# Optimal CD4 Count for Initiating HIV Treatment

## *Impact of CD4 Observation Frequency and Grace Periods, and Performance of Dynamic Marginal Structural Models*

*Fiona M. Ewings,*[a,b,c] *Deborah Ford,*[a] *A. Sarah Walker,*[a] *James Carpenter,*[a,d] *and Andrew Copas*[a]

**Background:** In HIV infection, dynamic marginal structural models have estimated the optimal CD4 for treatment initiation to minimize AIDS/death. The impact of CD4 observation frequency and grace periods (permitted delay to initiation) on the optimal regimen has not been investigated nor has the performance of dynamic marginal structural models in moderately sized data sets—two issues that are relevant to many applications.

**Methods:** To determine optimal regimens, we simulated 31,000,000 HIV-infected persons randomized at CD4 500–550 cells/mm$^3$ to regimens "initiate treatment within a grace period following observed CD4 first <$x$ cells/mm$^3$," $x = 200, 210, \ldots, 500$. Natural history and treatment response were simulated using previous model estimates from CASCADE data. Optimal treatment regimens for the observation frequencies and grace periods were defined by highest 10-year AIDS-free survival. To evaluate the performance of dynamic marginal structural models, we simulated 1000 observational studies (n = 3,000) with CD4-dependent treatment initiation.

**Results:** Decreasing the frequency of CD4 measurements from monthly to every 3, 6, and 12 months increased the optimal regimen from a CD4 level of 350 (10-year AIDS-free survival, 0.8657) to 410 (0.8650), 460 (0.8634), and 490 (0.8564), respectively. Under a regimen defined by $x = 350$ with annual CD4s, 10-year AIDS-free survival dropped to 0.8304. Extending the grace period from 1 to 3 or 6 months, with 3-monthly CD4s, maintained the optimal regimen at 410 for 3 months and increased it to 460 for 6 months. In observational studies with 3-monthly CD4s, the mean (SE) estimated optimal regimen was 402 (76), 424 (66), and 430 (63) with 1-, 3-, and 6-month grace periods; 24%, 15%, and 14% of estimated optimal regimens resulted in >0.5% lower AIDS-free survival compared with the true optimal regimen.

**Conclusions:** The optimal regimen is strongly influenced by CD4 frequency and less by grace period length. Dynamic marginal structural models lack precision at moderate sample sizes.

(*Epidemiology* 2014;25: 194–202)

Treatment of many diseases is determined by a person's time-varying characteristics; such a treatment regimen is termed dynamic. One example is HIV infection, where, to minimize time on treatment when side effects and resistance can develop, treatment initiation may be delayed until some immunodeficiency is apparent, indicated by low values of the biomarker, CD4 count. Dynamic treatment regimens can be defined by the observed CD4 threshold below which treatment is initiated. A large randomized trial comparing two CD4-based initiation regimens is ongoing[1]; in the meantime, researchers have attempted to answer this question across a broad range of regimens using observational data.[2–5] Dynamic treatment questions also arise in other contexts; for example, when and how to treat type 2 diabetes are determined by fasting blood glucose or HbA1c levels,[6] and diagnosis and hence treatment of hypertension are based on blood pressure.[7] Such questions usually imply the presence of a time-dependent confounder affected by prior treatment, like CD4 in HIV infection. In this situation, causal methods such as the *g*-formula,[8,9] *g*-estimation of structural nested models,[10] or dynamic marginal structural models[11,12] are required for unbiased estimation.

In any application, the optimal treatment regimen may depend on several factors, including observation frequency and the grace period (permitted delay from the time treatment is indicated, such as CD4 below a specified threshold, to initiation), but their effects have not been systematically investigated nor has the performance of dynamic marginal structural models been explored in

Submitted 6 March 2013; accepted 10 September 2013.

From the [a]Medical Research Council Clinical Trials Unit at University College London, London, United Kingdom; [b]MRC Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, London, United Kingdom; [c]Mwanza Intervention Trials Unit, National Institute for Medical Research, Mwanza, Tanzania; and [d]London School of Hygiene and Tropical Medicine, London, United Kingdom.

The authors report no conflicts of interest.

F.M.E. was funded by a Medical Research Council studentship.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

Correspondence: Fiona M. Ewings, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; and Mwanza Intervention Trials Unit, National Institute for Medical Research, PO Box 11936, Mwanza, Tanzania. E-mail: fiona.ewings@lshtm.ac.uk.

ISSN: 1044-3983/14/2502-0194
DOI: 10.1097/EDE.0000000000000043

realistically-sized data sets. We addressed these questions by simulating large randomized trials and realistically-sized observational studies using HIV infection as an example and assessed whether the observational dynamic marginal structural model analysis correctly identifies the optimal regimen defined by the randomized simulation. We outline the theory of dynamic marginal structural models, describe our simulation methods, present our results, and conclude with a discussion of our findings and recommendations.

## METHODS

### Definitions and Notation

We defined 31 regimens by "initiate treatment within $g$ months of observed CD4 first $<x$ cells/mm$^3$" where $x = 200, 210, \ldots, 500$ and $g$ is the grace period (the time period during which treatment initiation is still considered compliant with regimen $x$). Because we consider immediate initiation to be "within the first month," we denote this as $g = 1$. Thus, we consider true grace periods to be given by $g \geq 2$, in contrast to Cain et al,[12] who considered immediate treatment initiation as $m(= g - 1) = 0$. A benefit of a grace period is that, if a person's observed CD4 count rises slightly from the nadir before treatment initiation, then the person may still be considered compliant with some regimens with which they otherwise may not have been.

Let $T_x$ be the time from study entry to AIDS/death for a participant under regimen $x$. If we could observe $T_x$ for all participants and regimens, or if a large number of participants were randomized to each regimen $x$, then the optimal regimen $x$ would simply be that which minimizes the AIDS/death risk across all participants–reflected, for example, by the probability of remaining AIDS-/death-free 10 years after baseline. In practice, we observe only a subset of regimens for each participant, and the regimen(s) that each participant is observed to follow may be confounded by their prognosis. Under the assumption of consistency,[13] $T$, the time to AIDS/death under each observed regimen, $x$, is $T_x$. For those regimens $x$ with which a participant is not compliant throughout follow-up, $T_x$ is counterfactual.

We divide time into monthly intervals, $t = 1, 2, \ldots,$ and define $C_x(t)$ as an indicator for artificial censoring, which equals 0 if the participant's observed data are still compliant with regimen $x$ before time $t$ and equals 1 otherwise (eAppendix, http://links.lww.com/EDE/A753). Let $A(t)$ be an indicator for whether treatment was initiated before $t$, $Y(t)$ an indicator for whether AIDS/death was experienced before $t$, and $Q_x(t)$ an indicator for whether observed CD4 was $<x$ cells/mm$^3$ before $t$. Then[12]:

$$C_x(t) = 0 \text{ if and only if, for all } j \leq t,$$
$$A(j) = 0 \text{ when } Q_x(j-1) = 0, Y(t) = 0$$

and

$$A(j+g-1) = 1 \text{ when } Q_x(j-1) = 1,\ Y(t) = 0$$

(and compliance is not applicable if $Y(t) = 1$ since then that person is no longer at risk for the event). That is, participants are censored if they initiated treatment before becoming eligible under a given regimen (observed CD4 $\geq x$ cells/mm$^3$). No participants are censored during the grace period but are censored at the last ($g$th) month of the grace period if they have not initiated treatment by that time. Those who are not censored at that point (ie, initiated treatment within the grace period) will remain uncensored for the rest of their follow-up (eAppendix, http://links.lww.com/EDE/A753).

### Dynamic Marginal Structural Model Methods

A dynamic marginal structural Cox model may be defined for the time to AIDS/death by:

$$\lambda_{T_x}(t \mid x, V) = \lambda_0(t)\exp\{\alpha g(x,t) + \beta V\}$$

where $\lambda_0(t)$ is the baseline hazard, $V$ represents baseline covariates, and α and β are to be estimated. $g(x, t)$ is some function of the regimens $x$ including an interaction with time $t$, which is essential. For example, for a given participant, the regimens $x = 200$ and 350 are the same until CD4 is first observed <350 cells/mm$^3$; therefore, the hazard ratio between the two regimens cannot be constant over time.

All participants are considered to be compliant with all regimens initially; they are artificially censored from regimens when they first become noncompliant with that regimen (as above). In practice, estimation requires data expansion, duplicating each participant's data for the 31 regimens, while they remain compliant with each. The artificial censoring process is likely to be informative; we take account of this using inverse probability weighting. Under the assumption of no unmeasured confounders for censoring and outcome, and given baseline and time-updated covariates and treatment history, the weight for a participant on regimen $x$ at time $t$ is the inverse probability of remaining uncensored to $t$. Conditional on baseline and time-updated covariates and treatment history, this is the same as the probability of the observed treatment history to time $t$.[11,12,14] With a grace period, regimens require further specification; we defined them as: "initiate treatment within $g$ months of CD4 first $<x$ cells/mm$^3$, such that there is a uniform probability of starting in each of the months 1, 2, …, $g$."[12] Informally, participants who initiate treatment at any point during the grace period are upweighted to account for those censored at the end of the grace period because they did not initiate by that time. The (nonstabilized) weights are estimated as:

$$W_x(t) = \frac{I\left[C_x(t)=0\right]}{\prod_{j=1}^{t}\left\{p_A(j)^{I[j \leq q_x]}\right\}} \times \prod_{l=1}^{g}\left\{\begin{array}{l}\left\{\dfrac{1-1/(g+1-l)}{p_A(q_x+l)}\right\}^{I\left[t \geq q_x+l,\, A(q_x+l)=0\right]} \\ \times\left\{\dfrac{1/(g+1-l)}{1-p_A(q_x+l)}\right\}^{I\left[t \geq q_x+l,\, A(q_x+l-1)=0,\, A(q_x+l)=1\right]}\end{array}\right\} \qquad (1)$$

where $I[\cdot]$ is the indicator function, overbars indicate history, $q_x$ is the time CD4 is first observed $<x$ cells/mm$^3$, and $p_A(j) = \Pr\{A(j)=0 \mid \bar{A}(j-1)=0, Y(j)=0, \bar{L}(j-1)\}$ is the probability of not initiating treatment at time $j$, given covariate history and no initiation by time $j-1$. The first component of this equation is the probability of remaining uncensored while CD4 ≥ $x$ cells/mm$^3$, that is, off treatment. The second component spans the grace period $l = 1, ..., g$, that is, covering the $g$ months after treatment is indicated by the regimen and CD4 history. The denominator is based on the probabilities of observed treatment, that is, the probability of remaining off treatment while treatment-naive during the grace period, multiplied by the probability of initiating treatment when (if) treatment is initiated during the grace period. The numerator serves to upweight those participants who initiated during the grace period to account for those who are censored at the end of the grace period due to noninitiation of treatment and maintains the uniform distribution of treatment initiation over the grace period. We can estimate $p_A(j)$ from the data using a pooled logistic regression model. The treatment probabilities are independent of regimen $x$, so we fit this model on a data set with one observation per participant (per month). When we expand, as above, at any time, the weights are constant for each participant across all regimens with which they are still compliant. While stabilized weights may be used to increase precision, this is not guaranteed,[12] and their calculation is nontrivial when assuming uniform initiation across a grace period. We therefore used nonstabilized weights, truncated at 20.

We used weighted discrete-time survival regression (pooled logistic regression) to approximate the Cox model[15]:

$$\begin{aligned}\Pr\{Y(t+1)=0 \mid Y(t)=0, C_x(t)=0, x, V\} \\ = \operatorname{expit}\{\alpha g(x) + \beta V + \gamma f(t) + \delta g(x) f(t)\}\end{aligned} \tag{2}$$

where $g(x)$ and $f(t)$ are smooth functions of the regimens $x$ and time $t$, respectively. The parameters $\{\alpha, \beta, \gamma, \delta\}$ are estimated using weighted maximum pseudolikelihood, with robust standard errors. In real applications, administrative (end-of-study) censoring can be considered independent, but early loss to follow-up can be incorporated using independent inverse probability weighting.

## Simulation Study

We simulated CD4 trajectories monthly over 10 years for HIV-infected participants with observed baseline CD4s distributed uniformly in 500–550 cells/mm$^3$ (reflecting relatively early presentation to care). We used a piecewise-linear mixed-effects model for square root underlying CD4, with change point at treatment initiation and 1 year later. True CD4 was determined as the underlying CD4 plus Brownian motion (superior to standard mixed-effects model, based on data from CASCADE [Concerted Action on SeroConversion to AIDS and Death in Europe]).[16] Observed CD4 was the true CD4 plus measurement error. Correlations between CD4 at treatment initiation and subsequent (initial and after 1 year) slopes were negative, and slopes could be negative, particularly after 1 year and with high CD4 counts before initiating antiretroviral therapy (eAppendix, http://links.lww.com/EDE/A753). That is, we incorporated a potential penalty for early treatment initiation, meaning that CD4 levels could decline after treatment initiation (as observed in CASCADE data).

Let $u_i^Y(t)$ represent the probability of AIDS/death at a given time $t$; this was simulated dependent on true CD4 and treatment as follows:

$$u_i^Y(t) = \begin{cases} 0.582 - 0.266\sqrt{\mathrm{CD4}_i^T(t-1)} & \text{if } A_i(t-1)=0 \\ 0.763 - 0.415\sqrt{\mathrm{CD4}_i^T(t-1)} & \text{if } A_i(t-1)=1 \end{cases}$$

where CD4$^T(t-1)$ is the true CD4 at time $t-1$, and the parameters were chosen such that the probability of AIDS/death was 0.01 and 0.0005 off treatment at CD4s of 200 and 500 cells/mm$^3$, respectively, and correspondingly 0.006 and 0.0002 on treatment (based on previous work by the CASCADE collaboration and by Babiker et al[16]; eAppendix, http://links.lww.com/EDE/A753).

### Determining the Optimal Regimen via Simulated Randomized Trials

To determine optimal regimens for treatment initiation in these populations, we simulated large randomized trials, each with 31,000,000 persons randomly allocated to follow one of the 31 regimens. For each trial, the CD4 observation frequency was fixed at 1, 3, 6, or 12 months, and the grace period was also fixed at one of these values. For a given grace period >1 month, to achieve uniform treatment initiation across the grace period, participants identified for initiation (based on their randomized regimen and observed CD4) were randomly allocated to initiate in one of the grace period intervals, each with probability 1/$g$. We initially defined the optimal regimen to be that maximizing the 10-year AIDS-free survival estimated by Kaplan–Meier procedures; no underlying model was assumed to avoid implausible proportional hazards assumptions. We also applied a least-squares smoothing procedure (bandwidth 0.2) as the 10-year AIDS-free survival estimates did not always vary smoothly, despite the large sample size.

### Evaluating the Performance of Dynamic Marginal Structural Models via Simulated Observational Studies

To evaluate the performance of dynamic marginal structural models in realistically-sized data sets, under scenarios with a CD4 observation frequency of 3 months and grace periods of 1, 3, and 6 months, we simulated 1000 observational studies of 3000 persons, with the same populations as the randomized trials but with treatment initiation probability dependent on CD4. AIDS-free survival was estimated by weighted Kaplan–Meier analysis, with each member of the risk set for each regimen and at each time point weighted as per Equation 1 and pooled logistic regression (Equation 2; the inverse probability-weighted pooled logistic regression models approximating a Cox dynamic marginal structural model), with the optimal regimen determined as that yielding the highest 10-year AIDS-free survival. Note that the grace periods were a step in the data analysis, not in the data generation.

The following model was used to determine treatment initiation (probability given by $1-p_A(t)$), based on CASCADE data (eAppendix, http://links.lww.com/EDE/A753):

$$\log\frac{1-p_A(t)}{p_A(t)} = 4.62 - 0.412\sqrt{CD4^{O}(t)}$$

where $CD4^{O}(t)$ is the observed CD4 at time $t$.

While grace periods may occur in practice, due to time required to receive and act on CD4 results, a clinician may wish to know whether to initiate treatment immediately (1-month grace period). There may also be efficiency gains in permitting a grace period in observational analyses due to fewer censored treatment initiations, albeit at potentially increased risk of bias in answering the question of interest (impact of immediate treatment initiation). We evaluated this bias-variance trade-off, treating the optimum regimen identified from the simulated randomized trials as the true optimum value with a 1-month grace period, and estimated the bias, mean square error (variance + bias$^2$), and relative efficiency of various grace period lengths (relative efficiency = variance/variance under reference method, chosen as pooled logistic regression because this is typically used in practice,[5,12] with a 1-month grace period).[17]

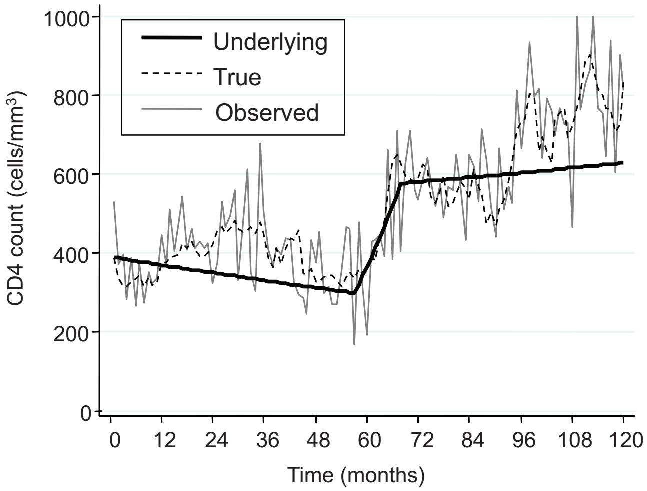

**FIGURE 1.** CD4 trajectory (underlying slope, true, and observed CD4) for an example participant in a simulated trial randomized to the regimen "initiate treatment within 1 month of observed CD4 first <200 cells/mm$^3$." True CD4 incorporates Brownian motion, and observed CD4 also incorporates measurement error (see text for more details).

## RESULTS

### Simulated Randomized Trials

For the simulated trial with monthly CD4s and a 1-month grace period, the CD4 trajectories for an example participant randomized to the regimen $x$ = 200 are shown in Figure 1 and illustrate the large measurement error in this biomarker. Baseline characteristics for the n = 1,000,000 simulated participants randomized to each of the regimens $x$ = 200, 350, and 500 were comparable across regimens (Table 1; similar

**TABLE 1.** Baseline Characteristics and Treatment for n = 1,000,000 Simulated Participants Randomized to Each of the Regimens $x$ = 200, 350, 500 (CD4s Observed Monthly, 1-month Grace Period)

| | Regimens Given by *x* | | |
|---|---|---|---|
| | 200 | 350 | 500 |
| Baseline | | | |
| Observed CD4 count (cells/mm$^3$) | 525 (513 to 538) | 525 (512 to 537) | 525 (512 to 538) |
| True CD4 count (cells/mm$^3$) | 525 (457 to 598) | 525 (457 to 598) | 525 (457 to 598) |
| Annual slope, square root scale | –1.10 (–1.44 to –0.76) | –1.10 (–1.44 to –0.76) | –1.10 (–1.44 to –0.76) |
| Treatment | | | |
| No. persons observed to initiate treatment; no. (%) | 783,766 (78) | 952,144 (95) | 995,522 (>99) |
| Time to initiation (months)[a] | 37 (21 to 62) | 10 (4 to 25) | 3 (2 to 5) |
| Observed CD4 count at initiation (cells/mm$^3$)[a] | 175 (154 to 189) | 313 (282 to 333) | 432 (379 to 469) |
| True CD4 count at initiation (cells/mm$^3$)[a] | 275 (240 to 313) | 425 (381 to 472) | 503 (441 to 565) |
| Initial annual slope after initiation, square root scale[a,b] | 3.42 (2.16 to 4.68) | 2.78 (1.52 to 4.04) | 2.51 (1.24 to 3.77) |
| Annual slope 1 year after initiation, square root scale[a,b] | 0.41 (0.30 to 0.52) | 0.01 (–0.10 to 0.12) | –0.17 (–0.31 to –0.03) |
| Percentage of follow-up time spent on treatment; % | 49 | 79 | 94 |

Unless otherwise stated, values are median (interquartile range).
[a] In those participants who were observed to initiate treatment.
[b] As assigned at treatment initiation.

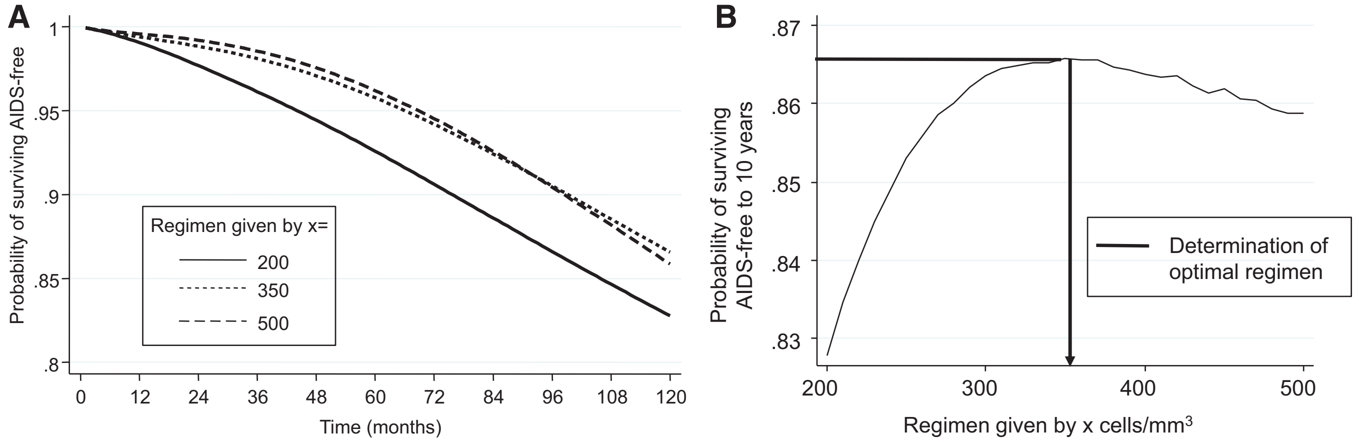

**FIGURE 2.** AIDS-free survival based on the simulated randomized trial with CD4s observed monthly and 1-month grace period. A, Over 10 years, for the three regimens *x* = 200, 350, and 500. B, By regimens *x* = 200, 210, ..., 500 and illustrating determination of the optimal regimen from the 10-year AIDS-free survival by regimen plot.

for other regimens, not shown). Treatment patterns were as expected between regimens (eg, those randomized to *x* = 500 initiated earliest). The median observed CD4 counts at initiation were 175 (interquartile range = 154–189), 313 (282–333), and 432 (379–469) cells/mm$^3$ for *x* = 200, 350, and 500, respectively, yet the corresponding true values were 275 (240–313), 425 (381–472), and 503 (441–565) cells/mm$^3$. Thus, the large measurement error illustrated in Figure 1 has a substantial impact, raising the true CD4 at initiation.

The 10-year AIDS-free survival was 0.8278, 0.8657, and 0.8587 for regimens *x* = 200, 350, and 500, respectively, indicating that, of these three regimens, *x* = 350 was optimal (Figure 2A). Indeed, across all 31 regimens, the optimal was *x* = 350 (Table 2 and Figure 2B).

Decreasing the CD4 observation frequency to every 3, 6, and 12 months increased the optimal regimen to 410, 460, and 490, respectively (1-month grace period); yet, despite increasing the optimal regimen's CD4 threshold, the 10-year AIDS-free survival dropped only modestly, from 0.8657 to 0.8650, 0.8634, and 0.8564, respectively. Under a 350 regimen (optimal when CD4s were observed monthly) with 3-, 6-, and 12-month CD4s, the 10-year AIDS-free survival dropped slightly more (from 0.8657 to 0.8616, 0.8528, and 0.8304, respectively). Under a 410-regimen (optimal when CD4s were observed every 3 months), the 10-year AIDS-free survival declined from 0.8650 to 0.8615 and 0.8484 with 6- and 12-monthly CD4s, respectively.

With monthly CD4s, extending the grace period from 1 to 3, 6, and 12 months maintained or increased the optimal regimen from 360 (10-year AIDS-free survival = 0.8657) to 360 (0.8644), 370 (0.8631), and 380 (0.8598), respectively. With 3-monthly CD4s, the optimal regimens under 3- and 6-month grace periods were 410 (0.8638) and 460 (0.8625), respectively. With 6-monthly CD4s, the optimal regimen under a 6-month grace period was 460 (0.8603).

Sensitivity analyses varying the initial off-treatment rate of CD4 decline (Table 3) showed little variation in the optimal regimen when CD4s were observed every 6 or 12

**TABLE 2.** Optimal Regimens as Determined from the Randomized Trials (Raw 10-year AIDS-free Survival) by CD4 Observation Frequency and Grace Period

| CD4 Observation Frequency (Months) | Grace Period (Months) | | | |
|---|---|---|---|---|
| | 1 | 3 | 6 | 12 |
| 1 | 350 (0.8657) | 360 (0.8644) | 370 (0.8631) | 380 (0.8598) |
| 3 | 410 (0.8650) | 410 (0.8638) | 460 (0.8625) | — |
| 6 | 460 (0.8634) | — | 460 (0.8603) | — |
| 12 | 490 (0.8564) | — | — | — |

Note: Simulations initially explored the impact of varying grace period for fixed CD4 observation frequency of 1 month and of varying CD4 observation frequency for 1-month grace period. Subsequent simulations focused on the most realistic scenarios: "—" indicates that a specific simulation was not performed.

**TABLE 3.** Optimal Regimens as Determined from the Randomized Trials (Raw 10-year AIDS-free Survival) by CD4 Observation Frequency and Rate of Initial (Off-treatment) CD4 Decline

| CD4 Observation Frequency (Months) | Initial (Off-treatment) CD4 Decline | | |
|---|---|---|---|
| | Slow | Regular | Fast |
| 1 | 310 (0.8731) | 350 (0.8657) | 360 (0.8601) |
| 3 | 380 (0.8720) | 410 (0.8650) | 460 (0.8592) |
| 6 | 460 (0.8705) | 460 (0.8634) | 460 (0.8569) |
| 12 | 490 (0.8671) | 490 (0.8564) | 500 (0.8471) |

Note: Estimated under 1-month grace period. "Regular" CD4 decline corresponds to 1-month grace period in Table 2; "slow" and "fast" decline correspond to the lower and upper quartiles, respectively, of the pretreatment CD4 trajectories in CASCASE data (eAppendix, http://links.lww.com/EDE/A753).

months. However, particularly when CD4s were observed every 3 months, populations with slower CD4 decline had a lower optimal regimen, and those with faster CD4 decline had a higher optimal regimen.

Applying the local smoothing, there were small changes in the estimated optimal regimens compared with those determined directly by Kaplan–Meier analysis: from 350 to 360 with monthly CD4s/1-month grace period, from 380 to 400 with monthly CD4s/12-month grace period, from 460 to 420 with CD4s observed every 3 months/6-month grace period, and from 460 to 470 with CD4s observed every 6 months/6-month grace period. These changes all occurred in regions where the AIDS-free survival was fairly constant.

## Simulated Observational Studies

In the observational studies (n = 3000 participants) with CD4 levels observed every 3 months, the mean estimated optimal regimen was 395 (standard error = 70), 418 (57), and 420 (60) with 1-, 3-, and 6-month grace periods, respectively, under Kaplan–Meier analysis, and 402 (76), 424 (66), and 430 (63), respectively, under pooled logistic regression (Figure 3). The corresponding percentages of estimated optimal regimens associated with AIDS-free survival that were more than 0.5% worse than under the true optimal regimen (from the trials) were 23%, 12%, and 16% under Kaplan–Meier, and 24%, 15%, and 14% under pooled logistic regression. Note that approximately 10% of simulations using Kaplan–Meier and 25% using pooled logistic regression estimated the optimal regimen at the maximum considered (*x* = 500).

Assuming the desired inference is under 1-month grace period, the bias was –15 (mean square error = 5,099; relative efficiency = 0.84), 8 (3,273; 0.55), and 10 (3,644; 0.61) with 1-, 3-, and 6-month grace periods, respectively, under Kaplan–Meier and –8 (5,874; 1.00 [reference]), 14 (4,537; 0.75), and 20 (4,333; 0.68), respectively, under pooled logistic regression. Bias was similar using local smoothing of the Kaplan–Meier estimates (data not shown). Simulation of large observational studies (n = 100,000) indicated no bias in

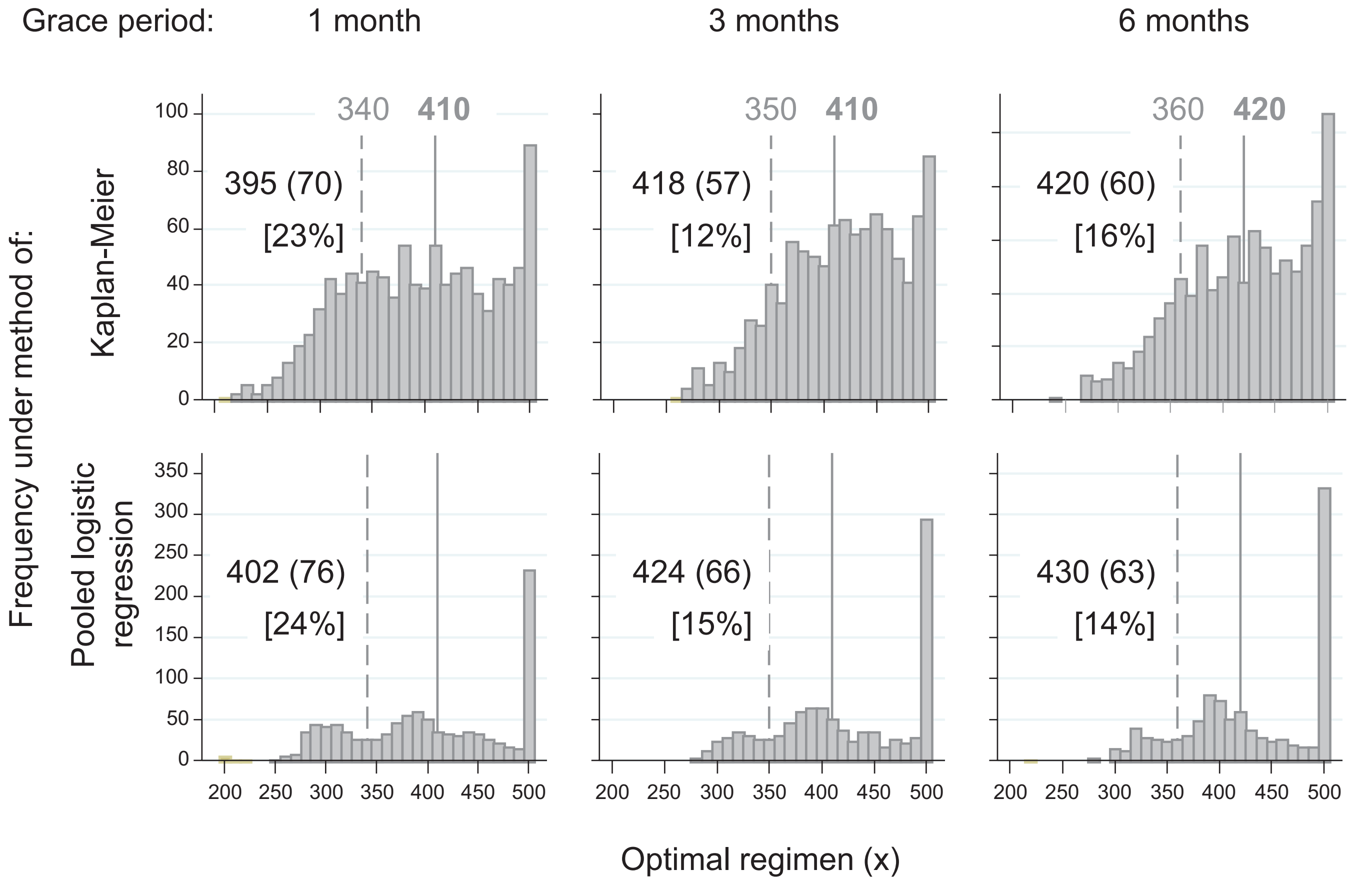

**FIGURE 3.** Optimal regimens as estimated by the simulated observational studies (n = 3000 participants in each of 1000 studies for each scenario) with CD4s observed every 3 months, with the optimal regimen estimated either by raw Kaplan–Meier analysis or pooled logistic regression, and with grace periods of 1, 3, or 6 months. Numbers in bold grey above the solid vertical lines are the optimal regimen as determined from the corresponding randomized trials (after smoothing), and numbers in grey above the dashed vertical lines indicate the highest regimen with >0.5% lower AIDS-free survival versus the true optimal regimen. Numbers in black are the mean (standard error) [percentage less than the dashed grey line] across the 1000 simulated observational studies.

the results as estimated by weighted Kaplan–Meier (results not shown).

## DISCUSSION

We have demonstrated via a simulation based on HIV infection that frequency of observation has a major impact on the optimal regimen for treatment initiation. Lengthening the grace period (permitted delay for treatment initiation) has less influence, at least for 3–6 months. In HIV infection, these findings have implications for the generalizability of results from studies in high-income to low-income countries; in the latter, CD4 levels are typically measured less frequently, although similar AIDS-free survival may be achieved, provided the optimal regimen (CD4 threshold for treatment initiation) is raised accordingly. Our results also highlight that it is essential for future studies using causal analysis methods to report their observation frequency (as observed in the data) and the grace period (as assumed for analysis) to enable comparisons and judgments as to the relevance of findings to different settings. Furthermore, it is likely that observation frequencies will vary within and between persons and cohorts; the impact of such variation warrants further exploration and suggests that reporting measures of the spread of observation frequencies within and between cohorts is important.

It is interesting to consider why CD4 observation frequency has a greater impact on the optimal regimen than does grace period length. There is an asymmetry due to regimens being defined by CD4 counts dropping below a threshold. For example, when CD4 counts were observed monthly, the same participants were identified for treatment initiation, regardless of whether a grace period of 1 or 12 months was permitted. However, if a CD4 count observed on the monthly schedule indicated treatment initiation but that CD4 count was not observed on the 12-monthly CD4 count schedule, then that participant would not have been identified for treatment initiation under the less-frequent CD4 observation until sometime later. Therefore, to identify such persons for treatment initiation, the optimal regimen would need to be higher. Measurement error also has a larger impact on observation frequency than it does on grace period. For example, if each random low CD4 count that happened to be observed on a monthly, but not yearly, schedule led the person to be started on treatment early, then the optimal regimen would be reduced with more frequent observation. Requiring a confirmatory CD4 count for initiation might reduce the impact of measurement error (although not random fluctuation). However, in practice, relatively few people have confirmatory counts before starting treatment, making implementation of such optimal regimens difficult. The delay to initiation assuming uniform treatment initiation across a grace period also has less effect than a similar length delay to next CD4 observation. This is because even with a 12-month grace period, for example, half of the (upweighted) participants are assumed to initiate within 6 months of reaching the threshold.

With sufficient data and under the necessary assumptions, the dynamic marginal structural model methods are unbiased,[11] as we showed in simulations of very large observational studies (n = 100,000 participants). However, where there are large natural fluctuations and measurement error in the biomarker that defines the regimens, and where the event rate is low, as in our example, the methods may lack precision in realistically-sized cohorts. This was exacerbated by the broadly constant AIDS-free survival at high CD4s, meaning that any single analysis may yield an estimate for CD4 for optimal initiation quite far from the true value. This was illustrated by the large number of simulations, suggesting that the optimal regimen was the maximum considered (CD4 = 500), particularly under the pooled logistic regression approach. Of note, the methods, particularly pooled logistic regression, performed better in further less-realistic simulations with a clearer “peak” in the optimal regimen (results not shown). Our findings reinforce the current view that large collaborative clinical cohorts are required to answer such causal questions. They also suggest that Kaplan–Meier estimates may be preferred and that interpretation should consider how the predicted outcome varies by regimen (Figure 2B) and recognize that precision may be low. However, if this curve is fairly flat, then the implications for clinical practice from not identifying the exact optimal regimen may not be severe, despite the lack of precision.

The recent extension of these methods to incorporate grace periods is an attempt to address the relatively limited data typically available, by decreasing censored treatment initiations and hence increasing power. However, the potential implications of these extended methods in realistic scenarios have not previously been systematically explored. Our findings suggest that observational studies in similar resource-rich settings, with CD4 observation frequencies of every 3 months, should use a 3-month grace period in analysis, as this increases precision with little bias. Longer grace periods were associated with increased bias. However, across the grace periods considered, the efficiency gains were perhaps smaller than might have been anticipated. Our results also suggest that a grace period of a length similar to the observation frequency would be reasonable in other settings.

Grace periods are typically referred to simply by their length, but this does not provide a full definition of the regimen. Previously, most researchers have upweighted only those observed to initiate treatment in the last interval of the grace period,[3,5,12] considering regimens of the form “initiate treatment in the last month of the grace period if not already on treatment.” However, we preferred to upweight all those who initiated during the grace period, considering regimens of the form “initiate treatment within the grace period such that there is a uniform probability of starting treatment in each of the months within the grace period” to avoid upweighting a potentially small and unrepresentative subset of participants. Cain et al[12] provide further discussion on this.

This simulation study has several limitations. First, alternative regimen definitions could be applied. For example, finer or coarser CD4 categorizations could be used, although our choice is considered practically and clinically a good compromise.[12] Regimens could include initiation at even higher CD4 thresholds, although in practice few people maintain such high CD4 levels after seroconversion.[18] Regimens could also incorporate different frequencies of CD4 measurement between and within individuals, possibly dependent on participants' covariates. Furthermore, regimens could be defined based on other time-independent or time-dependent covariates (such as HIV viral load), although estimation may be hampered by the limited data available for each regimen if many factors are incorporated. Identification of the optimal regimen may alternatively be based partly on the population-level benefits.[19]

Large measurement error and random fluctuation are inherent challenges in the application of these methods, causing large proportions of records with observed treatment initiation to be artificially censored from all regimens if the treatment initiations occurred at CD4 counts above the nadir, even when permitting a grace period of up to 6 months to censor fewer treatment initiations. If CD4 counts declined linearly while treatment-naive (ie, in the absence of Brownian motion and measurement error), this censoring would no longer occur. Alternative approaches, such as defining a regimen to require two CD4 counts below each threshold $x$ for treatment initiation, could perhaps reduce censored treatment initiations as an alternative or even parallel approach to include a grace period. However, if CD4 counts are not consistently confirmed in an observational study (as is typically the case), then enforcing this in analysis is unlikely to be beneficial and may leave only unrepresentative participants uncensored. A second approach could be to use CD4 history to better estimate current observed CD4 (eg, using smoothing), to potentially reduce the impact of measurement errors. However, given their size, some censoring of treatment initiations seems inevitable in our application.

A further limitation is that our simulation models incorporated a negative correlation between CD4 at treatment initiation and long-term slope thereafter, based on previous CASCADE data modeling. This correlation may be driven by persons who initiated treatment early but subsequently stopped. Therefore, our simulation may have underestimated the benefit of early treatment initiation at high CD4, assuming that treatment is continued once initiated. However, one could argue that the observed and modeled CD4 trajectories in the CASCADE data are likely to mimic what would happen in practice, regardless of underlying reasons. Of note, if a penalty for early treatment initiation had not been incorporated via this negative correlation, and CD4 never declined while on treatment, then it would always be optimal to initiate treatment immediately.

The determination of optimal treatment regimens is heavily dependent on time. In this example, sufficient time must be allowed for the biomarker to decrease and hence for differences in the outcome to emerge between the regimens. We simulated 10-year follow-up, but optimal regimens may have been different if longer follow-up was considered. For example, Figure 2A shows that the CD4 < 350 regimen is optimal only after 7 years, with the CD4 < 500 regimen optimal before this. In addition, other metrics, for example, a CD4-based or quality of life–based metric,[5,11] or restricted mean survival[20] may yield different optimal regimens.

Other approaches such as the $g$-formula or $g$-estimation of structural nested models could be used as alternatives to dynamic marginal structural models. $g$-Estimation of structural nested models has the potential to be more efficient than dynamic marginal structural models and has fewer parametric assumptions than the $g$-formula.[21] However, structural nested models are less robust to model misspecification and are not intuitive to apply. Dynamic marginal structural models more closely resemble standard methods and so their implementation and interpretation are more straightforward. However, these dynamic models require the assumption of positivity (ie, at all combinations of values of covariate and treatment histories which occur in the population, there is a nonzero probability of following each of the regimens under consideration).[13] In addition, the artificial censoring process required for the application of dynamic marginal structural models may result in reduced power. While the $g$-formula can easily incorporate highly complex dynamic regimens, it is computationally intensive and perhaps most useful when a small number of dynamic regimens are to be compared.

There are similarities between the application of the $g$-formula and our observational simulation studies. In our analyses, we a priori defined the covariate, treatment, and outcome distributions, conditional on covariate and treatment history, based on previous work with CASCADE data, while the $g$-formula estimates the parameters of these distributions from the data. Then, similar to the $g$-formula, we simulated a cohort using those distributions; however, while the $g$-formula would also incorporate the treatment regimen of interest, our approach incorporated this after expansion of the simulated cohort by censoring individuals when no longer compliant. The final step of our simulation studies was to estimate the outcome, as in the $g$-formula, except that inverse probability weighting was applied to account for the potentially informative censoring of noncompliant participants. In addition, estimation of the outcome is performed separately for each regimen under the $g$-formula, whereas the dynamic marginal structural models allowed us to model the outcome across all regimens at once.

In summary, causal analysis methods provide an opportunity to address many questions from observational studies, which could otherwise not be considered without the possibility of major bias due to time-dependent confounding. It is infeasible to conduct sufficient randomized controlled trials to address all these questions. However, we have shown that answers from causal analyses may depend strongly on their implementation, in ways that may not be obvious to a casual reader. This is a particular concern when comparing results between studies. Researchers conducting such analyses should be aware of these limitations, describe the full details of implementation, and present multiple sensitivity analyses to delineate the effect of their assumptions on the results.

## ACKNOWLEDGMENTS

*We are grateful to Abdel Babiker for his advice on developing the simulation study and to the CASCADE collaboration for allowing the use of their data to inform the simulation study.*

## REFERENCES

1. INSIGHT. START 001 international randomized trial. 2009. Available at: http://insight.ccbr.umn.edu/start/. Accessed August 26, 2010.
2. Kitahata MM, Gange SJ, Abraham AG, et al; NA-ACCORD Investigators. Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med*. 2009;360:1815–1826.
3. HIV-CAUSAL Collaboration. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries. *Ann Intern Med*. 2011;154:509–515.
4. When to Start Consortium. 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.
5. Shepherd BE, Jenkins CA, Rebeiro PF, et al. Estimating the optimal CD4 count for HIV-infected persons to start antiretroviral therapy. *Epidemiology*. 2010;21:698–705.
6. Nathan DM, Buse JB, Davidson MB, et al; American Diabetes Association; European Association for Study of Diabetes. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care*. 2009;32:193–203.
7. Williams B, Poulter NR, Brown MJ, et al; BHS guidelines working party, for the British Hypertension Society. British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary. *BMJ*. 2004;328:634–640.
8. Daniel R, De Stavola B, Cousens S. g-formula: estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata J*. 2011;11:479–517.
9. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011;3:119–143.
10. Robins J. Correcting for noncompliance in randomized trials using structural nested mean models. *Commun Stat*. 1994;23:2379–2412.
11. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Stat Med*. 2008;27:4678–4721.
12. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *Int J Biostat*. 2010;6:Article 18.
13. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664.
14. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
15. 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. *Stat Med*. 1990;9: 1501–1515.
16. Babiker AG, Emery S, Fätkenheuer G, et al; INSIGHT START Study Group. Considerations in the rationale, design and methods of the Strategic Timing of AntiRetroviral Treatment (START) study. *Clin Trials*. 2013;10(1 suppl):S5–S36.
17. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25:4279–4292.
18. Lodi S, Phillips A, Touloumi G, et al; CASCADE Collaboration in EuroCoord. Time from human immunodeficiency virus seroconversion to reaching CD4+ cell count thresholds <200, <350, and <500 cells/mm³: assessment of need following changes in treatment guidelines. *Clin Infect Dis*. 2011;53:817–825.
19. Eaton JW, Johnson LF, Salomon JA, et al. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. *PLoS Med*. 2012;9:e1001245.
20. Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30:2409–2421.
21. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. *Stat Med*. 2013;32:1584–1618.