Research Article

Statistics in Medicine

Received 21 January 2010, Accepted 29 June 2010 Published online 27 August 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4048

A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study

Jon Michael Gran,[a*†] Kjetil Røysland,[a] Marcel Wolbers,[b,c] Vanessa Didelez,[d] Jonathan A. C. Sterne,[e] Bruno Ledergerber,[f] Hansjakob Furrer,[g] Viktor von Wyl[f] and Odd O. Aalen[a]

When estimating the effect of treatment on HIV using data from observational studies, standard methods may produce biased estimates due to the presence of time-dependent confounders. Such confounding can be present when a covariate, affected by past exposure, is both a predictor of the future exposure and the outcome. One example is the CD4 cell count, being a marker for disease progression for HIV patients, but also a marker for treatment initiation and influenced by treatment. Fitting a marginal structural model (MSM) using inverse probability weights is one way to give appropriate adjustment for this type of confounding. In this paper we study a simple and intuitive approach to estimate similar treatment effects, using observational data to mimic several randomized controlled trials. Each 'trial' is constructed based on individuals starting treatment in a certain time interval. An overall effect estimate for all such trials is found using composite likelihood inference. The method offers an alternative to the use of inverse probability of treatment weights, which is unstable in certain situations. The estimated parameter is not identical to the one of an MSM, it is conditioned on covariate values at the start of each mimicked trial. This allows the study of questions that are not that easily addressed fitting an MSM. The analysis can be performed as a stratified weighted Cox analysis on the joint data set of all the constructed trials, where each trial is one stratum. The model is applied to data from the Swiss HIV cohort study. **Copyright © 2010 John Wiley & Sons, Ltd.**

Keywords: causal inference; survival analysis; time-dependent confounding; HIV/AIDS; observational studies

1. Introduction

When studying the effect of treatment on survival or time to AIDS diagnosis for patients with HIV infection, standard Cox models with time-varying covariates may give biased estimates in the presence of time-dependent confounders [1]. Time-dependent confounding can be present when a covariate, affected by past exposure, is both a predictor of the future exposure and the outcome.

An example of a time-dependent confounder when estimating treatment effects for HIV is the CD4 cell count, which, as an indicator of immune status, is a predictor of both treatment and outcome (AIDS or death), while at the same time influenced by treatment. To deal with this type of confounding, Robins *et al.* [2] introduced a new type of model, called

[a] *Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway*
[b] *Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Switzerland*
[c] *Oxford University Clinical Research Unit, Wellcome Trust Major Overseas Program, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam*
[d] *Department of Mathematics, University of Bristol, U.K.*
[e] *Department of Social Medicine, University of Bristol, U.K.*
[f] *Division of Infectious Diseases and Hospital Epidemiology, University of Zurich, Switzerland*
[g] *Division of Infectious Diseases, Bern University Hospital and University of Bern, Switzerland*
[*] *Correspondence to: Jon Michael Gran, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, P.O. Box 1122, Blindern 0317, Norway.*
[†] *E-mail: j.m.gran@medisin.uio.no*

Copyright © 2010 John Wiley & Sons, Ltd.

the marginal structural model (MSM). When fitting an MSM, time-dependent confounding is typically adjusted for using inverse probability of treatment (IPT) weighting. Each individual's probability of being treated is calculated conditioned on their observed covariates at each time point, which then are used to construct the IPT weights for that individual. The time-dependent confounding variables are no longer predictors of the exposure in the weighted analysis. The rest of the parameters in the MSM can therefore be estimated using a weighted time-dependent Cox analysis, adjusting only for baseline covariates.

Even though IPT weighting is an elegant way to adjust for time-dependent confounding, it has properties that make the weights unstable in certain situations. The main problem lies in the instability of the estimated weights at the time where individuals go from being off treatment to on treatment. When the conditional probability of initiating treatment is small, the denominator in the expression for the weight can be close to zero, making the estimated weights unstable. In other words, individuals with unusual covariate histories when starting treatment can be given very large weights. The fact that the individuals keep this weight constant for their remaining event history after initiating treatment adds to the problem.

In this paper we consider an alternative approach to time-dependent confounding, than the IPT weights used to fit an MSM. Our method seeks to estimate a similar treatment effect as the MSM, but now by looking at the causal or counterfactual effect of treatment in many mimicked randomized controlled trials, each trial being distinguished by the time of treatment start. This approach also allows us to investigate some questions that would not be that easy to answer with an MSM; such as estimating separate treatment effects for individuals with different CD4 counts at treatment start.

Where in the MSM the time-dependent confounding is typically adjusted for using weighting, we consider a method of many successive Cox analyses, comparing the event histories of individuals starting treatment and the ones not yet on treatment in different time intervals separately. Individuals not on treatment by the start of the trial are artificially censored at the time of later treatment start. The mode of analysis in this sequential Cox approach is related to the one proposed by Hernan *et al.* in [3, 4], by Lu in [5], and by van Houwelingen in [6]. In [3] they compare treatment regimes by artificially censoring individuals when they do not follow one of two defined regimes, whereas in [4] a randomized controlled trial is mimicked from an observational cohort study by constructing eight 'trials', each over a 2-year period, and then pooling all eight 'trials' into one single analysis. In [5] patients receiving treatment are matched with patients with similar history but not on treatment by certain times. One main difference is that we not only exclude individuals already on treatment in each 'trial', but also censor individuals at the time of later treatment start. When properly weighted for any dependent censoring, this mimics a trial where a patient is either on treatment or off treatment from the beginning and to the end. This means that, in every such trial, the treatment confounding is not time-dependent. There are also similarities between our method and the landmark analysis in [6]. Individuals at risk at some landmark point are analysed using only the information available at that moment, and multiple risk sets using different landmarks are created and analysed using a pseudo likelihood. The main differences are that it does not try to handle the same problem of time-dependent confounding, and that there are no artificial censoring performed to mimic randomized controlled trials. There are also differences in how the trials and the landmark data sets are analysed and combined.

In our analysis we construct a large number of mimicked randomized controlled trials, based on different time intervals of possible treatment start, and then analyse them simultaneously using composite likelihood inference. The idea is that the effect estimates in such a sequence of Cox analyses, performed on constructed subsets of the original data, would give appropriate adjustment for time-dependent confounding. The overall effect estimate from the composite likelihood analysis, aggregated over all possible intervals of treatment start, will serve as a causal effect of treatment, given some assumptions. The estimated parameter would not be identical to the one of an MSM, as it is adjusted for covariates values at the start of each mimicked trial. To make the effect estimates consistent, any selection bias due to the artificial censoring done when creating the mimicked randomized controlled trials (together with other possible dependent censoring) has to be accounted for. Such a bias could be accounted for using inverse probability of censoring (IPC) weights, which are more stable than IPT weights.

The methodology in this field is a frequent topic of discussion [7, 8]. Generally, when it comes to addressing problems of confounding, one can distinguish between the propensity score type of weighting methods and more regression-based strategies [8, 9]. The method addressed in this paper is more of the latter type, whereas the MSM, with its IPT weights, belongs to the first category. Owing to the differences in the parameters being estimated, the estimates from the two methods are not directly comparable. But, even though the interpretation is not entirely analogous, the treatment effect of interest is similar.

In Sections 2 and 3 we give a detailed description of the sequential Cox method, whereas we describe the Swiss HIV cohort data in Section 4. In Section 5 we present our results using the sequential Cox method to estimate the effect of treatment by highly active antiretroviral therapy (HAART) compared with no treatment on time to AIDS diagnosis or death in the Swiss HIV Cohort Study [10]. The same data have previously been analysed with an MSM by Sterne *et al*. [11]. A discussion follows in Section 6.

Copyright © 2010 John Wiley & Sons, Ltd.

2. A mimicked randomized controlled trial based on a specific starting interval

We now want to mimic a randomized controlled trial using individuals starting treatment in a certain time interval. We will model the hazard in this randomized controlled trial using a Cox proportional hazard model, and later combine models for all possible time intervals.

2.1. Notation and Cox model

Let us first consider individuals starting treatment in time interval k, which in our case would refer to a certain month of observation since the inclusion in the study. Now, when interval k is our reference interval, and we are seeking an unbiased estimate of the effect of treatment, we can consider the sub-population consisting of individuals that have not received treatment before interval k. In this sub-population, the individuals who initiate treatment in interval k form the treatment group, and the remaining individuals the control group. In order to get an unbiased estimate of the treatment effect, we censor individuals from the control group when they start treatment at a later time (see illustration in Figure 1). We adjust for covariate values at baseline (covariate values at interval 1), and for covariate values at interval k, but not for later covariate values as these would be affected by treatment. The two groups can now be compared using a standard Cox model.

More formally, let T_i^A be the initiation time of treatment for individual i, let T_i be the time of an event (AIDS or death) for individual i, let $N_i(t)=I(t\geqslant T_i)$, and let $Y_i(t)$ be the at-risk indicator for individual i at time t. $Y_i(t)$ takes the value 1 if individual i has not had an event or has been censored strictly before t, and 0 otherwise. Let $A_i(t)$ be the treatment indicator function for individual i at time t. $A_i(t)$ takes the value 0 when individual i is not on treatment, and 1 when on treatment. Let θ and β be the regression coefficients corresponding to treatment and other covariates, respectively. Let $s_0,\ldots,s_n$ be a partition of the interval $[0,\tau)$, where τ is the time of the end of follow up. Moreover, let $(s_{k-1},s_k]$ be the kth reference interval. The individuals included in the treatment group in the kth trial are those who initiate treatment in this time interval, i.e. $s_{k-1}<T_i^A\leqslant s_k$. The controls at time $t\leqslant s_k$ are those who have not started treatment at that time, i.e. $T_i^A>t$.

Now, let $X_i(s_k)$ be the vector of covariate values for individual i at time s_k. Cox regression on the resulting data after time s_k would now give an estimate of the effect of initiating treatment in the kth interval, where the intensity at time t for individual i, observed until m, is modelled by

$$\lambda_i^{(k)}(t|A_i,x_{i,k},\ldots,x_{i,m})=Y_i^{(k)}(t)\lambda_0^{(k)}(t)\exp(\theta A_i(t)+\beta' x_{i,k}),$$

for all $k=1,2,\ldots,K$, where $t\geqslant s_k$, and K is the reference interval including the last observations [12].

Let

$$x_{i,k}:=X_i(s_k),$$

$$Y_i^{(k)}(u):=\begin{cases}0, & T_i^A<s_k \text{ or } s_{k+1}\leqslant T_i^A<u\\ Y_i(u) & \text{else}\end{cases},$$

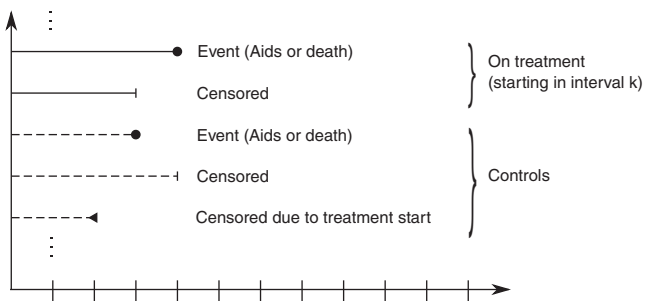

Figure 1. Illustration showing an example of the five types of individuals left in the constructed data set for a single sub-analysis used in the sequential Cox approach. The treatment group includes the individuals starting treatment in interval k, while the individuals not on treatment by interval k are the controls. Individuals can then experience an event, or be censored due to dropout or later treatment start. Notice the distinction between the two types of censoring in the control group.

Copyright © 2010 John Wiley & Sons, Ltd.

$$S_{\text{cox}}^{(0,k)}(u,\theta,\beta) := \sum_{i=1}^{n} Y_i^{(k)}(u)\exp(\theta A_i(u-)+\beta' x_{i,k}),$$

$$S_{\text{cox}}^{(1,k)}(u,\theta,\beta) := \sum_{i=1}^{n} \begin{bmatrix} A_j(u-) \\ x_{i,k} \end{bmatrix} Y_i^{(k)}(u)\exp(\theta A_i(u-)+\beta' x_{i,k}),$$

and note that $S_{\text{cox}}^{(1,k)}(u,\theta,\beta)$ is the gradient of $S_{\text{cox}}^{(0,k)}(u,\theta,\beta)$.

For the analysis starting in interval k, the partial score function becomes

$$U_{\text{cox}}^{(k)}(\theta,\beta,\tau) = \sum_{i=1}^{n} \int_{s_k}^{\tau} \left(\begin{bmatrix} A_i(u-) \\ x_{i,k} \end{bmatrix} - \frac{S_{\text{cox}}^{(1,k)}(u,\theta,\beta)}{S_{\text{cox}}^{(0,k)}(u,\theta,\beta)} \right) Y_i^{(k)}(u)\,\mathrm{d}N_i(u),$$

where $\begin{bmatrix} A_i(u-) \\ x_{i,k} \end{bmatrix}$ is a column vector with $A_i(u-)$ as the first component and the remaining components being the vector $x_{i,k}$. Ties can be handled using the Breslow approach [13].

2.2. *IPC weights*

Note that by using the partial score function of Section 2.1, that is, individuals are artificially censored when they start treatment at a later time, dependent censoring could be introduced into the analysis. Such a dependence could also be present in the observed data, and is likely to cause bias in an estimate of the treatment effect. One way to compensate for such a dependent censoring, and to adjust for this bias, is to apply stabilized IPC weights to the data, assuming no unmeasured confounders are present. The intention is to produce a weighted data set that will reflect most of the mechanisms in the original data set, but where the censoring now can be considered to be independent. Robins *et al.* [2] suggested a general method for doing this. We will follow his strategy, but rather than using pooled logistic regression as [1, 2, 11], we will estimate these weights using Aalen's additive regression model [12], as in [14–16]. The additive regression model is a flexible model where the parameter functions can vary freely with time. It also gives us simple expressions for the weights, making it a natural choice.

To derive our weights, let first $T_i^{(k),C}$ denote the time of what is first to occur, either the observed or the artificial censoring for the ith individual, and $C_i^{(k)}(t) := I(T_i^{(k),C} \leqslant t)$ be the indicator for whether individual i is censored by time t. In other words, we model the mixed process of two types of censoring, which should be fine as a tool for the main analysis. An alternative would be to estimate separate censor weights for the two types of censoring, and then using the product of these two as the overall censor weight.

In order to use the additive model, we assume that there exists a vector function $B^{(1,k)}$ such that the process

$$C_i^{(k)}(t) - \int_0^t Y_i^{(k)}(s) \begin{bmatrix} A_i(s-) \\ X_i(s-) \end{bmatrix}' \mathrm{d}B^{(1,k)}(s)$$

does not carry any information of the history of all the possible observations for person i before time $\max(t, s_k)$. $B^{(1,k)}$ is here the cumulative regression coefficients. Following [12, 4.2.1], we can provide an estimate $\hat{B}_i^{(1,k)}$ of the process $B_i^{(1,k)}$.

The equation

$$\hat{\Lambda}_i^{(1,k)}(t) := \int_0^t Y_i^{(k)}(s) \begin{bmatrix} A_i(s-) \\ X_i(s-) \end{bmatrix}' \mathrm{d}\hat{B}^{(1,k)}(s)$$

now gives a reasonable prediction of the cumulative hazard for individual i at time t, based on the baseline and time-dependent covariate history in the artificial 'at risk' period before t where $Y_i^{(k)} \neq 0$.

The same can obviously be without time-dependent covariates, hence, let $\hat{\Lambda}_i^{(0,k)}$ denote the predicted cumulative hazard for censoring individual i within the kth trial with respect to baseline covariates only.

The stabilized censor weight for individual i at time t is then

$$w_i^{(k)}(t) = \exp(\hat{\Lambda}_i^{(1,k)}(t) - \hat{\Lambda}_i^{(0,k)}(t)).$$

The corresponding partial score function based on the weighted data is now

$$\tilde{U}_{\text{cox}}^{(k)}(\theta,\beta,\tau) = \sum_{i=1}^{n} \int_{s_k}^{\tau} \left(\begin{bmatrix} A_i(u-) \\ x_{i,k} \end{bmatrix} - \frac{\tilde{S}_{\text{cox}}^{(1,k)}(u,\theta,\beta)}{\tilde{S}_{\text{cox}}^{(0,k)}(u,\theta,\beta)} \right) w_i^{(k)}(u-) Y_i^{(k)}(u)\,\mathrm{d}N_i(u),$$

Copyright © 2010 John Wiley & Sons, Ltd.

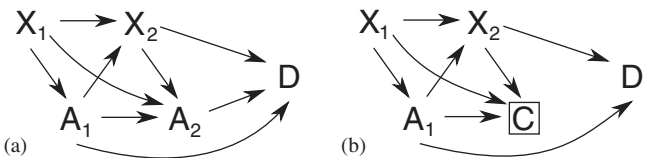

Figure 2. (a) A causal DAG representing the observational study in the simplified situation of only 2 months of follow up, assuming no unmeasured confounders. A_k is the treatment indicator in interval k, X_k is the vector of observed covariates at interval k, and D is the event AIDS or death and (b) A causal DAG representing one mimicked randomized controlled trial (the first one), based on the simplified system in (a), assuming no unmeasured confounders. C is now the artificial censoring of individuals starting treatment at a later time point.

where

$$\tilde{S}^{(0,k)}_{\text{cox}}(u,\theta,\beta)=\sum_{i=1}^{n} w_i^{(k)}(u-)Y_i^{(k)}(u)\exp(\theta A_i(u-)+\beta' x_{i,k}),$$

$$\tilde{S}^{(1,k)}_{\text{cox}}(u,\theta,\beta)=\sum_{i=1}^{n}\begin{bmatrix} A_i(u-) \\ x_{i,k}\end{bmatrix} w_i^{(k)}(u-)Y_i^{(k)}(u)\exp(\theta A_i(u-)+\beta' x_{i,k}).$$

2.3. Causal interpretation of the effect estimate

The estimate of θ from the partial score function in Section 2.2 is an estimate of the effect of treatment for individuals starting treatment in time interval k, or in other words, in our kth mimicked randomized controlled trial. This treatment effect θ could be interpreted as a causal effect of treatment given three main assumptions;

(i) there are no unmeasured confounders,
(ii) the model for estimating the hazard rate is correct, and
(iii) the model for estimating the censoring weights is correct.

We can illustrate our system with a directed acyclic graph (DAG). Assuming no unmeasured confounders, the diagram in Figure 2(a) is a causal DAG representing the observational study in the simplified situation of only 2 months of follow up, similar to the graph drawn in Figure 1 in [2]. A_k is the treatment indicator at interval k, X_k is the vector of observed covariates in interval k, and D is the event AIDS or death.

The diagram in Figure 2 (b), still assuming no unmeasured confounders, is a causal DAG representing the situation in one of our mimicked trials (the first one), based on the simplified system with only 2 months of follow up. The assumption of no unmeasured confounders is crucial, where an unmeasured confounder on treatment in the observational study now would be a predictor of the artificial censoring C, and then a source of selection bias. From the causal DAG in (b) we see that all back-doors paths from the only treatment variable A_1 to D are blocked by X_1, meaning that the chosen covariates are sufficient to adjust for confounding. The artificial censoring done when creating these mimicked randomized controlled trials corresponds to conditioning on not starting treatment in interval 2. This means blocking the node corresponding to treatment A_2, now denoted C. This is what is done when going from panel (a) to panel (b) in Figure 2. In panel (b) the situation is one of time-dependent censoring, but with no time-dependent confounding on treatment. Hence, our mimicked randomized controlled trials can be analysed with inverse probability weighting for censoring only.

3. Combining the partial score functions to make an overall effect estimate

Mimicked randomized controlled trials as the one in Section 2 can be constructed for all starting intervals k, and we now combine the partial score functions for all such trials (see Figure 3 for illustration) to estimate an overall effect of treatment using composite likelihood (sometimes referred to as pseudo likelihood) techniques [17–19]. This overall estimate of the treatment effect will still have a causal interpretation, assuming

(iv) that the effect of treatment is the same in every trial, and
(v) that the effect of treatment is the same for all covariate histories before the reference interval k given covariates at k.

Copyright © 2010 John Wiley & Sons, Ltd.

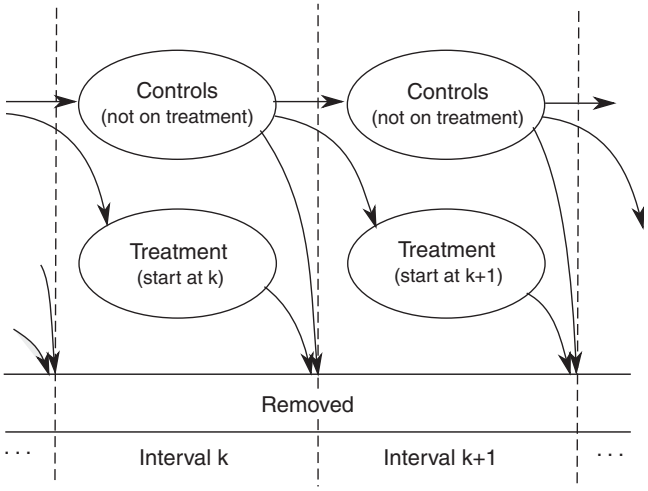

Figure 3. Illustration showing the movement between groups for individuals going from one sub-analysis (in interval k) to another (in interval $k+1$). For each interval individuals already on treatment are removed, together with individuals being censored, dying or developing AIDS without ever starting treatment, while the individuals still not on treatment are compared with the individuals starting treatment in that interval.

More specifically, we consider a pseudo-partial score function that is the sum of the partial score functions for each mimicked trial,

$$\begin{aligned}\tilde{U}(\theta,\beta,\tau) &= \sum_k \tilde{U}_{cox}^{(k)}(\theta,\beta,\tau)\\ &= \sum_k \sum_{i=1}^{n} \int_{s_k}^{\tau} \left(\begin{bmatrix} A_i(u-) \\ x_{i,k} \end{bmatrix} - \frac{\tilde{S}_{\text{cox}}^{(1,k)}(u,\theta,\beta)}{\tilde{S}_{\text{cox}}^{(0,k)}(u,\theta,\beta)} \right) R_i^{(k)}(u-) Y_i^{(k)}(u)\,\mathrm{d}N_i(u)\\ &= \sum_{i=1}^{n} \int_{s_k}^{\tau} \sum_k \left(\begin{bmatrix} A_i(u-) \\ x_{i,k} \end{bmatrix} - \frac{\tilde{S}_{\text{cox}}^{(1,k)}(u,\theta,\beta)}{\tilde{S}_{\text{cox}}^{(0,k)}(u,\theta,\beta)} \right) R_i^{(k)}(u-) Y_i^{(k)}(u)\,\mathrm{d}N_i(u), \end{aligned} \tag{1}$$

where again, assuming no unmeasured confounders, θ represent the causal effect of treatment, over all k time intervals.

The assumption that the treatment effect is the same in all trials could be relaxed, leaving the estimated effect to be interpreted as the aggregated causal effect of treatment over all possible times of treatment start (by the composite likelihood, which resembles a weighted average).

In order to fit our model with respect to θ and β, we compute the roots of (1). The estimates of θ and β will be our composite likelihood estimates. Consistency of composite likelihood estimators have been proven for special cases [17, 18]. We can carry out our analysis using standard statistical methods and software.

One way to perform this analysis would be to construct an extended version of the data set, a pseudo data set, which can be analysed using a standard stratified Cox model, stratifying on the reference interval k. This constructed pseudo data set consists of the sub-populations defined earlier for all values of k, where in each sub-population, all observations after time interval k, where the individuals have started treatment at a later time, are censored, and all time-dependent covariates are fixed at interval k. In other words, we create a large pseudo data set consisting of mimicked randomized controlled trials based on every reference interval k. In this pseudo data set, individuals would be present repeatedly as controls through different sub-populations.

The censor weights are calculated for each mimicked randomized controlled trial separately. Convergence problems will then typically be a problem when estimating regression coefficients used to calculate the censor weights for the smallest subsets (the ones with the latest reference intervals) due to the necessarily limited amount of data. This problem is avoided by introducing a ridge parameter in the model [12, p. 316].

When the pseudo data set is constructed from the original data and the censor weights are calculated, the analysis can be performed in any standard software which handles stratified weighted time-dependent Cox regression, such as R [20]. Note that, different from a usual stratified analysis, individuals will often be repeatedly used as controls in many strata.

Copyright © 2010 John Wiley & Sons, Ltd.

Since we can change the order of summation over sub-analyses and integration in (1), one should, for the unweighted case, be able to follow the proof for the asymptotic theory of ordinary Cox proportional hazard models.

A composite likelihood estimator will typically be less efficient than the estimator based on the true likelihood function [17]. The advantage in our setting is that the composite likelihood coincides with the likelihood from a particular stratified Cox analysis.

When analysing such a pseudo data set, the estimated standard error for the parameters, and then the p-values and confidence intervals, based on a standard stratified Cox model, would not be correct. It might be possible to estimate correct standard errors using sandwich procedures in the lines of van Houwelingen [6], but this would not be straightforward. We therefore estimate the standard error using the jackknife method [21], where each of the total n individuals in the data set are left out one at the time, constructing n jackknife samples. The standard error is then calculated based on the effect estimates from all these jackknife samples. Using this standard error we derive normal-based confidence intervals. Jackknife is used instead of standard bootstrapping because the bootstrap method leads to convergence problems due to the large number of covariates used in the analysis.

Using the jackknife, leaving out only one individual at a time, we are less likely to encounter such problems. Generally the jackknife can be viewed as an approximation to the bootstrap [21]; in our case when both were calculated, they gave similar estimates of the standard error.

4. The Swiss HIV cohort study

The Swiss HIV Cohort Study is an ongoing multi-center research project following up HIV-infected adults aged 16 or older [10]. Studying the effect of treatment by HAART, the data of interest are available from January 1996, when HAART became available in Switzerland. Thus, the baseline is the time of the first follow up visit after January 1996. Note that there is no strong clinical rationale for this choice of baseline, except that 1996 was when HAART was introduced. The arguably most relevant baseline, the time of HIV infection, is unknown for a huge proportion of individuals and therefore unavailable. However, with detailed clinical information available to use as covariates our chosen baseline should be reasonable. The same baseline was used in Sterne *et al.* [11]. Patients who died or refused further participation before 1996, who were on HAART or in clinical stage C at baseline, or whose treatment history before joining the cohort was uncertain were excluded. The data are organized in monthly intervals, with measures of CD4 count, HIV-1 RNA and haemoglobin levels in that month. In addition, indicator variables describe whether the individual was treated with HAART, or experienced a CDC stage B event (a disease associated with HIV but less severe than an AIDS defining disease) during that month. Once treatment is first initiated it is assumed that the individual remains on treatment from then on, as in [1, 11]. Time between visits varies (scheduled clinical follow-up is 6-monthly with additional laboratory measures taken every 3 months), and the last observation is carried forward for months without visits.

In total, 2161 individuals contributed to the data used in our analysis. The total observation time for one individual varied from 1 to 92 months. Two hundred and two of these individuals progressed to AIDS or death, and 717 were treated with HAART. The data set was also used in Sterne *et al.* [11]. See this paper for further details.

For the pseudo data set we construct in order to do our analysis on standard software, we now have 92 subsets of the original data set, one for each observation month, where each subset is exposed to some additional artificial censoring due to starting treatment. The full pseudo data set would correspond to 1 201 315 person-months of observation, a fairly large data set. Since the covariates for a single individual are not updated every month, and the last observation is carried forward, we can shorten the data to include only one row for each range of months where the covariate values are the same, adding the time intervals when these covariate values are valid. The data set is then reduced to 274 366 rows, and is used in the following analysis.

5. Results

For the Swiss HIV cohort data, the estimated hazard ratio considering time to AIDS or death for patients receiving HAART treatment versus receiving no antiretroviral treatment was 0.165, with a 95 per cent standard confidence interval of (0.079–0.343).

The analysis was carried out using a weighted stratified time-dependent Cox analysis on the constructed pseudo data set, stratifying on the reference month, and weighting for censoring. The included covariates were HAART, sex, risk group, and the following covariates included at baseline, at the reference month, and lagged at the reference month: CD4 group (grouped into 0–49, 50–99, 100–199, 200–349, 350–499, 500–749, $\geqslant$750 cells per μL), RNA group (grouped into <400, 400–1000, 1001–10 000, 10 001–100 000,>100 000 copies per mL), haemoglobin group (grouped into fifths), CDC B event, and previously experienced CDC B event. Ninety-five per cent standard confidence intervals were found

Copyright © 2010 John Wiley & Sons, Ltd.

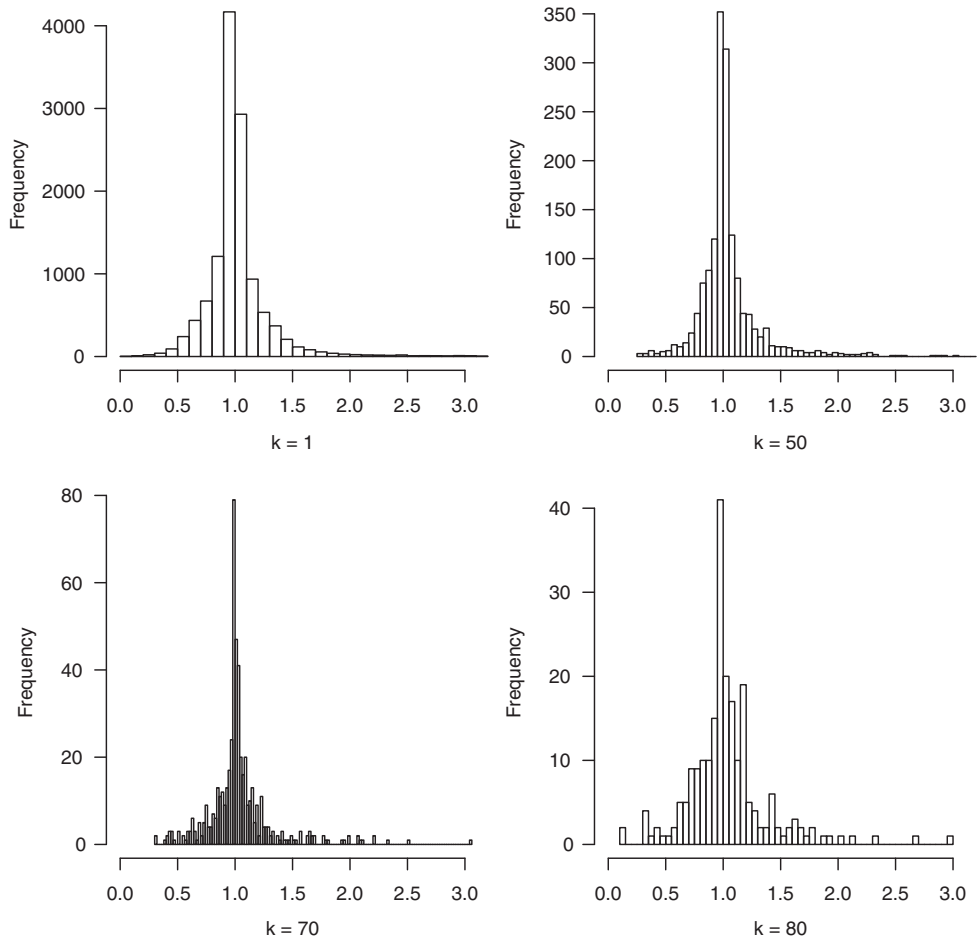

Figure 4. Histograms for estimated IPC weights in four chosen subsets ($k = 1$, 50, 70 and 80).

using jackknife estimated standard errors. The HAART covariate is a indicator of treatment, saying whether an individual receives treatment at a certain time. The lagged values are the values 3 months before the current value, corresponding to the scheduled time between study visits [11]. All programming was performed in the statistical package R, version 2.10.1 [20].

The bias due to censoring (both normal and artificial) was not substantial, as an unweighted model gave an estimated hazard ratio of 0.176 (95 per cent CI: 0.105–0.296). Histograms for estimated censor weights in four chosen subset ($k = 1$, 50, 70, and 80) are given in Figure 4. We see that the stabilized weights for the chosen subsets are around 1, and that most weights are in the region above 0 and below 2. The maximum value for the weights in the different subsets decreases with higher k, being namely 11.37, 5.26, 3.05, and 2.96 for the chosen subsets, respectively.

Table I lists the estimated hazard ratios for HAART versus no treatment, overall (IPC weighted and unweighted estimates), and for the subgroups of injecting drug users (IDU), non-IDU, baseline CD4 $\geqslant$200 and baseline CD4 <00. The results show a greater effect of treatment in the non-IDU group compared with the IDU group, with hazard ratios of HAART versus no treatment of 0.142 (95 per cent CI: 0.054–0.375) and 0.220 (95 per cent CI: 0.087–0.560), respectively. For the subgroups of individuals with baseline CD4 count greater and less than 200, the treatment effect was a lot better for the patients in the low CD4 group, with hazard ratios of 0.328 (95 per cent CI: 0.193–0.559) and 0.039 (95 per cent CI: 0.007–0.213), respectively.

Our estimated hazard ratio of 0.165 (0.079-0.343), when looking at the overall effect of HAART, is close to the results of Sterne *et al.* [11], where the hazard ratio for HAART versus no treatment was estimated to be 0.14 (0.07–0.29). When fitting an MSM close to the one used by Sterne *et al.*, leaving out the IPC weights and using only the IPT weights, we estimate the hazard ratio of HAART versus no treatment as 0.16 (95 per cent CI: 0.08–0.34), which again is close to our effect estimate without IPC weights, 0.176 (95 per cent CI: 0.105–0.296).

The sequential Cox method also allows us to estimate the treatment effect while grouping individuals by different levels of CD4 count at treatment start. Table II lists hazard ratios of four such analyses. The results show that there is a greater effect on individuals with low CD4 level at treatment start, for any of the chosen cut-off values between high

Copyright © 2010 John Wiley & Sons, Ltd.

Table I. Estimated hazard ratios for HAART versus no treatment, overall and for subgroups. IPC weights were included in all analyses unless otherwise stated. Analyses for subgroups were performed including interaction terms for grouping indicator variables and HAART. Confidence intervals are found using jackknife estimates.

	Hazard ratio	95 per cent CI
Overall effect (without IPC weights)	0.176	0.105–0.296
Overall effect	0.165	0.079–0.343
Non-IDU	0.142	0.054–0.375
IDU	0.220	0.087–0.560
Baseline CD4 $\geqslant$200	0.328	0.193–0.559
Baseline CD4 $<$200	0.039	0.007–0.213

Table II. Estimated hazard ratios for HAART versus no treatment for different groupings of CD4 count at treatment start. Analyses were performed including interaction terms for grouping indicator variables and HAART. Confidence intervals are found using jackknife estimates.

CD4 count at treatment start	Hazard ratio	95 per cent CI
CD4 $\geqslant$200	0.402	0.241–0.684
CD4 $<$200	0.066	0.019–0.229
CD4 $\geqslant$350	0.445	0.239–0.870
CD4 $<$350	0.080	0.057–0.325
CD4 $\geqslant$500	0.530	0.323–2.183
CD4 $<$500	0.122	0.076–0.374
CD4 $\geqslant$750	0.695	0.095–7.435
CD4 $<$750	0.155	0.078–0.363

and low CD4 count at treatment start, with similar trends as for the analysis for high and low CD4 at baseline. We see a greater effect when we lower the cut-off between the CD4 groups, and that the 95 per cent confidence intervals do not overlap for the effect estimates in the $\geqslant$200 and the $<$200 groups.

Using the sequential Cox approach one could also analyse selected mimicked randomized controlled trials separately. One example would be to divide into early and late treatment start, in terms of reference intervals. This was performed choosing a cut-off between early and late treatment start at 12 months since inclusion in the study. An additional CD4 decline covariate was investigated but had no impact. The results from this analysis show that the treatment effect is highest for the individuals with early treatment start (at observation time $\leqslant$ 12 months), compared with individuals with late treatment start (at observation time $>$12 months), with hazard ratios for HAART versus no treatment of 0.101 (95 per cent CI: 0.038–0.265) and 0.294 (95 per cent CI: 0.131–0.656).

6. Discussion

Looking at the overall effect of HAART, our results using a sequential Cox model are close to the results using an MSM, as was done in Sterne *et al.* Effect estimates in sub groups, such as non-IDU, IDU, and individuals with baseline CD4 $\geqslant$200 and $<$200, were also similar. We see that the difference between the estimated hazard ratios in models with and without IPC weights is similar using both methods, indicating that our extra artificial censoring due to later treatment start does not create much extra bias. Using both methods the estimated hazard ratio for HAART versus no treatment is lower when including censoring weights, but the vast majority of the correction in both analyses is due to time-dependent confounding.

One of the motivations behind the sequential Cox approach was to look at alternatives to IPT weighting. In the sequential Cox method the IPT weights are avoided, partly by using artificial censoring to censor individuals (not in treatment at the start of the mimicked trial) at later treatment start. It is to be expected that individuals with certain covariate histories are more likely to get artificially censored due to later treatment start than others, which would make the artificial censoring dependent on disease history. In addition, ordinary censoring could also be dependent. To adjust for this bias, both types of dependent censoring are accounted for using IPC weighting. Note that IPC weights are more stable than IPT weights. The problem of unstable IPT weights is based on the fact that the weights for individuals on treatment are calculated using the inverse of the probability of starting treatment. That way, an estimated small probability of starting treatment will give a large weight. IPC weights are only calculated using the probability of not being censored, hence, there are usually no situations which would involve dividing by a number close to zero. Thus, the same problem of unstable weights is not present.

Copyright © 2010 John Wiley & Sons, Ltd.

In summary, we made five main assumptions for our estimate of the treatment effect to be a causal estimate; these are mentioned in Sections 2 and 3: (i) the chosen covariates are sufficient to adjust for confounding, (ii) the model for estimating the hazard rate is correct, (iii) the model estimating the weights used to adjust for any dependent censoring is correct, (iv) the effect of treatment is the same in all mimicked trials, and (v) the effect is the same for all covariate histories before the start of the mimicked trials given covariates at the starting time. Assumptions (i)–(iii) are closely related to the assumption of no unmeasured confounders, which generally is not testable. Assumption (iv) can be checked or relaxed. A way to check it is to estimating separate treatment effects for individuals with different times of treatment start. For instance, instead of assuming a constant hazard over all mimicked randomized controlled trials, we could group individuals starting treatment early or late, as we did with a cut-off at 12 months in Section 5. The results showed a greater effect for the individuals starting treatment early, but the estimates had wide confidence intervals. We chose to only divide into two time intervals, one early and one late treatment group, principally to have a reasonable number of observations in both; but one could choose different groupings or resolutions. Alternatively, assumption (iv) can be relaxed, and the overall estimate interpreted as an aggregated effect over all the mimicked trials. By aggregated, we mean as a kind of weighted average based on the composite likelihood.

Avoiding some of the problems associated with estimating the IPT weights could potentially lead to lesser variance in the effect estimates. Quantifying this decrease of variance would be computationally difficult, with regards to both fitting an MSM using IPT weights and a sequential Cox model, as weights would need to be re-estimated for each new bootstrap sample.

Assessing sensitivity for these methods is clearly also of importance. Even though the parameter being estimated using the sequential Cox approach is not identical to the one in an MSM, they should not intuitively be very different. We therefore think that the general agreement between the results of our approach and those found by the MSM method [11] strengthens the general validity of both analyses. A second important aspect concerns the issue of whether there are unmeasured confounders. One possibility is to assume the existence of unmeasured confounders with a given amount correlation with both treatment and outcome. The effect of this could be analysed by simulating several data sets with such confounders and re-doing the analysis.

Using our sequential Cox approach one could end up creating a pseudo data set of considerable size. We found the pseudo data set constructed from the Swiss HIV Cohort data to be of appropriate size after performing a minor change in the representation of the data. Namely using only one row of data to represent successive months without covariate change, and then marking these rows with interval start and stop times. In an unweighted analysis this would lead to exactly the same results, but in a weighted analysis it would mean that the IPC weights for an individual are updated less frequently. For our data, the impact of this was minimal. In general, the sequential Cox method can be applied to constructed pseudo data sets with different resolution with regards to both data representation and interval length.

When comparing results from the sequential Cox approach and the MSM, even though the results are similar, one should be aware that there might be certain differences in the interpretation of the effect measure being estimated. With the sequential Cox method, we estimate the effect on the hazard (after the starting time of interval k) on starting treatment in interval k, given the covariate history up to interval k and given that treatment has not started before. Assuming that this effect is the same for all reference intervals we use the composite likelihood to get an estimated causal effect from the overall effect estimate. Alternatively, we relax this assumption and interpret the overall effect estimate as an aggregated, or weighted average, effect over all observed times of treatment start. This would not be exactly the same as for the MSM, which estimates the marginal effect on the hazard at time t, of having started treatment some time before t. A noticeable difference between the two methods is therefore that in the sequential Cox method we adjust for covariates both at baseline and at the starting time of the particular constructed trial, whereas in the MSM it is adjusted only for values at baseline. However, the opportunity to analyse how treatment effects depend on covariates values (such as CD4 or HIV-1 RNA) at the time treatment starts could be seen as an advantage. Analyses such as those presented in Table II, where treatment effects were estimated for different levels of CD4 at treatment start, cannot be easily done with a MSM. Estimates of such treatment effects are very interesting from a strategic standpoint. In two recent papers [22, 23] the issue of when to start treatment was also discussed from a different point of view, not directly comparable to our analysis. The current discussion is whether treatment should be started with CD4 counts above 350 or above 500, and randomized controlled trials are planned on this subject [22].

As in most models constructed for estimating the causal effects of treatment using data from observational studies, such as the sequential Cox method and the MSM, or other approaches such as G-estimation for Structural Nested Models [24, 25], one can say that the underlying idea is to mimic randomized controlled trials: to simulate experiments where there is no confounding related to treatment. The effect estimates based on such simulated experiments could be interpreted as causal, counterfactual, effects, given some model assumptions, such as the assumption of no unmeasured confounders. Note that the mimicked randomized controlled trials could be different, as is the case in the MSM and the sequential Cox approach. The MSM mimics a trial where the time of treatment start is random and independent

Copyright © 2010 John Wiley & Sons, Ltd.

of any covariate history, whereas the sequential Cox model mimics a sequence of trials comparing individuals starting treatment at a certain time with those who do not, conditioning on the covariates at that time.

Using the sequential Cox approach the construction of the mimicked randomized controlled trials is done directly and in an intuitive way, manipulating and creating subsets of the observed data set. Sets of valid treatment and control groups are constructed based on individuals starting treatment and individuals not on treatment in certain reference intervals. When such sets are constructed for all observed time intervals as reference, these sets can be analysed together, resulting in an overall effect measure. Once all the subsets of the original data are constructed, the sequential Cox method is easy to implement in standard software, using a stratified weighted Cox analysis.

Acknowledgement

This work was supported by the Research Council of Norway, contract/grant number: 170620/V30.

References

1. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570. DOI: 10.1093/aje/kwi216.
2. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560. DOI: 10.1097/00001648-200009000-00011.
3. Hernan MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic and Clinical Pharmacology and Toxicology* 2006; **98**:237–242. DOI: 10.1111/j.1742-7843.2006.pto_329.x.
4. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JAE, Robins JM. Observational studies analyzed like randomized experiments. An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008; **19**(6):766–779. DOI: 10.1097/EDE.0b013e3181875e61.
5. Lu B. Propensity score matching with time-dependent covariates. *Biometrics* 2005; **61**(3):721–728. DOI: 10.1111/j.1541-0420.2005.00356.x.
6. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* 2006; **34**:70–85. DOI: 10.1111/j.1467-9469.2006.00529.x.
7. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2005; **163**(3):262–270. DOI: 10.1093/aje/kwj047.
8. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies estimating a population mean from incomplete data. *Statistical Science* 2007; **22**(4):523–539. DOI: 10.1214/07-STS227.
9. Martens EP, Pestman WR, de Boer A, Belitser S, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International Journal of Epidemiology* 2008; 1–6. DOI: 10.1093/ije/dyn079.
10. Ledergerber B, Egger M, Opravil M, Telenti A, Hirschel B, Battegay M, Vernazza P, Sudre P, Flepp M, Furrer H, Francioli P, Weber R. Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. Swiss HIV Cohort Study. *Lancet* 1999; **353**(9156):863–868. DOI: 10.1016/S0140-6736(99)01122-8.
11. Sterne JAC, Hernan MA, Ledergerber B, Tilling K, Weber R, Sendi P, Rickenbach M, Robins JM, Egger M. Swiss HIV Cohort Study. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet* 2005; **366**:378–384. DOI: 10.1016/S0140-6736(05)67022-5.
12. Aalen OO, Borgan Ø, Gjessing HK. *Survival and Event History Analysis*: *A Process Point of View*. Springer: New York, 2008.
13. Breslow N. Analysis of survival data under the proportional hazards model. *International Statistical Review* 1975; **43**:45–48.
14. Satten GA, Datta S, Robins J. Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters* 2001; **54**:397–403. DOI: 10.1016/S0167-7152(01)00113-4.
15. Datta S, Satten GA. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics* 2002; **58**:792–802. DOI: 10.1111/j.0006-341X.2002.00792.x.
16. Gunnes N, Borgan Ø, Aalen OO. Estimating stage occupation probabilities in non-Markov models. *Lifetime Data Analysis* 2007; **13**:211–240. DOI: 10.1007/s10985-007-9034-4.
17. Lindsay B. Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, Prabhu NU (ed.). American Mathematical Society: Providence, RI, 1988.
18. Varin C, Vidoni P. A note on composite likelihood inference and model selection. *Biometrika* 2005; **92**:519–528. DOI: 10.1093/biomet/92.3.519.
19. Varin C. On composite marginal likelihoods. *Advances in Statistical Analysis* 2008; **92**:1–28. DOI: 10.1007/s10182-008-0060-7.
20. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2008; DOI: 10.1007/978-3-540-74686-7.
21. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall: New York, 1993.
22. Sterne JA, May M, Costagliola D, de Wolf F, Phillips AN, Harris R, Funk MJ, Geskus RB, Gill J, Dabis F, Miró JM, Justice AC, Ledergerber B, Fätkenheuer G, Hogg RS, Monforte AD, Saag M, Smith C, Staszewski S, Egger M, Cole SR, Brodt HR, Casabona J, Chêne G, del Amo J, Guest J, Hogg R, Justice A, Kitahata M, Lampe F, Mocroft A, Reiss P. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet* 2009; **373**(9672):1352–1363. DOI: 10.1016/S0140-6736(09)60612-7.
23. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC, Hogg RS, Deeks SG, Eron JJ, Brooks JT, Rourke SB, Gill MJ, Bosch RJ, Martin JN, Klein MB, Jacobson LP, Rodriguez B, Sterling TR, Kirk GD, Napravnik S, Rachlis AR, Calzavara LM, Horberg MA,

Copyright © 2010 John Wiley & Sons, Ltd.

Silverberg MJ, Gebo KA, Goedert JJ, Benson CA, Collier AC, Van Rompaey SE, Crane HM, McKaig RG, Lau B, Freeman AM, Moore RD. NA-ACCORD Investigators. Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine* 2009; **360**(18):1815–1826. DOI: 10.1056/NEJMoa0807252.
24. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**:319–336.
25. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2002; **2**(2):164–182.

Copyright © 2010 John Wiley & Sons, Ltd.