{\frac{1}{S_i^E} \sum_{w \in \mathcal{A}_i \cup \mathcal{C}_i} \tau_w S_{wi}^E h_{wi}^E - \frac{1}{S_i^C} \sum_{w \in \mathcal{A}_i} \tau_w S_{wi}^C h_{wi}^C}{\frac{1}{S_i^C} \sum_{w \in \mathcal{N}_i \cup \mathcal{C}_i} \tau_w S_{wi}^C h_{wi}^C - \frac{1}{S_i^E} \sum_{w \in \mathcal{N}_i} \tau_w S_{wi}^E h_{wi}^E}} \\ &= \frac{\pi_{1i}^E - \pi_{1i}^C}{\pi_{0i}^C - \pi_{0i}^E} \end{aligned}$$

where

$$\pi_{zi}^r = \text{P}(Z_i = z, T = i | T \geq i, R = r)$$

for $r = E, C, z = 0, 1$ is the proportion of survivors who die on treatment ($z = 1$) or who die off treatment ($z = 0$) within each arm. All the π_{zi}^r are directly estimable. Finally, it is convenient to write

$$\begin{aligned} ITT_i &= \frac{\pi_{+i}^E}{\pi_{+i}^C} \\ \lambda_i^E &= \text{P}(Z_i = 0 | T = i, R = E) = \pi_{0i}^E / \pi_{+i}^E \\ \lambda_i^C &= \text{P}(Z_i = 1 | T = i, R = C) = \pi_{1i}^C / \pi_{+i}^C \\ \text{so that } CACE_i^* &= \frac{(1 - \lambda_i^E) ITT_i - \lambda_i^C}{(1 - \lambda_i^C) - \lambda_i^E ITT_i} \end{aligned}$$

We propose estimating $CACE_i^*$ from natural estimates of ITT_i, λ_i^E and λ_i^C . In the presence of censoring, this estimate is valid provided that the natural estimates of ITT_i, λ_i^E and λ_i^C are valid, which occurs if censoring is independent of both survival and actual compliance. A likelihood-based approach would be valid if censoring was independent of survival given compliance-type, a condition that has been termed latent ignorability [26, 34].

A.4 One-way switching

In the setting described in the main text, the experimental group may stop treatment but the control group may not start treatment. W becomes simply the interval in which an individual would stop treatment if randomised to E , or $I + 1$ if the individual would remain treated throughout follow-up; $\mathcal{W} = \{1, 2, \dots, I + 1\}$; and $z_{wi}^C = 0$ for all i and w , while $z_{wi}^E = 1$ if $i < w$, 0 if $i \geq w$. The sets \mathcal{A}_i and \mathcal{D}_i are empty. It follows that $\lambda_i^C = 0$ and

$$CACE_i^* = \frac{(1 - \lambda_i^E)ITT_i}{1 - \lambda_i^E ITT_i}$$

as in equation (4).

B Hazard ratios under the simulation model

For $r = E, C$, define

$$\begin{aligned}
 S_{r0}(t) &= \text{P}(T \geq t, W < t | R = r) \\
 S_{r1}(t) &= \text{P}(T \geq t, W \geq t | R = r) \\
 f_{r0}(t) &= \lim_{\delta t \rightarrow 0} \text{P}(t < T < t + \delta t, W < t | R = r) / \delta t \\
 f_{r1}(t) &= \lim_{\delta t \rightarrow 0} \text{P}(t < T < t + \delta t, W \geq t | R = r) / \delta t
 \end{aligned}$$

For the frailty model of Section 7, we calculate

$$\begin{aligned}
 S_{E0}(t) &= \text{E}_U \left[\int_0^t \rho_W U e^{-\rho_W U s} e^{-\rho_T(\theta s + t - s)U} ds \right] \\
 &= \frac{\rho_W}{\rho_W + \rho_T(\theta - 1)} \{ \mathcal{L}(\rho_T t) - \mathcal{L}([\rho_W + \rho_T \theta]t) \} \\
 S_{E1}(t) &= \text{E}_U \left[\int_t^\infty \rho_W U e^{-\rho_W U s} e^{-\rho_T \theta t U} ds \right] \\
 &= \mathcal{L}([\rho_W + \rho_T \theta]t) \\
 S_{C1}(t) &= \text{E}_U [e^{-\rho_T U t} e^{-\rho_W U t}] \\
 &= \mathcal{L}([\rho_T + \rho_W]t) \\
 f_{C1}(t) &= \text{E}_U [\rho_T U e^{-\rho_T U t} e^{-\rho_W U t}] \delta t \\
 &= \rho_T \mathcal{L}'([\rho_T + \rho_W]t) \delta t \\
 f_{E1}(t) &= \text{E}_U [\rho_T \theta U e^{-\rho_T \theta U t} e^{-\rho_W U t}] \delta t \\
 &= \rho_T \theta \mathcal{L}'([\rho_T \theta + \rho_W]t) \delta t \\
 f_{E0}(t) &= \text{E}_U \left[\int_0^t \rho_T U e^{-\rho_T[\theta s + t - s]U} \rho_W U e^{-\rho_W U s} ds \right] \delta t \\
 &= \frac{\rho_T \rho_W}{\rho_T(\theta - 1) + \rho_W} \{ \mathcal{L}'(\rho_T t) - \mathcal{L}'([\rho_T \theta + \rho_W]t) \} \delta t
 \end{aligned}$$

where $\mathcal{L}'(k) = -\frac{d\mathcal{L}(k)}{dk} = \text{E}_U [U e^{-kU}]$. Using these, and noting $\pi_C(t) = \rho_T$, we can compute $ITT(t) = \frac{f_{E0} + f_{E1}}{S_{E0} + S_{E1}} / \rho_T$, $CACE(t) = \frac{f_{E1}}{S_{E1}} / \frac{f_{C1}}{S_{C1}}$, $\lambda(t) = f_{E0} / (f_{E0} + f_{E1})$ and $CACE^*(t) = \frac{ITT(t)(1 - \lambda(t))}{1 - \lambda(t)ITT(t)}$.

References

- [1] D. Schwartz and J. Lellouch. Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, 20:637–648, 1967.
- [2] McMahon AD. Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. *Statistics in Medicine*, 21:1365–1376, 2002.
- [3] G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25:305–327, 1997.
- [4] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.
- [5] I. R. White, S. Walker, A. G. Babiker, and J. H. Darbyshire. Impact of treatment changes on the interpretation of the Concorde trial. *AIDS*, 11:999–1006, 1997.
- [6] S. J. Pocock. *Clinical Trials: a practical approach*. Wiley, 1983.
- [7] P. Peduzzi, J. Wittes, and K. Detre. Analysis as-randomised and the problem of non-adherence: an example from the Veterans Affairs randomized trial of coronary artery bypass surgery. *Statistics in Medicine*, 12:1185–1195, 1993.
- [8] I. R. White, A. G. Babiker, S. Walker, and J. H. Darbyshire. Randomisation-based methods for correcting for treatment changes: ex-

- amples from the Concorde trial. *Statistics in Medicine*, 18:2617–2634, 1999.
- [9] R. G. Newcombe. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Statistics in Medicine*, 7:1179–1186, 1988.
- [10] A. Sommer and S. L. Zeger. On estimating efficacy from clinical trials. *Statistics in Medicine*, 10:45–52, 1991.
- [11] J. Cuzick, R. Edwards, and N. Segnan. Adjusting for non-compliance and contamination in randomized controlled trials. *Statistics in Medicine*, 16:1017–1029, 1997.
- [12] S. G. Baker. Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association*, 93:929–934, 1998.
- [13] T. Loeys and E Goetghebeur. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*, 59:100–105, 2003.
- [14] J. M. Robins and A. A. Tsiatis. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics – Theory and Methods*, 20(8):2609–2631, 1991.

- [15] J. M. Robins. Analytic methods for estimating HIV-treatment and co-factor effects. In D. G. Ostrow and R. C. Kessler, editors, *Methodological issues in AIDS behavioural research*. Plenum Press, New York, 1993.
- [16] J. M. Robins and S. Greenland. 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*, 89:737–749, 1994.
- [17] S. D. Mark and J. M. Robins. A method for the analysis of randomised trials with compliance information: an application to the multiple risk factor intervention trial. *Controlled Clinical Trials*, 14:79–97, 1993.
- [18] I. R. White and E. J. T. Goetghebeur. Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? *Statistics in Medicine*, 17:319–339, 1998.
- [19] A. S. Walker, I. R. White, and A. G. Babiker. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. *Statistics in Medicine*, 23:571–590, 2004.
- [20] Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. *Lancet*, 360:7–22, 2002.
- [21] D. B. Rubin. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17:371–385, 1998.

- [22] K. Hirano, G. W. Imbens, D. B. Rubin, and X.-H. Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000.
- [23] B. Jo. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine*, 21:3161–3181, 2002.
- [24] M. Branson and J. Whitehead. A score test for binary data with patient non-compliance. *Statistics in Medicine*, 22:3115–3132, 2003.
- [25] S. W. Lagakos, L. L.-Y. Lim, and J. M. Robins. Adjusting for early treatment termination in comparative clinical trials. *Statistics in Medicine*, 9:1417–1424, 1990.
- [26] C. E. Frangakis and D. B. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86:365–379, 1999.
- [27] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997.
- [28] P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73:671–678, 1986.
- [29] Concorde Coordinating Committee. Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet*, 343:871–81, 1994.

- [30] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 43:429–467, 1994.
- [31] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8:551–561, 1989.
- [32] Asbjorn Hrobjartsson and Peter C. Gotzsche. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine*, 344:1594–1602, 2001.
- [33] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.
- [34] G. Dunn, M. Maracy, C. Dowrick, J. L. Ayuso-Mateos, O. S. Dalgard, H. Page, V. Lehtinen, P. Casey, C. Wilkinson, J. L. Vazquez-Barquero, and G. Wilkinson. Estimating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. *British Journal of Psychiatry*, pages 323–331, 2003.

Table 1: Simulation results for log hazard ratio

θ	V [1]	ρ_w	True $\log CACE$	Bias in $\log ITT$	Bias in $\log CACE^*$		Standard error of $\log CACE^*$		Coverage (nominal 90%)	
					un- weighted	ITT- weighted	un- weighted	ITT- weighted		Model
0.6	0.1	0.5	-0.49	0.11	0.01	0.00	0.11	0.11	0.11	89%
0.6	0.1	1	-0.50	0.18	0.01	-0.01	0.12	0.14	0.13	88%
0.6	0.5	0.5	-0.45	0.11	0.01	0.01	0.11	0.12	0.11	89%
0.6	0.5	1	-0.46	0.18	0.01	0.01	0.13	0.14	0.14	90%
0.6	2	0.5	-0.39	0.13	0.02	0.03	0.13	0.13	0.14	90%
0.6	2	1	-0.42	0.21	0.03	0.05	0.16	0.17	0.21	90%
1	0.1	0.5	0	-0.01	-0.01	-0.01	0.10	0.10	0.10	88%
1	0.1	1	0	0.00	0.00	-0.01	0.11	0.12	0.12	90%
1	0.5	0.5	0	0.00	0.00	0.00	0.11	0.11	0.11	91%
1	0.5	1	0	0.00	0.01	0.00	0.12	0.13	0.13	91%
1	2	0.5	0	0.00	0.00	0.00	0.13	0.13	0.12	89%
1	2	1	0	0.00	0.00	0.00	0.15	0.15	0.15	90%
1.6	0.1	0.5	0.45	-0.08	-0.01	0.00	0.09	0.10	0.09	91%
1.6	0.1	1	0.45	-0.12	0.00	0.00	0.11	0.11	0.11	90%
1.6	0.5	0.5	0.39	-0.07	-0.01	-0.01	0.10	0.10	0.10	89%
1.6	0.5	1	0.41	-0.12	0.00	-0.01	0.11	0.12	0.12	90%
1.6	2	0.5	0.32	-0.07	-0.02	-0.02	0.12	0.12	0.11	89%
1.6	2	1	0.36	-0.14	-0.04	-0.05	0.13	0.13	0.14	88%

[1] $V = 0.1, 0.5, 2$ correspond to hazard ratios between non-compliers and compliers of 1.1, 1.6 and 3.6 in the absence of a treatment effect (see text).

Table 2: Concorde data

Time band (days)	Imm arm			Compliers			Non-compliers [1]			Total			λ [2]
	Events [3]	PY [4]	Rate [5]	Events [3]	PY [4]	Rate [5]	Events [3]	PY [4]	Rate [5]	Events [3]	PY [4]	Rate [5]	
0-	21	469	45	43	446	96	2	16	123	45	462	97	0.04
200-	45	447	101	45	372	121	8	51	155	53	424	125	0.15
400-	49	412	119	33	313	105	16	77	209	49	390	126	0.33
600-	56	361	155	30	252	119	13	92	141	43	344	125	0.30
800-	48	304	158	12	191	63	35	95	369	47	286	164	0.74
1000-	33	228	145	16	135	118	17	76	225	33	211	157	0.52
1200-	15	135	111	2	77	26	12	49	243	14	126	111	0.86
Total	267	2356	113	181	1787	101	103	456	226	284	2242	127	0.36

[1] Non-complier means having started open ZDV before the time of ARC, AIDS or death

[2] Probability of non-compliance among events in this time band

[3] ARC, AIDS or death

[4] Person-years

[5] Rate per 1000 person-years.

Table 3: Concorde results: global log hazard ratio, Imm vs. Def.

Model [1]	Log hazard ratio [2]	Standard error		P-value		95% CI	
		Conditional	Bootstrap	Conditional	Bootstrap	Conditional	Bootstrap
<i>Unadjusted</i>	-0.115	0.085		0.18		(-0.282, 0.052)	
<i>Adjusted for all treatment changes, unweighted</i>							
linear in log t	-0.294	0.150	0.152	0.05	0.05	(-0.589, 0.000)	(-0.598, -0.014)
<i>Adjusted for all treatment changes, ITT-weighted</i>							
constant	-0.187	0.143	0.143	0.19	0.19	(-0.468, 0.094)	(-0.472, 0.072)
linear in log t	-0.205	0.155	0.153	0.18	0.18	(-0.508, 0.098)	(-0.510, 0.084)
cubic spline	-0.203	0.153	0.156	0.18	0.19	(-0.504, 0.097)	(-0.555, 0.082)
<i>Adjusted for treatment changes lagged by 180 days, ITT-weighted</i>							
constant	-0.158	0.119	0.115	0.19	0.17	(-0.392, 0.076)	(-0.398, 0.046)
linear in log t	-0.167	0.124	0.124	0.18	0.18	(-0.411, 0.077)	(-0.424, 0.067)
cubic spline	-0.167	0.125	0.127	0.18	0.19	(-0.411, 0.077)	(-0.411, 0.086)
<i>Adjusted for treatment change when time on treatment > time off treatment, ITT-weighted</i>							
constant	-0.137	0.103	0.102	0.18	0.18	(-0.339, 0.064)	(-0.352, 0.049)
linear in log t	-0.142	0.106	0.108	0.18	0.19	(-0.350, 0.066)	(-0.367, 0.067)
cubic spline	-0.142	0.106	0.110	0.18	0.20	(-0.349, 0.066)	(-0.374, 0.061)

[1] Model for logit λ and log ITT [2] log ITT for unadjusted analysis, log CACE* for adjusted analyses.

Captions for Figures

Figure 1

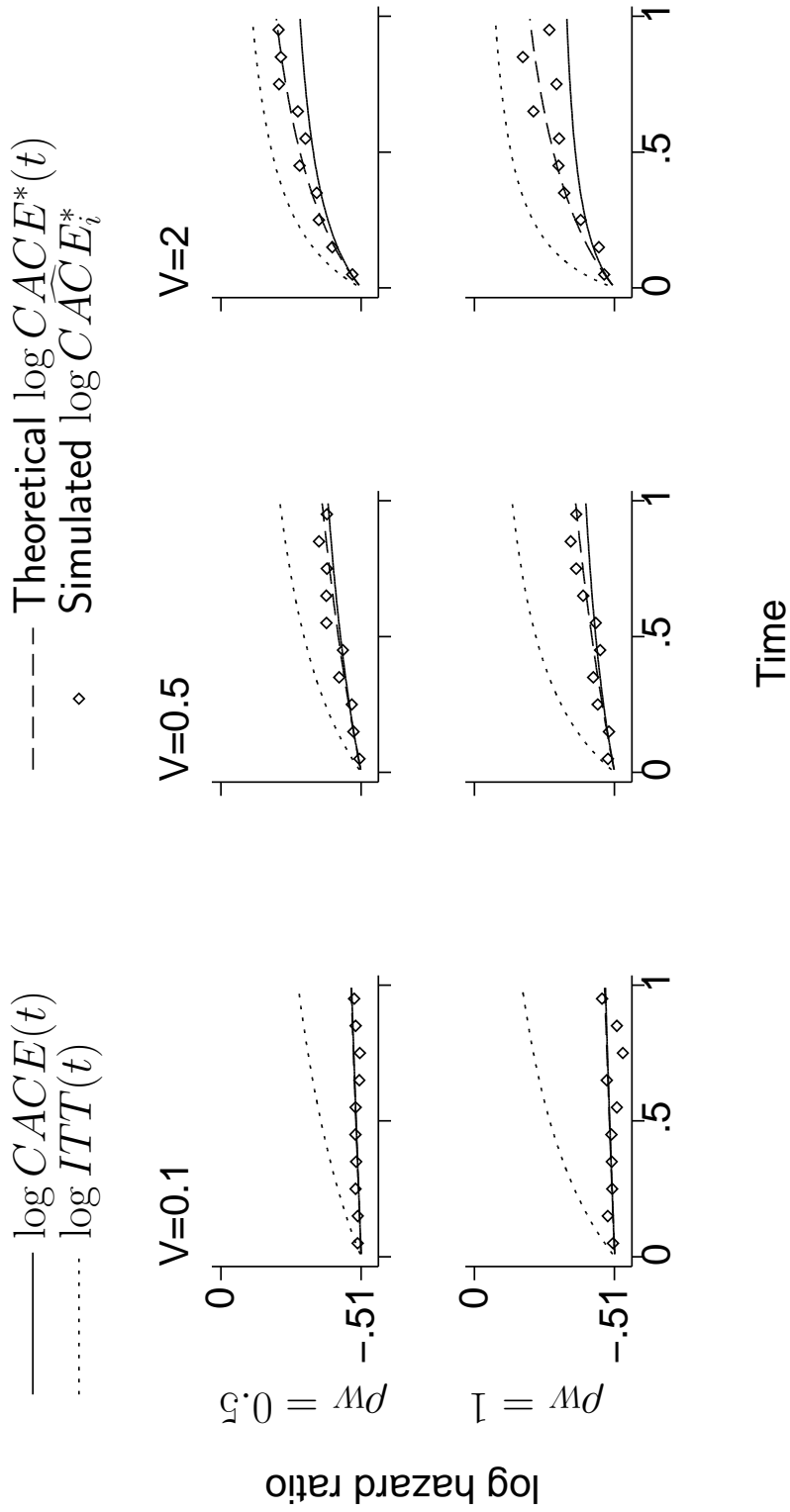
Log hazard ratios for the conditional exponential model with $\theta = 0.6$. True $CACE(t)$, unadjusted $ITT(t)$ and adjusted $CACE^*(t)$ are calculated theoretically. Median of adjusted \widehat{CACE}_i^* is derived from simulations.

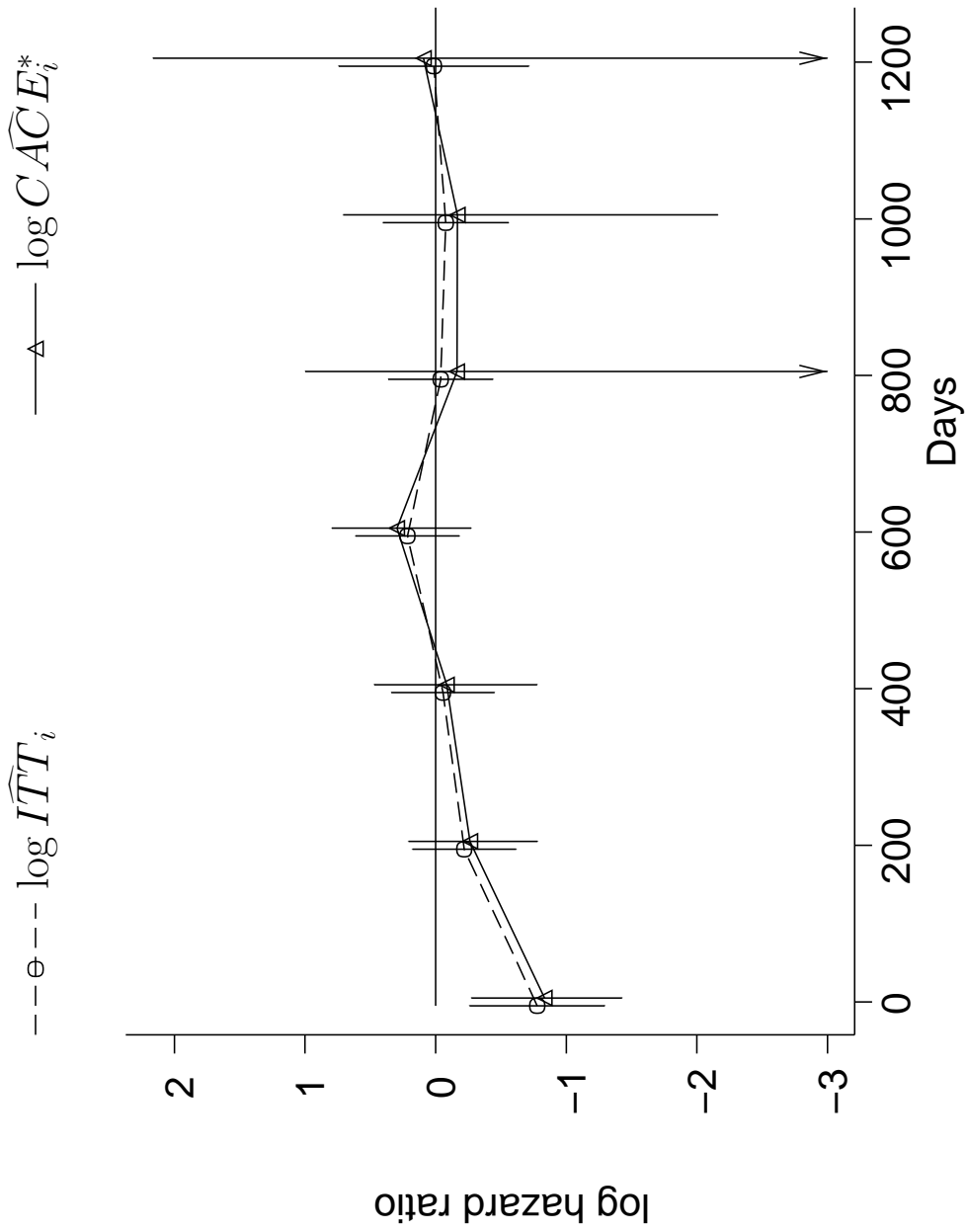
Figure 2

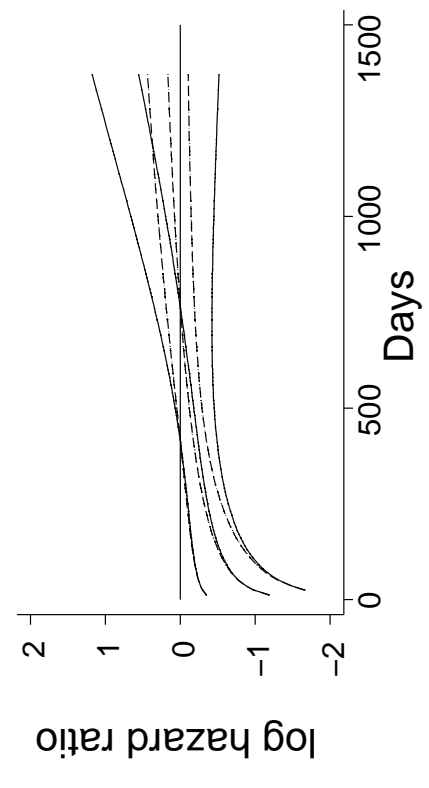
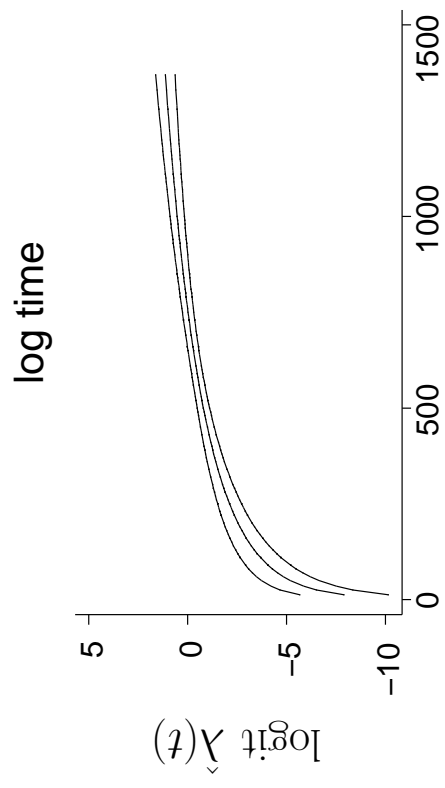
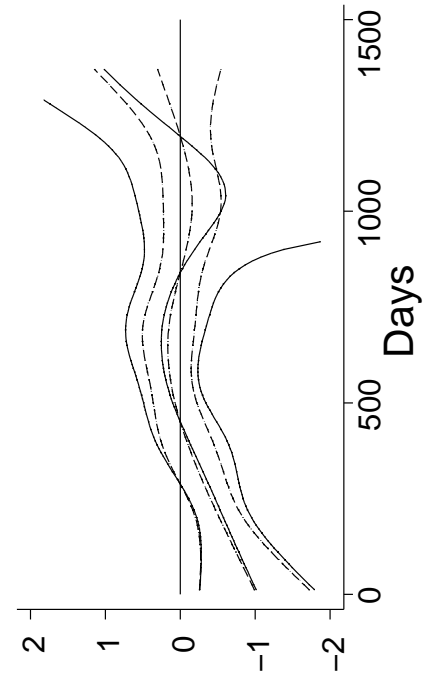
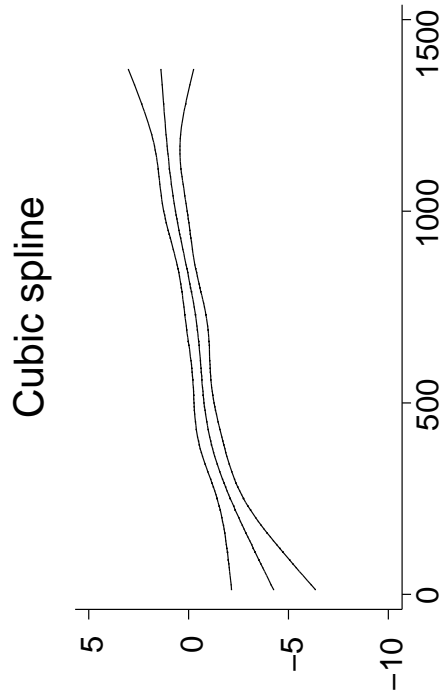
Concorde trial: \widehat{ITT}_i and \widehat{CACE}_i^* for Immediate versus Deferred zidovudine, estimated in 200-day intervals. 95% confidence intervals are conditional on $\hat{\lambda}_i$.

Figure 3

Results using the continuous-time approach in Section 6. Lower panels show $\log ITT(t)$ (dashed lines) and $\log CACE^*(t)$ (solid lines). $\log ITT(t)$ and $\text{logit } \lambda(t)$ are modelled as linear in log time (left-hand panels) or using cubic splines (right-hand panels).







Dear Mike,

Thank you for your letter of 21st July 2003 about this paper. We have now substantially revised it and we would like to resubmit it. Our replies to the referee's comments are listed below.

Best wishes,

Ian

Reply to referee

Thank you for your critical comments. In replying to them, we believe that we have substantially improved the paper. Most fundamentally, your comments about the apparent implausibility of our assumptions has led us to develop an entirely new justification for the method. The method itself is unchanged.

In this revision we have also improved the notation. We have changed T to E for the experimental treatment arm; changed RR , \widetilde{RR} and RR^* to ITT , $CACE$ and $CACE^*$; dropped the subscript on α ; and worked in terms of $\lambda = \frac{\alpha\pi_{E0}}{\pi_E}$ instead of $\nu = \lambda/(1 - \lambda)$.

Major comments

- 1 *The assumptions are implausible.* We agree that the assumptions appeared too restrictive. We have done new work to justify the method. Instead of basing our method on two assumptions that would never be exactly true, it is now based on one 'extended exclusion restriction' assumption that is at least reasonable (that randomised allocation does

not affect hazard at time t in individuals who would not receive treatment at time t regardless of allocation). The approximation is made explicit, and we can now show that the approximation is valid under broader circumstances than we thought previously.

As a result we have completely rewritten Section 3. The theoretical justification, based on the principal stratification idea of Frangakis and Rubin (2002), appears in a new Appendix A.

Methods far cruder than our approximate method are currently being used in practice: as an example, we now cite the analysis of the Heart Protection Study (Sections 1 and 6). We believe that our method combines sufficient theoretical justification with sufficient simplicity to be able to replace such ad hoc methods.

- 2 *Lack of attention to special cases.* We do not agree that the case when there are no compliant patients in the treated group is of special interest: such a trial would provide no evidence about treatment efficacy, no matter how sophisticated the statistical methods used. However, we do agree that we overlooked a second “boundary condition” (see our reply to minor point 11 below). We have corrected and extended the discussion of special cases in Sections 2 and 7.
- 3 *Lack of motivation for the ITT-weighted approach.* We have extended the argument in Sections 7 and 9. In particular, we have explored the discrepancy between the two methods in the simulation study. Up to 13% of simulated data sets have statistically significant results on unweighted analysis but not on ITT analysis, and we argue that this

unnecessarily generates extra multiplicity and extra potential pitfalls for unwary analysts.

Minor comments

- 1 Agreed: we have inserted the word “approximate” in the title.
- 2 Agreed: “our method” is now singular throughout the abstract.
- 3 Agreed: we have inserted a definition of ‘efficacy’ here.
- 4-6 Agreed: we have re-written these paragraphs.
- 7 Agreed: we have removed this equation.
- 8 Agreed: we have clarified the approximation as follows: “Firstly, for risk ratios near 1, a Taylor series expansion of equation (4) about $\log ITT = 0$ gives $\log CACE^* \approx \log ITT / (1 - \lambda)$.”
- 9 *Clarify equality $\pi_{T0} = \pi_C$ and distinguish $\alpha_T = 1, \neq 1$.* We now state the assumption $\alpha < 1$, which we believe is the only case of interest, at the start of Section 2. We include a new paragraph to discuss the case $\hat{\alpha} = 1$.

We have added a brief explanation of how we derive the expression $ITT = \alpha + (1 - \alpha)CACE^*$ – we do not agree that this would be clearer if we stated the assumption as $\pi_{E0} = \pi_{C1}$.
- 10 Agreed, but we have now dropped the subscript T from α throughout.
- 11 We have much improved our treatment of the boundary conditions. We have separated breaches of the first boundary condition into cases $d_{E1} >$

0, = 0. We have justified our assertion that the best estimate if $d_{E1} > 0$ is $+\infty$ – “the profile likelihood increases with *CACE* and the maximum likelihood estimate is $+\infty$ ”. We have commented on the case $d_{E1} = 0$: “the data provide no information about *CACE*”. We have also inserted the second boundary condition $(n_C - d_C)/n_C > (n_{E0} - d_{E0})/n_E$: we wrongly neglected it before, because we are imagining short intervals with few events, in which the second boundary condition will almost certainly hold.

- 12 Agreed: we now start the second paragraph of Section 3 with “We consider the case where individuals in the control arm are never treated.”
- 13 Agreed: our statement that $\nu = d_{T1i}/d_{T0i}$ was wrong and should have been $\nu = d_{T0i}/d_{T1i}$. In our new notation this becomes $\hat{\lambda}_i = d_{E0i}/d_{E+i}$.
- 14 Agreed that these lines were unhelpful: they have been removed and appendix B has been tidied up to link better with the text.
- 15 We used 90% instead of 95% confidence intervals to give greater relative precision in estimating empirical coverages. We have now stated this in Section 7.
- 16 *Checking boundary conditions in the simulation study.* We previously reported checks of the first boundary condition without identifying them as such. We have now clarified this: “Boundary condition (1) discussed in Section 2 is that $1 - \hat{\lambda}ITT_i > 0$. When this was violated, \widehat{CACE}_i^* was taken as $+\infty$. This occurred in 1.7% of cases overall, rising to 5.6% in the last time interval.” We did not check boundary condition

(2) and we have inserted the following justification: “Boundary condition (2) was not checked since it is very unlikely to be violated when, as here, interval-specific event rates are low.” The previous version of the paper also stated that $C\widehat{A}C\widehat{E}_i^*$ “was taken as 0 when $\hat{\nu}_i = +\infty$ ”: this is now redundant as $\hat{\lambda} = 1$ in this case, so $C\widehat{A}C\widehat{E}^* = 0$ follows directly except in the case $1 - \hat{\lambda}\widehat{ITT}_i \leq 0$ which was discussed before.

17 Agreed: we have run this sentence on to the previous one.

18 *Link function.* The link function was correctly stated as log because the argument ν was the odds of having stopped treatment. Our new notation uses the probability of having stopped treatment so the link function is indeed now the logit.

How is censoring incorporated into the estimation of $\nu(t)$? This was and is addressed in Section 3, but the text was rather obscure. We have expanded it as follows: “In the presence of censoring, the estimate of $CACE_i$ is valid provided that (1) the natural estimate of ITT_i is valid, which happens if censoring is independent of event time, and (2) the natural estimate of λ_i is valid, which happens if censoring in the treated arm is independent of actual treatment, conditional on event time. We therefore make different assumptions in the two arms. In the treated arm, we assume that censoring is independent of both events and compliance, whereas in the control arm we only assume that censoring is independent of events”.

Other changes

1. Sections 6 and 7 have been swapped round.
2. We have rewritten substantial parts of Section 1, 2, 3 and 9.
3. Parts relating to the two old approximate assumptions have been deleted and replaced with new material as appropriate: old Section 6.1 has been deleted, old Figure 1 has been dropped with associated text in old Section 6, and the assumptions for Concorde in Section 8 have been rewritten.