Comparing results from multiple imputation and dynamic marginal structural models for estimating when to start antiretroviral therapy

Bryan E. Shepherd,[a][*][†] Qi Liu,[a] Nathaniel Mercaldo,[a] Cathy A. Jenkins,[a] Bryan Lau,[b] Stephen R. Cole,[c] Michael S. Saag[d] and Timothy R. Sterling[a]

Optimal timing of initiating antiretroviral therapy has been a controversial topic in HIV research. Two highly publicized studies applied different analytical approaches, a dynamic marginal structural model and a multiple imputation method, to different observational databases and came up with different conclusions. Discrepancies between the two studies' results could be due to differences between patient populations, fundamental differences between statistical methods, or differences between implementation details. For example, the two studies adjusted for different covariates, compared different thresholds, and had different criteria for qualifying measurements. If both analytical approaches were applied to the same cohort holding technical details constant, would their results be similar? In this study, we applied both statistical approaches using observational data from 12,708 HIV-infected persons throughout the USA. We held technical details constant between the two methods and then repeated analyses varying technical details to understand what impact they had on findings. We also present results applying both approaches to simulated data. Results were similar, although not identical, when technical details were held constant between the two statistical methods. Confidence intervals for the dynamic marginal structural model tended to be wider than those from the imputation approach, although this may have been due in part to additional external data used in the imputation analysis. We also consider differences in the estimands, required data, and assumptions of the two statistical methods. Our study provides insights into assessing optimal dynamic treatment regimes in the context of starting antiretroviral therapy and in more general settings. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: HIV/AIDS; causal inference; survival analysis; multiple imputation; dynamic marginal structural models

1. Introduction

HIV attacks an infected person's immune system, making them prone to opportunistic infections, AIDS, and eventually death. CD4 cell count is a marker of the strength of the immune system, and it declines during the natural course of HIV infection. With the introduction of effective combination antiretroviral therapy (ART) in the late 1990s, HIV has become a treatable chronic disease: HIV-infected persons who take ART greatly reduce the amount of virus circulating in their body, raise their CD4 count, and can remain generally healthy. A fundamental question in the treatment of HIV has been as follows: When, specifically at what CD4 count, should an HIV-infected person start ART? It has long been clear that patients should start ART before their CD4 drops to levels at which the risk of AIDS or death becomes high (e.g., $< 200 \, \text{cells/mm}^3$). However, because ART must be taken for life and because there are short-term and long-term toxicities associated with the medications and the potential for generating resistance

[a]Vanderbilt University, Nashville, TN, U.S.A.
[b]Johns Hopkins University, Baltimore, MD, U.S.A.
[c]University of North Carolina, Chapel Hill, NC, U.S.A.
[d]University of Alabama at Birmingham, Birmingham, AL, U.S.A.
*Correspondence to: Bryan E. Shepherd, Department of Biostatistics, Vanderbilt University, 2525 West End, Suite 11000, Nashville, TN 37203, U.S.A.
†E-mail: bryan.shepherd@vanderbilt.edu

mutations through non-adherence, there has been reluctance to start treatment immediately if CD4 count is relatively high (e.g., >500 cells/mm^3).

Most of the early data addressing the when-to-start question came from observational studies looking at the association between CD4 at ART initiation and the subsequent time until AIDS or death (e.g., [1]). It is well known, however, that these studies are poor at addressing the when-to-start question, not just because of the inherent limitations of observational studies (e.g., confounding) but also because they could be subject to lead time bias [2]. Briefly, suppose a researcher wanted to compare two treatment rules: start ART when CD4 drops below 500 (immediate therapy) versus start when CD4 drops below 350 (defer therapy). An ideal study would randomize patients at the time their CD4 first falls below 500 to immediate or deferred therapy and *start follow-up for both treatment arms at the time of randomization*. A study that compares immediate versus deferred therapy but only compares the time from ART initiation until event inappropriately ignores the time it took for CD4 to drop from 500 to 350 (i.e., lead time due to left truncation) for subjects in the deferred therapy group. Such an approach also ignores events that may have happened among patients not on ART after their CD4 dropped below 500 but before reaching 350 (i.e., unseen events). These problems are similar to those arising from 'prevalent sampling', which in our case involves only including those who started ART [3].

The when-to-start ART question prompted the development of new statistical methods for observational data to deal with lead time bias and other issues. We focus on two specific methods, a multiple imputation (MI) approach [4] and a dynamic marginal structural model (dMSM) approach [5]. Briefly, the MI approach uses observational data beginning at ART initiation and augments these data by imputing lead time and unseen events in the deferred therapy arm using additional, historic data. In contrast, the dMSM approach uses observational data both before and after ART initiation, starts follow-up when CD4 drops below a specific constant threshold, expands data when it is consistent with more than one treatment initiation rule, censors data when it becomes inconsistent with particular treatment initiation rules, and uses inverse probability weighting to remove potential bias. More detailed descriptions of the two approaches are given in the Section 2.

In April 2009, two papers were published, one in the *Lancet* and the other in the *New England Journal of Medicine*, which addressed the when-to-start ART question using the MI and dMSM methods, respectively, and they came to different conclusions [6,7]. Using data from 24,444 ART initiators from North America and Europe (the ART-CC cohort) and 21,247 patients from the pre-ART era, Sterne and colleagues applied the MI approach and obtained results suggesting that it was optimal to start ART at CD4 > 350 but that there was little benefit to starting at higher CD4 counts. In contrast, using data from 17,517 patients in North America (the NA-ACCORD cohort), Kitahata and colleagues applied an analysis based on the dMSM approach and concluded that it was best to start ART at CD4 > 500.

The conflicting results led to confusion among HIV clinicians and policymakers who were left wondering what to believe – particularly, given the complexity of both studies' analyses. In addition, the NA-ACCORD analysis was criticized for incorrectly applying the dMSM approach – most notably for incorrectly handling periods of time during which a patient's data were consistent with multiple treatment rules [8], although subsequent sensitivity analyses suggested that this issue had little impact on results [9].

There have been additional methodological developments and investigations to estimate the optimal timing of treatment with observational data (e.g., [10–14]). There have also been other estimates of the optimal timing of ART initiation using variations of the dMSM approach applied to observational HIV data; these studies generally obtained results more consistent with the ART-CC analysis [15–17]. A randomized trial was conducted in Haiti, which demonstrated that it is better to start at CD4 between 200 and 350 compared with <200 [18]. In many clinicians' eyes, the question of when to start ART was answered in favor of starting as early as possible, given little evidence of harm in observational studies and because of strong evidence suggesting that early ART initiation prevents transmission of HIV to others [19,20]. However, until recently, this view has been challenged [21]. A large randomized trial comparing the rules start ART with CD4 < 350 vs. >500 was recently stopped early because of evidence of protection in the >500 arm [22,23]. There were many key differences between this trial and the observational studies. However, these latest results have generally pushed the field towards starting ART as early as possible.

Despite recent developments, it is important to better understand differences between the NA-ACCORD and ART-CC results. In particular, a better understanding is important for the credible future use of these types of methods – for questions pertaining to HIV infection or to other diseases. The two research groups applied different statistical methods to different patient populations and got different

results. What would happen if the two statistical methods were correctly applied to data from the same cohort – would they produce similar results? There are fundamental differences between the two approaches that could be the cause of the discrepancy, including differences between the estimands, required data, and assumptions. In addition, there were many technical differences between the implementation of the two approaches. For example, the studies compared different CD4 thresholds, included different covariates, and handled censoring differently. What is the impact of these technical decisions on estimates? What would happen if the two statistical methods were applied to data from the same cohort holding technical details constant?

In this study, we used data from a separate cohort, the Center for AIDS Research Network of Integrated Clinical Systems (CNICS), together with data from the Vanderbilt Comprehensive Care Clinic (VCCC), and applied the MI and dMSM approaches – holding technical details constant – to address the question of when to start ART. Rather than try to reproduce the exact ART-CC and NA-ACCORD analyses, we attempted to perform analyses in manners consistent with the original methodological publications. In Section 2, we describe and compare central aspects of the methods. In Sections 3 and 4, we describe our implementation of the methods and present results using the CNICS + VCCC data. In Section 5, we provide a brief simulation study, and in Section 6, we discuss findings and provide general conclusions.

2. Methods: estimands, data, and assumptions

In this section, we provide a brief overview of the MI and dMSM approaches, focusing on estimands, required data, and assumptions. We also compare these fundamental features between the two methods.

2.1. The multiple imputation approach

The MI approach considers HIV-positive, treatment-naïve, event-free persons whose CD4 (measured in the past m' months) is in a specific range, $(x_{\text{defer}}, x_{\text{immediate}}]$, for example, (350, 500], and compares the time to event (death or the composite outcome AIDS/death) for those who start ART immediately (denoted $R=0$) versus those who defer ART initiation until CD4 is in $(x_a, x_{\text{defer}}]$, for example, (200, 350] (denoted $R=1$). Let T be the time from when CD4 is in $(x_{\text{defer}}, x_{\text{immediate}}]$ ('baseline') to event, C be the time from baseline to censoring, Y be the minimum of T and C, and Δ be the indicator of an event. Of interest is to compare the hazard of death between those with $R=1$ and $R=0$, specifically, to estimate α in a Cox proportional hazards model with the hazard of event equaling $\lambda_0(t)\exp(\alpha R)$.

The MI approach was developed under the context that data are only available among ART initiators and from the time of ART initiation. Therefore, the time from baseline to ART initiation ('lead time') is unknown for those in the defer therapy arm. Furthermore, there are a certain number of individuals, denoted as n^*, who had a CD4 count in $(x_{\text{defer}}, x_{\text{immediate}}]$ but had an event prior to starting ART and were therefore not included in the database. The MI approach augments the incomplete observational data from ART initiators by imputing lead time and unseen events. To impute lead time and unseen events, data from the pre-ART era are used, where one can model the natural history of HIV without the impact of effective therapies. Specifically, let Y_0 be the observed time from ART initiation until the first of event or censoring. If $R=0$, then $Y=Y_0$; however, if $R=1$, then $Y=Y_0+Y_L$, where Y_L is the lead time, or the time from baseline to starting ART. The MI approach imputes Y_L for those with $R=1$, thereby making Y available for all patients. The MI approach also imputes n^*, the number of patients with unseen events, and then imputes Y for these n^* patients and assigns them $R=1$ and $\Delta=1$. The hazard ratio, $\exp(\alpha)$, is then estimated using the complete data, some of which are imputed. The process is repeated multiple times, and estimates are combined using standard MI formulas [24].

For writing assumptions, we use potential outcomes notation, with $T(r)$ and $C(r)$ denoting the potential times to event and censoring, respectively, if $R=r$. To obtain an unbiased estimate of the hazard ratio, the MI approach makes the following key assumptions:

> A1 Independent censoring: $T(r)$ is independent of $C(r)$.
> A2 No confounding: $(T(r), C(r))$ are independent of R.
> A3 The imputation model is correct.

The independent censoring and no confounding assumptions A1 and A2 are standard for time-to-event analyses and could be relaxed by conditioning on other covariates; however, the original MI approach and its implementation by the ART-CC implicitly assumed A1 and A2 without conditioning on covariates.

Assumption A3 is that models of event rates and CD4 decline in the pre-ART era are an accurate reflection of what they would have been in the absence of treatment in the post-ART era. Stated alternatively, A3 states that lead time and event rates in the pre-ART era are exchangeable with that in the post-ART era.

The estimand, required data, key assumptions, and necessary models for the MI approach are summarized in Table I.

2.2. The dynamic marginal structural model approach

The estimand for the dMSM approach is the hazard ratio comparing the rules: (i) start ART within m months of CD4 first dropping below $x_{\text{immediate}}$, for example, 500 (immediate therapy; $R=0$) versus (ii) start ART within m months of CD4 first dropping below x_{defer}, for example, 350 (defer therapy; $R=1$). To estimate this hazard ratio, the dMSM approach uses longitudinal data of CD4 count, ART use, other covariates, and events from treatment-naïve, event-free patients beginning at enrollment in care (i.e., before ART initiation).

The dMSM approach attempts to analyze the observational data in a manner that mimics a randomized trial. For example, if one were comparing the rules start when CD4 drops below 500 (immediate therapy arm) versus start when CD4 drops below 350 (deferred therapy arm), a trial might randomize patients at their first CD4 observation below 500 to either immediate or deferred therapy. With the observational data, anyone who started ART when their CD4 first dropped below 500 (within a grace period of m months) would be in the immediate group, whereas anyone who did not start at that time would be in the deferred group. If patients in the deferred group started ART within m months of their CD4 first dropping below 350, then their data would be consistent with the rule 'start within m months of CD4 first dropping below 350'. Similarly, if they had an event before their CD4 dropped below 350 and they had not yet started ART, then their data would also be consistent with the 'start within m months of CD4 first dropping below 350' rule. However, if they started ART too early (e.g., before their CD4 dropped below 350, but not within m months after it dropped below 500) or too late (e.g., they failed to start ART within m months after it first dropped below 350), then their data would no longer be consistent with either treatment rule. A patient's CD4 and treatment history may be consistent with both ART initiation rules for some (or all) of their follow-up time; in this case, the dataset is expanded using duplicate records (one in the immediate therapy arm and one in the deferred therapy arm) for the period of time in which their

Table I. Comparing the estimands, necessary data, assumptions, and models of the MI and dMSM methods.	
Multiple imputation method	Dynamic marginal structural model method
Estimand: hazard ratio comparing	
(1) Start ART with CD4 measured within the last m' months in $(x_{\text{defer}}, x_{\text{immediate}}]$	(1) Start ART within m months of CD4 first dropping below $x_{\text{immediate}}$
versus	versus
(2) Start ART with CD4 measured within the last m' months in $(x_{\text{a}}, x_{\text{defer}}]$	(2) Start ART within m months of CD4 first dropping below x_{defer}
Necessary data	
CD4 at ART initiation, time from ART initiation to event or last visit	All CD4 measurements since enrollment (before and after ART initiation), covariates (both time invariant and varying), date of ART initiation, dates of events, dates of last visits
External data from pre-ART era: all CD4 measurements, dates of events, dates of last visits	
Key assumptions	
A1. Independent censoring: $T(r)$ is independent of $C(r)$	B1. Independent censoring: $\{\overline{D}_K(r)\}$ is independent of $C_t(r)$ conditional on $\overline{L}_t$, $\overline{A}_{t-1}$, and $D_t=0$
A2. No confounding: $(T(r), C(r))$ are independent of R	B2. No confounding: $\{\overline{D}_K(r)\}$ is independent of A_t conditional on $\overline{L}_t$, $\overline{A}_{t-1}=0$, and $D_t=0$
A3. The imputation model is correct	
Models to be fit	
(1) External data from pre-ART era: joint model of probability of event and time to CD4 dropping below x_{defer}	(1) Probability of starting ART
(2) Hazard of event using data imputed from model 1	(2) Probability of being censored
	(3) Hazard of event weighted by inverse probability weights derived from models 1 and 2

ART, antiretroviral therapy; MI, multiple imputation; dMSM, dynamic marginal structural model.

data are consistent with both rules. A patient contributes to whatever arm their data are consistent with until their data are no longer consistent with that treatment rule, at which time they are artificially censored from that arm. Details of this approach in a general setting with more than two treatment rules are well illustrated elsewhere [12]. To account for potential bias induced by this artificial censoring, the dMSM approach uses inverse probability weighting to assign more weight to observations that, based on their covariate history, were more likely to have been censored, but were not. Hazard ratios or other metrics to compare the immediate and deferred therapy arms can then be estimated, and sandwich variance estimates are used to account for correlation induced by duplicating records.

Let $t=0$ be the time when CD4 is first measured in the range $(x_\text{defer}, x_\text{immediate}]$, and let $t=1, \ldots, K$ be the subsequent months, with K denoting the maximum follow-up time. A_t is the indicator of being on ART at month t, D_t is the indicator of an event at month t, L_t are covariates at month t, and C_t is the indicator of being censored (due to loss to follow-up (LTFU) or end of study) at month t. $\bar{A}_t$ and $\bar{L}_t$ denote treatment and covariate history up until, and including, month t. In addition to the standard causal assumptions of consistency and positivity [25], the dMSM approach makes the following key assumptions:

B1 Independent censoring: $\{\overline{D}_K(r)\}$ is independent of $C_t(r)$ conditional on $\bar{L}_t$, $\bar{A}_{t-1}$, and $D_t=0$ for all $t=0, 1, \ldots, K$.

B2 No confounding: $\{\overline{D}_K(r)\}$ is independent of A_t conditional on $\bar{L}_t$, $\bar{A}_{t-1}=0$, and $D_t=0$ for all $t=0, 1, \ldots, K$.

The estimand, required data, key assumptions, and necessary models for the dMSM approach are summarized in Table I.

2.3. Comparing the MI and dMSM approaches

The MI and dMSM approaches have fundamental differences in their estimands, the data required, and their assumptions, which we highlight in Table I and in this section.

First, there are differences between the two approaches' hazard ratio estimands. The dMSM approach compares starting ART within m months of CD4 first dropping below $x_\text{immediate}$ versus x_defer, whereas the MI approach compares starting ART with most recent CD4 (measured within the last m' months) in the range $(x_\text{defer}, x_\text{immediate}]$ versus $(x_\text{a}, x_\text{defer}]$. Even with $m=m'$, these estimands are different. For example, there is no time period connected to the MI estimand – a patient could have CD4 in the range $(x_\text{defer}, x_\text{immediate}]$ for years and as long as they started ART while in range, then their data are consistent with the treatment rule. Therefore, this regimen is clinically ambiguous, as a clinician/patient does not know when their CD4 will drop such that they are no longer compliant with the regimen. In contrast, using the dMSM estimand, such a person would be censored from the immediate therapy arm after m months and would actually contribute most of their follow-up time to the deferred treatment arm (up until the point that they started ART, at which time they would be artificially censored). This event-free time, which would favor deferring therapy in the dMSM approach, would be ignored in the MI approach. With that said and acknowledging the presence of long-term non-progressors, most HIV-infected persons do not stay in the same CD4 stratum for years without ART.

As another example of differences between estimands, the dMSM approach could include a person who started ART within m months of CD4 first measured in the range $(x_\text{defer}, x_\text{immediate}]$ in the immediate treatment arm even if they had a subsequent pre-ART CD4 below x_defer – even substantially below. Note that this person would be included in both the immediate and deferred arms using the dMSM approach, whereas with the MI approach this person would be included, at most, in the deferred treatment group. (In fact, it is arguable that for persons whose CD4 decrease precipitously such that they never have a measurement in $(x_\text{a}, x_\text{defer}]$, their potential outcomes for $R=1$ under the MI approach do not exist.) The supporting information contains a table showing a few hypothetical CD4 and ART histories and details into which treatment arms patients would be placed under the two approaches, highlighting differences.

Second, there are differences in the data required by the two methods. The dMSM approach incorporates a patient's pre-ART follow-up into the analysis, whereas the MI approach only uses post-ART follow-up. Because of this, the MI approach requires imputing lead time and unseen events using models based on external data. If complete follow-up data are available (as is the case with our CNICS + VCCC analysis dataset described in the succeeding discussion), the MI approach somewhat unnaturally discards pre-ART data. In contrast, in practice, the dMSM approach necessitates discretizing time into periods, such as monthly intervals, which can result in some loss of information. In addition, the dMSM

approach has stricter inclusion criteria (a CD4 count in $(x_{\text{defer}}, x_{\text{immediate}}]$ prior to receiving ART) than the MI approach (a CD4 count in $(x_{\text{a}}, x_{\text{immediate}}]$ prior to receiving ART). The causal estimands are interpreted in the context of patients meeting inclusion criteria, so differences between the MI and dMSM inclusion criteria may have implications beyond simply which records are used for the analysis.

Finally, there are differences in the assumptions between the two approaches. Writing the MI approach in the same notation as the dMSM approach may be helpful for comparing assumptions. Everyone in the defer arm under the MI approach can be thought of as starting ART at time k, where k is imputed and varies between patients, and setting $A_j = 1$ for $j \geq k$. Imputing unseen events in the defer arm can be thought of as imputing n^*, k, and then setting $D_k = 1$ and $A_j = 0$ for $j \leq k$. In the immediate therapy arm, $A_0 = 1$.

With this notation, assumptions A1 and A2 can be re-written as the following:

A1 Independent censoring: $\{\overline{D}_K(r)\}$ is independent of $\{\overline{C}_K(r)\}$ for all $t = 0, 1, \ldots, K$.
A2 No confounding: $\{\overline{D}_K(r)\}$ is independent of A_0.

Assumptions A1 and A2 of the MI approach can then be compared with assumptions B1 and B2, respectively, of the dMSM approach. With regard to independent censoring, note that B1 is an assumption of conditional independence and is therefore weaker than A1 . With regard to assumptions of no confounding, A2 differs from B2, not just that B2 is conditional on covariates but also that the event vector is assumed to be conditionally independent of A_t at all t, whereas A2 only assumes independence (unconditional) from A_t at time 0. Therefore, in some ways, A2 is both a stronger and weaker assumption than B2. Although pedagogically helpful, the comparisons with this notation are not perfect. As discussed previously, the treatment initiation rules, r, differ between the MI and dMSM approaches.

Assumption A3 for the MI approach remains the same. No similar assumption is needed for the dMSM approach.

Further discussion of these assumptions and their plausibility with the CNICS + VCCC data is given in Section 6.

3. Implementation details

In this section, we go through details of the implementation of both the MI and dMSM approaches to the CNICS + VCCC data. In general, we tried to be consistent with what was performed in the NA-ACCORD and ART-CC publications while being true to the original methodological publications. We also describe implementation of analyses with technical details held constant between the two approaches. All analyses were performed in R version 3.0.2. Analysis scripts are posted at http://biostat. mc.vanderbilt.edu/ArchivedAnalyses

3.1. Multiple imputation approach

The MI approach requires data from the pre-ART era to impute lead times and unseen events in the deferred therapy arm. Data from the pre-ART era are limited and were used by the ART-CC in their analysis. We also used the same data, specifically, data from patients in the Multicenter AIDS Cohort Study, the Swiss HIV Cohort Study, the ANRS CO4 French Hospital Database on HIV, the ANRS CO3 Aquitaine Cohort, the Amsterdam Cohort Studies, the South Alberta Clinic, and the Concerted Action on Seroconversion to AIDS and Death in Europe collaboration. Rather than using the original data, we simplified our analyses by using the pre-ART parameter estimates that the ART-CC obtained from these data and used to impute lead times and unseen events.

For a comparison between two CD4 thresholds, $x_{\text{immediate}}$ and x_{defer}, where $x_{\text{defer}} < x_{\text{immediate}}$, patients were considered eligible for inclusion if they started ART with a CD4 count at ART initiation between $x_{\text{defer}} - (x_{\text{immediate}} - x_{\text{defer}})$ and $x_{\text{immediate}}$. In all of our analyses, $x_{\text{immediate}}$ and x_{defer} were separated by 100 cells, so this simplified to a person being eligible if they started ART with CD4 in the range $(x_{\text{defer}} - 100, x_{\text{immediate}}]$ (e.g., (300, 500] if comparing $x_{\text{immediate}} = 500$ vs. $x_{\text{defer}} = 400$). CD4 at ART initiation was the measurement closest to but no more than $m' = 3$ months prior to ART initiation; patients without a CD4 measurement in that window were excluded.

Lead time was added to all patients in the deferred arm by randomly drawing from a generalized gamma distribution [26] with parameters estimated from the pre-ART era data. The generalized gamma distribution is a flexible family that contains many of the most commonly used distributions for time-

to-event outcomes and hazard functions and was used in the ART-CC's implementation of the MI approach. The number of unseen events to be added to the deferred arm was randomly selected from a negative binomial distribution, and the corresponding time of the event was drawn from a generalized gamma distribution, again using parameter estimates from the pre-ART era data. Unadjusted hazard ratios, comparing the time to event in the immediate versus deferred arms, and their standard errors were then computed with a Cox regression model using the augmented data. This process was repeated for 20 imputation replications.

For our initial analyses, we incorporated the technical details of the ART-CC paper as summarized in Table II. For patients who did not experience an event, we added 2 months to the patients' follow-up time (corresponding to approximately 50% of the mean time between follow-up visits; time between follow-up visits was similar across sites). Patients with an AIDS-defining event prior to ART initiation were excluded. For our sensitivity analyses, we investigated the impact of the technical details on estimates by perturbing them one at a time from a base model (shown in bold in Table II, except with time kept continuous).

3.2. Dynamic marginal structural model approach

Data were initially discretized into 30-day intervals (labeled as months). ART initiations, CD4 measurements, AIDS-defining events, or deaths that occurred in the range (30, 60], for example, were assigned to month 2, taking care to maintain the temporal ordering of measurements, exposures, and events. For example, if CD4 was measured on day 35 and ART was started on day 40, then both were assigned to month 2; in contrast, if CD4 was measured on day 40 and ART was started on day 35, then the CD4 measurement on day 40 was assigned as a month 3 measurement (unless superseded by a later month 3 measurement), and CD4 from month 1 was carried forward as the month 2 measurement. Similarly, a CD4 measured during the same month but after an AIDS-defining event was assigned to the next month; and an AIDS-defining event that occurred during the same month as, but prior to, ART initiation was counted as an event prior to ART initiation.

For a comparison between two CD4 thresholds, $x_{immediate}$ and x_{defer}, patients were considered eligible for inclusion if they had a CD4 count in $(x_{defer}, x_{immediate}]$ prior to receiving ART. Time 0 was the time of this first qualifying CD4 measurement. Patients with an AIDS-defining event on or before time 0 were excluded. Note that for the analyses with death as the outcome, any AIDS-defining events after time 0 were ignored (i.e., they were not part of the starting rules). Similar to the NA-ACCORD analyses, a grace period of $m = 6$ months was used.

Table II. Differences between the technical details of the ART-CC and NA-ACCORD analyses.	
ART-CC analysis	NA-ACCORD analysis
$x_{immediate}$ versus x_{defer}	$x_{immediate}$ versus x_{defer}
550 vs. 450	>500 vs. <500
525 vs. 425	500 vs. 350
500 vs. 400, …	
200 vs. 100	
Excluded patients who died in first 2 weeks of ART	**Did not exclude patients who died**
Added 50% of the mean time between visits to patients' last visit	**Did not add time to patient's last visits**
Administratively censored after 6 years	**Did not administratively censor**
Excluded IDU; analyzed separately in secondary analysis	**Included IDU**; included as covariate in secondary analysis
CD4 had to be measured within $m' = 3$ months of ART initiation to qualify	Had to start ART within $m = 6$ **months of CD4** to be in adherence
Did not incorporate covariates	Adjusted for CD4 at entry, age, and sex in primary analysis
Non-informative censoring implicitly assumed for patients lost to follow-up	Inverse probability of censoring weights for patients lost to follow-up
No stratification	Analyses stratified by site and year of entry
Time left continuous	**Time discretized**

Bold text indicates the detail chosen for analyses that kept technical details constant across approaches (reported in Figure 3).
ART, antiretroviral therapy; IDU, injection drug users.

Inverse probability weights were used to account for potential bias that may have been induced by artificial censoring. Each subject's time-varying weight was the inverse predicted probability of them not being artificially censored from a particular treatment rule. Artificial censoring only occurred when a patient started or failed to start ART. Therefore, the probability of not being artificially censored for a given rule was simply the probability of starting/not starting ART conditional on not yet starting ART and time-updated covariates, including CD4 count. During a grace period (e.g., for the immediate therapy arm, the first 6 months after which a CD4 count drops below $x_{\text{immediate}}$) the probability of being censored was 0; following [12], our weights during the grace period were chosen to be consistent with rules that the distribution of the time of starting ART within the grace period was roughly uniform. Weights were multiplied together in the standard manner [28]. The probability of starting ART conditional on not yet being on ART and covariates was estimated using logistic regression in the complete, non-replicated data. The logistic regression model included the most recent CD4 measurement, sex, site, age at time 0, calendar year at time 0, and months since time 0 expanded using restricted cubic splines with four knots. These covariates were used in the NA-ACCORD study and were chosen because of their availability and to minimize confounding.

Loss to follow-up was defined in all analyses as no clinic visit (including laboratory measurement) within 1 year of the database closing date, defined as the most recent visit date for each cohort. Censoring for lost patients occurred at the last visit date. In NA-ACCORD analyses, inverse probability of censoring weights were incorporated to account for possible bias due to LTFU. The probability of being LTFU conditional on not yet being lost and covariates was estimated using logistic regression with the complete, non-replicated data. The logistic regression models included the most recent CD4 measurement, sex, site, age at time 0, calendar year at time 0, whether or not a patient was on ART during that month, and months since time 0 expanded using restricted cubic splines with four knots. Weights were the inverse of the predicted probability of not being LTFU, multiplied together in the standard manner.

Stabilized weights can be constructed for dMSMs with grace periods, but they are somewhat complicated, and their benefits are uncertain [12]. Therefore, in our analyses, we used unstabilized weights. To avoid extreme weights due to near violations of the positivity assumption, in primary analyses, we truncated our weights at the 99th percentile (i.e., any weight larger than the 99th percentile was assigned to the 99th percentile). This truncation level was arbitrary, but chosen *a priori*; the sensitivity of results to different truncation levels was explored by considering estimates with truncation at the 95th, 97.5th, and 99.5th percentiles, as well as no truncation [28].

Weighted pooled logistic regression was used to estimate the effect of immediate versus deferred therapy. Initial analyses based on the technical details of the NA-ACCORD analysis incorporated the covariates CD4 at time 0, age at time 0, and sex. NA-ACCORD analyses were also stratified by site and calendar year; our logistic regression models included months since time 0 (expanded using restricted cubic splines with four knots), site, and calendar year (kept continuous). (Stratification in a Cox model typically implies assuming separate baseline hazards for stratification variables, which for pooled logistic regression corresponds to including an interaction between months since time 0 and the stratification variables. Such a model was computationally infeasible, so we simply adjusted for the stratification variables.) In sensitivity analyses, we investigated the impact of the technical details on estimates by perturbing them one at a time from a base model (shown in bold in Table II). Standard errors were computed using robust, sandwich variance estimators.

Our analyses were not an attempt to exactly mimic those of the NA-ACCORD given specific implementation criticisms [7], subsequent methodological developments [11,12], and uncertainties regarding the original analyses (e.g., it is unclear whether truncated weights were used in the NA-ACCORD analysis). Our definition of LTFU differed from that used by either cohort, although it was consistent between analyses; a sensitivity analysis examined the impact of defining LTFU as a gap in care of >1 year [29]. The supporting information contains an example, including intermediary model results, for deriving estimates under the MI and dMSM approaches.

3.3. Analyses holding technical details constant

We performed an additional set of analyses holding many technical details constant between the MI and dMSM approaches; the selected details for these analyses are shown in bold in Table II. In these analyses, we did not include weights for LTFU, incorporate covariates (not in our weighting model nor in our outcome model), or stratify by site and year of entry. For the MI approach, time in the augmented

data was discretized into 30-day intervals, and we used pooled logistic regression to approximate hazard ratios [27]. For these specific analyses, we computed the difference between estimates (on the logarithmic scale) from the MI and dMSM approaches and obtained 95% Wald confidence intervals (CIs) using the standard error estimated from 200 bootstrap replications; the ratio of the hazard ratios was estimated by exponentiating this log-difference. Five imputation replications were used for the MI approach inside each bootstrap replication.

4. Results

4.1. CNICS and VCCC cohorts

The CNICS and VCCC cohorts have been described elsewhere [30,31]. Briefly, the combined dataset included 12,708 patients from nine primarily urban clinics in the USA (Baltimore, Birmingham, Boston, Cleveland, Nashville, Raleigh, San Diego, San Francisco, and Seattle). The combined cohort was 79% male, 40% were African American, and 17% were likely infected through injection drug use. The median age at entry was 36 years (interquartile range [IQR] 29–43 years), the median follow-up was 3.4 years (IQR 1.2–6.7), and the median CD4 at clinic entry was 380 (IQR 213–569). Approximately 76% of patients initiated ART sometime during their follow-up. The median CD4 at start of ART was 294 (IQR 150–438), and the median time from clinic entry until ART start for those who started ART was 3 months (IQR 1–17 months). CD4 was measured on average every 3–4 months at each site.

In the combined cohorts, 1817 patients had at least one AIDS-defining event, and 1499 patients died during follow-up. Figure 1 shows the association between CD4 at ART initiation and death or the composite outcome of AIDS/death. Persons who started ART with a lower CD4 were more likely to experience events; trends are clear for CD4 less than approximately 400 but less clear above 400. As
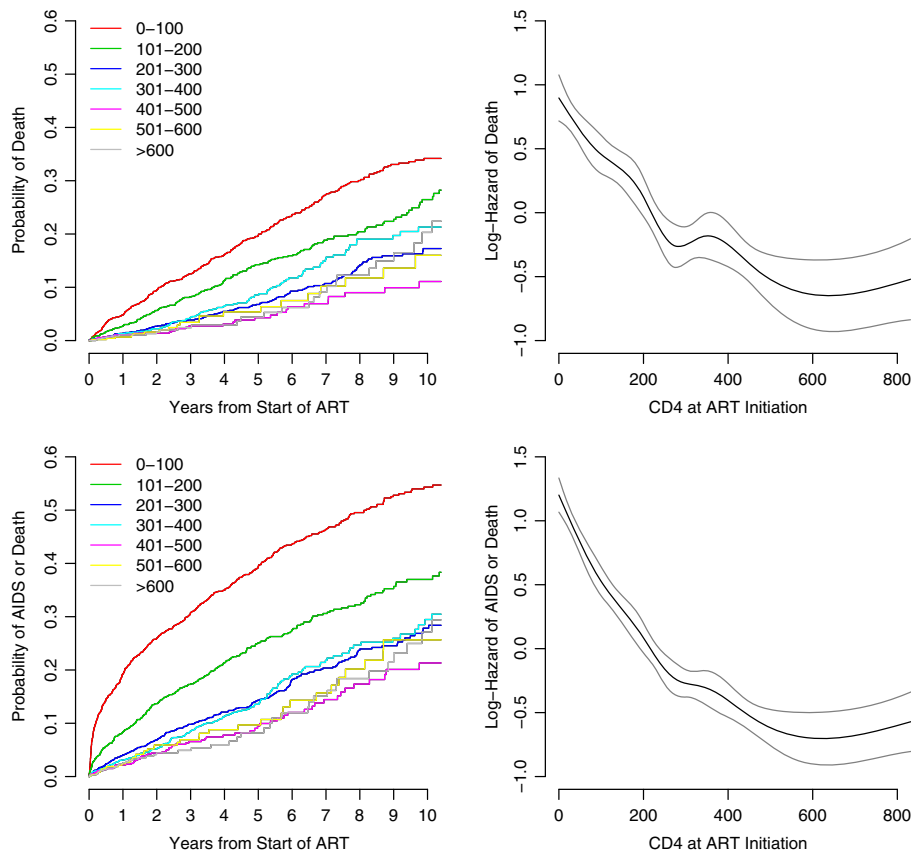

Figure 1. Association between CD4 at antiretroviral therapy (ART) initiation and death (top panels) and AIDS/death (bottom panels). The left panels show Kaplan–Meier estimates after ART initiation by CD4 strata; the right panels show the log-hazard of event (and 95% confidence intervals) as a function of CD4 at ART initiation, estimated from a Cox proportional hazards model with CD4 expanded using restricted cubic splines with seven knots.

mentioned previously, because of potential confounding and lead time bias, these results may be misleading if interpreted as evidence for an optimal time to start ART.

4.2. Results using multiple imputation and dynamic marginal structural model approaches

Figure 2 shows estimated hazard ratios using the MI and dMSM approaches, implemented using technical details similar to those of the ART-CC and NA-ACCORD analyses (outlined in Table II). For both the mortality and AIDS/mortality composite endpoint, the MI approach suggested that it is better to start ART while CD4 is in the range 101–200 compared with 0–100 and that starting in the range 201–300 is better than starting between 101 and 200, with positive hazard ratios and 95% CI that did not include the null of 1. In contrast, there was little evidence with the MI approach to conclude that starting with CD4 ranging from 201 to 300 is worse than starting at 301–400 or that 301–400 is worse than 401–500. The dMSM of the NA-ACCORD analysis compared different CD4 strata (<500 vs. <350 and <500 vs. ≥500); for both comparisons, with our data, CIs were wide and crossed 1. For purpose of comparison, the corresponding hazard ratios reported in the ART-CC and NA-ACCORD papers are also shown in Figure 2. Our hazard ratio estimates tended to be smaller than those of both studies and more variable, particularly when compared with those reported in the NA-ACCORD analysis.

We next performed analyses holding technical details constant across approaches. Selected details are bolded in Table II. It is important to note that in many cases, our analysis choices were not based on HIV medicine but rather ease of performing the analysis under both techniques. For example, the dMSM approach requires discretizing time, and the MI approach does not incorporate baseline covariates. The primary purpose of this analysis is to compare results, not necessarily to choose the optimal CD4 for ART initiation.

Figure 3 shows the results of both analysis methods holding technical details constant. The top row shows the estimated hazard ratio for both methods for each comparison; the bottom row directly compares the two approaches by showing the ratio of the hazard ratios for each comparison and 95% bootstrap CIs. Point estimates were not identical but generally tended to be of similar magnitude. The exception was the comparison of the thresholds 400 vs. 300 for the composite AIDS/death outcome, where the dMSM approach resulted in a larger hazard ratio (favoring starting at 400) than that of the MI approach, and the 95% CI of the ratio of the hazard ratios did not include 1. The hazard ratio
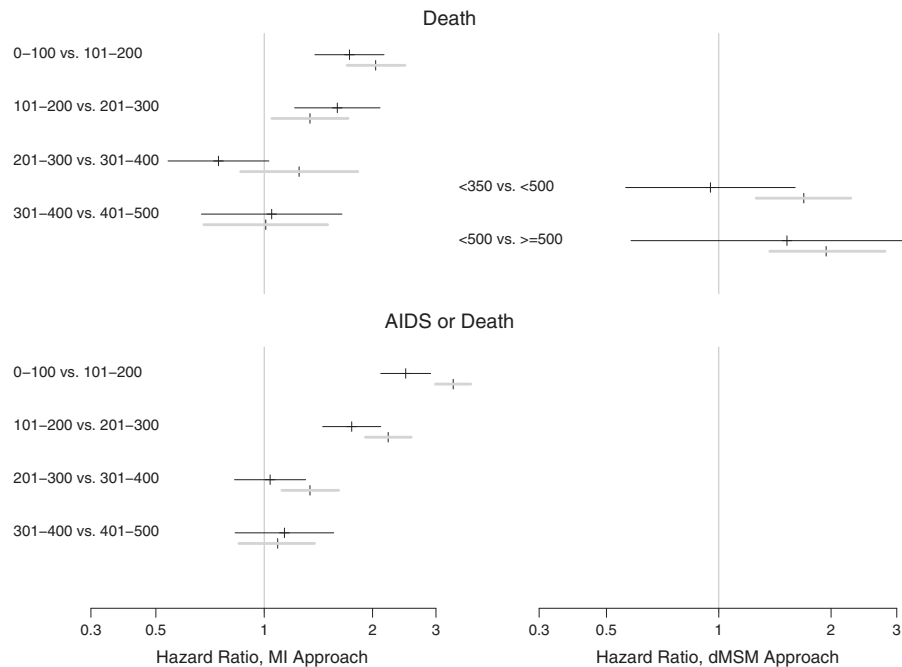

Figure 2. Estimated hazard ratios and 95% confidence intervals for deferring antiretroviral therapy (ART) versus immediate ART (reference) at the given CD4 thresholds/ranges using the multiple imputation (MI) approach (left panel) and the dynamic marginal structural model (dMSM) (right panel). The top and bottom rows correspond to estimates with death and the first of AIDS or death, respectively, as the endpoint. Technical details are similar to those of the ART-CC and NA-ACCORD analyses. The gray lines are estimates from the original publications.
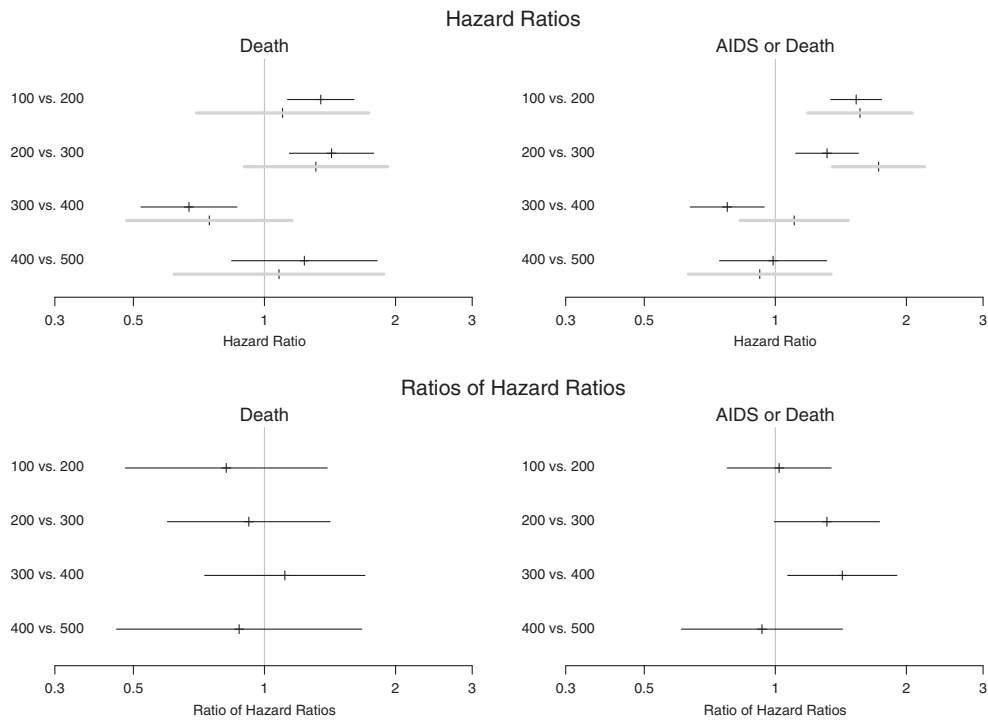
Figure 3. Top row: estimated hazard ratios and 95% confidence intervals (CIs) for the deferred versus the immediate (reference) arms at CD4 thresholds using the multiple imputation (MI) (black) and dynamic marginal structural model (gray) approaches while holding technical details constant (described in Table II). Outcomes are death (left column) and first of AIDS or death (right column). Bottom row: estimated ratios of hazard ratios and 95% CI, with the MI approach as the reference.

comparing 300 vs. 200 for the composite endpoint also tended to be higher for the dMSM approach, although the 95% CI of the ratio of the hazard ratios for this comparison included 1. CIs were wider using the dMSM approach. For the AIDS/death outcome, both methods suggested that it is better to start before CD4 drops at or below 300 than 200 (and 200 before 100). With death as the outcome, the MI approach yielded the same conclusion, whereas with the dMSM approach, CIs included 1. For comparing thresholds of 400 vs. 300, the MI approach actually favored starting when CD4 first drops below 300, whereas such a conclusion would not be supported by the dMSM analysis.

Figure 4 demonstrates the sensitivity of results to changes in technical details with death as the outcome. The topmost row of the top panel shows the basic analysis for the MI approach, using the same technical details used in Figure 3 except with continuous, not discrete time. The following rows show estimates after varying one of the technical details. For example, estimates in the second row were derived using the same technical details as those in the first row except excluding any deaths that happened within 2 weeks of ART initiation; estimates in the third row were derived in the same manner as those in the first row except that 2 months were added to the follow-up time of those without death; and so forth. The bottom panel provides similar information for the dMSM approach (with the topmost row being identical to the hazard ratio estimates of Figure 3). From this plot, we can see that the technical details influenced estimates, and generally more so for the dMSM approach particularly when the grace period for starting ART changed from 6 to 3 months. However, the impact of any one technical component was generally minor.

Finally, the dMSM approach uses inverse probability weights that, in practice, are often truncated to avoid unstable estimates. Our analyses truncated weights at the 99th percentile. The sensitivity of results to this choice is illustrated in Figure 5. Although largely similar across truncation levels, estimates and CIs were influenced by the choice of truncation percentile with the direction of point estimates changing for some comparisons.

5. Simulations

We performed a limited set of simulations to assess whether we could estimate the optimal CD4 threshold, and obtain similar estimates, using the MI and dMSM approaches when the truth was known and
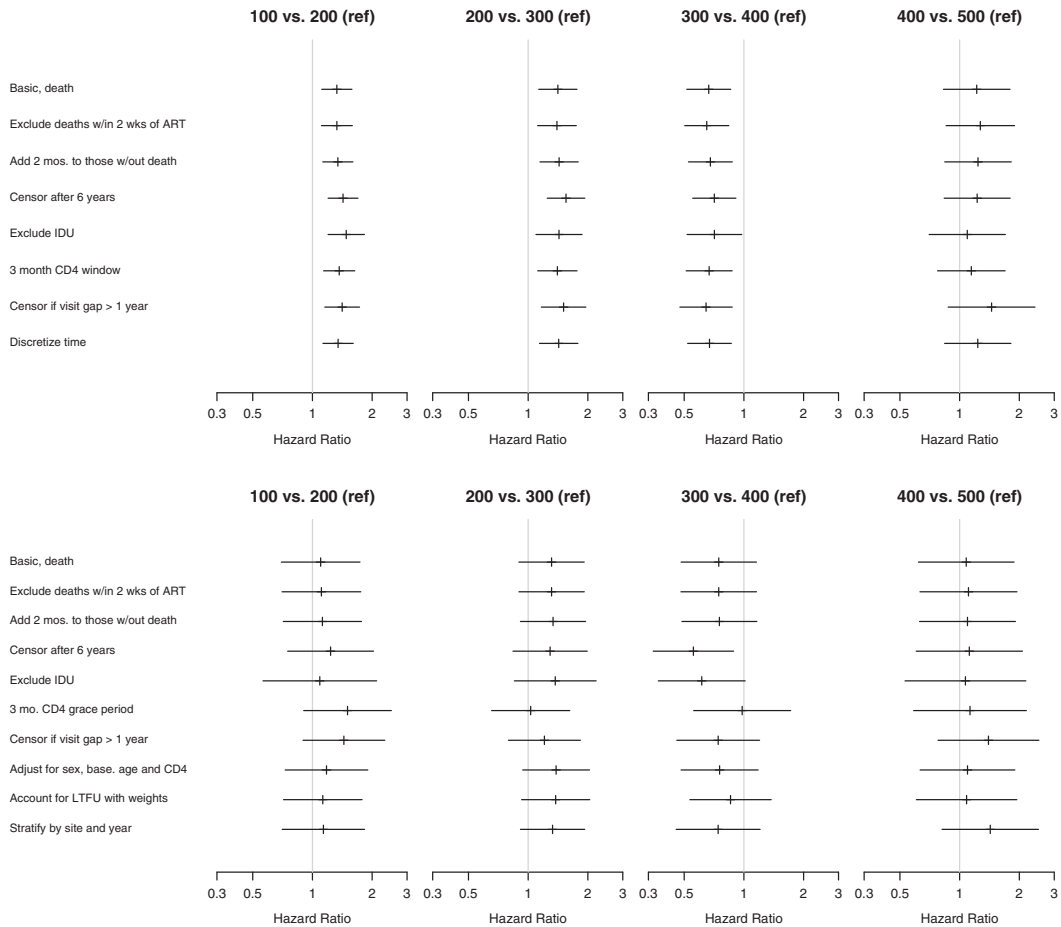
Figure 4. Sensitivity of mortality results (hazard ratio estimates and 95% confidence intervals) to technical details. The upper panel is the multiple imputation (MI) approach, and the bottom panel is the dynamic marginal structural model (dMSM) approach. The basic model for the MI approach is identical to that in bold in Table II except time is left continuous. The basic model for the dMSM approach is identical to that described in Table II.
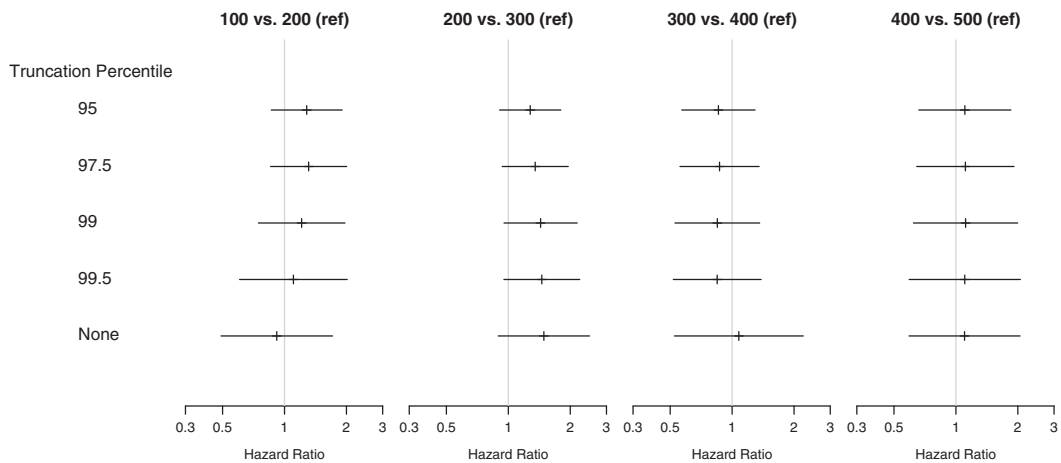

Figure 5. Sensitivity of mortality results (hazard ratio estimates and 95% confidence intervals) obtained using the dynamic marginal structural model approach to different choices for truncation percentile for the weights. These models incorporate covariates and weights for loss to follow-up but otherwise keep all technical details the same as those in bold in Table II.

with properly specified models. In our simulations, we generated data under three different scenarios in which the optimal time to initiate ART was when CD4 dropped below approximately 700, 500, and 350. In each scenario, we simulated CD4, ART use (yes/no), and event status at 3-month intervals for up to

10 years of follow-up for 15,000 patients; this was repeated for 500 simulation replications. CD4 was set to decline when a patient was not on ART and increase during ART use (quickly at first and plateauing later). The probability of starting ART was simulated such that it was higher for lower CD4. Events were also simulated such that their probability was higher at lower CD4. To simulate data under different optimal CD4 thresholds for initiating ART, we allowed the event probability to increase with increasing time on ART. The 10-year survival probability as a function of different CD4 thresholds for starting ART for the three different data generation scenarios is shown in Figure 6A. Simulation details regarding data generation and analysis are in the supporting information. The complete code is posted at http://biostat.mc.vanderbilt.edu/ArchivedAnalyses

With each of the 500 simulation replications for each of the three data generation scenarios, we obtained estimated hazard ratios and 95% CI comparing various ART initiation rules using both the MI and dMSM approaches. Figure 6B summarizes these quantities across all simulation replications and scenarios. The average hazard ratio (exponential of the mean log-hazard ratio) and average CI width (exponential of the mean log-hazard ratio plus/minus half the average width of the log-HR's 95% CI) are plotted. Note that the quantities plotted in the figure are not point estimates and CIs, but rather summaries of point estimates and the width of CIs across simulation replications.

In general, estimates from the MI and dMSM approaches were in harmony with the underlying data generation mechanism. Hazard ratios tended to be greater than 1 for all comparisons of adjacent CD4 thresholds (with the higher threshold as the reference) when data were generated such that the optimal CD4 for initiating ART was 700. When data were generated with the optimal CD4 of 500, HR estimates tended to be greater than 1 comparing thresholds below 500 and then approximately 1 thereafter. When data were generated with the optimal CD4 of 350, HR tended to be above 1 for comparisons below 350 and below 1 for comparisons above the 350. These results are as would be expected from Figure 6A.

Confidence intervals from the dMSM approach tended to be wider than those using the MI approach. For the MI approach, the mean standard error tended to be slightly larger than the standard deviation of the estimates, particularly for comparisons with the lower CD4 thresholds (supporting information). This suggests that CIs for the MI approach were actually conservative. For the dMSM approach, the mean standard error was similar to the standard deviation of estimates except for comparisons of 600 vs. 700 and 550 vs. 650 in the data generation scenario with the optimal CD4 of 700, where the mean standard error (0.66 and 0.55) was smaller than the standard deviation of estimates (0.81 and 0.64); as seen in Figure 6A, there are very few events at higher CD4 levels under this data generation scenario, so the poor performance of the dMSM standard errors based on large sample theory is not particularly surprising.

Estimates from the MI and dMSM approaches applied to the same dataset tended to be fairly similar. Figure 6C shows the ratio of the hazard ratios for the 500 simulations (specifically, the exponential of the mean difference between the log estimates ±1.96 times its standard deviation). The discrepancy between the MI and dMSM estimates was more variable when examining the larger CD4 thresholds that had fewer events, but on average, results were very similar with the average ratio of the hazard ratios close to 1.

These simulations were designed such that there were no confounders, so that both methods were correctly specified. Certainly, performance of the methods would be worse under improperly specified models.

6. Discussion

We applied two separate methods for estimating when to start ART to a single dataset holding technical details constant. The methods applied were an MI approach [4] and a dMSM [5] and were the basis of conflicting results in two highly publicized prior studies [6,7]. When applied to the same dataset with technical details held constant, the two methods produced fairly similar results with largely overlapping CIs. However, for certain comparisons, conclusions may have differed depending on the method employed (Figure 3). The MI approach generally resulted in narrower CIs, which may be due in part to the use of additional external data but may also be due to the dMSM's stricter inclusion criteria and use of artificial censoring. When applied to large simulated datasets, results were also similar between the MI and dMSM methods, although in any single simulation, estimates could be somewhat different between approaches.

As highlighted in Section 2.3, there are differences between the two methods that might lead to different answers even when applied to data from the same cohort. First, there are differences between
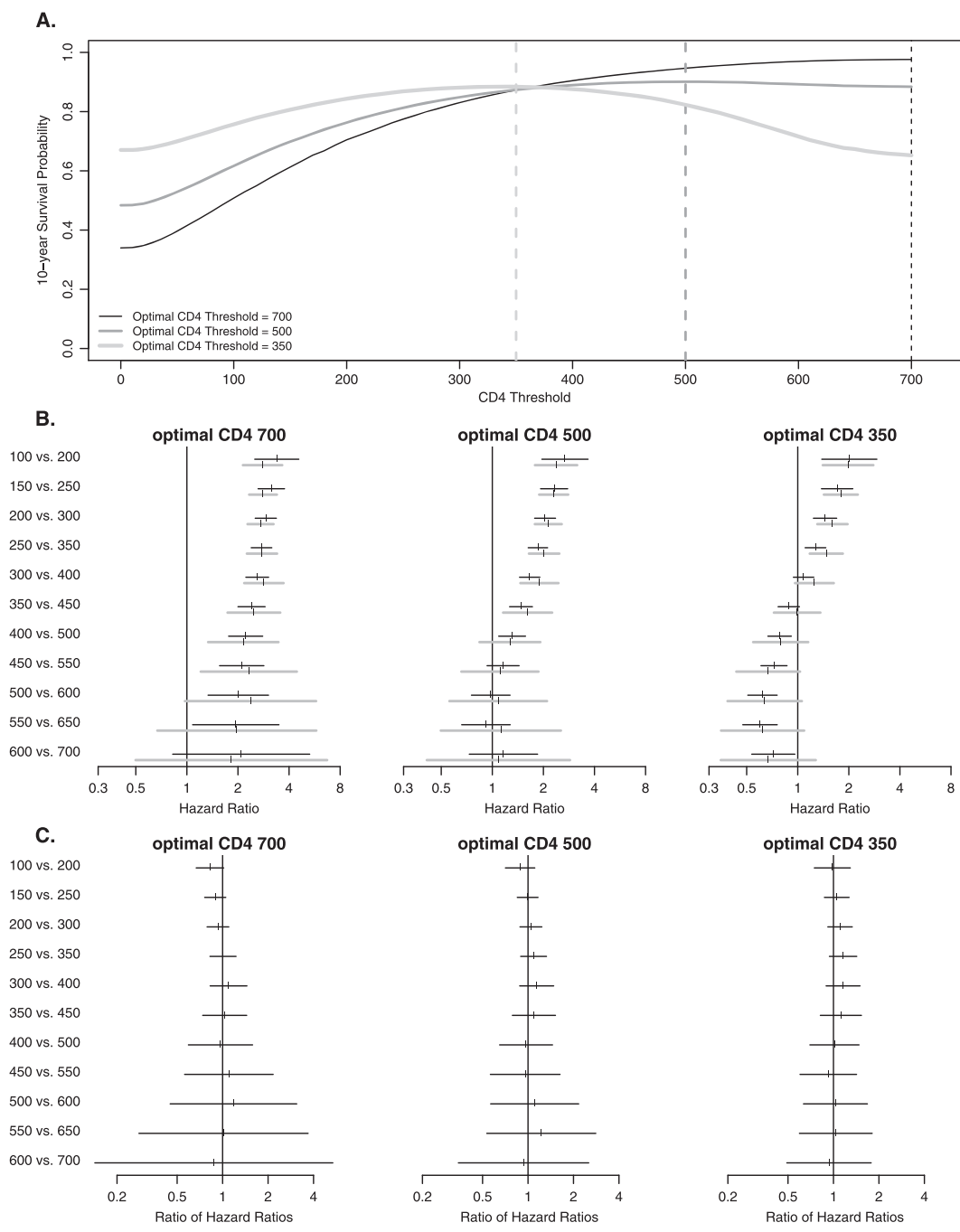
Figure 6. (A) Ten-year survival probability for initiating antiretroviral therapy when CD4 drops below the given CD4 threshold for three data generation scenarios. The optimal CD4 count (with maximal 10-year survival probability) is approximately 700, 500, and 350 for the various data generation scenarios. (B) The average hazard ratio (exponential of the mean log-hazard ratio) and average confidence interval width (exponential of the mean log-hazard ratio plus/minus half the average width of the log-hazard ratio's 95% confidence interval) when applying the multiple imputation (MI) (black) and dynamic marginal structural model (gray) approaches to 500 simulation replications under the three data generation scenarios. (C) Ratio of the hazard ratio, with the MI approach as the reference: exponential of the mean difference of the log-hazard ratios ±1.96 standard deviations.

the two approaches' estimands. The dMSM method addresses a more clinically relevant, and therefore preferable, estimand.

Second, there are differences between the two approaches' assumptions. Both approaches make assumptions of no unmeasured confounders. However, the MI analysis makes no attempt to incorporate potential confounding variables. In contrast, the MI approach only assumes no unmeasured confounding at time 0, whereas the dMSM approach assumes no unmeasured confounding at all time points.

The MI approach also assumes that historical data from the pre-ART era accurately capture CD4 changes and rates of events among persons who do not start ART. This is potentially a strong assumption and could be violated by differences over time in the circulating virus, temporal differences in care of patients not on ART, and cohort heterogeneity (the pre-ART historical data were a mix of North American and European cohorts; the CNICS and VCCC cohorts are US cohorts). In essence, the two approaches are filling in lead time with different sets of data, so it is not surprising that the two approaches obtain different results. The MI approach used pre-ART era data, whereas the dMSM approach used data from patients in the study before they started ART. For this reason, the dMSM approach might be favored. However, there are certainly differences between patients who start and delay starting ART – the dMSM approach attempts to control for these differences using inverse probability weighting – and the validity of this approach comes down to correctly incorporating all confounders. In contrast, in the pre-ART era, nobody started ART, so there may be less selection bias as the entire cohort is used to estimate lead times. In addition, the dMSM approach requires discretization of time, which may result in some information loss (although apparently not too much in our analysis) and extra effort preparing data for analysis. It is worth mentioning, however, that the choice of the time unit is somewhat arbitrary, and setting aside computing time, we could have discretized time to weeks or even days to minimize its impact.

While both statistical methods were novel, the MI approach was specifically designed for the when-to-start ART question and appears less widely applicable given that in many settings, historical data may not be available. In contrast, the dMSM and subsequent modifications appear more applicable across a wide range of scientific problems. With that said, the MI approach was more efficient and seemed to be less sensitive to arbitrary analysis decisions, suggesting there may be some benefits to extensions that might make it more applicable. Possible extensions of the MI approach to estimate a causal parameter similar to that of the dMSM approach and to permit imputation without using historical data and to incorporate covariates would be worthwhile. Although dMSMs appear to be the preference for selecting optimal dynamic regimes, it is important to recognize their limitations, some of which have been seen in our application. dMSMs are complicated and require substantial data assembly; simple coding mistakes and arbitrary choices like at what level to truncate weights can significantly impact results – sensitivity analyses are therefore important. Large sample sizes are also needed to have reasonable precision, and the assumption of no unmeasured confounders is often questionable.

Both of these methods could be used to estimate metrics other than the hazard ratio for comparing treatment initiation rules. Alternative estimands could include contrasts of the predicted survival at specific times after study entry or survival quantiles. It should also be recognized that a better way to select the optimal CD4 count for starting ART would be to simultaneously compare multiple treatment rules, rather than to perform multiple pairwise comparisons as performed in this study and the original publications/methods. For example, one could estimate the predicted survival probability at 5 years for different CD4 initiation thresholds (e.g., 200–700 in 10-unit increments) and then select the CD4 with the highest predicted survival as the optimal CD4 for starting ART [11,12,15]. It might also be preferable to compare all CD4 initiation thresholds with a single reference threshold, rather than with adjacent thresholds, as performed in this study. Our decision to present hazard ratios for pairwise comparisons with adjacent CD4 thresholds was based on our desire to keep the analyses as consistent as possible with those of the original publications. There are other methodological approaches that also could have been employed. The parametric g-formula is well suited for investigating dynamic treatment regimes [32,33], and its use for estimating the optimal CD4 to initiate ART has been demonstrated [13]. Targeted maximum likelihood estimation has also been used in similar problems [10,34,35]. More generally, there is a rapidly growing literature on the estimation of optimal dynamic treatment regimes (e.g., [36–38]).

Our results are neither an indictment nor an endorsement of the analyses presented by the ART-CC and NA-ACCORD [6,7]. The purpose of this study was not to validate either analysis or to come up with yet another estimate of when to start ART using observational data. As a final aside, this exercise highlights the importance of reproducible research [39]. From the original papers, it is impossible to deduce all analysis decisions necessary to exactly replicate the methods employed. In this era of web-based supplements, it is easy to post analysis code so that analyses are transparent – even when the data cannot be made publically available. Although posting one's code takes a bit of courage as it may reveal suboptimal coding and even errors, we believe that such a step is important for assessing scientific methodology and results, and ultimately, medical research.

Acknowledgements

References

1. Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, D'Arminio Monforte A, de Wolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B, Sterne JA, ART Cohort Collaboration. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *Lancet* 2002; **360**:119–129.

2. Pomerantz RJ. Initiating antiretroviral therapy during HIV infection: confusion and clarity. *Journal of the American Medical Association* 2001; **286**:2597–2599.

3. Wang MC, Brookmeyer R, Jewell NP. Statistical models for prevalent cohort data. *Biometrics* 1993; **49**:1–11.

4. Cole SR, Li R, Anastos K, Detels R, Young M, Chmiel JS, Munoz A. Accounting for leadtime in cohort studies: evaluating when to initiate HIV therapies. *Statistics in Medicine* 2004; **23**:3351–3363.

5. 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.

6. Sterne JA, May M, Costagliola D, de Wolf F, Phillips AN, Harris R, Funk MJ, Geskus RB, Gill J, Dabis F, Miro JM, Justice AC, Ledergerber B, Fatkenheuer G, Hogg RS, Monforte AD, Saag M, Smith C, Staszewski S, Egger M, Cole SR. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet* 2009; **373**:1352–1363.

7. Kitahata MM, Gange SJ, Abraham A, 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, Silverberg MJ, Gebo KA, Goedert JJ, Benson CA, Collier AC, Van Rompaey SE, Crane HM, McKaig RG, Lau B, Freeman AM, Moore MD. Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine* 2009; **360**:1815–1826.

8. Hernan MA, Robins JM. Early versus deferred antiretroviral therapy for HIV [letter]. *New England Journal of Medicine* 2009; **361**:822–823.

9. Kitahata MM, Gange SJ, Moore RD. Early versus deferred antiretroviral therapy for HIV [authors reply]. *New England Journal of Medicine* 2009; **361**:823–824.

10. van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics* 2007; **3** Article 3.

11. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* 2008; **27**:4678–4721.

12. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernan MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *International Journal of Biostatistics* 2010; **6** Article 18.

13. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernan MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences* 2011; **3**:119–143.

14. Ewings FM, Ford D, Walker S, Carpenter J, Copas A. Optimal CD4 count for initiating HIV treatment: impact of CD4 observation frequency and grace periods, and performance of dynamic marginal structural models. *Epidemiology* 2014; **25**:194–202.

15. Shepherd BE, Jenkins CA, Rebeiro PF, Stinnette SE, Bebawy SS, McGowan CC, Hulgan T, Sterling TR. Estimating the optimal CD4 count for HIV-infected persons to start antiretroviral therapy. *Epidemiology* 2010; **21**:698–705.

16. Cain LE, Logan R, Robins JM, Sterne JA, Sabin C, Bansi L, Justice A, Goulet J, van Sighem A, de Wolf F, Bucher HC, von Wyl V, Esteve A, Casabona J, del Amo J, Moreno S, Seng R, Meyer L, Perez-Hoyos S, Muga R, Lodi S, Lanoy E, Costagliola D, Hernan MA. When to initiate combination antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries; an observational study. *Annals of Internal Medicine* 2011; **154**:509–515.

17. Jonsson-Funk M, Fusco JS, Cole SR, Thomas JC, Porter K, Kaufman JS, Davidian M, White AD, Hartmann KE, Eron JJ. Timing of HAART initiation and clinical outcomes in human immunodeficiency virus type 1 seroconverters. *Archives of Internal Medicine* 2011; **171**:1560–1569.

18. Severe P, Jean Juste MA, Ambroise A, Eliacin L, Marchand C, Apollon S, Edwards A, Bang H, Nicotera J, Godfrey C, Gulick RM, Johnson WD, Pape JW, Fitzgerald DW. Early versus standard antiretroviral therapy for HIV-infected adults in Haiti. *New England Journal of Medicine* 2010; **363**:257–265.

19. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, Hakim JG, Kumwenda J, Grinsztejn B, Pilotto JH, Godbole SV, Mehendale S, Chariyalertsak S, Santos BR, Mayer KH, Hoffman IF, Eshleman SH, Piwowar-Manning E, Wang L, Makhema J, Mills LA, de Bruyn G, Sanne I, Eron J, Gallant J, Havlir D, Swindells S, Ribaudo H, Elharrar V, Burns D, Taha TE, Nielsen-Saines K, Celentano D, Essex M, Fleming TR. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine* 2011; **365**:493–505.

20. Franco RA, Saag MS. When to start antiretroviral therapy: as soon as possible. *BMC Medicine* 2013; **11** 1:147.

21. Lundgren JD, Babiker AG, Gordin FM, Borges AH, Neaton JD. When to start antiretroviral therapy: the need for an evidence base during early HIV infection. *BMC Medicine* 2013; **11** 1:148.

22. Babiker AG, Emery S, Fatkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, Vjechaj MJ. Considerations in the rationale, design and methods of the Strategic Timing of AntiRetroviral Treatment (START) study. *Clinical Trials* 2013; **10**:S5–S36.
23. INSIGHT START Study Group, Lundgren JD, Babiker AG, Gordin F, Emery S, Grund B, Sharma S, Avihingsanon A, Cooper DA, Fätkenheuer G, Llibre JM, Molina JM, Munderi P, Schechter M, Wood R, Klingman KL, Collins S, Lane HC, Phillips AN, Neaton JD. Initiation of antiretroviral therapy in early asymptomatic HIV infection. *New England Journal of Medicine* 2015; **373**:808–822.
24. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.
25. Cole SR, Frangakis CE. Consistency statement in causal inference: a definition or an assumption? *Epidemiology* 2009; **20**:3–5.
26. Cox C, Chu H, Schneider MF, Munoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine* 2007; **26**:4352–4374.
27. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine* 1990; **9**:1501–1515.
28. Cole SR, Hernan M. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**:656–664.
29. Shepherd BE, Blevins M, Vaz LME, Moon TD, Kipp AM, Jose E, Ferreira FG, Vermund SH. Impact of definitions of loss to follow-up on estimates of retention, disease progression, and mortality: application to an HIV program in Mozambique. *American Journal of Epidemiology* 2013; **178**:819–828.
30. Kitahata MM, Rodriguez B, Haubrich R, Boswell S, Mathews WC, Lederman MM, Lober WB, Van Rompaey SE, Crane HM, Moore RD, Bertram M, Kahn JO, Saag MS. Cohort profile: the Centers for AIDS Research Network of Integrated Clinical Systems. *International Journal of Epidemiology* 2008; **37**:948–955.
31. Lemly DC, Shepherd BE, Hulgan T, Rebeiro P, Stinnette S, Blackwell RB, Bebawy S, Kheshti A, Sterling TR, Raffanti SP. Race and sex differences in antiretroviral therapy use and mortality among HIV-infected persons in care. *Journal of Infectious Diseases* 2009; **199**:991–998.
32. 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.
33. Taubman SL, Robins JM, Mittleman MA, Hernan MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology* 2009; **38**:1599–1611.
34. van der Laan MJ. Targeted maximum likelihood based causal inference: part I. *International Journal of Biostatistics* 2010; **6** Article 2.
35. van der Laan MJ. Targeted maximum likelihood based causal inference: part II. *International Journal of Biostatistics* 2010; **6** Article 3.
36. Murphy SA. Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B* 2003; **58**:331–366.
37. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 2013; **100**:681–694.
38. Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science* 2014; **29**:640–661.
39. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. *American Journal of Epidemiology* 2006; **163**:783–789.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.