STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2325–2337
Published online 4 June 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3629

A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease

W. M. van der Wal[1,*,†], M. Prins[2,3], B. Lumbreras[4] and R. B. Geskus[1,3]

[1] *Academic Medical Center, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, P.O. Box 22660, 1100 DD Amsterdam, The Netherlands*
[2] *Academic Medical Center, Division of Infectious Diseases, Tropical Medicine and AIDS, University of Amsterdam, The Netherlands*
[3] *Amsterdam Health Service, The Netherlands*
[4] *Department of Public Health, University Miguel Hernández, CIBER in Epidemiology and Public Health, Spain*

SUMMARY

Progression of a chronic disease can lead to the development of secondary illnesses. An example is the development of active tuberculosis (TB) in HIV-infected individuals. HIV disease progression, as indicated by declining CD4+ T-cell count (CD4), increases both the risk of TB and the risk of AIDS-related mortality. This means that CD4 is a time-dependent confounder for the effect of TB on AIDS-related mortality. Part of the effect of TB on AIDS-related mortality may be indirect by causing a drop in CD4. Estimating the total causal effect of TB on AIDS-related mortality using standard statistical techniques, conditioning on CD4 to adjust for confounding, then gives an underestimate of the true effect. Marginal structural models (MSMs) can be used to obtain an unbiased estimate. We describe an easily implemented algorithm that uses G-computation to fit an MSM, as an alternative to inverse probability weighting (IPW). Our algorithm is simplified by utilizing individual baseline parameters that describe CD4 development. Simulation confirms that the algorithm can produce an unbiased estimate of the effect of a secondary illness, when a marker for primary disease progression is both a confounder and intermediary for the effect of the secondary illness. We used the algorithm to estimate the total causal effect of TB on AIDS-related mortality in HIV-infected individuals, and found a hazard ratio of 3.5 (95 per cent confidence interval 1.2–9.1). Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: marginal structural models; G-computation; causal effect; HIV; AIDS

1. INTRODUCTION

During the progression of a chronic disease, secondary illnesses often develop. For instance, declining CD4+ T-cell count (CD4), an important indication of HIV disease progression, leads to the development of opportunistic infections in HIV-infected individuals. *Active tuberculosis*

[*] Correspondence to: W. M. van der Wal, Academic Medical Center, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, P.O. Box 22660, 1100 DD Amsterdam, The Netherlands.
[†] E-mail: w.m.vanderwal@amc.uva.nl

Received 6 June 2008
Accepted 24 April 2009

(TB), the disease that may develop after infection by *Mycobacterium tuberculosis*, is such an opportunistic infection. In HIV-infected individuals, TB is associated with an increased risk of AIDS-related death [1–5]. It remains largely unknown whether the association between TB and AIDS-related mortality is only due to TB and death both being more likely at low CD4, or whether TB in itself has a causal effect on AIDS-related mortality. TB may have both a direct and an indirect effect on AIDS-related death, the latter by causing an additional CD4 drop. We denote the sum of the direct and indirect effect as the *total* causal effect.

With CD4 being both a time-dependent confounder as well as intermediary for the effect of TB, analysis of the total causal effect of TB using standard models that condition on CD4 by including it as a covariate, leads to bias [6]. Bias can occur because the indirect effect of TB through intermediary CD4 is lost, as illustrated in Section 2. Moreover, estimates from such conditional models are typically not true causal effect estimates, and can induce selection bias, as further explained in Section 2. In such a situation *marginal structural models* (MSMs) [7] can be used to estimate the total causal effect, including the indirect effect. Effect estimates obtained from MSMs, with all confounders measured and adjusted for and with no model misspecification, are comparable to estimates obtained from randomized experiments. In our study, the use of an MSM provides for a quantification of the total causal effect of TB on AIDS-related mortality, which is only obtainable from observational data, since randomization of TB is impossible.

At present, inverse probability weighting (IPW) [7] is the most commonly used method for fitting MSMs. Another method to fit MSMs is G-computation [8]. IPW is typically easier to implement than G-computation. G-computation often requires more complicated, iterative procedures, as noted in Section 4. However, we will show that when estimating the effect of a secondary disease, correcting for a marker for primary disease development, a simplified form of G-computation can be used. This simplified form is as easily implemented as IPW. IPW uses a model for the exposure allocation given confounding variables, to correct for confounding. G-computation uses a model from which the outcome of interest is predicted given exposure and confounding variables. Therefore, in situations where exposure is more difficult to predict than the outcome, the use of G-computation is potentially preferable to the use of IPW.

In our TB example, we assume that not the measured value of CD4, but some unobservable '*true*' value, separate from short-term fluctuations and measurement error, is a marker for HIV disease progression. True CD4 is a confounder and intermediary for the effect of TB on mortality. The true CD4 trajectory may depend on individual baseline parameters and the moment of onset of TB only. The use of such baseline parameters leads to an easily implemented G-computation algorithm for estimating the total causal effect of a secondary illness on survival, correcting for primary disease development. This simplified G-computation algorithm will be described in detail below, making it accessible to a wide range of potential users.

2. BASIC CAUSAL STRUCTURE

First, we will clarify the causal relations that exist between our exposure of interest (TB), the marker for disease progression that is a time-dependent confounder and intermediary for the effect of the exposure (CD4), and the outcome of interest (AIDS-related mortality). We assumed the following causal structure, over time since HIV seroconversion (Figure 1). At a given time t, TB status has a causal effect on AIDS-related death status Y. At the same time point, CD4 has both a causal effect on death status and on TB status. At a previous time $t-1$ the same causal

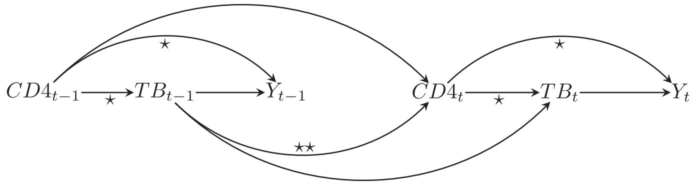

Figure 1. Assumed causal structure of the interaction between true CD4, TB status and AIDS-related mortality over time after HIV seroconversion.

structure exists. Suppose we want to estimate the effect of TB on AIDS-related mortality using Cox proportional hazards regression. A standard Cox model regressing the hazard of AIDS-related death on time-dependent TB status only would produce a biased estimate since CD4 is a time-dependent confounder for the effect of TB on death (as indicated by ⋆ in Figure 1). Conditioning on CD4 to correct for the time-dependent confounding, by including it in the Cox model, will not result in a correct estimate of the total causal effect of TB, for the following reasons:

1. When TB has a causal effect on subsequent CD4 development (as indicated by ⋆⋆ in Figure 1), TB has an indirect effect on death through CD4. This indirect effect would be ‘adjusted away’ by conditioning on CD4 [6, 8].
2. The estimate would be a conditional estimate, which is unequal to the marginal causal effect due to non-collapsibility [9].
3. Even if CD4 itself were not a confounder, but there exists an unmeasured confounder for the effect of CD4 on AIDS-related mortality such as HIV RNA level, selection bias will arise by conditioning on CD4 [10]. This is a form of Berkson’s bias [11].

MSMs do not suffer from these drawbacks [6–10]. Therefore, we will use an MSM to estimate the total causal effect of TB on AIDS-related mortality, adjusting for confounding by CD4. We will fit the MSM using G-computation.

3. INTRODUCTION TO G-COMPUTATION

To illustrate the basic principle of G-computation to fit an MSM, consider first the case of a point treatment study—a study in which we estimate the effect of an exposure, given at a fixed time point. Let A indicate the exposure variable. Then within an individual i we can indicate all *potential outcomes*—the outcomes that might have been observed with exposure set to all possible levels a—as $Y_i|\text{DO}[A_i=a]$. ‘DO’ indicates Pearls ‘*DO*’ operator [12], which can be interpreted as ‘to intervene and actively set exposure to level a’. Note that only one of these potential outcomes, or *counterfactuals*, can be observed simultaneously—the outcome that is observed under the exposure level that has actually been given to individual i. The distribution of potential outcomes over all individuals, with A set to all possible levels a, is indicated as

$$f(y|\text{DO}[A=a]) \tag{1}$$

If this distribution were known, we could evaluate the causal effect of A on Y by applying a chosen contrast to it, such as the risk ratio, for instance investigating whether, on average, exposure increases the risk of death within individuals. However, we only observe the distribution of outcomes for the exposures that are actually received by the individuals in the sample. Evaluating the causal effect of A on Y using this latter distribution, the observed data, would give a biased estimate when there exists a confounder L that affects both the exposure allocation as well as the outcome.

To evaluate the causal effect of A on Y in the presence of a confounder L, we can use the following approach [8]. The first step is to factorize (1) over the distribution of possible values l of L as

$$f(y|\mathrm{DO}[A=a])=\int_l f(y|\mathrm{DO}[A=a], L=l)f(l)\,\mathrm{d}l \tag{2}$$

In discrete terms, (2) can be thought of as the weighted sum of distributions of potential outcomes for all exposure levels, given specific observed values l of L—weighted by the frequencies with which those specific values l occur. In other words, within subgroups of individuals who have the same value of L, there are distributions of potential outcomes, and those distributions are summed together, weighted by the sizes of those subgroups. This practice is known in epidemiology as *standardization* (see e.g. [13], Chapter 14). Note that (2) is still unobservable, since individuals only receive one specific exposure level.

To rewrite (2) into an observable quantity, we need the assumption that $Y_i|A_i=a,\ L_i=l$, the observed outcome for an individual i with actual value l on the confounder L and who has received exposure level a, is always equal to $Y_i|DO[A_i=a],\ L_i=l$, the potential outcome with exposure *set* to a for an individual with actual value of l on the confounder L. This assumption is known as the consistency assumption [6]. We also need to assume that there are no other confounders than L. Under these two assumptions (2) is equal to

$$\int_l f(y|A=a, L=l)f(l)\,\mathrm{d}l \tag{3}$$

which, in discrete terms, can be thought of as the weighted sum of distributions of the outcomes for observed exposure levels, given specific observed values l of L—weighted by the frequencies with which those specific values l occur. In other words, within subgroups of individuals who have the same value of L, there are distributions of outcomes for all observed exposure levels, and those distributions are summed together, weighted by the sizes of those subgroups.

Only in simple cases can (3) be computed directly, to obtain an estimate of (1). However, we can estimate (1) in a different manner. We regress the outcome Y on both exposure A and confounder L, fitting a 'data model' $Y\sim(A,L)$ on the observed data. This data model describes the distribution $f(y|A=a, L=l)$. We then work backwards from (3) to (2)—using the data model, we predict potential outcomes $Y_i|\mathrm{DO}[A_i=a],\ L=l_i$ for each individual i under all possible a, with l_i his or her observed value of L. The distribution of L in the sample is assumed to be representative of the population distribution of L. In this manner, the distribution of potential outcomes $f(y|\mathrm{DO}[A=a])$ is approximated. When exposure is a continuous variable, values a can be randomly sampled. To improve precision, the sampling of a followed by imputation of y can then be done multiple times.

4. PROPOSED SIMPLIFIED G-COMPUTATION ALGORITHM

When dealing with a longitudinal confounder $L(t)$ such as true CD4, and an exposure varying over time $A(t)$ such as TB status, factorization of the potential outcome distribution over the observed distribution of $L(t)$ becomes much more complicated. The temporal structure of causal effects has to be taken into account [14]. In general, this leads to G-computation algorithms in which drawing of $A(t)$, $L(t)$ and outcome $Y(t)$ has to be done at multiple time points [15]. We will show that when estimating the causal effect of a secondary illness when a marker for primary disease progression is both a confounder as well as intermediary for the effect of the secondary illness, a factorization over estimated baseline parameters that describes the development of the marker can be used. This will yield a G-computation algorithm as outlined in Section 3 for a point treatment, which is considerably less complex.

HIV disease progression, as indicated by true CD4 in our study, is a continuous process. Only a subset of the CD4 process is observed, with error, namely the CD4 measurements. We chose to model CD4 with a longitudinal random effects model. With L representing CD4 and A representing TB, we parameterize true CD4, separate from measurement error and short-term fluctuations, via

$$\sqrt{L_i(t)} = \xi_i + \eta_i t + \beta_2 A_i(t-1) \tag{4}$$

The square root transformation is used to obtain more normally distributed random effects and error terms [16]. Random effects ξ_i and η_i are normally distributed with mean $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ and covariance matrix

$$\Sigma = \begin{bmatrix} V_1 & V_{12} \\ V_{12} & V_2 \end{bmatrix}$$

The model includes a fixed effect for TB, β_2. We assume that CD4 is affected by the TB status *right before* t, as in Figure 1, with time points t and $t-1$ infinitely close together, since CD4 can vary continuously. $A(t-1)$ is the TB status 1 day before t, approximating the TB status right before t. We also assume that TB has a lasting effect on mortality. Therefore, TB status switches at most once from 0 to 1, at the first occurrence of TB after HIV seroconversion. Individual true CD4 development is then determined by individual baseline parameters ξ_i and η_i, in addition to the moment of onset of TB. The potential outcome distribution can then be factorized as

$$f(y|DO[\bar{A}=\bar{a}]) = \int_{\xi,\eta} f(y|DO[\bar{A}=\bar{a}], \xi, \eta) f(\xi,\eta)\, \mathrm{d}\xi\, \mathrm{d}\eta = \int_{\xi,\eta} f(y|\bar{a}, \xi, \eta) f(\xi,\eta)\, \mathrm{d}\xi\, \mathrm{d}\eta \tag{5}$$

with Y now indicating the AIDS-related death time. $\bar{A}$ indicates the TB history and $\bar{a}$ indicates a specific course of that TB history. For example, the histories 'never TB' or 'TB 2 years after seroconversion' are possible values of $\bar{A}$. The middle part of (5) can be thought of as the weighted sum of distributions of potential outcomes for all possible TB histories given specific combinations of ξ and η, weighted by the frequencies with which those combinations of ξ and η are observed.

Integral (5) can be approximated analogously to the approach described for a point treatment in Section 3. With the observed data, we can fit a data model, regressing AIDS-related mortality on a function of the baseline parameters ξ and η (e.g. fitted CD4), and time-dependent TB status. Using the data model, we can then predict mortality for individuals, as a function of chosen TB development scenarios, to approximate the distribution of potential outcomes. Using this

DOI: 10.1002/sim

distribution, the causal effect of TB on mortality can be evaluated. The details of the algorithm are as follows:

G-computation algorithm

1. Longitudinal CD4 modelling.
   (a) Fit CD4-model (4) on the observed data, and save individual best linear unbiased prediction (BLUP) estimates [17] of the individual CD4 parameters, $\hat{\xi}_i$ and $\hat{\eta}_i$.
   (b) t_j denotes the observed event times (days since HIV seroconversion), sorted in ascending order.
   (c) Indicate observed TB status

$$A_i(t_j) = \begin{cases} 0, & t_j < T_{\text{TB},i} \\ 1, & t_j \geqslant T_{\text{TB},i} \end{cases}$$

   given $T_{\text{TB},i}$, the time of first observed TB after HIV seroconversion, for $t_j \leqslant T_i^f$, with T_i^f the individual end time (time of death or censoring).
   (d) Compute CD4 $L_i(t_j)$ with (4) given $\hat{\xi}_i$, $\hat{\eta}_i$ and $A_i(t_j - 1)$, for $t_j \leqslant T_i^f$.
2. Fit the data model.
   (a) Fit the data model using imputed CD4 $L_i(t_j)$ and observed TB history $A_i(t_j)$ for $t_j \leqslant T_i^f$, and individual survival information. We chose the Cox model

$$\lambda_i(t) = \lambda_0(t) \exp\{\delta_1 \sqrt{L_i(t)} + \delta_2 A_i(t)\} \tag{6}$$

   (b) Compute the Breslow estimate of the cumulative baseline hazard, $\Lambda_0^B(t_j)$.
3. Draw TB times and compute CD4 development given drawn TB times.
   (a) Calculate the median[‡] T_m of the observed event times t_j.
   (b) Draw for each individual a random TB time $T_i^\star$ from the uniform distribution on $[0, T_m]$. $A_i^\star(t_j)$ is the corresponding TB status at t_j.
   (c) Compute CD4 $L_i^\star(t_j)$, given $\hat{\xi}_i$, $\hat{\eta}_i$ and $A_i^\star(t_j - 1)$, for $t_j \leqslant T_m$, using (4).
   (d) Compute the cumulative distribution of event times[§] at t_j, using $F_i^\star(t_j) = 1 - \exp[-\sum_{r=1}^{j} \Delta\Lambda_0^B(t_r) \exp\{\delta_1 \sqrt{L_i^\star(t_r)} + \delta_2 A_i^\star(t_r)\}]$, for $t_j \leqslant T_m$, with $\Delta\Lambda_0^B(t_j)$ the observed increment in $\Lambda_0^B(t_j)$ at time t_j.
4. Simulate potential outcomes and fit MSM.
   (a) Use the calculated values $F_i^\star(t_j)$ to simulate a random event time $T_i^\star = \min(t_j : F_i^\star(t_j) \geqslant u_i)$, with u_i drawn from the uniform distribution on [0, 1].
   (b) Fit the MSM; we chose the Cox model

$$\lambda_i(t) = \lambda_0(t) \exp\{\theta A_i(t)\} \tag{7}$$

[‡] To avoid excessive extrapolation, we draw TB times, compute CD4 development and simulate potential outcomes only within the interval $[0, T_m]$, see also Section 7.

[§] We use a discrete approximation of the cumulative hazard given by $\Lambda_i(t) = \int_0^t \exp\{\delta_1 \sqrt{L_i^\star(s)} + \delta_2 A_i^\star(s)\} d\Lambda_0^B(s)$, see, for instance, [18, p. 34].

DOI: 10.1002/sim

5. Enhance precision by performing steps 3 and 4 multiple times.
   (a) Step 3 is performed p times, and step 4 is performed q times after each execution of step 3, yielding pq estimates $\hat{\theta}_{kl}$, with $k=1\dots p$ and $l=1\dots q$.
   (b) Averaged over both loops, our estimate of the total causal effect of TB on AIDS-related death is $\hat{\theta}=(1/pq)\sum_{k=1}^{p}\sum_{l=1}^{q}\hat{\theta}_{kl}$.

Note that we did not predict outcomes for multiple TB histories within each individual followed by the evaluation of the effect of TB on death directly. We predicted outcomes for only one simulated TB history within each individual and then evaluated the effect of TB on death, but repeated this to obtain an estimate of θ. This latter approach is easier to program, whereas in our example the former approach required to much computer memory.

To compute a confidence interval for θ, we used the bootstrap with bias correction [19], with 500 bootstrap repetitions. The entire algorithm as outlined above was applied to each bootstrap sample. We included the baseline covariates gender, age at seroconversion and study site in both (4) and (6) while analyzing the observed data, but did not use these baseline covariates in the simulation study below for simplicity. We chose the number of repetitions for both loops as $p=5$ and $q=5$, to increase precision while still being practical when performing a simulation study. However, when analyzing the observed data using our algorithm, we used $p=10$ and $q=10$ to improve precision, since the number of events and TB cases were substantially lower than in the simulated data.

5. SIMULATION STUDY

We assessed the performance of our G-computation algorithm and standard methods in a simulation study. Survival data were generated using the algorithm described in the Appendix. We generated data sets of 400 individuals, using the parameters in Table I. The effect of TB on mortality was either both direct and indirect through CD4, indirect only or direct only, with hazard ratios (HRs) corresponding to the total effect of 6.0 2.2 and 2.7, respectively. These parameter values resulted in simulated data with approximately 85 (indirect only and direct only) or 100 (both direct and indirect) events and 110 TB cases on average. The number of TB cases and events in the simulated data are larger than in the observed data, to avoid the problem of monotone likelihood that can occur in simulations with low event rates and highly predictive covariates [20]. The effect of TB on death was estimated using a standard Cox model including TB only, a standard Cox model including TB and CD4, and using our G-computation algorithm. Confidence intervals obtained from the standard models are Wald-based, estimated using the `survival` package in `R`. In each setting 500 simulation runs were performed.

Table II contains the results of our simulation study. *Mean bias* refers to the mean of the differences between parameter estimates and the real causal effect $\delta_1\beta_2+\delta_2$. When the *standardized mean bias* differs no more than 1.96 from 0, we attribute the difference to random sampling fluctuation. RMSE is the root of the mean squared difference between parameter estimates and the real value. *Coverage* is the proportion of simulation runs for which the 95 per cent confidence interval for a particular estimate contains the real value, 0.95 is desirable. *Average length* is the average length of the 95 per cent confidence interval.

Table I. Parameter values used in the simulation study, generating data with a secondary illness having an effect on death which is both direct and indirect through primary disease progression, indirect only or direct only. See appendix for a description of the parameters.

TB effect	β_0	β_1	β_2	V_1	V_2	V_{12}	γ_0	γ_1	λ_0	δ_1	δ_2	$T_{\max}$ (years)
Direct+indirect	26	−2	−4	5	0.25	−0.5	3	−0.2	1	−0.2	1	10
Indirect only	26	−2	−4	5	0.25	−0.5	3	−0.2	1	−0.2	0	10
Direct only	26	−2	0	5	0.25	−0.5	2	−0.2	1	−0.2	1	10

Table II. Simulation results: evaluation of our G-computation algorithm and standard methods.

	Estimate			95 per cent CI	
Condition/Method	Mean bias	Stand. mean bias	RMSE	Coverage	Average length
TB-effect direct+indirect					
Cox, no CD4	0.17	17.5	0.28	0.89	0.89
Cox, including CD4	−0.80	−61.6	0.85	0.21	1.17
G-computation	0.01	1.3	0.21	0.96	1.04
TB-effect indirect only					
Cox, no CD4	0.26	23.4	0.36	0.83	1.02
Cox, including CD4	−0.82	−55.1	0.89	0.33	1.40
G-computation	0.01	0.8	0.24	0.95	1.27
TB-effect direct only					
Cox, no CD4	0.24	22.7	0.34	0.85	0.98
Cox, including CD4	−0.01	−0.5	0.25	0.97	1.02
G-computation	0.01	1.1	0.28	0.95	1.01

When TB has both a direct and an indirect effect, the standard Cox model including only TB produces an estimate with a mean bias of 0.17 and a coverage of 0.89. This is an overestimate because of the confounding by CD4, with both TB and AIDS-related death being more likely at low CD4. The Cox model including TB and CD4 produces an estimate with a mean bias of −0.80, and a coverage of 0.21. The large underestimation of the true effect occurs because the large indirect effect is lost by conditioning on CD4. Our G-computation algorithm produces an unbiased estimate. The absolute bias of 0.01 can be attributed to chance as indicated by the standardized mean bias. The corresponding coverage of 0.96 is adequate.

When TB has an indirect effect only, both standard methods yield estimates significantly biased in the expected directions. The estimate from our G-computation algorithm is apparently unbiased, and the coverage of 0.95 is adequate.

When TB has a direct effect only, the estimate from the Cox model including only TB is biased, because of the confounding by CD4. However, the Cox model including both TB and CD4 produces an apparently unbiased estimate. The corresponding coverage of 0.97 is larger than 0.95, although not significantly different from 0.95 ($p=0.12$). With a direct effect only, our G-computation algorithm produces an unbiased estimate and adequate coverage of 0.95. The RMSE of 0.28 is larger than the value of 0.25 for the second Cox model, which is a significant difference when tested with a permutation test ($p=0.037$), indicating that the G-computation algorithm is slightly less efficient in this situation.

6. DATA EXAMPLE

After assessing the performance of our G-computation algorithm using simulation, we used it to estimate the causal effect of TB on AIDS-related mortality in HIV-positive injecting drug users.

We used data from two separate cohort studies, which were described in Van Asten *et al.* [21]. We included 276 HIV-infected injecting drug users, 78 from Amsterdam, the Netherlands and 198 from Valencia, Spain. We used time from HIV seroconversion as the principal time scale. For the Valencia cohort, the midpoint between last negative and first positive HIV test was used as an estimate of the HIV seroconversion date. Geskus [22] estimated HIV seroconversion dates for the Amsterdam cohort. TB diagnoses, both pulmonary and extrapulmonary, were recorded during follow-up. Individuals who developed TB received TB medication, which was taken under supervision of health-care workers. Hence, in our study we examined the causal effect of TB after HIV seroconversion on AIDS-related mortality, in the presence of TB medication. TB after HIV seroconversion occurred in 24 individuals during follow-up. Median follow-up length was 4.4 years (lower quartile = 3.0, upper quartile = 6.2), with a median of 2.1 (lower quartile = 1.1, upper quartile = 3.1) CD4 measurements per year. To avoid using data from the period during which HAART was available (after 1997), and to ensure that the end of follow-up was at most one year after the last known visit at which CD4 was measured, we determined end of follow-up as follows:

1. Subjects who did not die before January 1, 1997, were censored one year after their last known visit or at January 1, 1997, whichever came first.
2. For subjects who died before January 1, 1997, and died less than one year after their last known visit, the death date was used as end date.
3. Subjects who died before January 1, 1997, and died more than one year after their last known visit were censored at one year after their last visit.

During follow-up we observed 18 AIDS-related deaths, defined as a death indicated in the medical records as being AIDS- or HIV-related or as caused by an AIDS-defining illness.

Results obtained from analyzing the observed data are displayed in Table III, with estimated HRs for the effect of TB on AIDS-related mortality and 95 per cent confidence intervals. Using a standard Cox proportional hazards model regressing the hazard of AIDS-related death on TB, a significant HR of 9.4 was found. This is probably an overestimate, because death and TB are both more likely at low CD4. When including gender, age at seroconversion and study site in the model, to correct for possible confounding by baseline covariates, a similar HR of 9.7 was obtained. When also including CD4 as a time-dependent covariate, adjusting for the confounding by CD4, a non-significant HR of 1.4 was obtained. This is likely to be an underestimate of the

Table III. Estimates of the causal effect of TB on AIDS-related mortality.

	Hazard ratio = exp(parameter)		Parameter	
Method	Estimate	95 per cent CI	Estimate	95 per cent CI
Standard Cox, no covariates	9.4	3.6–25	2.24	1.3–3.2
Standard Cox, baseline covariates	9.7	3.7–26	2.27	1.3–3.2
Standard Cox, CD4 and baseline covariates	1.4	0.4–4.8	0.34	−0.92–1.6
G-computation	3.5	1.2–9.1	1.25	0.22–2.2

true effect because the indirect effect is lost, as explained in Section 2. Using our G-computation algorithm we estimated a HR of 3.5 (95 per cent CI 1.2–9.1), corresponding to the total causal effect of TB on AIDS-related mortality.

7. DISCUSSION

We described an easily implemented algorithm that uses G-computation to fit an MSM, estimating the causal effect of a secondary illness on progression to an event such as death. This algorithm can be used when a marker for primary disease development is a confounder as well as intermediary for the effect of the secondary illness. Our algorithm can be used as an alternative to IPW, for instance when the experimental treatment assumption (ETA) does not hold. The ETA is necessary when using IPW, and states that all levels of the exposure have a probability different from 0 or 1 of occurring at all possible values of the confounding variable. In addition, IPW requires a correctly specified model that allows prediction of the exposure given the value of the confounding variable. On the other hand, G-computation requires a correctly specified data model from which the outcome can be predicted given the exposure and the confounder. Depending on the situation, either IPW or G-computation may be a more feasible option.

Using our G-computation algorithm we inferred from our data that active TB has a significant causal effect on the hazard of AIDS-related death in HIV-infected injecting drug users. We estimated a HR for the total causal effect of TB of 3.5 (95 per cent CI 1.2–9.1), adjusted for confounding by CD4 count. This estimate is similar to the estimate obtained by López-Gatell *et al.* [23], using IPW to fit an MSM, of 4.0 (95 per cent CI 1.2–14). We also fitted an MSM to our data using IPW, and obtained an estimate of 6.3 (95 per cent CI 1.4–12.5). The difference between our G-computation and our IPW estimate is attributable to the fact that both procedures use non-saturated models, and therefore do not necessarily produce equal estimates. Note that we compared mortality for TB versus no TB. The latter group included both subjects with and without an AIDS diagnosis.

In our study, the G-computation algorithm extrapolates CD4 over time, given a drawn TB time, to generate a potential event time. The amount of extrapolation beyond the period during which CD4 measurements were made within each individual should be limited. The estimated baseline parameters describing CD4 development may only be appropriate within or close to the observed follow-up period. In our study, we chose to predict CD4 given a drawn TB time, and generate event times, within $t=[0, T_m]$, with t time since HIV seroconversion, and T_m the median observed event time, to avoid excessive extrapolation of CD4. Note that one might choose to censor predicted CD4 at the observed end times (observed censoring or observed death). That approach has the advantage that the computation of CD4 and the prediction of event times need to be done only within the observed follow-up period for each individual. However, informative censoring would be introduced. Simulated event times would be associated with the end times used to censor those simulated event times, because both simulated event times and observed event times are affected by the baseline parameters describing CD4 development.

Our G-computation algorithm is easily adapted to handle other outcome types, such as when estimating the causal effect of TB on subsequent CD4 development. Our algorithm cannot be used to estimate the effect of a non-randomized treatment when treatment decisions are based on the most recently measured value of a marker for primary disease progression, including short-term fluctuation and measurement error. In that case, development of the confounder over time cannot be adequately described using baseline parameters alone. However, when making treatment decisions,

DOI: 10.1002/sim

the long-term development of the marker over time is often taken into account. In that situation, it might not be necessary to correct for short-term fluctuations of the marker. Our algorithm can then be used to estimate the treatment effect.

APPENDIX A

Simulated data used in our simulation study was generated as follows. Time-dependent CD4 is modelled as

$$\sqrt{u_i(t)} = \xi_i + \eta_i t + \beta_2 A_i(t^-) \tag{A1}$$

similar to (4), with $t^- = \lim_{\Delta \downarrow 0} t - \Delta t$, the time point *right before t*. Parameters ξ_i and η_i are drawn from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\beta}, \Sigma)$ with mean $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ and covariance matrix

$$\Sigma = \begin{bmatrix} V_1 & V_{12} \\ V_{12} & V_2 \end{bmatrix}$$

The hazard of TB is modelled using an exponential model with a constant baseline hazard,

$$\lambda_{\text{TB},i}(t) = \gamma_0 \exp\{\gamma_1 \sqrt{u_i(t)}\} \tag{A2}$$

The hazard of death is also modelled with an exponential model with a constant baseline hazard,

$$\lambda_i(t) = \lambda_0 \exp\{\delta_1 \sqrt{u_i(t)} + \delta_2 A_i(t)\} \tag{A3}$$

Before TB, time-dependent CD4 is described as

$$\sqrt{u_i(t)} = \xi_i + \eta_i t \tag{A4}$$

substituting $A_i(t^-) = 0$ in (A1). Then the distribution of TB times, as derived from (A2) and (A4) is given by

$$F_{\text{TB},i}(t) = 1 - \exp\left\{ -\frac{\gamma_0 \exp(\gamma_1 \xi_i)}{\gamma_1 \eta_i} [\exp(\gamma_1 \eta_i t) - 1] \right\}, \quad \gamma_1 \eta_i \neq 0 \tag{A5}$$

The inverse function of (A5) is

$$F^{-1}_{\text{TB},i}(y) = \frac{1}{\gamma_1 \eta_i} \ln \left\{ 1 - \frac{\gamma_1 \eta_i}{\gamma_0 \exp(\gamma_1 \xi_i)} \ln(1-y) \right\}, \quad \gamma_1 \eta_i > 0 \tag{A6}$$

which is used to generate a TB time $T_{\text{TB},i} \in [0, \infty)$ by substituting y with a random value from the uniform distribution on $[0, 1)$. Given $A_i(t) = 0$ (which means that $A_i(t^-) = 0$ also) we can derive from (A3) and (A4) the distribution of event times

$$F_i(t) = 1 - \exp\left\{ -\frac{\lambda_0 \exp(\delta_1 \xi_i)}{\delta_1 \eta_i} [\exp(\delta_1 \eta_i t) - 1] \right\}, \quad \delta_1 \eta_i \neq 0 \tag{A7}$$

with inverse function

$$F_i^{-1}(y) = \frac{1}{\delta_1 \eta_i} ln \left\{ 1 - \frac{\delta_1 \eta_i}{\lambda_0 \exp(\delta_1 \xi_i)} \ln(1-y) \right\}, \quad \delta_1 \eta_i > 0 \tag{A8}$$

Equation (A8) is used to generate a potential event time $T_i^\star \in [0,\infty)$, given time-dependent CD4 and no TB, by substituting y with a random value from the uniform distribution on $[0,1)$. When $T_i^\star <= T_{\mathrm{TB},i}$ take event time $T_i = T_i^\star$. When $T_i^\star > T_{\mathrm{TB},i}$ however, an event time given occurrence of TB has to be generated.

Given $A_i(t)=1$ and $A_i(t^-)=1$, we can derive from (A1) and (A3) the distribution of event times

$$F_i^\dagger(t) = 1-\exp\left\{-\frac{\lambda_0 e^{\delta_1(\xi_i+\beta_2)+\delta_2}}{\delta_1\eta_i}(e^{\delta_1\eta_i t} - e^{\delta_1\eta_i T_{\mathrm{TB},i}})\right\}, \quad \delta_1\eta_i \neq 0, t \geqslant T_{\mathrm{TB},i} \tag{A9}$$

which has the inverse function

$$(F_i^\dagger)^{-1}(y) = \frac{1}{\delta_1\eta_i} ln\left\{-\frac{\delta_1\eta_i}{\lambda_0 e^{\delta_1(\xi_i+\beta_2)+\delta_2}}\ln(1-y) + e^{\delta_1\eta_i T_{\mathrm{TB},i}}\right\}, \quad \delta_1\eta_i > 0 \tag{A10}$$

Equation (A10) is used to generate a residual event time $T_i^\dagger \in [T_{TB,i},\infty)$, by substituting y with a random value from the uniform distribution on $[0,1)$. Given occurrence of TB, take $T_i = T_i^\dagger$.

Finally, a censoring time $T_{C,i}$ is drawn from the uniform distribution on $[0, T_{\max}]$ for a chosen $T_{\max}$. Compute the end time $T_i^f = \min(T_i, T_{C,i})$ and collect event time T_i when $T_i = T_i^f$. Collect TB time $T_{\mathrm{TB},i}$ when $T_{\mathrm{TB},i} \leqslant T_i^f$. With time t representing days since seroconversion, we have chosen to draw a CD4 measurement every 140 days from $t=0$ up to T_i^f, using (A1) to compute CD4, with $A_i(t^-)=0$ for $t \leqslant T_{\mathrm{TB},i}$ and $A_i(t^-)=1$ for $t > T_{\mathrm{TB},i}$. Normally distributed random noise with mean 0 and standard deviation 1 is added to $\sqrt{u_i(t)}$.

The total causal effect of TB on mortality is now given by $\delta_1\beta_2+\delta_2$, substituting (A1) in (A3).

REFERENCES

1. Del Amo J, Pérez-Hoyos S, Hernández Aguado I, Díez M, Castilla J, Porter K. Concerted action on seroconversion to AIDS and death in Europe (CASCADE) collaboration. Impact of tuberculosis on HIV disease progression in persons with well documented-time of HIV seroconversion. *Journal of Acquired Immune Deficiency Syndromes* 2003; **33**:184–190.
2. Selwyn PA, Hartel D, Lewis VA, Schoenbaum EE, Vermund SH, Klein RS, Walker AT, Friedland GH. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *The New England Journal of Medicine* 1989; **320**:545–550.
3. Perneger TV, Sudre P, Lundgren JD, Hirschel B. Does the onset of tuberculosis in AIDS predict shorter survival? Results of a cohort study in 17 European countries over years. AIDS in Europe Study Group. *British Medical Journal* 1995; **311**:1468–1471.
4. Whalen C, Horsburgh CR, Hom D, Lahart C, Simberkoff M, Ellner J. Accelerated course of human immunodeficiency virus infection after tuberculosis. *American Journal of Respiratory and Critical Care Medicine* 1995; **151**:129–135.
5. Badri M, Ehrlich R, Wood R, Pulerwitz T, Maartens G. Association between tuberculosis and HIV disease progression in a high tuberculosis prevalence area. *The International Journal of Tuberculosis and Lung Disease* 2001; **5**:225–232.
6. Robins JM, Greenland S, Hu FC. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 1999; **94**:687–700.
7. Hernán MA, Brumback BA, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**:561–570.
8. Robins JM. Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*. Springer: New York, 1997; 69–117.

DOI: 10.1002/sim

9. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**:29–46.
10. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.
11. Berkson J. Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin* 1946; **2**:47–53.
12. Pearl J. *Causality*: *Models*, *Reasoning and Inference*. Cambridge University Press: New York, 2000.
13. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford University Press: New York, 1993.
14. Wasserman L. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome: comment. *Journal of the American Statistical Association* 1999; **94**:704–706.
15. Neugebauer R, van der Laan MJ. G-computation estimation for causal inference with complex longitudinal data. *Computational Statistics and Data Analysis* 2006; **51**:1676–1697.
16. Taylor JMG, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* 1998; **17**:2381–2394.
17. Robinson GK. That BLUP is a good thing: the estimation of random effects. *Statistical Science* 1991; **6**:15–51.
18. Thernau TM. *A Package for Survival Analysis in S*. Mayo Foundation: 1999. Online document: http://mayoresearch.mayo.edu/mayo/research/biostat/upload/tr53.pdf.
19. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
20. Loughin TM. On the bootstrap and monotone likelihood in the Cox proportional hazards regression model. *Lifetime Data Analysis* 1998; **4**:393–403.
21. van Asten L, Langendam M, Zangerle R, Hernández Aguado I, Boufassa F, Schiffer V, Brettle RP, Robertson JR, Fontanet A, Coutinho RA, Prins M. Tuberculosis risk varies with the duration of HIV infection: a prospective study of European drug users with known date of HIV seroconversion. *AIDS* 2003; **17**:1201–1208.
22. Geskus RB. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Statistics in Medicine* 2001; **20**:795–812.
23. López-Gatell H, Cole SR, Hessol NA, French AL, Greenblatt RM, Landesman S, Preston-Martin S, Anastos K. Effect of tuberculosis on the survival of women infected with human immunodeficiency virus. *American Journal of Epidemiology* 2007; **165**:1134–1142.

DOI: 10.1002/sim