Practical and theoretical considerations of the application of marginal structural models to estimate causal effects of treatment in HIV infection

Fiona Marie Ewings

Submitted to University College London
for the degree of Doctor of Philosophy

University College London, &
Medical Research Council Clinical Trials Unit

Declaration

I, Fiona Marie Ewings, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

Acknowledgements

I am extremely grateful to my supervisors Andrew Copas, Sarah Walker and James Carpenter, and particularly to my unofficial supervisor Debbie Ford.

I would like to thank the CASCADE collaboration for allowing me to use their data, and the Medical Research Council for my studentship.

I am indebted to my family and friends, especially my parents and Alex, for their unwavering support throughout the last few years.

Abstract

Standard marginal structural models (MSMs) are commonly applied to estimate causal effects in the presence of time-dependent confounding; these may be extended to history-adjusted MSMs to estimate effects conditional on time-updated covariates, and dynamic MSMs to estimate effects of pre-specified dynamic regimes (Cain et al., 2010). We address methods to assess the optimal time for treatment initiation with respect to CD4 count in HIV-infected persons, and apply these to CASCADE cohort data. We advocate the application of all three types of MSM to address such causal questions and investigate gaps in the literature concerning their application.

Of importance is the construction of suitable inverse probability weights. We have structured this process as four key decisions, defining a range of strategies; all demonstrated a beneficial effect of ART in CASCADE. We found a trend towards greater treatment benefit at lower CD4 across a range of models.

Via large simulated randomised trials based on CASCADE data, longer grace periods (permitted delay in treatment initiation) and in particular less-frequently observed CD4 indicated higher optimal regimes (earlier treatment initiation at higher CD4), although similar AIDS-free survival rates may be achieved at these higher optimal regimes. In realistically-sized observational simulations, the optimal regime estimates lacked precision, mainly due to broadly constant AIDS-free survival rates at higher CD4. Optimal regimes estimated from dynamic MSMs should be interpreted with regard to the shape of the outcome-by-regime curve and the precision. In our clinical setting, we found that allowing a 3-month grace period may increase precision with little bias under the interpretation of no grace period; under longer grace periods, the bias outweighed the efficiency gain. In our CASCADE population, immediate treatment was preferable to delay, although estimation was limited by relatively short follow-up. Comparison across the MSM approaches offers additional insights into the methodology and clinical results.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The motivation for this thesis lies in the application of marginal structural models and their extensions to estimate optimal dynamic treatment regimes. The clinical motivation arises from the field of HIV infection, namely the contentious question of when to initiate treatment in HIV-infected persons. We begin by introducing the concept and estimation methods of causality, followed by a definition of dynamic treatment regimes and an outline of the methods for their optimisation. We give an overview of the treatment of HIV-infection (section 1.5) and an introduction to the CASCADE data which are used throughout the thesis (section 1.6). Finally, we provide an outline for the rest of the thesis (section 1.7).

1.1 Causality

The causal effect of an intervention on an individual is defined as the difference in the outcome of interest under that intervention compared to the outcome in the absence of the intervention (Rubin, 1974). Our interest lies in the receipt of a treatment compared to no treatment, but "intervention" could, for example, also refer to other medical procedures or environmental exposures, and may be compared to standard practice or a control. For example, for an individual i, let $A_i = 1$ if the individual receives a particular treatment and $A_i = 0$ otherwise, and let $Y_i(A_i)$ represent some outcome of interest under treatment A_i. Then we may be interested in for example the causal effect $Y_i(1) - Y_i(0)$. Clearly, it is not possible to observe both these outcomes in the same individual, and so they are referred to as "potential outcomes". These concepts were introduced by Neyman et al. (1923) for randomised experiments, developed by Rubin (1974) for non-randomised studies, and later formalised and referred to as Rubin's Causal Model (Holland, 1986). If there are no (classically) missing data, then for each subject one potential outcome will be observed, while the other remains counterfactual, and so it is clearly not possible to

calculate the causal effect in one person. In general, interest lies in the average causal effects in a population, and looking at average causal effects (hereafter, simply causal effects) allows us statistically to overcome the issue of counterfactuals (Rubin, 1974). Note that counterfactual variables can be considered a form of missing data and the methods can be applied similarly (see section 1.2.6).

In general, treatment may be initiated or stopped over time and so $A_i(t)$ may be time-dependent. We describe different patterns of treatment as "treatment regimes". For example, $\overline{A}_i(t) := \{A_i(0), ..., A_i(t)\} = \{0, ..., 0\}$ indicates the regime of no treatment up until time t, $\overline{A}_i(t) = \{1, ..., 1\}$ indicates immediate and continuous treatment to time t, whereas $\overline{A}_i(t) = \{0, ..., 0, 1, ..., 1\}$ represents treatment initiation at some intermediate time s, $0 < s < t$ (and continuous thereafter to time t). From here on, we assume that patients are a random sample from a large population with a common distribution and hence drop the subscript i for subject.

In a randomised controlled trial (RCT), the balance created by randomisation means that we can simply compare the average outcomes in those randomised to receive treatment compared to those not, in a standard intention-to-treat (ITT) analysis. In the presence of non-compliance to randomised regime, an ITT analysis will still provide an unbiased estimate of effectiveness (the expected effect of the randomised strategy in an equivalent population of compliers and non-compliers), but may be biased for efficacy (the effect in those persons who would follow exactly the randomised treatment regime). Further, while analysis by ITT is generally seen as conservative, this is not true for trials in which the outcome is safety or which aim to demonstrate equivalence (Toh et al., 2010).

In the absence of evidence from an RCT we may turn to observational studies, but these are prone to confounding. That is, there may exist variables which are simultaneously predictors of (future) treatment and risk factors for the outcome of interest. We could use standard methods such as a suitable model for the outcome of interest with adjustment for the confounders, but if there exist time-dependent confounders $L(t)$ which are predicted by past treatment, then standard methods will be biased for the estimation of causal effects (Hernán et al., 2005). This is sometimes referred to as confounding by intermediate variables, since the covariates lie on the causal pathway between treatment and outcome (Figure 1.1; Robins (1989a)).

For the estimation of efficacy, either from an RCT which suffers from non-compliance, or an observational study in which there exist time-dependent confounders which are predicted by treatment history, causal methods are required (Hernán et al., 2006).

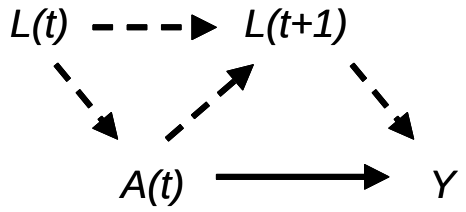


Figure 1.1: The time-dependent confounder $L(t+1)$ lies on the causal pathway between treatment $A(t)$ and the outcome Y.

1.2 Methods for estimating causal effects

The field of causal inference has advanced enormously in the last couple of decades. Robins (1986) first introduced the g-computation algorithm estimator to compare the causal effects of different treatment regimes on the time to an event of interest in an observational setting, and extensions of these methods led to the development of g-estimation of structural nested models (SNMs, section 1.2.2). The prefix "g" stands for "generalised" and is used to indicate methods which permit unbiased estimation of any form of hypothetical intervention, even in the presence of time-dependent confounders which are affected by previous treatment. The term "structural" arose from the disciplines of economics and social sciences but is synonymous with "causal". Robins (1998) was also responsible for the development of another methodology for investigating causal effects: estimation of marginal structural models (MSMs) using inverse probability of treatment weighting (section 1.2.3). Below, we outline and briefly compare these three approaches. Non-parametrically, these methods will lead to identical results (Daniel et al., 2011), but in realistic scenarios with a number of time-points and/or more complex treatment regimes, parametric methods are required.

1.2.1 G-computation formula

G-computation originated in an observational setting to compare the causal effects of different treatment regimes on the time to an event of interest (Robins, 1986). Let $T_{\overline{A}}$ represent a potentially counterfactual time-to-event outcome under a treatment regime $\overline{A}$ for a given patient. Then the g-null hypothesis of no effect of treatment on the time to the event of interest is given by:

$$\Pr(T_{\overline{A}_1} > t) = \Pr(T_{\overline{A}_2} > t) \text{ for all treatment regimes } \overline{A}_1 \text{ and } \overline{A}_2.$$

The g-computation formula expresses $\Pr(T_{\overline{A}} > t)$ in terms of the conditional probabilities of the event given treatment and covariate history, and of the current covariates given treatment

and covariate history (Lok et al., 2004). For example, consider a study with clinic visits at times $t = 0, 1, 2, ...$ where $t = 0$ represents baseline (time of entry into study). For each patient, at each time-point t, a covariate vector $L(t)$ is measured and treatment $A(t)$ is determined, and an overall time T to an event of interest is observed (in the absence of any censoring). As above, let overbars represent history, and let lower case letters represent realisations of random variables. Then, under certain assumptions, the g-computation formula is given by (Lok et al., 2004):

$$\Pr(T_{\overline{A}} > t+1) = \sum_{l_0} ... \sum_{l_t} \Bigg[\Pr\left(T > t+1 | \overline{L}(t) = \overline{l}(t), \overline{A}(t) = \overline{a}(t), T > t\right) \times \prod_{k=0}^{t} \left\{ \begin{array}{c} \Pr\left(T > k | \overline{L}(k-1) = \overline{l}(k-1), \overline{A}(k-1) = \overline{a}(k-1), T > k-1\right) \\ \times \Pr\left(L(k) = l(k) | \overline{L}(k-1) = \overline{l}(k-1), \overline{A}(k-1) = \overline{a}(k-1), T > k\right) \end{array} \right\} \Bigg]$$

where the summation over $l_0, ..., l_t$ is over all possible values $\overline{l}$ of the covariate history. For continuous $L(t)$, this summation is replaced with an integral, as in Daniel et al. (2011). This equation is sometimes equivalently referred to simply as the g-formula (Daniel et al., 2011; Taubman et al., 2009). Therefore, the g-formula expresses $\Pr(T_{\overline{A}} > t+1)$ in terms of:

$$\Pr\left(T > t+1 | \overline{L}(t) = \overline{l}(t), \overline{A}(t) = \overline{a}(t), T > t\right)$$

which is the probability of remaining event-free beyond time $t+1$, given covariate and treatment history to time t and remaining event-free to time t;

$$\Pr\left(T > k | \overline{L}(k-1) = \overline{l}(k-1), \overline{A}(k-1) = \overline{a}(k-1), T > k-1\right)$$

which for $k = 0, ..., t$ is the probability of remaining event-free beyond time k, given covariate and treatment history to time $k-1$ and remaining event-free to time $k-1$; and

$$\Pr\left(L(k) = l(k) | \overline{L}(k-1) = \overline{l}(k-1), \overline{A}(k-1) = \overline{a}(k-1), T > k\right)$$

which is the probability of the covariates $L(k) = l(k)$ measured at time k, given covariate and treatment history to time $k-1$ and remaining event-free to time k. As Daniel et al. (2011) outline, the g-formula is the appropriate generalisation of standardisation (estimation of expected outcome in a population under a hypothetical time-independent intervention, given

time-independent covariates) to a scenario with time-dependent covariates and treatment.

Robins (1986) developed an algorithm to aid the computation of this formula which requires knowledge, or estimation from the data, of the conditional distributions, such as implemented by Taubman et al. (2009) and Young et al. (2011). Briefly, there are three main steps which must be applied for each of the treatment regimes under consideration. The first step is to use the data to estimate the parameters of the conditional distributions of (*a*) *each* of the current covariates, and (*b*) the outcome, given covariate and treatment history. Secondly, Monte Carlo simulation is used to simulate a cohort based on the estimated distributions and under the given treatment regime. In the simple example of initiating treatment immediately, this would mean setting the treatment indicator variable(s) in the conditional distribution models equal to 1 for all time, and similarly equal to 0 for the scenario of never initiating treatment. Thirdly, the simulated cohort is used to estimate the outcome, which can be interpreted as an estimate for the outcome under that specific treatment regime. Once this is repeated for each treatment regime, these estimates can be compared across regimes.

A disadvantage of the g-formula is the number of parametric assumptions required and hence increased risk of bias. In addition, this approach may suffer from the "g-null paradox", whereby under certain situations and given enough data, the null hypothesis (of no effect of treatment for example) will be rejected even when true. This is discussed further by Daniel et al. (2011) and Robins et al. (1999).

1.2.2 G-estimation of structural nested models

To address the limitations of the g-computation formula, Robins (1989b) developed semi-parametric accelerated failure time (AFT) structural nested models (SNMs), which directly model the causal effect of treatment received at a given time on subsequent outcome, given treatment and covariate history (Robins, 1994). Lok et al. (2004) showed that AFT SNMs may be considered as a reparameterisation of the g-computation formula and estimated using maximum likelihood estimation, but this is not straightforward and cannot be computed easily using standard software (see also Walker et al. (2004)). Alternatively, the parameters of AFT SNMs can be estimated using a technique called g-estimation, which controls for confounding by intermediate variables. Conceptually, for each time, the procedure estimates the association between the treatment at that time and the counterfactual underlying true but unknown time to event under no treatment, after adjusting for treatment and covariate history, but without adjusting for subsequent treatment and covariate values (Robins et al., 1992). It does not

consider identical treatments received at different times (that is, with different treatment and covariate histories) to be the same, since the time-varying confounding means that these are not comparable (Robins et al., 1992).

Consider just the single parameter case. In the absence of censoring, each subject's observed time to event T under observed treatment $\overline{A}(T) = \{A(0), ..., A(T)\}$ may be related to the potentially counterfactual event time T_0 which would have been observed had the subject never received treatment, using:

$$T_0 = \int_0^T \exp\{\psi A(t)\}\, dt \tag{1.1}$$

(Robins and Tsiatis, 1991). In this equation, $\exp\{\psi\}$ is the factor by which time is "stretched" when on treatment compared to not. For example, if ψ is estimated as $-\log(2)$, then for a patient who initiates treatment immediately, their time to event is doubled compared to that which would have been observed had they remained off treatment for all time.

The parameters of this AFT SNM can be estimated using g-estimation as follows. For a chosen estimate $\widetilde{\psi}$ of ψ, it is possible to calculate $T_0(\widetilde{\psi})$ from the observed data $\{T, \overline{A}(T)\}$ using 1.1 and inserting the chosen $\widetilde{\psi}$ for ψ. A "g-test" is constructed and applied to test the hypothesis that $\widetilde{\psi}$ is equal to the true value ψ. The g-estimate $\widehat{\psi}$ of ψ is that for which the g-test has p-value equal (or closest) to 1.

In an RCT, by the nature of randomisation, at the true value of ψ the randomised treatment is independent of T_0, and a test of this hypothesis constitutes the g-test (that is, the randomised group is the intermediate variable). Therefore g-estimation is able to directly exploit the randomisation, and in such circumstances these methods are known as "randomisation-respecting" or "randomisation-based" and preserve the ITT p-value (White et al., 1999).

In an observational study, one approach would be to formulate a model for treatment, given treatment and covariate history, which incorporates T_0. For example, consider a study with clinic visits at times $t = 0, 1, 2, ...$ in which we are interested in the causal effects of a treatment which once initiated is continued. A possible model for treatment initiation, given treatment and covariate history, and incorporating $T_0(\widetilde{\psi})$, might be:

$$\text{logit}\ \Pr\Big\{A(t) = 1 \mid A(t-1) = 0, L(t), T_0(\widetilde{\psi}), T > t\Big\} = \beta(t) + \gamma L(t) + \theta T_0(\widetilde{\psi})$$

where $\text{logit}(p) = \log\Big(\frac{p}{1-p}\Big)$ and $\beta(t)$, γ and θ are unknown parameter vectors. Under the assumption of no unmeasured confounders, the treatment received at a given time t is independent of T_0 at the true value of ψ, given the treatment and covariate history. Therefore, a test for

$\theta = 0$ corresponds to a test for the independence of $A(t)$ and T_0 given treatment and covariate history, and this forms the g-test (Witteman et al., 1998).

Test-based $(1-\alpha)\%$ confidence intervals for ψ can be found based on those values of $\widetilde{\psi}$ for which the g-test fails to reject at the $\alpha\%$ level. In practice, a simple way to find the g-estimate and associated confidence interval is to perform a grid or interval search (White et al., 1999).

Under certain assumptions (see section 1.2.4), g-estimates are unbiased under the null hypothesis of no effect of treatment, but are only valid under administrative censoring and even then require an additional step of artificial re-censoring (Robins and Tsiatis, 1991). If a patient is censored at a time C, then for a given $\widetilde{\psi}$, $T_0(\widetilde{\psi})$ is censored at:

$$R_0 = \int_0^C \exp\left\{\widetilde{\psi} A(t)\right\} dt$$

which is a function of $A(t)$ and therefore may depend on the underlying prognosis of the patient. As White et al. (1999) explain, even if censoring on the T_0-scale is non-informative, it may be informative on the R_0-scale. For example, patients with the same T_0 are more likely to be censored the more treatment they receive, assuming treatment is beneficial. This problem can be addressed by artificially re-censoring $T_0(\widetilde{\psi})$ by a function of $T_0(\widetilde{\psi})$ and C which is observed for all patients. White et al. (1999) provide an example of this for an RCT and Witteman et al. (1998) outline an example in an observational setting.

SNMs have been applied to repeated measure outcomes (Robins, 1994), but the extension to Cox proportional hazards (PH) models has been limited. Greenland et al. (2008) attempted to interpret their results from an AFT SNM in terms of hazards, by assuming that the underlying event (hazard) rate was constant given baseline covariates, so that the inverse rate (hazard) ratios were equal to the event time ratios. White et al. (1999) attempted to translate from AFT modelling into a PH interpretation by constructing artificial datasets based on the parameters from the AFT model that would have been observed under a desired treatment scenario. They used these data to estimate "corrected" hazard ratios, though the properties of such estimators "are unclear" (Loeys et al., 2005). Cox PH SNMs have been constructed directly in a randomisation-based setting, firstly for all-or-nothing treatment (Loeys and Goetghebeur, 2003), and then for the more general case of time-constant (but could be categorical or continuous, perhaps time-averaged) treatment (Loeys et al., 2005), but the estimation of these models is not straightforward. While the AFT SNMs construct potential survival times under no treatment for each patient, the approach of Loeys et al. (2005) uses a PH model to relate

the treatment-regime-specific observed survival time in the treatment arm to the counterfactual survival time which would have been observed had that patient (counter to fact) been randomised to no treatment. The authors assumed that patients randomised to the no treatment arm cannot access the treatment. Although the estimation of Cox PH SNMs is challenging, an advantage is that artificial re-censoring is not required as for the AFT models. Further, hazard ratios are more commonly used and understood in practice. However, as far as we are aware, the PH methods have not been developed for time-dependent treatment and so far have not been applied to allow for treatment changes in both treatment and control arms; the extension to observational studies is not trivial.

1.2.3 Estimation of marginal structural models using inverse probability of treatment weighting

MSMs model the marginal distributions of potential outcomes relating to different treatment histories, rather than modelling the joint distribution of such variables (Fewell et al., 2004). That is, they directly model the outcomes that would have been observed had all patients been subject to the same treatment history (Robins and Tsiatis, 1991). The beauty of MSMs is that they are natural extensions of standard methods and can relatively easily be applied to any outcome of interest, leading to an explosion in the application of MSMs in the last decade. For example, Toh et al. (2010) applied MSMs with a survival outcome to estimate the causal effect of postmenopausal hormone therapy on the risk of invasive breast cancer, Cole et al. (2005) used repeated measures MSMs to estimate the effect of treatment on the biomarker CD4 count in HIV-infected persons and Bodnar et al. (2004) applied logistic MSMs to look at the effect of iron supplements during pregnancy on the odds of anaemia at delivery, to name but a few.

We will be interested in a time-to-event outcome and hence the estimation of Cox PH MSMs. As above, let $T_{\overline{A}}$ be the potentially counterfactual time to event under a treatment regime $\overline{A} = \{A(0), A(1), ...\}$ and let V be a vector of baseline covariates where $V \subseteq L(0)$. Then for each possible $\overline{A}$, a Cox PH MSM is given by:

$$\lambda_{T_{\overline{A}}}\{t|A(t), V\} = \lambda_0(t) \exp\{\alpha A(t) + \beta V\}$$

where $\lambda_0(t)$ is the baseline hazard, α and β are unknown parameters and $\exp\{\alpha\}$ can be interpreted causally as the hazard ratio of the outcome of treatment versus no treatment at time t, given V (Hernán et al., 2000). (Note that other specifications exist, for example incorporating functions of $\overline{A}(t)$ rather than just the treatment received at time t.) Since at least some of these

outcomes will remain unobserved, it is not possible to fit this model directly. However, the parameters of MSMs can be consistently estimated (under certain assumptions, see section 1.2.4) using inverse probability of treatment weighting. Briefly, for a Cox PH MSM, each individual still in the risk set at each event time is weighted by an estimate of the inverse probability of the observed treatment received by that person at that time, given their observed treatment and covariate history. This weighting addresses the bias due to time-dependent confounding of intermediate variables and will be discussed further in chapter 2.

1.2.4 Assumptions

In any study, whether randomised or observational, measurement error may be present and there exist methods to address this. However, here we assume that the data are measured without error (or minimally). We also assume that there is no interaction between patients (known at "SUTVA", the stable unit treatment value assumption; Little and Rubin (2000)) and that any missing data are missing at random (Hernán and Robins, 2006). For causal inference in observational studies, we also require the following assumptions (Cole and Hernán, 2008):

- Consistency: this states that the potentially counterfactual outcome under a particular treatment regime is equal to the observed outcome if the individual was observed to follow that regime.
- No unmeasured confounders between treatment and the outcome (otherwise known as exchangeability).
- No misspecification of the models.

Further, MSMs require the assumption of positivity (or the experimental treatment assignment assumption), that is that there is a non-zero probability of receiving each treatment regime for all combinations of covariate and treatment history. This is discussed further in section 2.2.2.

With time-to-event data, right-censoring is common. If information on prognostic factors for censoring is available, then censoring-weighted estimators can be used to correct for the potential bias due to this censoring under the (untestable) assumption that there is no residual confounding (Robins and Finkelstein, 2000). This will be addressed further in section 2.2.4. The assumption of no unmeasured confounders between outcome and censoring can be explored via sensitivity analyses by considering the potential effects of an imaginary unmeasured confounder; Scharfstein et al. (2001) developed such methods for discrete time and Scharfstein and Robins (2002) extended these methods to allow for continuous time. More recently, the methods of

Rotnitzky et al. (2007) allow for competing censoring mechanisms by introducing a "censoring bias function". However, we assume that the available data are sufficient to describe the censoring processes and these methods are not addressed further in this thesis.

1.2.5 Comparison of methods

A considerable advantage of the use of MSMs to estimate causal effects, using inverse probability of treatment weighting, is their resemblance to standard models and hence relative ease of implementation. A potential difficulty associated with the application of MSMs is the need for positivity. If the data are "close" to non-positivity (for example, if at some levels of the covariates, treatment is nearly always given), then large weights may arise (Cole and Hernán, 2008). Similarly, if there are many time-points or treatment is strongly correlated with baseline covariates, then the weights may become large. These problems may be attenuated to some extent by stabilisation of the weights (see section 2.2.3), truncation of the very largest weights (Cole and Hernán, 2008), or addressed using doubly robust estimators (Bang and Robins, 2005). Doubly robust estimators are not discussed further in this thesis.

While the g-computation formula can be applied to highly complex pre-defined interventions (such as "avoid smoking, exercise at least 30 minutes daily *and* consume at least 5g of alcohol daily"; Taubman et al. (2009)), it is computationally intensive and best suited to a small number of interventions (Daniel et al., 2011). In addition, this approach is at risk of the g-null paradox (see section 1.2.1).

Although g-estimation of SNMs may benefit from greater efficiency than inverse probability treatment weighting of MSMs and fewer parametric assumptions than g-computation (Daniel et al., 2011), it is perhaps less robust to model misspecification and is not intuitive nor easy to apply. Further, if there is right-censoring of survival times, g-estimation of SNMs requires artificial re-censoring in order to break any dependency of the censoring time on treatment, which may be related to the underlying prognosis of the patient (section 1.2.2). In practice, other authors have found that this method may suffer from low power (White et al., 1999). Young et al. (2009) performed a simulation study to illustrate and compare MSMs versus SNMs and found that, compared to the g-estimators, the inverse probability weighted estimators were similarly or less biased, and were more efficient.

1.2.6 Relation to missing data problems

Counterfactual variables may be considered as a missing data problem; they are monotonely missing data. Inverse probability weighting methods have been developed and applied similarly in the missing data paradigm (Robins et al., 1995). Drawing on other methods from the field of missing data, one could perhaps consider implementing multiple imputation in a potential outcomes setting, where counterfactual outcomes are multiply imputed using the observed outcomes and measured confounders. Under certain scenarios, inverse probability weighting methods resemble those of multiple imputation, but these methods are not addressed further in this thesis.

1.2.7 Effect modification by baseline covariates

The methods described above can all easily be adapted to incorporate an interaction between treatment and a baseline covariate to investigate effect modification. A number of papers describe this for MSMs (Bodnar et al., 2004; Hernán et al., 2006; Robins et al., 2000), but, to our knowledge, it has rarely been applied in practice in the setting of antiretroviral therapy for HIV-infected persons (our clinical example, introduced in section 1.5). The only example we are aware of is a series of papers by Cole and colleagues (2007; 2005; 2003), looking at whether there is a differential effect of treatment by sex or CD4 count at study entry on a range of different outcomes. Loeys et al. (2005) outline how to adapt their causal PH SNM to allow for an interaction between treatment and a baseline covariate, but we are not aware of this having been applied in practice.

1.3 Dynamic treatment regimes

There are many situations in medical practice in which treatment decisions are made based on the current well-being of the patient, perhaps to minimise time spent on potentially toxic treatments or to optimise resources. For example, treatment may be given until a desired level of recovery is achieved, delayed until a certain stage of disease progression is reached, or given intermittently, perhaps based on some observed biomarker. Such treatment regimes which are in response to a patient's time-dependent measurements are known as "dynamic" (Hernán et al., 2006). Moodie et al. (2007) and Murphy (2003) view dynamic regimes as a function or list of decision rules, which are based on treatment and covariate history. These types of treatment regimes have also been referred to as "individualized treatment rules" (Petersen, Deeks, and van der Laan, 2007) or "adaptive strategies" (Murphy, 2003). Note that while treatment regimes

may change over time, they may not necessarily be dynamic; for example, "take drug X for Y weeks then drug Z" is an example of a time-varying but non-dynamic regime. Although clinical trials most often compare non-dynamic treatment regimes, dynamic treatment regimes may be more common in practice (Cain et al., 2010). Dynamic treatment regimes can be considered as interactions between treatment and time-dependent covariates (Hernán et al., 2002). We may naturally wish to identify optimal dynamic treatment regimes, defined by Petersen, Deeks, and van der Laan (2007) as "the treatment rule that produces, on average, the best patient outcome at a given time-point". Dawid and Didelez (2010) recommend approaching the assessment of strategies as a decision theory problem.

1.3.1 Classes of dynamic treatment regimes

Theoretically, dynamic treatment regimes may be a complex function of all covariate and treatment history; consider such a large class of regimes R. In practice, a more limited set of well-defined regimes may be preferable. For example, consider the simple question of when to initiate treatment for a chronic disease, where once treatment is initiated it is continued for life (for example, in the motivating clinical example of HIV infection introduced in section 1.5). In order to preserve time off treatment, treatment may be delayed until a certain disease stage is reached and if so then one could pre-specify a limited set of regimes, defined by treatment initiation dependent on different stages of disease. This pre-defined set of regimes R^* is a subset of the larger class R. It is possible to imagine the equivalent RCT which in theory could be conducted to determine the optimal choice in terms of an outcome of interest from this pre-defined set of regimes R^*: patients would be enrolled at some starting point, perhaps onset of the disease, and then randomised to one of the regimes in the set R^*. Comparison of the outcome across the patients would inform the optimal regime from this set R^*. Such a pre-defined set may be of most use to inform policy makers.

In contrast, rather than pre-defining a limited set of regimes in advance, consider the presentation of a patient to clinic, where the natural question arising is whether to initiate treatment at that time, or delay. This decision may be based on the covariate and treatment history of that patient to that time-point. The RCT comparison in this situation would be based on a series of randomisations at each successive clinic visit, to immediately initiate or defer treatment. Van der Laan and Petersen (2004) refer to the optimisation of such regimes as estimating "optimal history-adjusted static treatment regimes" or "statically-optimal dynamic treatment regimes"; we will use the former nomenclature. Following the optimal history-adjusted static

regime over time maps to a specific type of optimal dynamic treatment regime, defined by following at each time-point the first action of the optimal history-adjusted static treatment regime. It may be argued that optimisation of such regimes would be of most interest to clinicians who wish to determine the best immediate course of action for a patient who has presented to clinic.

1.4 Methods for estimating optimal dynamic treatment regimes

1.4.1 G-computation formula

The g-computation formula may easily be applied to optimise dynamic treatment regimes (Taubman et al., 2009; Young et al., 2011). The method as described in section 1.2.1 can be directly extended at the second stage to incorporate dynamic treatment regimes, and the optimal treatment regime is identified as that with the best outcome across all the regimes.

1.4.2 G-estimation of structural nested models

SNMs can easily be extended to handle simple dynamic treatment regimes by incorporating interactions (Hernán et al., 2006). For example, Hernán et al. (2005) discuss the extension to a two-parameter model in an observational study for evaluating how the effect of treatment received at a given time is modified by a time-dependent covariate. We are aware of only one such application in practice: White et al. (1999) grouped HIV-infected persons by their CD4 count at treatment initiation ($\leq$ or $>$ 350 cells/mm^3) and used a bivariate model to obtain separate estimates for the effect of treatment by whether it was initiated at low or high CD4 count, and thus providing (albeit limited) information of the timing of treatment initiation in such patients (see further detail on this clinical example in section 1.5). Each parameter requires a separate test; the authors used logrank and Gehan-Wilcoxon (or Breslow) tests (Breslow, 1970; Gehan, 1965). However, the authors found this method was not robust and suffered from a lack of power. Murphy (2003) and Robins (2004) developed semi-parametric methods for structural nested mean models for optimisation of more complex dynamic treatment regimes, which have been compared and reconciled by Moodie et al. (2007). Rosthoj et al. (2006) applied the methods of Murphy (2003) to investigate dosing strategies for patients on anticoagulant treatments, and discussed problems met in their implementation.

1.4.3 Estimation of marginal structural models using inverse probability of treatment weighting

MSMs have been deemed "less useful" for estimation of causal effects of dynamic treatment regimes since they are not directly applicable to such questions (Hernán et al., 2002, 2006). Two extensions to standard MSMs have been suggested: history-adjusted MSMs (HAMSMs) for the estimation of optimal history-adjusted static treatment regimes (Van der Laan et al. (2005); Petersen, Deeks, Martin, and van der Laan (2007)) and dynamic MSMs for the optimisation of pre-defined dynamic treatment regimes (Cain et al., 2010; Hernán et al., 2006; Robins et al., 2008).

HAMSMs can be viewed as a series of "trials" at each time in the visit schedule, where the aim at each new "baseline" visit is to optimise subsequent outcome. In its most basic form, a history-adjusted model may just estimate the causal effect of the treatment received at "baseline" and adjust for the "baseline" covariates (Writing Committee for the CASCADE Collaboration, 2011), but may also look at subsequent treatment received during each "trial", with adjustment for subsequent time-dependent confounders using inverse probability of treatment weights as for a standard MSM. That is, a standard MSM is assumed at each time-point (Petersen, Deeks, Martin, and van der Laan, 2007), and a common model is formulated which considers each time-point in turn, in a static way, resulting in regimes in terms of treatment at each time-point with respect to time-dependent covariates. A potential criticism of these methods is that while the models need to be sufficiently flexible to allow time-dependent treatment effects, this could result in incompatibilities and implausible conclusions (Robins et al., 2007). Further details on these methods are given in chapter 3.

Dynamic MSMs are another extension of standard MSMs for simple but perhaps more pragmatic dynamic regime classes. In their most basic form, they depend upon the availability of a suitable time-dependent covariate upon which the dynamic treatment regimes may be defined in advance. Hernán et al. (2006) introduced these methods for just two dynamic regimes and they have since been extended to many regimes (Cain et al., 2010; Robins et al., 2008). The key idea behind this approach is that all patients are considered to follow all of the pre-defined treatment regimes initially and are censored from each regime if they become noncompliant. Of course, this artificial censoring process is likely to be informative but inverse probability weighting can be used to address this. Since the censoring process will depend entirely on treatment and the time-dependent covariate by which the dynamic regimes are defined, the weights required are the inverse probability of treatment weights as employed by the standard

MSMs. These methods will be further discussed in chapter 4.

1.4.4 Comparison of methods

While the methods of Murphy (2003) and Robins (2004) using SNMs and the approach of Petersen, Deeks, Martin, and van der Laan (2007) with HAMSMs are useful for optimising complex treatment regimes, such as intermittent treatment dependent on a number of factors, this can be a disadvantage if simpler treatment regimes are desired which may perhaps be more readily translated into clinical practice. Thus, the g-computation formula and dynamic MSMs may be more useful as they can estimate the causal effects of pre-specified treatment regimes, from the set R^* (section 1.3.1). In theory, SNMs may be used to estimate potentially large classes of dynamic treatment regimes from the larger set R, however in practice this is typically restricted by the number of interactions which can be included given the available data.

The SNMs discussed above require a correct model but use all the available data (except if re-censoring is required in the presence of right-censored data). Conversely, the censoring required under the dynamic MSMs means that data after artificial censoring is discarded, therefore this approach may be less efficient, but does not impose a structural model for the effect of treatment across regimes. This is the usual bias-variance trade-off frequently encountered in statistical modelling (Hernán et al., 2006).

1.4.5 Consistency across marginal structural models

Further to the estimation of treatment effects using a standard MSM, under the assumption of constant treatment effect regardless of the time on treatment, it is possible to estimate the cumulative effects of having received immediate and continuous versus no treatment. For example, if we are interested in a time-to-event outcome, then it would be possible to estimate the event-free survival under immediate versus no treatment initiation. These treatment regimes of immediate versus no treatment could also be incorporated into the set of regimes considered under a dynamic MSM and we might expect the results to be consistent with those from the standard MSM. Similarly, a basic HAMSM considering treatment initiation or deferral given "baseline" covariates at successive "trials" will yield information on the benefit of treatment at different values of the "baseline" covariates. If these covariates are also those by which the dynamic regimes are defined, then we might expect consistency in the conclusions drawn from the history-adjusted and dynamic MSMs.

1.5 Treatment of HIV-infection

We now introduce the clinical topic of interest throughout the thesis.

Thanks to the rapid development of a range of antiretroviral therapies, HIV has been transformed from a disease with poor prognosis to a manageable condition with much improved long-term survival (Ewings et al., 2008). However, successful treatment requires a number of concurrent drugs which may have a variety of side effects. Further, long-term treatment may result in the development of drug resistance. For these reasons, initiation of treatment is often delayed until some immunodeficiency is evident, but there are arguments for starting treatment earlier to potentially minimise the long-term damage of HIV. A key biomarker used to monitor the degree of immunosuppression in HIV-infected persons is CD4 cell count, which typically declines after infection and low levels predict poor prognosis. Previous guidelines in the UK recommended initiating treatment around CD4 counts of 200 cells/mm^3, while current guidelines in the UK recommend initiating treatment around 350 cells/mm^3 (Gazzard and on behalf of the BHIVA Treatment Guidelines Writing Group, 2008), but there may be benefits of initiating still earlier at higher CD4 counts. For example, a recent study in more than 40,000 treatment-naïve persons with high CD4 counts ($\geq$ 350 cells/mm^3) found that the mortality rate was higher than that in the general population, and was greatest at lower CD4 counts within this range, therefore offering support for further exploration of treatment in such persons (Study Group on Death Rates at High CD4 Count in Antiretroviral Naive Patients, 2010). Recently, there have been a number of observational studies investigating the question of when to start treatment in patients with HIV infection, but the findings have left experts divided.

A large study in approximately 17,500 persons found that in those with CD4 counts in the ranges $351 - 500$ or > 500 cells/mm^3, immediate treatment initiation was associated with a reduction in the risk of death compared to delaying treatment (Kitahata et al., 2009). Subsequent to these findings, the US guidelines were amended at the end of 2009 to recommend earlier treatment initiation, though the panel members were not able to reach agreement regarding initiation of treatment at CD4 counts > 500 cells/mm^3 (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2009). However, there were a number of concerns raised about this study relating to potential sources of bias (Arribas et al., 2009; Buchbinder and Jain, 2009; Hernán and Robins, 2009).

A different approach relying on historical data was used by the When to Start Consortium (2009). In contrast, they found that deferring treatment until CD4 count was in the range $251 - 350$ cells/mm^3 was associated with a higher risk of AIDS or death than initiating when

CD4 count was in the range $351 - 450$ cells/mm^3 but did not see a benefit of earlier treatment initiation with CD4 count > 450 cells/mm^3.

While a separate study found that treatment initiation at CD4 counts < 500 cells/mm^3 was beneficial, the authors cautioned that due to the low absolute AIDS and death rate at CD4 counts ≥ 350 cells/mm^3, the benefits of treatment should be balanced against the implications of long-term therapy, such as side-effects and the risk of developing drug resistance (Writing Committee for the CASCADE Collaboration, 2011). Of note, this study used data from the CASCADE collaboration; we will be using a subset of these data (see section 1.6).

More recently, the HIV-CAUSAL Collaboration (2010) applied the methods of Cain et al. (2010) and found that treatment initiation when CD4 counts were around 500 cells/mm^3 improved AIDS-free survival compared to waiting until CD4 counts dropped lower, but mortality rates did not vary greatly when treatment was initiated > 300 cells/mm^3.

Aside from AIDS-defining illnesses, the implications of other serious adverse events have more recently been recognised. For example, Lichtenstein et al. (2010) found higher risk of cardiovascular disease at low CD4 counts, therefore raising the question of whether early treatment may help reduce the risk of events other than those traditionally associated with HIV infection.

A large international randomised controlled trial (START; INSIGHT (2009)) is currently underway to determine whether immediate initiation of treatment in patients with CD4 counts ≥ 500 cells/mm^3 is superior to deferral of treatment initiation until CD4 count drops to < 350 cells/mm^3 in terms of mortality and HIV- and non-HIV-related morbidity, but this will not be completed until 2016. Therefore our interest lies in the question of when to initiate treatment with respect to CD4 count in HIV-infected individuals, which is an example of a dynamic treatment regime. Further, while the START trial compares just two regimes, in practice some intermediary regime may be preferable and indeed the application of observational methods to this problem may suggest other possible regimes for consideration in future trials, or potentially yield additional information worth further exploration such as effect modification by baseline covariates.

1.6 CASCADE

Throughout the thesis, we used data from CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe), an ongoing collaboration of cohorts of HIV-infected persons with well-estimated dates of infection (CASCADE Collaboration, 2009). CASCADE annually pools participant data, including information on demographics, vital status, AIDS events, treatment

use, and CD4 cell count and HIV RNA measurements. Data collection and follow-up varies across the cohorts, which are a mixture of interval and clinical cohorts (Lau et al., 2007). For each participant, the HIV seroconversion date was estimated as the date of laboratory evidence of seroconversion, if available, otherwise as the midpoint of the last negative and first positive HIV antibody tests, no more than 3 years apart. The data used here were collated in July 2008 on 19,615 participants from 22 cohorts across 12 countries.

1.6.1 Data used for our analyses

Entry to our analysis

Participants were eligible to enter our analysis at their first CD4 count ≥ 500 cells/mm^3 at least 1 year but no more than 5 years after seroconversion and after 1 January 1996, provided still treatment-naïve and AIDS-free at this point (Figure 1.2). We refer to the time of entry to the study as baseline. These criteria are quite stringent but necessary to capture the population of interest and avoid bias, for the reasons as follows. Firstly, the inclusion of participants from the time of a high CD4 cell count at least 1 year after seroconversion ensured that we captured participants at a "peak" CD4 cell count before the decline associated with long-term infection. This is the population in whom our question "when to start treatment" has meaning, since patients with low CD4 counts shortly after seroconversion do not have the opportunity to start treatment at high CD4 counts. While the initial methods we used (standard MSMs, section 1.2.3) could be applied without the rather stringent restriction of a first CD4 ≥ 500 cells/mm^3, thus enabling us to include a greater number of patients and from an earlier starting time, we wished to demonstrate a treatment effect in the subset of patients which are included in our later analyses (to answer the question of when to start, using dynamic regime MSMs), which in this case do require such restrictions if we desire all patients to initially be eligible for all regimes (although this was not enforced by all researchers; see discussion in chapter 5). Secondly, we excluded patients who had been infected for over 5 years at analysis entry since, in the absence of treatment, it is unusual for an individual to remain alive, AIDS-free and with high CD4 counts for over 5 years after infection (Lodi et al., 2011). These excluded patients were a mixture of (*i*) those who were enrolled late into the cohort with missing earlier CD4 counts and (*ii*) those who initially had CD4 counts < 500 cells/mm^3 with a blip to ≥ 500 cells/mm^3 at some time later during infection. The inclusion of type (*i*) patients may have led to survivorship bias, since these are a select group who have survived long enough to enter the cohort. Our question of "when to start treatment" is less applicable to type (*ii*) patients, for whom a different

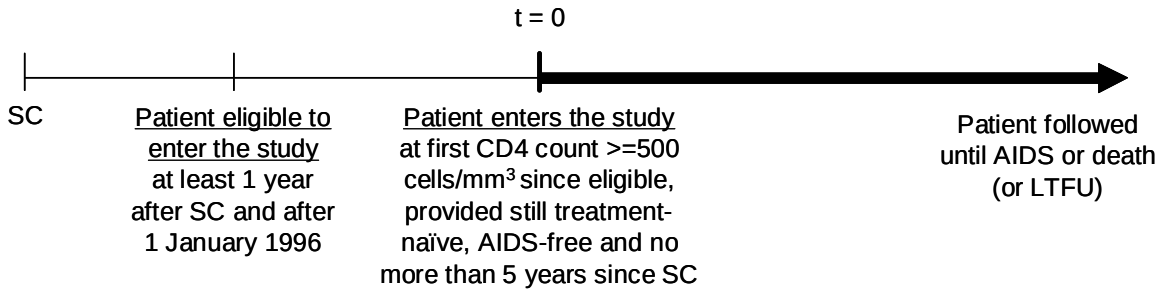


Figure 1.2: Timeline showing eligibility and entry of participants into the study. SC=seroconversion. LTFU=lost to follow-up.

treatment management strategy is likely to be preferable. Lastly, effective treatment was only available from 1 January 1996, therefore we only allowed entry after this time to avoid classic survivorship bias (only a select subset of participants with good outcome surviving long enough without treatment to enter the analysis). We further excluded participants aged <16 years old at HIV seroconversion and a relatively small number of people who initiated treatment with a suboptimal or unknown treatment regime (see definition below).

Treatment

Effective antiretroviral therapy (hereafter, ART) was defined as any regimen consisting of at least three antiretroviral drugs from at least two classes, or containing abacavir or tenofovir. We were interested in the initiation of ART, therefore we ignored subsequent treatment interruptions. This approach assumes that once a participant initiates treatment, they remain on it and so has an ITT "flavour" (Hernán et al., 2006). Therefore participants who have stopped ART will be counted as having started on treatment, and so potentially diluting estimated treatment effects as estimates of efficacy. However, treatment may be stopped for a number of reasons including for example due to toxicity, which is an inherent part of the chosen treatment path. Estimating the effect of "ever having started" (effectiveness) is an estimate of likely population level effect assuming that the cohort participants are representative of the wider population of interest in terms of patterns of treatment discontinuation. Further, only 6% of the total follow-up time post-treatment initiation was spent off ART.

Outcome

Our outcome of interest was time to first diagnosis of AIDS or death (CDC, 1992); reaching CD4 count < 200 cells/mm^3 was not considered an AIDS event. Data on serious non-AIDS

events were not available.

Baseline covariates

Baseline covariates included sex, age at and year of HIV seroconversion, route of HIV transmission, an indicator for short HIV test interval ($\leq$30 days between last negative and first positive HIV tests, or laboratory evidence of seroconversion) as a proxy for seroconversion illness (Tyrer et al., 2003), length of time HIV-infected at baseline (approximated by time from estimated seroconversion to study entry), baseline CD4 count and baseline HIV RNA (if available within -6 to $+1$ months relative to study entry date). Cohorts were grouped by country, and countries with less than 100 participants were combined. In the analyses, continuous baseline covariates were treated as linear, route of HIV transmission was categorised as injecting drug use (IDU) versus other, and a missing indicator for the availability of baseline HIV RNA was included.

Follow-up and censoring

We split time into one-monthly intervals from entry into this analysis (at the first CD4 count ≥ 500 cells/mm^3 at least one year but no more than five years after seroconversion and after 1 January 1996), given by $t = 1, 2, ...$ Follow-up ended when the patient progressed to AIDS or death, or was censored. We defined three types of censoring, with indicators $C_x(t)$, $x = 1, 2, 3$:

1. lost to follow-up (LTFU), which was defined as when a patient had no CD4 count in the 12 months prior to the last CD4 measurement within their cohort. Censoring occurred at the earliest of the patient's last known alive date or 12 months after their last CD4 count.

2. irregular CD4 counts, where there was a gap between measurements of over 12 months, with subsequent CD4 counts recorded. Censoring occurred at 12 months after the last CD4 count before the (first) gap.

3. administrative, which included all remaining patients who were alive, AIDS-free and not otherwise censored. We used the last alive date as the date of censoring.

Of note, there is a temporal ordering to these censorings: a patient had to remain in follow-up with regular CD4 counts in order to be administratively censored, and must not have been LTFU in order to be censored due to irregular CD4 counts. The reasoning behind censoring type (2) is that we require time-updated data in order to reliably estimate the weights via the treatment prediction model. Patients did not re-enter the risk set after a gap between CD4

count measurements of more than 12 months in CD4 count measurements because there was concern that the reasons for the gap may not be adequately captured by the available data.

Time-dependent covariates

Time-dependent covariates $L(t)$ were taken as the latest recorded strictly before time t; the notation will be formalised in section 2.2.1. This included the latest CD4 count provided within the last 12 months. By allowing a CD4 count to be valid for up to 12 months, we essentially carried the last observation forward for this time, so the CD4 count for 12 consecutive months was constant if there was no interim measurement. For modelling purposes, CD4 counts > 1000 cells/mm^3 were truncated to 1000 cells/mm^3 since the inherent variability at such high CD4 counts (which are within the normal range for HIV-uninfected adults) means there is little to distinguish such values biologically (Samet et al., 2001).

In addition to CD4 count, we considered a number of other time-varying covariates to be included in $L(t)$, broadly following Writing Committee for the CASCADE Collaboration (2011):

- CD4 count decrease from time $t-1$ (artificially zero if the last observation was carried forward due to no recent CD4 count)
- time in months since last CD4 count, defined as $t-$*date of last CD4 count*
- nadir CD4 count prior to time t
- number of previous CD4 counts prior to time t
- number of previous HIV RNA measurements prior to time t (with all the following variables set to zero if none)
- last HIV RNA (observation carried forward indefinitely if no subsequent measurements recorded)
- time in months since last HIV RNA
- peak HIV RNA prior to time t.

CD4 count decrease was categorised as large increase (> 100 cells/mm^3), small increase (≤ 100 cells/mm^3), no change, small decrease (≤ 100 cells/mm^3) or large decrease (> 100 cells/mm^3) since it was heavily weighted on zero due to no change when the last CD4 count value was carried forward. HIV RNA was categorised using the 10, 25, 50, 75 and 90^{th} percentiles (corresponding to ≤ 500, $> 500-2910$, $> 2910-11820$, $> 11820-37743$, $> 37743-97809$ and > 97809 copies/ml, respectively), with an additional category for no previous measurement.

1.6.2 Sample characteristics

Participants and baseline characteristics

Of the initial 19,615 CASCADE participants, 115 and 9 patients had estimated seroconversion date (as defined in section 1.6) after treatment initiation and progression to AIDS, respectively, 451 did not have any available CD4 counts and 976 had no CD4 counts before progression to AIDS. In order to meet our analysis entry criteria (outlined in section 1.6.1) of a CD4 count $\geq$ 500 cells/mm^3 at least 1 year but no more than 5 years after seroconversion and after 1 January 1996, the following participants were excluded:

- 55 and 2388 due to treatment initiation or AIDS, respectively, before 1 January 1996
- 601 due to no CD4 count available after 1 January 1996
- 4375 due to no CD4 count at least 1 year but no more than 5 years after seroconversion
- 6672 due to no (treatment-naïve) CD4 count $\geq$ 500 cells/mm^3 within the above window.

A further 6 patients aged $<$ 16 years at seroconversion, 539 who initiated suboptimal or unknown treatment (357 of whom initiated in 1996-97), and 46 who had less than one month of follow-up were excluded, leaving 3382 adults for our analysis. The numbers of participants within cohorts and respective countries are summarised in Table 1.1. The majority (55%) of patients were from French cohorts; the patients from the smallest cohorts in Australia, Canada, Denmark, the Netherlands and Norway were combined.

The median (interquartile range, IQR) age at seroconversion was 31 (26, 37) years, and the majority of participants were male (80%) and infected through sex between men (61%) (Table 1.2). The median (IQR) year of seroconversion was 2000 (1995, 2003) and time between seroconversion and entry to our analysis was 1.3 (1.1, 1.9) years. Only 8% of participants were identified as HIV-infected close to seroconversion. At baseline (entry to our analysis), the median (IQR) CD4 count was 641 (560, 788) cells/mm^3 and, of those who had a measurement available (2671, 79% of patients), the baseline HIV RNA was 4.1 (3.5, 4.7) $\log_{10}$ copies/ml and $\leq$ 500 copies/ml in 289 (11%) patients (a broadly similar rate of viraemic control as seen in previous CASCADE analyses of untreated patients; Madec et al. (2005)).

Country and cohort(s)	Type of cohort[1]	N patients (%)	Median (IQR) [range] follow-up, years	Median (IQR) months between CD4 counts	Month/year of last CD4 date[2]
France		1862 (55%)	2.2 (1.1, 4.1) [0.2, 12.0]	3.2 (2.2, 4.5)	
Aquitaine Cohort	Clinical	136			March 2008
French Hospital Database	Clinical	1569			April 2008
Lyon Primary Infection Cohort	Clinical	19			June 2008
PRIMO Cohort	Interval	129			August 2008
SEROCO Cohort	Interval	9			July 2008
Germany		107 (3%)	1.1 (0.7, 1.9) [0.2, 9.0]	3.1 (2.6, 4.0)	
German Cohort	Clinical	107			May 2008
Italy		194 (6%)	2.2 (1.1, 4.2) [0.3, 11.6]	4.1 (2.5, 6.1)	
Italian Seroconversion Study	Clinical	194			February 2008
Spain		276 (8%)	1.8 (1.1, 3.7) [0.2, 12.0]	4.1 (3.2, 5.6)	
Badalona IDU Hospital Cohort	Clinical	18			February 2008
Barcelona IDU Cohort	Clinical	36			February 2008
Madrid Cohort	Clinical	136			March 2008
Valencia IDU Cohort	Clinical	86			May 2008
Switzerland		143 (4%)	2.9 (1.7, 5.3) [0.3, 12.4]	3.5 (2.8, 5.8)	
Swiss HIV Cohort	Interval	143			July 2008
UK		571 (17%)	3.3 (1.4, 6.1) [0.2, 12.0]	3.2 (2.4, 4.4)	
UK Register of HIV Seroconverters	Clinical	571			May 2008
Others		229 (7%)	3.1 (1.3, 6.0) [0.2, 12.4]	3.4 (2.6, 4.6)	
Sydney Primary HIV Infection Cohort, Australia	Interval	12			September 2005
Southern Alberta Clinic, Canada	Clinical	53			July 2008
Copenhagen Cohort, Denmark	Clinical	45			December 2005
Amsterdam Cohort Study among drug users, Netherlands	Interval	15			December 2007
Amsterdam Cohort Study among homosexual men, Netherlands	Interval	27			June 2006
Oslo and Ulleval Hospital Cohorts, Norway	Clinical	77			June 2008

Table 1.1: Contributing cohorts. [1] Interval cohorts have scheduled visits at specified intervals; clinical cohorts collate data collected as part of routine follow up (Lau et al., 2007). [2] Used for defining loss to follow up (see section 2.4.1).

	France	Germany	Italy	Spain	Switzerland	UK	Others	Overall
	n=1862	n=107	n=194	n=276	n=143	n=571	n=229	n=3382
Sex, female	444 (24%)	9 (8%)	64 (33%)	49 (18%)	45 (31%)	33 (6%)	35 (15%)	679 (20%)
Age at seroconversion, years	31 (26, 38)	33 (28, 38)	31 (27, 36)	28 (24, 33)	33 (27, 40)	31 (26, 37)	33 (28, 39)	31 (26, 37) [16, 77]
Year of seroconversion	2000 (1996, 2003)	2004 (2003, 2005)	1996 (1994, 2000)	1997 (1994, 2001)	2001 (1997, 2004)	1998 (1995, 2001)	1998 (1995, 2002)	2000 (1995, 2003) [1991, 2007]
Route of HIV transmission								
Sex between men (MSM)	1070 (57%)	92 (86%)	69 (36%)	147 (53%)	53 (37%)	506 (89%)	140 (61%)	2077 (61%)
Sex between men & women (MSW)	605 (32%)	10 (9%)	59 (30%)	20 (7%)	54 (38%)	43 (8%)	43 (19%)	834 (25%)
Injecting drug use (IDU)	80 (4%)	1 (1%)	60 (31%)	101 (37%)	23 (16%)	14 (2%)	42 (18%)	321 (9%)
Other/unknown	107 (6%)	4 (4%)	6 (3%)	8 (3%)	13 (9%)	8 (1%)	4 (2%)	150 (4%)
Identified close to seroconversion[1]	161 (9%)	39 (37%)	8 (4%)	3 (1%)	4 (3%)	34 (6%)	29 (13%)	278 (8%)
Length of time HIV-infected at study entry, years	1.3 (1.1, 1.8)	1.2 (1.1, 1.6)	1.6 (1.3, 2.9)	1.4 (1.2, 2.4)	1.4 (1.1, 1.7)	1.3 (1.1, 2.0)	1.3 (1.1, 1.8)	1.3 (1.1, 1.9) [1.0, 5.0]
Baseline CD4 count, cells/mm^3	640 (560, 781)	632 (552, 781)	687 (571, 822)	706 (583, 872)	641 (538, 824)	629 (555, 760)	641 (550, 780)	641 (560, 788) [500, 2189]
Baseline HIV RNA available	1574 (85%)	106 (99%)	126 (65%)	144 (52%)	139 (97%)	400 (70%)	182 (79%)	2671 (79%)
$\log_{10}$ copies/ml	4.1 (3.4, 4.7)	4.3 (3.7, 4.7)	3.9 (3.2, 4.5)	4.1 (3.6, 4.7)	4.0 (3.4, 4.7)	4.3 (3.7, 4.8)	4.2 (3.5, 4.6)	4.1 (3.5, 4.7) [0.9, 6.9]
$\leq$500 copies/ml	186 (12%)	3 (3%)	18 (14%)	9 (6%)	20 (14%)	36 (9%)	17 (9%)	289 (11%)

Table 1.2: Demographics of CASCADE patients included in our analyses. Values are n (%) for categorical variables, and median (interquartile range) [range, in last column] for continuous variables. [1] Defined as last negative and first positive HIV tests within 30 days, or laboratory evidence of seroconversion.

There were some differences in the patient characteristics across the countries, reflecting underlying differences in the HIV-infected populations targeted by the different seroconverter cohorts. Germany and the UK had few female patients (8 and 6%, respectively), and this was reflected in a greater proportion of patients reporting the route of HIV transmission as sex between men in those countries (86 and 89%, respectively). On average, the patients from Germany seroconverted later (median year 2004) and those from Italy and Spain earlier (1996 and 1997, respectively); this tied in with a relatively large proportion of Italian and Spanish patients reporting the route of HIV transmission as injecting drug use (IDU; 31 and 37%, respectively, compared to 9% overall). These individuals were also less likely to have an available baseline HIV RNA measurement (65 and 52%, respectively). A large percentage of German patients were identified as HIV-infected close to seroconversion (36%). The median baseline CD4 count ranged from 629 cells/mm^3 in UK patients to 706 cells/mm^3 in Spanish patients, and the percentage of patients with baseline HIV RNA $\leq$500 copies/ml ranged from 3% in Germany to 14% in Italy and Switzerland.

Follow-up

A total of 686 (20%) patients were censored due to irregular CD4 counts (resulting in the censoring of 74 events); of these, 626 patients (19 events) would subsequently have been censored due to LTFU. A further 1652 (49%) patients (with otherwise regular CD4 counts during follow up) were censored due to LTFU (no CD4 count in the 12 months before the last CD4 in the cohort); of these, 240 patients (34 events) were censored at 12 months after their last CD4 date and 1412 patients were censored at their last alive date. The large number of patients censored at their last alive date (which ranged from July 1996 to March 2008) came mainly from France ($n = 911$) and the UK ($n = 279$). After these censorings, 157 (5%) AIDS or death events were observed (103 AIDS and 54 deaths). The remaining 705 (21%) patients were considered to be administratively censored. The median follow-up time was 2.3 years (IQR 1.1, 4.6; maximum 12.4; Table 1.1). Overall, 1082 (32%) patients were observed to initiate treatment during follow-up, at median (IQR) [range] 17 (5, 33) [0, 123] months after baseline and at CD4 count 432 (296, 576) [12, 1998] cells/mm^3. The median (IQR) time between CD4 count measurements was 3.3 (2.4, 4.6) months, though varied by country (Table 1.1). The last CD4 count was carried forward in 75% of patient-months. There were no HIV RNA data available at all for 7% of patients and no prior HIV RNA data was available in 6% of patient-months. The time-dependent covariates are summarised over all follow-up time in Table 1.3. Of note, the median (IQR) CD4 count over all follow-up was relatively high, at 595 (468, 767) cells/mm^3.

	All follow-up	Treatment-naïve follow-up (1082 patients initiated treatment)
Number of patient-months follow-up	133 568	88 545
Follow-up time, months	27 (13, 55)	17 (5, 33)
CD4 count, cells/mm^3	595 (468, 767)	591 (477, 751)
[5^{th} and 95^{th} percentiles]	[303, 1116]	[326, 1090]
CD4 count decrease		
Large increase (>100 cells/mm^3)	5798 (4%)	2521 (3%)
Small increase (≤100 cells/mm^3)	7604 (6%)	3976 (4%)
No change[1]	104 982 (79%)	72 711 (82%)
Small decrease (≤100 cells/mm^3)	8241 (6%)	5001 (6%)
Large decrease (>100 cells/mm^3)	6943 (5%)	4336 (5%)
Time since last CD4 count, months	2.1 (1.0, 3.8)	2.3 (1.0, 4.0)
Nadir CD4 count, cells/mm^3	482 (345, 610)	530 (417, 661)
Number of previous CD4 counts	6 (3, 14)	4 (2, 8)
Number of previous HIV RNAs	6 (3, 13)	4 (2, 8)
Last HIV RNA, $\log_{10}$ copies/ml[2,3]	3.7 (2.3, 4.4)	4.1 (3.4, 4.6)
Time since last HIV RNA, months[2]	2.1 (1.0, 3.9)	2.4 (1.0, 4.2)
Peak HIV RNA, $\log_{10}$ copies/ml[2]	4.6 (4.0, 5.0)	4.3 (3.8, 4.8)

Table 1.3: Summary of time-dependent covariates over all follow-up and treatment-naïve follow-up. Values are n (%) for categorical variables and median (interquartile range) for continuous variables unless otherwise indicated. [1] By definition, there was no change in CD4 count if the last value was carried forward (as for 75% of observations over all follow-up and 78% of observations over treatment-naïve follow-up). [2] Of the patient-months with prior HIV RNA data available (94% over all follow-up and 90% over treatment-naive follow-up). [3] If no subsequent measurements available, last HIV RNA measurement carried forward regardless of the length of time.

1.7 Scope of the thesis

Our interest ultimately lay in the application of dynamic MSMs to optimise pre-specified dynamic treatment regimes, defined by time-dependent covariates, but these rely on having appropriately estimated inverse probability weights. We begin in chapter 2 with the estimation of causal treatment effects using a standard MSM, in order to investigate the construction of such weights. This process is not straightforward and there currently exists limited guidance for researchers. We illustrate and discuss the complexities of obtaining a suitable set of weights. We propose a simple and transparent algorithm for the construction of the weights, framed as a series of decisions, which must inevitably be subjective. We applied our algorithm to the CASCADE data to explore the implications of those decisions on the overall conclusions relating to the effect of antiretroviral therapy on the risk of AIDS or death in HIV-infected persons.

In chapter 3, we extend the standard MSMs in the most straightforward way to incorporate effect modification by a time-dependent covariate. We firstly considered the estimation of the effect of immediate versus deferred treatment initiation given current CD4 count, which addresses the clinically-relevant question regularly faced by health care providers and patients regarding whether to initiate or defer treatment with respect to current CD4 count. We then used history-adjusted MSMs to estimate the effects of treatment initiation immediately versus never, given current CD4 count. We compared these results to those under the immediate versus deferred treatment scenario, and also to the results from the standard MSMs; although these approaches address different questions, these comparisons may help improve understanding of the causal effects of treatment. The results from these history-adjusted MSMs could then be used to determine the optimal history-adjusted static regime for a patient, given their time-dependent covariate history.

We move to dynamic MSMs in chapter 4 to consider the optimisation of pre-specified dynamic treatment regimes, defined by CD4 count in our application to the treatment of HIV-infected persons. Although history-adjusted and dynamic MSMs share similar concepts, they are applied to different questions. Once again, while these different questions will of course give different answers, one might expect some consistency across the results with respect to treatment initiation in relation to CD4 count in HIV-infection. The inverse probability weights which are required for history-adjusted and dynamic MSMs are constructed in a similar way as for the standard MSM, hence the importance of the first step in determining adequate weights for the standard MSM before proceeding to more complex methods. We aimed to use this sequential application of all three types of MSM to enhance our understanding of the causal

effects of interest.

There have been recent developments in the application of dynamic MSMs to incorporate permitted delays in treatment initiation ("grace periods"; Cain et al. (2010)). These have rarely been applied in practice (Cain et al., 2010; HIV-CAUSAL collaboration, 2011; Shepherd et al., 2010) and their implications have not previously been investigated; we attempted to address this gap in the literature in chapter 4. In addition, we aimed to contribute to the debate outlined above in section 1.5 regarding the optimal timing of treatment initiation with respect to CD4 count in HIV-infected individuals.

Finally, in chapter 5, we compare and summarise the results across the chapters, draw some conclusions, discuss limitations and outline potential future work.

1.8 Summary of main contributions of the thesis

We advocate the application of all three types of MSM to address dynamic causal questions, and comparison across the approaches offers additional insights into the methodology and clinical results.

For the crucial step of construction of suitable inverse probability weights, we have structured this process as four key decisions, defining a range of strategies; all demonstrated a beneficial effect of ART in CASCADE. We found a trend towards greater treatment benefit at lower CD4 across a range of models.

Via large simulated randomised trials based on CASCADE data, longer grace periods (permitted delay in treatment initiation) and in particular less-frequently observed CD4 indicated higher optimal regimes (earlier treatment initiation at higher CD4), although similar AIDS-free survival rates may be achieved at these higher optimal regimes. In realistically-sized observational simulations, the optimal regime estimates lacked precision, mainly due to broadly constant AIDS-free survival rates at higher CD4. Optimal regimes estimated from dynamic MSMs should be interpreted with regard to the shape of the outcome-by-regime curve and the precision. When our desired inference is under the absence of a grace period, we found in our clinical setting that allowing a 3-month grace period may increase precision with little bias; under longer grace periods, the bias outweighed the efficiency gain. In our CASCADE population, immediate treatment was preferable to delay, although estimation was limited by relatively short follow-up.

Chapter 2

Standard marginal structural models

2.1 Introduction

In chapter 1, we introduced a number of methods for the estimation of causal effects. The aim of this chapter is to explore the estimation of causal effects using marginal structural models (MSMs) via inverse probability weighting. As discussed in section 1.7, the construction of appropriate weights may be a complex process. This has been addressed to some extent by previous authors (Cole and Hernán, 2008), but the majority of previous approaches are somewhat opaque and perhaps not easily implementable by many researchers (see section 2.3.1; Brookhart and van der Laan (2006); Mortimer et al. (2005); Petersen, Deeks, Martin, and van der Laan (2007)). We aim to contribute to this area by approaching the construction of the weights as a series of decisions, and use these to propose a range of plausible model building strategies. We apply these methods to estimate the causal effects of treatment on time to AIDS or death in HIV-infected persons in our population of patients from CASCADE, and assess the implications of these decisions.

2.2 Methodology

2.2.1 Notation

We wish to estimate the effect of treatment on time to the first occurrence of an AIDS-defining illness or death, assuming for now that there is no censoring (relaxed in section 2.2.4). We discretise time into small intervals (months) so that treatment and event probabilities can be calculated within those intervals and therefore aid computation; the weighted logistic regression models which we introduce in section 2.2.2 approximate weighted Cox proportional hazards regression models, provided event probabilities within each time interval are small (D'Agostino

et al., 1990), and are easier to implement using standard software.

Adapting and extending the notation introduced in chapter 1, let T be the time to the first AIDS event or death, and let $Y(t)$ be an indicator for whether an AIDS event or death occurred prior to time t, that is $Y(t) = 1$ if $T < t$ and $Y(t) = 0$ if $T \geq t$. Similarly, let $A(t) = 0, 1$ be an all-or-nothing but time-dependent indicator for whether treatment was initiated prior to time t, that is $A(t) = 1$ if treatment was initiated before time t, and 0 otherwise (including if treatment was initiated at time t). As indicated previously, we are interested in treatment initiations and ignore subsequent treatment discontinuations, therefore if $A(t) = 1$ then $A(s) = 1$ for $s > t$. Let $L(t)$ represent the latest time-dependent covariates measured prior to time t. As before, we use overbars to indicate history, so $\overline{A(t)} = \{A(0), A(1), \ldots, A(t)\}$ and $\overline{L(t)} = \{L(0), L(1), \ldots, L(t)\}$, and V represents a vector of baseline covariates.

For illustration, Figure 2.1 shows two examples of the data which may be collected. In example (a), the patient's time-dependent covariates L were measured between times $t-2$ and $t-1$; these measurements would then be used for $L(t-1)$. The patient initiated treatment between $t-1$ and t, meaning that $A(t-1) = 0$ while $A(t) = 1$, and the patient experienced the event between times t and $t+1$, therefore $Y(t) = 0$ and $Y(t+1) = 1$. Example (b) is included to illustrate what happens if these measurements and events take place at given time-points, which may be considered *at*, rather than *between*, clinic visits. The time-dependent covariates L were measured at time $t-2$ for this patient, therefore by our definition this informs $L(t-1)$. Similarly, this patient initiated treatment at time $t-1$ and therefore $A(t-1) = 0$ and $A(t) = 1$, and this patient experienced the event at time t, so $Y(t) = 0$ with $Y(t+1) = 1$.

Of note, we have used the end of each time interval to label the outcome, treatment and covariates (that is, *prior* to time t rather than *including* time t). These choices are unlikely to affect the findings from our work, but clearly it is important to apply the definitions consistently throughout to ensure temporality (time-dependent covariates predicting treatment, and both time-dependent covariates and treatment predicting outcome). We have applied these definitions since, in our work, the covariates L are typically measurements such as CD4 count or HIV RNA, for which bloods are taken and the results available at some later time. Other choices may be more appropriate in situations where the covariates L consist of measurements whose results are known immediately, such as blood pressure or weight.

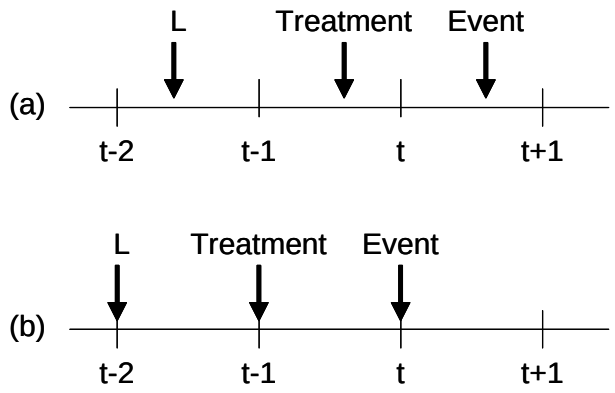


Figure 2.1: Examples of measurement of time-dependent covariates (L), treatment and events. See text for how these are used to determine the values at each time-point. Both examples are likely to occur in any observational data (including in interval and clinical cohorts).

2.2.2 Marginal structural Cox proportional hazards models

We might attempt to estimate the effect of current treatment using a Cox proportional hazards (PH) model given by:

$$\lambda_T \{t|A(t), V\} = \lambda_0(t) \exp \left\{\alpha' A(t) + \beta' V\right\} \tag{2.1}$$

If, given V, treatment was unconfounded, then $\widehat{\alpha'}$ would be an unbiased estimate for the causal effect of ever versus never having initiated treatment. However, as discussed in chapter 1, if treatment is confounded by time-updated covariates (an example of confounding by indication), then $\widehat{\alpha'}$ will be a biased estimate for the causal effect of treatment, whether or not we adjust for those time-dependent covariates in addition to V (Hernán et al., 2002).

Recall that $T_{\overline{A}}$ represents the time to event under a particular treatment regime $\overline{A}$. For a given patient, $T_{\overline{A}}$ will remain unobserved for the regimes which that patient did not follow and, under the assumption of consistency (see below), will equal the observed T for the treatment regime(s) $\overline{A}$ that they did follow. Then, as introduced in section 1.4.3, the Cox PH MSM of interest is given by:

$$\lambda_{T_{\overline{A}}} \{t|A(t), V\} = \lambda_0(t) \exp \{\alpha A(t) + \beta V\} \tag{2.2}$$

As outlined previously, since at least some of these outcomes will remain unobserved, we cannot fit this model directly. However, if we use inverse probability weights to create a pseudo-population in which we upweight patients who are and are not on treatment at each time-point to account for the patients with the same covariate history but who are not on the same treatment path, then we have removed the dependence of treatment on the measured time-dependent covariates and so treatment is no longer confounded by those covariates (Hernán et al., 2000).

Then the standard Cox PH model of equation 2.1 applied to the pseudo-population (via inverse probability of treatment weighting) will yield an unbiased estimate $\hat{\alpha}$ for the causal effect of ever versus never treated, under certain assumptions.

Assumptions

As outlined in section 1.2.4, for the application of MSMs to estimate causal effects in observational studies, we require the following assumptions (Cole and Hernán, 2008):

- Consistency: formally, this states that the potentially counterfactual outcome $T_{\overline{A}}$ under treatment regime $\overline{A}$ is equal to the observed outcome T if the patient was observed to follow regime $\overline{A}$.
- No unmeasured confounders between treatment and the outcome (otherwise known as exchangeability). That is, at each time t, the treatment received at that time is independent of the time to event, given treatment and covariate history. Formally, this means that for each $\overline{A}$, we assume that $T_{\overline{A}}$ is independent of $A(t)$ given $\overline{A}(t-1)$ and $\overline{L}(t-1)$ (Hernán et al., 2001).
- No misspecification of the models.
- Positivity, that is that there is a non-zero probability of receiving each treatment regime for all combinations of treatment and covariate history. Formally, letting $f(\cdot)$ represent the probability density function, we assume that $f\left\{\overline{A}(t-1), \overline{L}(t-1)\right\} > 0$ implies $f\left\{A(t)|\overline{A}(t-1), \overline{L}(t-1)\right\} > 0$ for all $A(t), \overline{A}(t-1), \overline{L}(t-1)$ (Hernán et al., 2002).

The first three assumptions cannot be tested from the data, though the second and third can be explored by considering a broad range of potential confounders and different model specifications. Note that we also require temporality, that is $L(t-1)$ is measured prior to $A(t)$, which we have by the conservative construction of our data (section 2.2.1).

Since current guidelines (outside the USA) recommend HIV-infected persons to initiate treatment around CD4 counts of 350 cells/mm^3 (Gazzard and on behalf of the BHIVA Treatment Guidelines Writing Group, 2008; WHO, 2010), we might for example expect all patients with CD4 count < 300 cells/mm^3 to be on treatment. This would violate the positivity assumption and we would not be able to estimate the causal effects of treatment in this CD4 count range. Cole and Hernán (2008) refer to such cases as "structural zeroes" since by definition in these circumstances there would be zero probability of never having started treatment when CD4 count < 300 cells/mm^3. In practice, it is unlikely that we will see such consistent treatment

patterns. "Random zeroes" due to chance are permitted since the use of a model essentially "borrows" from the remaining data, although the presence of random zeros increases the chance of bias due to non-positivity (Cole and Hernán, 2008).

Aside from non-positivity concerns, if there exist patients who remain treatment-naïve with very low CD4 counts, these patients are an unusual subset. If it was possible to identify such "treatment refusers" from the outset, then we might consider excluding those patients altogether, on the grounds that they do not constitute our population of interest and further we may be worried that we have not captured all potential confounders to adequately describe the treatment behaviour of these patients. However, this could result in bias; we cannot identify these patients from the outset. Alternatively, if we observed a patient to reach CD4 count < 100 cells/mm^3 without initiating treatment then we may be tempted to censor the patient at that time, in order to attempt to restrict to our population of interest, namely patients who would consider taking treatment. However, such a censoring process is dynamic and cannot be appropriately accounted for, via weighting of MSMs or otherwise, without addressing the dynamic element. In chapter 3, where we consider the start of each treatment-naïve month of follow-up as a "trial" for immediate treatment initiation versus deferral, we will be able to exclude "trials" where the "baseline" CD4 count is < 100 cells/mm^3. Further, in chapter 4 such patients will implicitly be censored from regimes defined by earlier treatment initiation at higher CD4 counts. However, we cannot easily address this issue further with standard MSMs.

2.2.3 Inverse probability of treatment weights

In general, the treatment weights are not known, therefore we must estimate them from the data. However, even if the true weights are known, it has been shown that appropriately estimated weights are more efficient (Hernán et al., 2001; Moodie, 2009). The inverse probability of treatment weight for a particular patient at time t is defined by the inverse probability of that patient having received their observed treatment to t, given their baseline and time-updated covariates and previous treatment. In practice, we split time into suitable intervals denoted $t = 1, 2, ...$ and use pooled logistic regression, treating each person-time interval as an observation and estimate the treatment probabilities up to each time t as follows.

Following the notation of Hernán et al. (2001), define:

$$p_A(t) := \Pr\left\{A(t) = 0 | \overline{A}(t-1) = 0, Y(t) = 0, \overline{L}(t-1)\right\} \tag{2.3}$$

for $t = 1, 2, ...$ where $A(0) = 0$, since by definition all patients are treatment-naïve at base-

line, and $L(0)$ are the time-dependent covariates measured at baseline and include the time-independent covariates V. We perform a pooled logistic regression, on patients previously treatment-naïve, with outcome of treatment initiation, to obtain the probabilities of treatment initiation in the time intervals $t = 1, 2, ...$ given time-updated covariates, and hence obtain the estimates $\widehat{p}_A(t)$ for non-initiation of treatment. We are then able to estimate the probability of each patient's observed treatment to time t, given baseline covariates, time-dependent covariate history and past treatment, as follows:

$$\widehat{q}_A(t) = \begin{cases} \prod_{k=1}^{t} \widehat{p}_A(k) & \text{if patient did not initiate treatment up to time } t \\ \{1 - \widehat{p}_A(k)\} \prod_{l=1}^{k-1} \widehat{p}_A(l) & \text{if patient initiated treatment in } [k-1, k), \text{ for } k \leq t \end{cases}$$

and we estimate the weights using:

$$\widehat{W}(t) = \frac{1}{\widehat{q}_A(t)}.$$

For example, consider four patients with the same covariate history prior to time $t-1$, who were all treatment-naïve prior to time $t-1$, and three of these patients remained off treatment to time t but the fourth patient initiated treatment prior to time t. Then in this subset of patients, the probability of initiating treatment prior to time t (given off treatment prior to time $t-1$) is $1/4$. At time t, the first three patients who did not initiate treatment are assigned weight $\frac{1}{1-1/4} = 4/3$, so these three patients count for themselves and also the fourth patient who is no longer following that treatment regime of not initiating treatment prior to time t. Conversely, the fourth patient who did initiate treatment is assigned weight $\frac{1}{1/4} = 4$, and therefore counts for him/herself plus the three patients who did not follow that regime of initiating treatment prior to time t.

In practice, a select few patients may have large weights and these would dominate the analysis thus leading to large standard errors. Therefore, we usually stabilise the weights to increase the efficiency (Hernán et al., 2000). In theory, this can be done by replacing the numerator of 1 in $\widehat{W}(t)$ with any function of treatment $A(t)$ but which is not a function of the time-dependent covariates (Hernán and Robins, 2006). In practice, we typically use a function of time-independent variables by defining $p_A^*(t)$ analogously to $p_A(t)$ as in equation 2.3, except replacing $\overline{L}(t-1)$ with V, and similarly estimate $\widehat{q}_A^*(t)$. Then the stabilised weights are given by:

$$\widehat{SW}_A(t) = \frac{\widehat{q}_A^*(t)}{\widehat{q}_A(t)}.$$

Informally, the denominator is the probability of treatment given treatment history and *time-updated* covariates (including baseline), whereas the numerator is the probability of treatment given treatment history and *baseline* covariates only. The informal reasoning behind this is that we can adjust more efficiently for the baseline covariates in the outcome model instead, rather than via the weighting. We must adjust for the covariates V in the outcome model, since the stabilised weights only remove the time-dependent confounding conditional on V (Cole and Hernán, 2008). To further help control the weights, truncation may be performed (Cole and Hernán, 2008). For further discussion on the weight estimation, see section 2.3.

2.2.4 Censoring

As mentioned in section 1.2.4, right-censoring is common with time-to-event data therefore some patients will be censored before we observe the event. Under the assumption that the censoring process is independent of T, conditional on covariate and treatment history (no unmeasured confounders), then we can easily adapt our weighting method of above to estimate inverse probability of censoring weights and therefore account for censoring. By doing so, we are attempting to estimate the effect of treatment in the absence of any censoring (Hernán et al., 2001). For illustration, assume there is just one type of censoring and let $C(t) = 0, 1$ represent whether censoring has occurred prior to time t. Define for $t = 1, 2, ...$:

$$p_C(t) := \Pr\left\{C(t+1) = 0 | \overline{A}(t), \overline{L}(t), C(t) = 0, Y(t+1) = 0\right\}$$

and again analogously for $p_C^*(t)$ with $\overline{L}(t)$ replaced by V. We consider $C(t+1)$, that is, censoring in the interval $[t, t+1)$, rather than $C(t)$, to correspond with the interval used for the outcome estimation (see section 2.2.5). Estimation of the stabilised weights $\widehat{SW}_C(t)$ then follows as above for the treatment weights. In practice, there may be a number of different reasons for censoring (as in section 1.6.1, for example). These methods can be applied to different censoring types to estimate separate weights, which can then be combined, or extended analogously to treat different censoring types as a range of outcomes in a multinomial logistic regression. While in theory any number of different types of censoring may be incorporated in this way, in practice this may be limited by the data available, and the analyst should check that this does not cause excessive variability in the weights.

The overall weights are given by the joint probability of observed treatment and remaining uncensored, assuming these are independent processes given the measured confounders. Therefore in the presence of censoring, we amend the treatment weight estimation to also condition on

$C(t) = 0$ in equation 2.3 and similarly for $p_A^*(t)$, and obtain the overall weights by multiplying the (amended) treatment and censoring weights together to estimate the overall weights:

$$\widehat{SW}(t) = \widehat{SW}_A(t) \times \widehat{SW}_C(t).$$

2.2.5 Estimation of treatment effect

Finally, for $t = 1, 2, ...$, we estimate:

$$p(t) := \Pr\left\{Y(t+1) = 1 | Y(t) = 0, C(t+1) = 0, \overline{A}(t), V\right\}$$

using for example a pooled logistic regression of the form:

$$\text{logit}\left\{p(t)\right\} = \log\left\{\frac{p(t)}{1-p(t)}\right\} = \alpha A(t) + \beta V + \gamma f(t) \tag{2.4}$$

weighted using the overall stabilised weights $\widehat{SW}(t)$, where $f(t)$ is some function of time. That is, the log-likelihood function which we seek to maximise is given by:

$$\begin{aligned} & \sum_i \left\{ I\left[Y_i(t+1) = 1\right] SW_i(t) \log p_i(t) + I\left[Y_i(t+1) = 0\right] SW_i(t) \log\left(1 - p_i(t)\right) \right\} \\ = & \sum_i SW_i(t) \log\left\{\frac{\exp\left\{I\left[Y_i(t+1) = 1\right](\alpha A_i(t) + \beta V_i + \gamma f(t))\right\}}{1 + \exp\left\{\alpha A_i(t) + \beta V_i + \gamma f(t)\right\}}\right\} \end{aligned}$$

where $I[\cdot]$ is an indicator equal to 1 if $\cdot$ is true, and 0 otherwise, and i indexes the patients in the study; the parameters to be estimated are α, β and γ. Using the weights $\widehat{SW}(t)$ for the estimation of the outcome in the interval $[t, t+1)$ means that we do not adjust for treatment initiations in that interval, to ensure temporality (treatment initiation in that interval could be in response to the event).

In general, it is not possible to efficiently estimate an intercept for every time interval, therefore we use a function of time $f(t)$, perhaps categorical or a spline (Hernán et al., 2000). Assuming small event probabilities per time interval, the resulting odds ratios can be interpreted as hazard ratios (D'Agostino et al., 1990). Under the assumption of no unmeasured confounders for treatment and outcome, we obtain an unbiased estimate $\hat{\alpha}$ for the effect of ever versus never treated on the time to AIDS or death. We use robust variance estimators to allow for correlated observations induced by the use of time-dependent weights (Cook et al., 2002; Zeger and Liang, 1986), implemented using Stata's `robust` command which uses the sandwich variance estimator.

This model assumes a constant effect of treatment over time, which may not be plausible

in certain scenarios. For example, we might expect a greater benefit of treatment the longer time spent on it; this could be incorporated with a covariate capturing time on treatment. In addition, it is possible to investigate treatment effect modification by baseline covariates V by incorporating interactions with treatment, for example using:

$$\text{logit}\left\{p(t)\right\} = \alpha A(t) + \beta V + \gamma f(t) + \delta A(t)V.$$

Of note, if we wish to look at such interactions with baseline covariates then it is not essential to incorporate those baseline covariates into the model for the numerator of the weights, but since they will be in the outcome model including them in the model for the numerator may potentially increase efficiency.

Usual model fitting techniques and methods for checking goodness of fit can and should be applied.

2.3 Estimation of the weights in practice

We now discuss the estimation of the inverse probability of treatment weights in practice, but in the presence of censoring the same principles and methods can be applied to estimate inverse probability of censoring weights.

2.3.1 Bias-variance trade-off

Adequate specification of the treatment prediction model is necessary for consistent estimation of causal treatment effects via an MSM, but in practice determination of such a model may not be straightforward. Cole and Hernán (2008) outline three main steps for constructing the weights: firstly, they recommend checking the positivity assumption for the confounders which are suspected to be most influential. Secondly, they suggest investigation of the assumption of no unmeasured confounders by considering a broad range and specification of measured potential time-dependent covariates in the weight estimation, checking for sensitivity in the estimated treatment effect. Lefebvre et al. (2008) recommend including in the treatment model confounders for outcome and treatment, and risk factors for the outcome, but not predictors of treatment alone (that is, which are not also associated with the outcome); they found via simulations that the bias in using an incorrect treatment model was not significant and was outweighed by the gain in efficiency. The third step of Cole and Hernán (2008) is to assess model specifications by looking at the distribution of the weights, namely the mean and spread.

At each time-point, the mean of the stabilised weights should be close to one. To see why, let $f(A, C)$ be the probability density function for treatment A and censoring C. Then the stabilised weights are given by $\frac{f(A,C|V)}{f(A,C|\overline{L}(t))}$ and so the stabilised weights should have mean one since, under the assumption of no unmeasured confounders, $E\left\{\frac{f(A,C|V)}{f(A,C|\overline{L}(t))}\right\} = E\left[E\left\{\frac{f(A,C|V)}{f(A,C|\overline{L}(t))}|\overline{L}\left(t\right)\right\}\right] = 1$ (Hernán and Robins, 2006). For simplicity, we usually simply look at the mean over all time intervals. As Cole and Hernán (2008) indicate, this could lead to the selection of weights which fit reasonably well over all time-points rather than a set of alternative weights which fit better at most time-points but poorly at a few. For this reason, we may also wish to check the weights at each time-point.

Large weights may arise due to a few patients who for some reason do not follow typical treatment patterns (and are thus most informative with regard to confounding), leading to situations close to non-positivity. Further, large weights may arise due to model misspecification, particularly for continuous covariates since the treatment model may predict very low or high probabilities of treatment initiation at the extremes of the range, therefore any patients who do or do not initiate treatment, respectively, will receive large weights. Even in the absence of bias due to non-positivity or model misspecification, large weights which are merely a consequence of sampling variation may dominate the analysis leading to large standard errors and unstable estimates, so some truncation of the weights may be prudent.

Whilst we would expect well-estimated weights to have small standard deviation or range, Cole and Hernán (2008) note that the "best" weights, with respect to these conditions of small standard deviation and mean one, would be equal to one for all patients and time-intervals, but this would not adjust for time-dependent confounding at all. They describe this process of simultaneously attempting to address the assumptions of positivity, no unmeasured confounders and no model misspecifications as a bias-variance trade-off. This balance can be explored by looking at progressive truncation of the weights (Cole and Hernán, 2008), though Cole et al. (2005) caution that the most extreme weights contain the most information with respect to confounding therefore weight truncation is not ideal for model checking.

Therefore, even within the extent of existing guidance, there are a number of subjective decisions to be made in terms of this balance of bias and variance which may legitimately be approached differently by different researchers. The potential for different choices primarily lies with determining what size change is important when investigating the sensitivity in the estimated treatment effect to different model specifications and what is deemed a reasonable weight distribution. Of note, "traditional" model building approaches such as stepwise back-

wards selection are not appropriate since they focus on determining predictors of treatment rather than confounders of treatment and outcome, and do not consider the efficiency of the treatment effect estimator.

There have been a number of attempts to formalise the treatment model selection procedure. Firstly, the approach of Mortimer et al. (2005) requires predefining a candidate set of treatment models (they used 10) and optimising the bias-variance trade-off using a cross-validation approach with a residual sum of squares (RSS) criterion. In particular, the first step of their method involves splitting the data by 90% to 10% into training and test sets, respectively. Each of the candidate treatment models are fit to the training set and that which minimises the Akaike information criterion (AIC, given by $2k - 2\ln(L)$, where k is the number of parameters and L is the maximised value of the likelihood function) is labelled X. From each of the candidate models, inverse probability weights are estimated and the corresponding MSM estimate of the parameter of interest is obtained. The outcome is then predicted for each observation in the test set based on each of the MSM estimates. Ideally, the best MSM estimate would be that which minimises the mean counterfactual RSS, but of course those are not all observed. Therefore, a modified RSS is employed, whereby the observed RSS is weighted just as in an MSM. That is, the modified RSS is given by:

$$Q = \frac{\frac{1}{n}\text{observed RSS}}{\Pr\left(\text{treatment} \mid \text{time-updated covariates}\right)}$$

where n is the number of observations in the test dataset. The model X determined above is used to apply this weighting. This process should be iterated a large number of times (Mortimer et al. (2005) did so 10,000 times) and the overall Q is taken to be the average. The best treatment model is then chosen as that with minimal overall Q.

A potential limitation of this approach is the requirement for a restricted set of treatment models to be chosen at the outset. Further, although the authors note that the distribution of the weights should always be checked and recommend that sensitivity analyses should be performed in order to give confidence in the chosen weights, they do not provide any additional recommendations on how this may be done. In particular, if this procedure leads to a model with weights which are not deemed suitable for some reason, then it is unclear how the analyst should proceed; indeed, it is not entirely clear how to assess the suitability of the weights.

In a paper on HAMSMs (see chapter 3), Petersen, Deeks, Martin, and van der Laan (2007) used a somewhat different cross-validation approach with a "deletion/substitution/addition algorithm" in order to select their treatment model. Once again, this process involves fitting

models of various sizes and complexities, assessing performance in independent samples, and selecting that which optimises the bias-variance trade-off. While the range of model possibilities under Petersen, Deeks, Martin, and van der Laan (2007) was quite extensive, there was still some lack of transparency in the process, which would not be straightforward to implement. Brookhart and van der Laan (2006) also used a cross-validation approach, to minimise the mean square error.

The ultimate aim of these methods is to select a treatment model which best balances bias versus variance. While a parsimonious model may offer relatively low variance and avoid bias due to positivity, it may inadequately control for the time-dependent confounding in the treatment model. Further, treatment model misspecification may result in bias or inflated variance. While the approaches above methodically select an optimal treatment model based on a (finite sample) bias-variance trade-off, they are not easy to apply and are unlikely to be adopted by many researchers. The processes are not transparent; a suitable stepwise approach to the model selection process (which focuses on controlling for confounding between treatment and outcome, not determining predictors of treatment like in "traditional" model building approaches) may offer insights into the data at hand and potential issues with particular variables or models. Further, there are a number of other factors, such as truncation of the weights, which are not addressed by these methods. Lastly, as mentioned above, it is unclear how to proceed if the approaches discussed above yield weights that are for some reason deemed unsuitable. We sought to address the various subjective decisions that an analyst may be faced with when attempting to determine a suitable treatment model and propose an informal, transparent approach to the construction of the weights, as a series of decisions.

The positivity assumption

As mentioned above, the first step of Cole and Hernán (2008) for constructing the weights is to check the positivity assumption for the key confounders. In practice, this can be explored by examining the treatment initiation patterns across different categories of the confounder, to see whether patients do and do not initiate treatment at all levels of the confounder. This could simply be done for one key confounder (for example, CD4 count), or across different levels of multiple confounders (for example, CD4 count and HIV RNA). Even if it is thought that structural zeroes are unlikely, then this may help identify random zeroes.

If there is concern about violations of the positivity assumption, then one option for the analyst may be to collapse categories of the confounder, or model the confounder continuously in order to smooth over the random zeroes. An alternative, and somewhat more extreme option,

would be to restrict the sample to exclude groups of patients for whom there exists limited variability in the treatment pattern (for example, if the vast majority of patients initiated treatment when CD4 < 200 cells/mm^3 then we exclude the small subset of patients who did not initiate treatment with such a low CD4 count, on the grounds that those patients do not constitute the population in whom we wish to estimate treatment effects).

Petersen et al. (2010) provide more detail about diagnosing and responding to violations in the positivity assumption, and, in the presence of such violations, the authors recommend a systematic approach to the trade-off between the desired inference (unbiased) and identifiability (precision).

2.3.2 Key decisions

While Cole and Hernán (2008) give a broad outline of the principles behind the treatment model building process, there is still scope for a number of different approaches. One such difference lies in the starting point for analysis. Let L_{key} be a small number of covariates which are known a priori to be important confounders of the relationship between treatment and outcome; the first step of Cole and Hernán (2008) is to investigate the positivity assumption with respect to these variables. Let L_{pot} denote the remaining potential confounders which may be considered for inclusion in the treatment model. One possible approach to the model building process might be to begin with a treatment model consisting of time, baseline covariates V and L_{key} only and then consider the addition of each of the other potential confounders of L_{pot} in turn with reference to some pre-defined criteria for identifying which covariates are important confounders and so should be included in the treatment model. Iterating this process until no further covariates meet the criteria for inclusion would yield the final treatment model. Alternatively, the analyst could start from a "full" treatment model including the potential confounders L_{pot}, in addition to time, V and L_{key}, and the reverse procedure applied to identify covariates which are not important confounders and so can be removed from the treatment model.

The pre-defined criteria for identifying important confounders may incorporate a number of factors, such as the distribution of the weights, the estimated treatment effect and/or its standard error; this is different to a standard model selection procedure such as backwards elimination which would only consider the significance or otherwise of the variables in the treatment model. Note that covariates which are solely risk factors for the outcome, and not confounders for treatment, may be adjusted for directly in the outcome model, even if they are time-dependent.

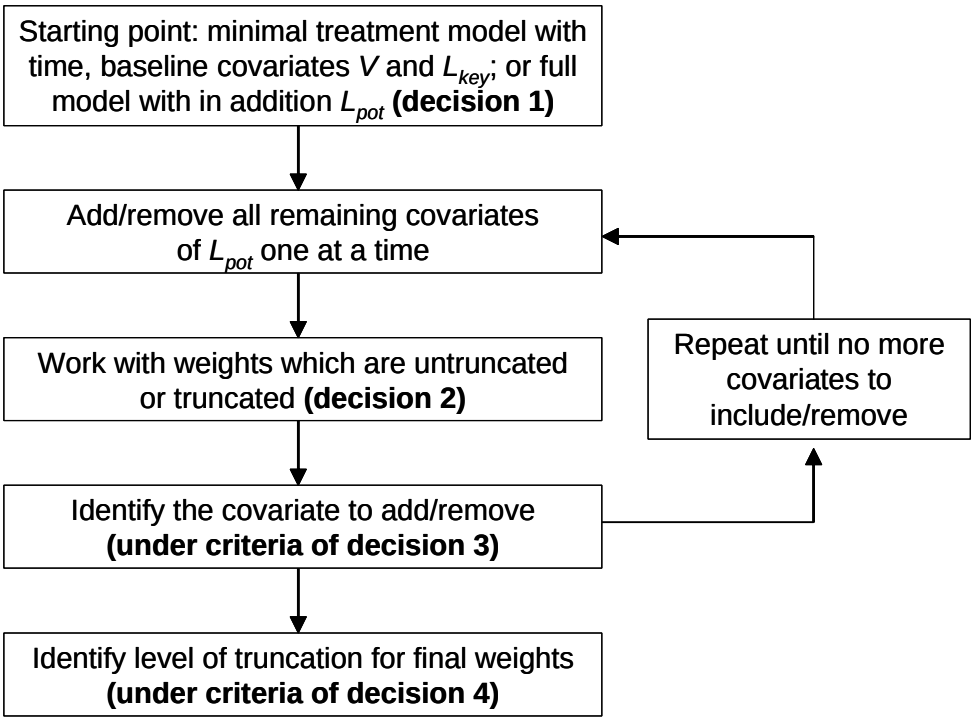


Figure 2.2: Flow chart for treatment model building process. See text for further details on the decisions.

In the absence of bias in the estimated treatment effect, progressive truncation of the weights will result in weights which have mean closer to one and smaller standard deviation. In practice, truncation may result in bias due to poorer control of confounding, but conversely it may help protect against bias due to non-positivity or model misspecification. Therefore it is prudent to check sensitivity of the estimated treatment effect to weight truncation.

One possible approach to the treatment model building process, covering the aspects discussed above, is outlined in Figure 2.2, with the key decisions highlighted. While there are other possible approaches and scope for variation, these decisions cover a number of key aspects and will illustrate the potential differences that may arise under alternative but equally plausible strategies. We now consider each of the decisions in detail.

Decision 1: starting point

As suggested above, the analyst may choose to start with a minimal model consisting of time, baseline covariates V and the key confounders L_{key}, or a "full" model consisting of the potential confounders L_{pot}, in addition to time, V and L_{key}. Backwards selection could be done, but if for example L_{pot} contains a large number of variables and there are concerns about non-positivity, then it may be preferable to start with the minimal model. This relates to the forward/backward part of the algorithm proposed by Brookhart and van der Laan (2006).

Decision 2: working weights

The weights used during the treatment model building process may be truncated or untruncated. If there is concern about non-positivity and it is thought that some truncation may be performed on the final weights, then one may argue that it is preferable to work with truncated weights throughout; the extent of the truncation could be determined from a priori set criteria. Conversely, one may argue that it is preferable to work with untruncated weights during the model building process to ensure that all potential confounders are identified, although one may wish to still consider truncation at the end under the criteria of decision 4.

Decision 3: covariate selection

Firstly consider the case where we start with a minimal model and identify covariates for inclusion. At each stage, we will have a basic model M to which we are considering adding each of the remaining covariates of L_{pot} which are not yet in the treatment model; let model M_i denote model M but with in addition the i^{th} covariate of L_{pot}. If the addition of covariate i to the treatment model moves the estimated treatment effect away from the unweighted estimate, then this could be due to better control of confounding (or similarly improved model specification) or problems relating to bias. Bias could be due to finite-sample bias and being close to non-positivity (Cole and Hernán, 2008), or to selection bias arising from collider-stratification. This latter issue is discussed by Greenland (2003); it is difficult to think how this might arise in our application and it would not be possible to detect empirically, but if this source of bias was suspected then it may perhaps indicate that we should not include the covariate i in the model (Cole and Hernán, 2008). Problems with positivity can be identified by looking at the distribution of the weights: mean weights far from one indicate problems with positivity, or perhaps model misspecification. Preference may also be given to models with a small standard error (as in Cole and Hernán (2008)) though note it is likely that the unweighted treatment effect estimate will also have small standard error since there is no additional variability introduced by the weights. Of note, if a covariate is a statistically significant predictor of treatment but does not result in much change in the estimated treatment effect, then that covariate is unlikely to be a (strong) confounder of treatment and outcome. Further to the findings of Lefebvre et al. (2008), such a variable should not be included in the treatment model since the potential bias due to an incorrect treatment model would be minimal compared to the gain in efficiency. Therefore when comparing models M with M_i to determine whether to include covariate i, there are three factors to be considered: the proximity of the mean of the weights to one, the

size of the treatment effect estimate and the size of the standard error of the treatment effect estimate (which is related to the standard deviation and range of the weights, therefore we do not consider these separately). We quantified these factors as follows:

1. proximity of the mean of the weights to one by looking at the absolute change in the mean of the weights towards one for model M_i compared to M

2. movement of the estimated treatment effect relative to the unweighted estimate by looking at the absolute and relative change for model M_i compared to M

3. absolute and relative increase in the standard error of the estimated treatment effect for model M_i compared to M.

These factors were informally considered by Cole and Hernán (2008) when comparing different treatment models. These could be combined and parameterised as follows: the i^{th} covariate of L_{pot} is eligible for inclusion if, compared to model M, it (i) moves the mean of the weights $>\mu$ closer to one or (ii) changes the treatment effect estimate relative to the unweighted estimate by $> \phi_p\%$ (and $>\phi$ for some small ϕ, to avoid very small differences being counted) or reduces the standard error by $> \sigma_p\%$ (and $> \sigma$) but with mean of the weights $\leq \mu_m$ further from one. Increasing μ, ϕ_p or σ_p, or decreasing μ_m, will make the criteria more stringent and hence may lead to a smaller treatment model. When identifying eligible covariates, all three criteria are of equal importance. If more than one covariate is eligible, then we pick the one for inclusion as that which most improves the weights (or impairs the least if none improve, since we may wish to include a covariate which has a large impact on the treatment effect estimate even at the expense of slightly poorer weights). That is, we prioritised criterion 1 over criteria 2 and 3, in order to focus on obtaining well-behaved weights. With our logistic regression models, the treatment effect estimate and associated standard error should be considered on the log odds scale since the standard error of the odds ratio will be related to the size of the odds ratio.

Interactions and stratification Interactions between key covariates can be considered under these criteria in the same way as the addition of a covariate. Taking this one step further, if there are known strata such as different centres or countries, then stratification on that factor can be considered similarly, since stratification can be considered as incorporating interactions between the stratification factor and all other variables. Therefore although there would be a large increase in the number of parameters, the unstratified model is still nested within the stratified model. So letting M represent the unstratified model and M_i represent the stratified

model, then the same criteria as above can be applied to determine whether separate treatment models should be estimated within levels of the stratification factor (e.g. centre or country).

Process beginning with a "full" model Under decision 1 where the analyst begins with a "full" model, the reverse of this process can be applied to determine removal of covariates. That is, if M' represents the current model, then compare with each of the models M'_i which are identical to M' but with the i^{th} remaining covariate of L_{pot} removed. Applying the reverse of the same criteria as above, we can determine whether or not to remove covariate i from M' in favour of the smaller model M'_i.

A combination of these steps, akin to the stepwise backwards procedure often used in traditional model selection, could be applied. For ease, this was not applied here, but we would anticipate seeing broadly similar results as under the range of strategies considered below.

Decision 4: degree of weight truncation

As discussed above, weight truncation may induce bias due to poorer control of confounding but conversely may help protect against bias due to non-positivity or model misspecification. Similarly to Cole and Hernán (2008), we propose considering progressive truncation to investigate the effects on the estimated treatment effect and distribution of the weights. In addition we propose specific criteria to determine what level of truncation to choose, reflecting either a desire for a simpler model if there are concerns about positivity or a more complex model in order to better control for confounding. Assuming no bias in the initial treatment effect estimate, progressive truncation will result in estimates which are increasingly biased but more precise (Cole and Hernán, 2008), so we will typically see a decrease in the mean and standard deviation of the weights, which translates to a treatment effect estimate closer to the unweighted estimate and smaller standard error of that estimate.

It will be necessary to propose a set of truncations to consider. There is little to inform the specific levels of truncation in this set and it should be recognised that different choices may lead to different conclusions, since clearly the decision whether to progress from one level of truncation to the next will depend on the (relative) levels of truncation. If very extreme weights are seen then it may be prudent to perform some minimal truncation by default.

For each final treatment model, we propose the following two possible criteria to determine whether to proceed with additional truncation (and stop when decide not to truncate any further):

(a) truncate if it leads to a reduction (that is, weakening) in the estimated treatment effect of

$> \varphi$ and $> \varphi_p\%$, provided no worsening of the mean of the weights in terms of absolute distance from one (and assuming reduction of the standard error)

(b) truncate if it leads to a reduction in the standard error by $> \rho$ and $> \rho_p\%$, provided the reduction in the estimated treatment effect is $\leq \varphi_m\%$

Rule (a) favours truncation if that is associated with a large change in the treatment effect, since this large change could indicate problems with positivity and truncation may help protect against the potential bias due to non-positivity. This will typically be a conservative approach. In contrast, rule (b) will only truncate if that offers benefits in terms of increased precision and is not associated with a large change in the treatment effect; since "the extreme weights encode the greatest amount of confounding" (Cole et al., 2005), the argument is that truncation may lead to inadequate control for confounding.

2.3.3 Strategies

In theory, numerous permutations of the decisions above could be combined to form a large number of strategies; in order to have a manageable but varied set of strategies, we may wish to consider a limited combination of decisions. For example, if we were most concerned about positivity (perhaps from initial investigations or through consultation with clinicians) then we might favour the following options, which we might suspect would lead to a smaller model and therefore avoid large models where there may be a higher risk of non-positivity:

- start with the minimal model; recognise that some weights are likely to be unreasonably large and so work with truncated weights; choose the parameters $\{\mu, \phi, \sigma\}$ to favour a smaller model; apply criterion (a) to favour greater truncation to avoid bias due to non-positivity.

Conversely, if we suspect that our a priori specified set of covariates really are likely to be important confounders for treatment and outcome and we are most concerned about adequate control for confounding, then we might instead prefer the following options to tend towards a larger model to allow maximum control for confounding:

- start with a large model; work with untruncated weights which may better capture the confounding; choose the parameters $\{\mu, \phi, \sigma\}$ to favour a larger model; apply criterion (b) to favour less truncation.

Strategy	Decision			
	1. Where to start?	2. Working weights?	3. Covariate selection procedure? Parameterise $\{\mu, \phi, \sigma\}$ to favour:	4. Degree of weight truncation? Favour:
I	(a) minimal model	(a) truncated	(a) smaller model	(a) greater truncation
II	(a) minimal model	(a) truncated	(a) smaller model	(b) less truncation
III	(a) minimal model	(a) truncated	(b) larger model	(b) less truncation
IV	(a) minimal model	(b) untruncated	(b) larger model	(b) less truncation
V	(b) "full" model	(b) untruncated	(b) larger model	(b) less truncation

Table 2.1: Summary of the treatment model building strategies.

These two strategies are shown as I and V, respectively, in Table 2.1. In addition, we propose to evaluate three other intermediate strategies, yielding a set of strategies which provide direct comparisons for each of the four decisions. We would anticipate different combinations of the decisions to be intermediaries of this set. While other approaches or decisions are possible, we believe the chosen strategies provide a realistic yet varied set of approaches to the weight construction.

2.3.4 Model checking using centre or country

As discussed above, if there exist known strata such as centre or country, then we may wish to stratify the treatment model on that factor by fitting separate treatment models for each level of that covariate. Such a variable may also be exploited in a different way: while we might expect different event rates across the different strata, we may expect to see consistent treatment effects across the strata, assuming that the treatment is homogeneous. This may not be the case for complex interventions such as behavioural therapy for example, but where there exist fairly standard drug regimens and guidelines across the strata, the assumption of homogeneity is likely to be reasonable. This can easily be investigated by incorporating interactions between treatment and the centre or country covariate in the outcome model. However, if there exist interactions between treatment and other baseline covariates, these could induce spurious interactions between treatment and the stratification factor if the baseline covariates differ across the strata, therefore such interactions should also be considered.

2.4 Application to CASCADE

The CASCADE data were introduced in section 1.6.

2.4.1 Methods

We firstly demonstrated that CD4 count is a time-dependent confounder, by fitting pooled logistic regression models for (*i*) AIDS or death on time (2-yearly categories), V, time-updated treatment (ever versus never initiated) and time-updated CD4 count to show that CD4 count predicts time to AIDS or death and (*ii*) treatment initiation on time (five knot spline with knots at the 5, 25, 50, 75 and 95^{th} percentiles of 0.1, 0.6, 1.3, 2.7 and 5.8 years), V and time-updated CD4 count to show that CD4 count predicts treatment initiation. We used a linear model for mean CD4 count at time t, adjusting for t (2-yearly categories), V, treatment at time $t-1$ and CD4 count at time $t-2$, to show that treatment predicts subsequent CD4 count and hence CD4 count is on the causal pathway between treatment and outcome.

To investigate the positivity assumption, we tabulated treatment initiations firstly by CD4 count alone and secondly also by HIV RNA.

Model fitting

As a preliminary treatment model, we included time (five knot spline as above), V and $L_{key} = \{\text{CD4 count}\}$. We considered a variety of functional forms for CD4 count, including categorical (by 50 cells/mm^3) and three, five and seven knot splines (with knots at the 10, 50 and 90^{th} percentiles; 5, 25, 50, 75 and 95^{th} percentiles; and equally spaced between 2.5 and 97.2^{th} percentiles, respectively, broadly following Harrell (2001)). All weights were stabilised using the baseline covariates V; the outcome model included the same covariates V plus an indicator for treatment and time in 2-yearly categories.

Since our preliminary model indicated some large weights (perhaps due to positivity or model misspecification), we decided to apply by default minimal truncation of the outer 0.1% of all final stabilised weights. By truncation of the outer p% of weights, we mean replacing those which are $< p^{th}$ or $> (100-p)^{th}$ percentiles with the p^{th} and $(100-p)^{th}$ percentiles, respectively. In a slight abuse of phrase, we will refer to this as "p% truncation". A common (though arbitrary) practice is to truncate the weights at a maximum of 10 (see for example HIV-CAUSAL Collaboration (2010)); 0.1% truncation roughly corresponded with a similar order of truncation across most models.

We used clustered sandwich estimators to estimate robust standard errors, since the esti-

mated weights induce correlation within patients. These estimators may be conservative, therefore we also calculated bootstrap confidence intervals for the main results using nonparametric resampling with 1000 replications.

The additional time-dependent covariates to be included in $L(t)$ were introduced in section 1.6.1. CD4 count decrease and HIV RNA were categorised as previously. The remaining time-dependent covariates were included as five knot splines (with knots at the 5, 25, 50, 75 and 95^{th} percentiles). In models where HIV RNA-related variables were included but the absence of any previous measurements was not captured by those variables, we also included a missing indicator for availability. For example, if the number of previous HIV RNA measurements was included in a model, then such an indicator was not required since it was captured by the value zero of the number of previous HIV RNA measurements. However, if for example the only HIV RNA related variable included was the last value, then the missing indicator was included. An indicator for whether the last CD4 count was carried forward (maximum 12 months; termed LOCF) was also considered for inclusion.

Since different guidelines or typical clinical practice across the different countries may impact on treatment decisions, we considered firstly an interaction between country and the key confounder CD4 count, and secondly separate treatment models for each country. Since only one German patient was observed to progress to AIDS or death, we combined the German patients with the "Other" category. Further, there were no Italian patients who met the criteria for being identified as HIV-infected close to seroconversion and were subsequently observed to initiate treatment, therefore we omitted this variable from the treatment model for Italy.

Application of the strategies

We applied the model building process and five strategies of section 2.3. Under decision 2, where working weights were truncated, we used 0.5% truncation since we did not want to be overzealous with the weight truncation at this stage; a different choice may have yielded different results but this choice will serve to illustrate the potential differences that may arise.

Under decision 3, we used one of the following two parameterisations, which were directly compared (that is, holding the other conditions the same) under strategies II and III, respectively:

(a) the i^{th} covariate of L_{pot} was eligible for inclusion if, compared to model M, it (i) moved the mean of the weights > 0.01 closer to one or (ii) moved the treatment effect estimate away from the unweighted estimate by $> 10\%$ (and > 0.05) or reduced the standard error

by $> 10\%$ (and > 0.05) but with mean of the weights ≤ 0.005 further from one.

(b) the i^{th} covariate of L_{pot} was eligible for inclusion if, compared to model M, it (i) moved the mean of the weights closer to one at all (practicably, say > 0.001) or (ii) moved the treatment effect estimate away from the unweighted estimate by $> 5\%$ (and > 0.05) or reduced the standard error by $> 5\%$ (and > 0.05) but with mean of the weights ≤ 0.01 further from one.

To put these parameterisations into context, the preliminary treatment model yielded weights with mean 1.133 and treatment effect estimate on the log-scale of -2.28 (standard error 0.40). Therefore, under criterion (a), covariate i of L_{pot} would be eligible for inclusion if it reduced the mean of the weights to < 1.123, or resulted in treatment effect estimate < -2.51 or with standard error < 0.35 provided the mean of the weights was ≤ 1.138. Under criterion (b), the variable would be eligible if it reduced the mean of the weights to < 1.132, or yielded treatment effect estimate < -2.39 or standard error < 0.35 but with mean of the weights ≤ 1.143.

These criteria impose the direction of change of the estimated treatment effect to be away from the unweighted estimate. The reason for this is that, based on prior knowledge about HIV treatment, we know the direction of the causal effect relative to the unweighted estimate (and supported by the preliminary treatment model with CD4 count alone). However, there may be concern that these criteria lead to a causal estimate that is too strong. Therefore we also introduced a sixth strategy (labelled strategy Ib) which was the same as the original strategy I (now labelled Ia) but with the following parameterisations for decision 3, which did not specify the direction of change of the estimated treatment effect relative to the unweighted estimate:

(c) the i^{th} covariate of L_{pot} was eligible for inclusion if, compared to model M, it (i) moved the mean of the weights > 0.01 closer to one or (ii) changed the treatment effect estimate relative to the unweighted estimate by $> 10\%$ (and > 0.05) or reduced the standard error by $> 10\%$ (and > 0.05) but with mean of the weights ≤ 0.005 further from one.

Under decision 4, we considered the following progressive truncations: 0.1, 0.5, 1, 2, 5 and 10%. As mentioned above, a common practice is to truncate weights at a maximum of 10; in our analyses, truncations of around 0.1 or 0.5% yielded weights with similar order maximum weights. We used the following two parameterisations of the criteria of section 2.3:

(a) truncate if it leads to a reduction (that is, weakening) in the estimated treatment effect of > 0.01 and $> 10\%$, provided no worsening of the mean in terms of absolute distance from one.

Strategy	Decision			
	1. Where to start?	2. Working weights? Truncated:	3. Covariate selection procedure? Parameterise $\{\mu, \phi, \sigma\}$ to favour:	4. Degree of weight truncation? Favour: (default 0.1%)
Ia	(a) minimal model	(a) 0.5%	(a) smaller model ($\mu = 0.01, \phi_p = 10$, $\sigma_p = 10, \mu_m = 0.005$)	(a) greater ($\varphi = 0.01, \varphi_p = 10$)
Ib	(a) minimal model	(a) 0.5%	(c) smaller model (as above but no direction for treatment effect)	(a) greater (as above)
II	(a) minimal model	(a) 0.5%	(a) smaller model (as top)	(b) less ($\rho = 0.01$, $\rho_p = 10, \varphi_m = 10$)
III	(a) minimal model	(a) 0.5%	(b) larger model ($\mu = 0.001, \phi_p = 5$, $\sigma_p = 5, \mu_m = 0.01$)	(b) less (as above)
IV	(a) minimal model	(b) -	(b) larger model (as above)	(b) less (as above)
V	(b) "full" model	(b) -	(b) larger model (as above)	(b) less (as above)
VI	HIV RNA and interaction with CD4 count			
VII	"Traditional" model-building approach			

Table 2.2: Summary of the treatment model building strategies applied to the CASCADE data. See text for more details.

(b) truncate if it leads to a reduction in standard error by > 0.01 and $> 10\%$, provided the reduction in the estimated treatment effect is $\leq 10\%$.

These were directly compared (that is, keeping the other criteria constant) under strategies I and II. Putting these parameterisations into context by applying them to the results from the preliminary treatment model, criterion (a) would lead to truncation if it yielded a treatment effect estimate > -2.05 provided the mean of the weights remained ≤ 1.133, whereas criterion (b) would lead to truncation if it yielded a standard error of < 0.36 provided the treatment effect estimate was ≤ -2.05.

None of the predefined strategies led to inclusion of any HIV RNA data, which we felt could be an important confounder and appeared to have a differing impact on treatment initiation by CD4 count, therefore we additionally considered a model with CD4 count, HIV RNA and their interaction (labelled strategy VI). Lastly, for comparison we also employed a "traditional" model building strategy of stepwise backwards selection (remove if $p > 0.05$, re-enter if $p < 0.01$) for comparison with our defined pre-strategies (labelled strategy VII). The strategies are summarised in Table 2.2.

Standard error estimation

We used robust standard errors throughout, but since these may be conservative we also bootstrapped with resampling stratified by country (1000 repetitions; grouped Italy with Others since few patients in Italy). We assumed fixed weights for all strategies; for two strategies (Ia and II) we also did a separate set of bootstraps re-estimating the weights each time to incorporate the uncertainty associated with estimating the weights.

Censoring

As indicated in section 2.3, the same process for the construction of the treatment model can be applied for that of the censoring process(es). Therefore we applied the same first six strategies as outlined above to construct inverse probability weights for the three different censoring mechanisms, starting with the same covariates of time, V and CD4 count as previously. We used the same L_{pot} for consideration for inclusion for all three censoring types except under censoring type 2 (irregular CD4 counts). By the definition of that censoring, the last CD4 count and most likely last HIV RNA measurement would be 12 months previously therefore we omitted the variables relating to time since last CD4 count and HIV RNA measurement, and also the indicator for LOCF (true by definition). Lastly, the usual CD4 decrease variable (by definition equal to zero when LOCF) was amended to take the value of the decrease in CD4 count when that variable was last measured 12 months previously. We also applied the "traditional" model building approach as an additional strategy.

The treatment weights from each strategy were multiplied together with the three sets of censoring weights from the matching strategy to form the overall weights for each strategy. Since there was no strategy VI (with CD4 count by HIV RNA interaction) considered for the censoring weights, the overall strategy VI weights were obtained using the censoring strategy IV weights. The degree of truncation of these overall weights was decided according to the relevant criteria for each strategy (decision 4).

Treatment effect modification

We investigated treatment effect modification by baseline covariates, by incorporating interactions between treatment and all the baseline covariates (except country) in the outcome model and applying a stepwise backward selection procedure (remove if $p > 0.05$, re-enter if $p < 0.01$). We allowed for non-linearity in continuous baseline covariates using splines (three knots at the 10, 50 and 90^{th} percentiles; Harrell (2001)), which were tested for non-linearity and included if

$p < 0.05$, otherwise linear. We lastly examined the interaction between treatment and country (combining German and Italian patients in the "Other" category since there were few events among those patients) as a model checking procedure (see section 2.3.4).

AIDS-free survival

Throughout, we report hazard ratios, which estimate the effect of ever versus never having received treatment, assuming a constant treatment effect regardless of the time spent on treatment. The comparable RCT would consist of sequential randomisations at each time-point. However, it is possible that the benefit of treatment may change with the time spent on treatment. Therefore we replaced the treatment indicator in the outcome models with time on treatment, categorised as < 0.5, $0.5- < 2$ and ≥ 2 years. From these models, we were able to estimate the standardised (by baseline covariates) survival curves for immediate versus no treatment. To do this, we estimated the predicted conditional probabilities of survival at each time t given survival through to time $t-1$, and multiplied across time to obtain the survival estimates. We did this firstly assuming all patients initiated treatment at baseline to represent immediate treatment, and secondly with the time on treatment set to zero for all time to represent no treatment; survival was estimated at every time-point for all patients regardless of when events or censoring was observed (Toh et al., 2010). This allowed us to plot the survival curves over time for immediate versus no treatment. We obtained 95% confidence intervals using bootstrap stratified by country (1000 repetitions).

2.4.2 Results

Demonstration of time-dependent confounding by CD4 count

Compared to < 200 cells/mm^3, current CD4 counts of $200-349$, $350-499$ and ≥ 500 cells/mm^3 were associated with a 80% (95% confidence interval 63, 89), 88% (78, 93) and 94% (90, 97) lower odds of AIDS or death, respectively. Therefore, CD4 count is a risk factor for the outcome, with lower CD4 counts associated with poorer outcome, as we would expect. Compared to < 200 cells/mm^3, CD4 counts of $200-349$, $350-499$ and ≥ 500 cells/mm^3 were associated with a 85% (77, 90), 97% (95, 98) and 99% (98, 99) lower odds of initiating treatment, respectively. Therefore low CD4 count predicted subsequent treatment, and we have demonstrated that CD4 count is a time-dependent confounder for AIDS or death (see section 1.1). Being on treatment predicted a 25 (22, 28) cells/mm^3 higher CD4 count in the next month, thus demonstrating that CD4 count is affected by prior treatment.

CD4 count, cells/mm^3	Number of patient-months	Initiated treatment?	
		No	Yes
< 50	29 ($< 1\%$)	27 (93.1%)	2 (6.9%)
$50 - 99$	34 ($< 1\%$)	25 (73.5%)	9 (26.5%)
$100 - 149$	87 ($< 1\%$)	59 (67.8%)	28 (32.2%)
$150 - 199$	217 ($< 1\%$)	167 (77.0%)	50 (23.0%)
$200 - 249$	763 (1%)	682 (89.4%)	81 (10.6%)
$250 - 299$	1640 (2%)	1535 (93.6%)	105 (6.4%)
$300 - 349$	3411 (4%)	3301 (96.8%)	110 (3.2%)
$350 - 399$	4885 (6%)	4782 (97.9%)	103 (2.1%)
$400 - 449$	6592 (8%)	6513 (98.8%)	79 (1.2%)
$450 - 499$	7093 (8%)	7032 (99.1%)	61 (0.9%)
$500 - 549$	10828 (13%)	10696 (98.8%)	132 (1.2%)
$550 - 599$	8973 (10%)	8875 (98.9%)	98 (1.1%)
$600 - 649$	8268 (10%)	8216 (99.3%)	52 (0.6%)
$650 - 699$	6155 (7%)	6106 (99.2%)	49 (0.8%)
$700 - 749$	5431 (6%)	5396 (99.4%)	35 (0.6%)
$750 - 799$	4739 (5%)	4710 (99.4%)	29 (0.6%)
$800 - 849$	3607 (4%)	3589 (99.5%)	18 (0.5%)
$850 - 899$	2837 (3%)	2830 (99.8%)	7 (0.3%)
$900 - 949$	2068 (2%)	2060 (99.6%)	8 (0.4%)
$950 - 999$	1962 (2%)	1957 (99.8%)	5 (0.3%)
≥ 1000	6626 (8%)	6605 (99.7%)	21 (0.3%)

Table 2.3: Pattern of treatment initiation across patient-months by CD4 count. Values are number of (previously treatment-naïve) patient-months and either (column) percentage of patient-months over all (previously treatment-naïve) follow-up for column 2 or (row) percentage of patient-months within that CD4 count category for columns 3 and 4.

Investigation of the positivity assumption

Treatment initiation was more likely at lower CD4 counts, as we would expect, but treatment initiations did and did not occur over a broad range of CD4 counts (Table 2.3). However, at high CD4 counts, the probability of treatment initiation was very low ($< 1\%$ for CD4 counts ≥ 600 cells/mm^3). For CD4 counts < 50 cells/mm^3, there were a surprisingly low number of treatment initiations; these results were driven by a small number of patients who either were not observed to initiate treatment or delayed treatment initiation despite having very low CD4 counts. We shall see that there were potential problems with the weights at lower CD4 counts, possibly due to non-positivity, but also perhaps due to residual unmeasured confounding in these "treatment refusers" or model misspecification. However, without resorting to the rather drastic approach of excluding these patients altogether, we cannot address these problems without moving to a dynamic modelling framework.

Looking at treatment initiations also by HIV RNA, but with broader CD4 count categories (Table 2.4), we can see that at low CD4 counts, participants were more likely to initiate treatment if they also had a high HIV RNA.

HIV RNA, copies/ml	CD4 count, cells/mm^3 < 200	$200 - 349$	$350 - 499$	≥ 500	Total
None available	1/7 (14%)	4/140 (3%)	5/582 (1%)	37/7507 (< 1%)	47/8236 (1%)
≤ 500	2/35 (6%)	5/180 (3%)	7/1029 (1%)	65/7280 (1%)	79/8524 (1%)
>500-2910	2/12 (17%)	10/425 (2%)	10/1874 (1%)	31/8800 (< 1%)	53/11111 (< 1%)
>2910-11820	1/39 (3%)	38/1088 (3%)	27/4480 (1%)	55/13682 (< 1%)	121/19289 (1%)
>11820-37743	10/49 (20%)	59/1619 (4%)	43/4979 (1%)	89/12929 (1%)	201/19576 (1%)
>37743-97809	24/77 (31%)	64/1096 (6%)	66/3257 (2%)	67/7278 (1%)	221/11708 (2%)
>97809	49/148 (33%)	116/1266 (9%)	85/2369 (4%)	110/4018 (3%)	360/7801 (5%)
Total	89/367 (24%)	296/5814 (5%)	243/18570 (1%)	454/61494 (1%)	1082/86245 (1%)

Table 2.4: Treatment initiations by CD4 count and HIV RNA. Values are n/N (%) where n=number of treatment initiations and N=number of (previously treatment-naïve) patient-months. HIV RNA categorised by 10, 25, 50, 75 and 90^{th} percentiles.

Functional form for CD4 count

6717 (8%) observed treatment-naïve CD4 counts were > 1000 cells/mm^3 (median 1158, maximum 2367 cells/mm^3) and therefore truncated to 1000 cells/mm^3. Figure 2.3 illustrates the odds of initiating treatment over the range of CD4 counts, compared with CD4 count of 450 cells/mm^3 (approximate median CD4 count at treatment initiation) for different functional forms of CD4 count. In general, as we would expect, the probability of treatment initiation was higher at lower CD4 counts, however the categorical plot clearly shows a decline in the probability of treatment at lower CD4 counts, which is not captured by any of the splines. As discussed above, this is likely to be due to a small subset of patients who repeatedly refused treatment. While the three knot spline displayed some evidence of poor fit at higher CD4 counts, the five knot spline appeared to capture the data well for the majority of the CD4 count range, and for parsimony we favoured this over the seven knot spline, which is similar. In order to attempt to address the sharp drop off in treatment initiation at CD4 counts < 100 cells/mm^3, we truncated CD4 counts < 100 cells/mm^3 to 100 cells/mm^3 ($n = 66$; median 50, minimum 5 cells/mm^3) and re-fit this "blunted" five knot spline, which essentially forced a constant probability of treatment initiation within that range. While this is somewhat arbitrary, it provided us with a better treatment prediction model in that it reduced somewhat the predicted probabilities of treatment for CD4 counts < 100 cells/mm^3 (Figure 2.3, green dashed line).

Naïve estimation of treatment effect

Table 2.5 illustrates the estimated treatment effects based on an unadjusted model (model only with treatment indicator and time), a model adjusted for baseline covariates, and a model adjusted for baseline covariates and time-dependent CD4 count. Under the unadjusted model, the point estimate suggests a benefit of treatment (HR=0.91) but this is not statistically significant

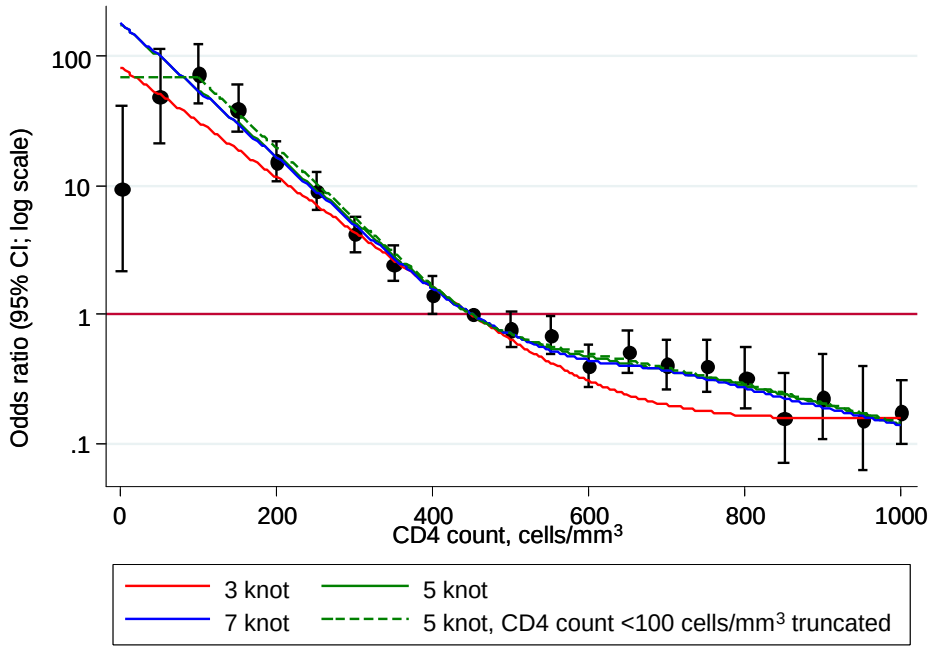


Figure 2.3: Treatment initiation by CD4 count, with CD4 count categorical or modelled as a three, five or seven knot spline.

Model	Hazard ratio for treatment effect (95% CI)
Unadjusted	0.91 (0.63, 1.32)
Adjusted for baseline covariates[1]	0.91 (0.61, 1.36)
Adjusted for baseline covariates and time-dependent CD4	2.58 (1.16, 5.71)

Table 2.5: Naïve estimation of treatment effect. [1] This result is the same as the unweighted estimate in the first row of Table 2.6.

(95% CI 0.63, 1.32). Adjusting for the baseline covariates does not materially alter the results. However, adjusting in addition for time-dependent CD4 count changes the results considerably (HR 2.58, 95% CI 1.16, 5.71), suggesting that treatment is harmful in terms of time to AIDS or death. This result is biased since it does not appropriately adjust for the time-dependent confounder CD4 which is also predicted by treatment history.

Estimation of the inverse probability of treatment weights

The overall results based on the preliminary treatment model of section 2.4.2, with CD4 count the only time-dependent covariate included, are shown in Table 2.6. The mean and maximum of the estimated weights were large at 1.133 and 1508, respectively; minimal truncation of the outer 0.1% controlled the weights well bringing the mean and maximum down to 1.052 and 26, respectively. This resulted in a more moderate estimated treatment effect, with odds ratio (OR) of 0.33 compared to 0.10 with untruncated weights. This indicated that there may be

issues with non-positivity in our dataset, and this observation led to the decision to perform 0.1% truncation by default regardless of the strategy. Compared to the unweighted treatment effect estimate, the OR was considerably further from one (0.33 compared to 0.91 unweighted), demonstrating control of confounding.

Strategy Ia The complete treatment model selection process for strategy Ia is illustrated in Table 2.7. At the first stage in strategy Ia, nadir CD4 count and number of previous CD4 counts met the criteria for inclusion under decision 3a (together with time-dependent CD4 count), since both brought the mean weights at least 0.01 closer to one; the latter was chosen since the mean of the weights was slightly closer to one (1.002 compared to 0.997 with nadir CD4 count and 1.013 without either, after 0.5% truncation according to decision 2a of strategy I). At the next step, only time since last CD4 count was eligible for inclusion; although the mean weights increased slightly, from 1.002 to 1.005, the OR moved further from one, from 0.56 to 0.52, suggesting perhaps better control of confounding. No subsequent variables were identified, therefore yielding a final treatment model with number of previous CD4 counts and time since last CD4 count, in addition to time, CD4 count and baseline covariates (Table 2.6).

Strategy Ib Strategy Ib was the same as Ia except that it did not specify direction of change of the estimated treatment effect in the model selection process. At the first step, two additional covariates (number of previous HIV RNA measurements and time since last HIV RNA) were identified, since they moved the estimated treatment effect $> 10\%$ towards the null, with mean of the weights within the permitted limits. However, since the covariate which most improves the mean of the weights is selected, number of previous CD4 counts was chosen as in strategy Ia. At the second stage, no additional covariates beyond time since CD4 count were identified, therefore was included as in strategy Ia. In contrast to strategy Ia, nadir CD4 count was identified for inclusion at the third step, since it moved the estimated treatment effect on the log-scale from -0.65 to -0.56. No further variables were identified, therefore yielding a final treatment model the same as that under strategy Ia but including also nadir CD4 count. Compared to strategy Ia, the estimated treatment effect was more moderate with an OR of 0.57 (SE 0.14) versus 0.52 (0.15). The means of the weights were a similar distance from one (0.994 under strategy Ib compared to 1.005 under strategy Ia; Table 2.6).

Strategy	Estimated weights				Estimated treatment effect			
	Time-dependent variables included	Truncation	Mean (SD)	Range	OR (SE[2])	95% CI Robust[2]	95% CI BS[3]	Log OR (SE[2])
Unweighted	-	-	-	-	0.91 (0.18)	0.61, 1.36	-	-0.10 (0.21)
Preliminary	(Time, V and CD4 count only)	None	1.133 (7.406)	0.03, 1508	0.10 (0.04)	0.05, 0.22	-	-2.28 (0.40)
		0.1%	1.052 (1.205)	0.04, 26	0.33 (0.13)	0.16, 0.70	-	-1.10 (0.38)
Ia	Time since last CD4 count, number of previous CD4 counts	0.5%	1.005 (0.706)	0.04, 6	0.52 (0.14)	0.31, 0.89	0.30, 0.90	-0.65 (0.27)
Ib	Time since last CD4 count, number of previous CD4 counts, nadir CD4 count	0.5%	0.994 (0.635)	0.04, 5	0.57 (0.15)	0.34, 0.95	0.33, 0.97	-0.56 (0.26)
II / III	As Ia	0.1%	1.025 (0.936)	0.03, 16	0.39 (0.13)	0.21, 0.75	0.21, 0.75	-0.93 (0.33)
IV	Time since last CD4 count, number of previous CD4 counts, nadir CD4 count, LOCF; stratified by country	0.1%	1.020 (0.895)	0.02, 11	0.60 (0.20)	0.30, 1.17	0.31, 1.15	-0.52 (0.34)
	[not stratified by country]	*0.1%*	*1.025 (0.982)*	*0.03, 19*	*0.40 (0.13)*	*0.21, 0.76*		*-0.91 (0.33)*
V	Time since last CD4 count, nadir CD4 count, LOCF; stratified by country	0.1%	1.020 (0.847)	0.02, 19	0.54 (0.18)	0.28, 1.05	0.28, 1.03	-0.62 (0.34)
	[not stratified by country]	*0.1%*	*1.020 (0.873)*	*0.03, 15*	*0.43 (0.14)*	*0.22, 0.82*		*-0.85 (0.33)*
VI	As IV except included CD4 count by HIV RNA interaction[1] and not stratified by country	0.1%	1.011 (1.287)	0.02, 28	0.39 (0.13)	0.21, 0.74	0.20, 0.78	-0.94 (0.32)
VII	Time since last CD4 count, number of previous CD4 counts, LOCF, HIV RNA, peak HIV RNA, time since last HIV RNA, number of HIV RNA measurements	0.5%	0.971 (0.813)	0.03, 8	0.43 (0.13)	0.23, 0.78	0.23, 0.78	-0.85 (0.31)

Table 2.6: Results from the strategies: treatment models, weights and estimated treatment effects. All treatment models included time as a 5 knot spline, the baseline covariates V and CD4 count as a spline as discussed in the text. SD=standard deviation. OR=odds ratio. SE=standard error. CI=confidence interval. LOCF=last (CD4) observation carried forward. [1] CD4 count and RNA categorical. [2] Robust SE calculated using clustered sandwich estimator, except for unweighted models since no weights to induce correlations. [3] Nonparametric bootstrap, 1000 replications.

Stage	Time-dependent covariate tested	Mean of weights		Treatment effect (log-odds scale)			SE of treatment effect			Eligible for inclusion[2]	Chosen for inclusion[2]
		Absolute	Absolute change towards 1[1]	Absolute	Change away from unweighted[1]		Absolute	Change[1]			
					Absolute	%		Absolute	%		
	-	1.013	-	-0.59	-	-	0.29	-	-	-	-
1	CD4 decrease	1.009	0.004	-0.57	0.02	3	0.28	-0.01	3	-	-
	Time since last CD4	1.006	0.007	-0.63	-0.04	7	0.28	-0.01	5	-	-
	Nadir CD4	0.997	0.010	-0.46	0.13	21	0.28	-0.01	3	Yes	-
	Number of previous CD4s	1.002	0.011	-0.58	0.01	1	0.26	-0.03	9	Yes	Yes
	Number of previous HIV RNAs	1.005	0.008	-0.49	0.10	17	0.28	-0.01	3	-	-
	Last HIV RNA[3]	0.968	-0.018	-0.56	0.02	4	0.31	0.02	8	-	-
	Time since last HIV RNA[3]	0.987	0.000	-0.49	0.09	16	0.28	-0.02	5	-	-
	Peak HIV RNA[3]	0.966	-0.021	-0.51	0.08	14	0.32	0.03	11	-	-
	LOCF	1.005	0.008	-0.59	-0.01	1	0.29	-0.01	2	-	-
2	CD4 decrease	1.006	-0.004	-0.60	-0.02	3	0.27	0.01	4	-	-
	Time since last CD4	1.005	-0.003	-0.65	-0.07	11	0.27	0.00	2	Yes	Yes
	Nadir CD4	0.996	-0.002	-0.53	0.05	8	0.26	0.00	1	-	-
	Number of previous HIV RNAs	0.994	-0.004	-0.59	-0.01	2	0.27	0.00	1	-	-
	Last HIV RNA[3]	0.965	-0.033	-0.61	-0.03	5	0.31	0.04	16	-	-
	Time since last HIV RNA[3]	0.987	-0.011	-0.52	0.06	11	0.27	0.01	2	-	-
	Peak HIV RNA[3]	0.978	-0.020	-0.63	-0.05	8	0.33	0.06	24	-	-
	LOCF	1.002	0.001	-0.59	-0.01	1	0.27	0.01	3	-	-
3	CD4 decrease	1.010	-0.005	-0.68	-0.03	5	0.27	0.00	0	-	-
	Nadir CD4	0.994	-0.001	-0.56	0.08	13	0.26	-0.01	3	-	-
	Number of previous HIV RNAs	0.994	-0.001	-0.67	-0.02	4	0.27	0.00	2	-	-
	Last HIV RNA[3]	0.971	-0.024	-0.72	-0.07	11	0.31	0.04	13	-	-
	Time since last HIV RNA[3]	0.992	-0.003	-0.59	0.06	9	0.27	0.00	0	-	-
	Peak HIV RNA[3]	0.972	-0.023	-0.70	-0.05	8	0.31	0.04	15	-	-
	LOCF	1.005	0.000	-0.64	0.01	1	0.27	0.00	1	-	-

Table 2.7: Demonstration of the treatment model building process for Strategy Ia. [1] Relative to the reference model. Reference model for Stage 1 is the basic model with time, baseline covariates and time-dependent CD4 only (results shown in the first row; these covariates are included in all subsequent models). Reference model for Stage 2 is that chosen at Stage 1, namely including number of previous CD4s. Reference model for Stage 3 is that chosen at Stage 2, namely including number of previous CD4s and time since last CD4. The procedure did not identify any further variables for inclusion. [2] See section 2.4.1 (sub-section entitled Application of the strategies) for the criteria determining eligibility and selection of covariates. [3] Also including an indicator for availability of HIV RNA data.

Strategy II Strategy II differed from Ia only in the degree of truncation performed at the end of the modelling process (decision 4), therefore used the same treatment model. However, strategy Ia led to 0.5% truncation (notably, the level at which the modelling was performed according to decision 2a) whereas strategy II suggested no truncation, but our default 0.1% truncation was applied. As we would expect, greater truncation under strategy Ia led to weights with mean closer to one (1.005 versus 1.025 after 0.5% (Ia) and 0.1% (II) truncation, respectively) and a more moderate estimated treatment effect (OR 0.52 versus 0.39, respectively; Table 2.6).

Strategy III Under strategy III, a number of covariates met the criteria for inclusion at the first stage (CD4 decrease, time since last CD4 count, nadir CD4 count, number of previous CD4 counts, number of previous HIV RNA measurements and LOCF) but number of previous CD4 counts was selected as under strategy Ia, and the subsequent covariate selection was as that of strategy Ia. Therefore strategy III yielded the same treatment model as strategy II, indicating that decision 3 relating to preference for a smaller or larger model did not make a difference in practice in this example.

While the CD4 count by country interaction was highly statistically significant ($p < 0.0001$), strategies Ia, Ib, II and III did not support inclusion of this interaction, nor of separate treatment models by country, according to decision 3. For example, under strategy Ia, the mean of the weights was slightly increased with separate treatment models by country (from 1.005 to 1.010, after 0.5% truncation) with no clear strengthening of the estimated treatment effect (OR 0.52 compared to 0.58 with separate treatment models).

Strategy IV Strategy IV differed from III in that the modelling process was performed using untruncated weights (decision 2). At the first step, a number of covariates were identified as eligible for inclusion under decision 3b (CD4 count decrease, time since last CD4 count, nadir CD4 count, number of CD4 counts, number of previous HIV RNA measurements and LOCF), of which time since last CD4 count was selected. At the second step, nadir CD4 count, number of previous CD4 counts and LOCF met the criteria for inclusion; nadir CD4 count was selected. At the third stage, only number of previous CD4 counts was eligible and so was included. At the fourth stage, LOCF was additionally identified and included; no further variables were subsequently identified. Thus strategy IV yielded a more complex treatment model than the previous strategies, incorporating nadir CD4 count and LOCF, in addition to time since last CD4 count and number of previous CD4 counts. Further, this strategy supported separate treatment models by country under decision 3b. Decision 4b indicated only the default

0.1% truncation. Compared to strategies II/III, the mean and maximum of the weights were slightly smaller under strategy IV (1.020 versus 1.025, and 11 versus 16, respectively; Table 2.6). Correspondingly, the estimated treatment effect was closer to the unweighted estimate (0.60 versus 0.39).

Strategy V The first four strategies all started with a minimal model with time, V and CD4 count only; in contrast, strategy V began with a "full" model including all potential time-dependent confounders (see section 2.4.1; decision 1b), with the remaining decisions reflecting those of strategy IV. At the first stage, CD4 count decrease, time since last CD4 count, time since last HIV RNA, peak HIV RNA and LOCF were identified for removal under decision 3b; peak HIV RNA was selected. The subsequent iterations led to the successive removal of last HIV RNA, CD4 count decrease, number of previous CD4 counts, time since last HIV RNA and number of previous HIV RNA measurements, thus yielding a model with time since last CD4 count, nadir CD4 count and LOCF. Once again, this strategy indicated separate treatment models by country under decision 3b, and decision 4b indicated only the default 0.1% truncation. Therefore this model was the same as that under strategy IV, except it did not include number of previous CD4 counts, and the results were similar, with mean of the weights 1.020 and estimated treatment effects 0.54 (versus 1.020 and 0.60, respectively, under strategy IV; Table 2.6).

There was clear overlap in the different treatment models across the strategies: all contained time since last CD4 count and there was no variable which made an appearance in only one strategy. Across all models, there was evidence of nonlinearity for all the included covariates (test of the spline components, all $p \leq 0.05$). In contrast to I-III, strategies IV and V indicated separate treatment models by country according to decision 3. For comparison, the results from strategies IV and V, but with one overall treatment model across all countries, are also given in Table 2.6. The results from strategies II/III, IV and V with one overall treatment model were fairly similar (mean of the weights 1.025, 1.025 and 1.020; ORs 0.39, 0.40 and 0.43, respectively). However, allowing separate treatment models in strategies IV and V changed the results fairly dramatically (mean of the weights both 1.020; ORs more modest at 0.60 and 0.54, respectively).

Strategy VI Interestingly, none of the treatment models included any HIV RNA-related variables as main effects. High HIV RNA is a known predictor for faster pre-treatment disease progression (Mellors et al., 1996) and Figure 2.4 illustrates that high HIV RNA was associated

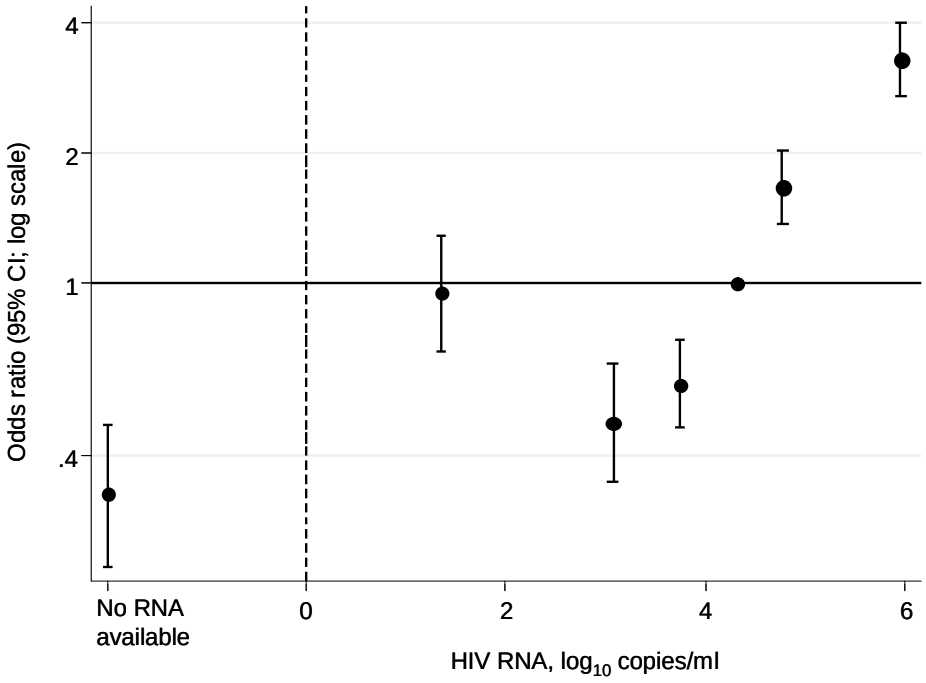


Figure 2.4: Treatment initiation by HIV RNA. CI=confidence interval.

with higher probability of treatment initiation; we have seen that there may be a differential association by CD4 count (see Table 2.4). Of note, no previous HIV RNA measurement was associated with low probability of treatment initiation, probably related to fewer clinic visits. Therefore we considered an additional treatment model, based on that of strategy IV (the largest model), but in addition incorporating an interaction between CD4 count and HIV RNA (both categorical). A few of the weights were exceptionally large (> 10000) likely due to positivity issues; after 0.1% truncation, HIV RNA met the criteria for inclusion under decision 3b since it reduced the mean of the weights to 1.007 (compared to 1.025 in the model with categorical CD4 count alone). Although the interaction did not meet the criteria for inclusion (after 0.1% truncation, the mean of the weights increased slightly to 1.011 and the treatment effect was little changed), it was highly statistically significant ($p = 0.0006$). This formed our model under strategy VI. We did not consider separate treatment models by country under this approach due to limited numbers of patients in each CD4 count/HIV RNA category within each country. The estimated OR was very similar to that under strategy IV without stratification by country (0.39 [SE 0.32 on the log-odds scale] versus 0.40 [0.33]; Table 2.6), perhaps indicating that HIV RNA is not an important confounder.

Strategy VII The standard model building approach with stepwise backwards selection removed CD4 count decrease and nadir CD4 count ($p = 0.05$ and 0.07, respectively), yielding a

model with time since last CD4 count, number of previous CD4 counts, LOCF, last HIV RNA, peak HIV RNA, time since last HIV RNA, and number of HIV RNA measurements. The untruncated weights were somewhat unwieldy, with mean 42 and maximum > 1000000. Applying decision 4a to favour truncation, 0.5% truncation was preferred, giving mean weights of 0.971 (maximum 8) and OR 0.43 (Table 2.6).

Summary To summarise, we have derived four treatment models from our six original strategies: one each from Ia/II/III, Ib, IV and V. In addition, we have two models which are more complex: one incorporating an interaction between CD4 count and HIV RNA (strategy VI), and one from a "traditional" model building approach (strategy VII). Therefore we have six treatment models in total. All strategies resulted in the default 0.1% weight truncation, except strategies Ia, Ib and VII with 0.5% truncation.

The (robust) standard errors of the treatment effect estimates were very similar for strategies II-VII (ranging from 0.31 to 0.33 on the log odds scale). The standard errors from strategies Ia and Ib, where there was greater truncation, were somewhat smaller at 0.27 and 0.26, respectively. In contrast, the standard error from the preliminary model, which had a large mean of the weights, was larger at 0.40 with no truncation and 0.38 after 0.1% truncation. These standard errors were all larger than that from the unweighted model (0.20) by the nature of being based on weighted estimation. Where estimated, bootstrap confidence intervals were very similar to the confidence intervals based on a robust standard error (Table 2.6) and the medians of the bootstrapped estimates were broadly similar to the overall point estimates. Where estimated, the bootstrap confidence intervals which re-estimated the weights were fairly similar (although slightly larger as expected) to those which assumed fixed weights ((0.30, 0.92) and (0.20, 0.76) for strategies Ia and II/III, respectively, versus (0.30, 0.90) and (0.21, 0.75), respectively).

Predictors of treatment initiation

For illustration, the treatment model from strategies Ia, II and III is summarised in Table 2.8 and the model used for the numerator of the stabilised weights (the same across all strategies except estimated separately by country under strategies IV and V) is summarised in Table 2.9. Of note, while we have given standard errors and p-values here for reference, these were deliberately omitted by Petersen, Deeks, Martin, and van der Laan (2007) in order to emphasise that only the point estimates are relevant in terms of contributing to the estimated weights.

In both models, higher baseline HIV RNA, no available baseline HIV RNA, earlier year of HIV seroconversion and shorter time HIV-infected at baseline were associated with higher prob-

Variable	Odds ratio	Standard error	p
Time-dependent covariates			
CD4 count, cells/mm^3	[1]	-	< 0.0001
Number of previous CD4 counts	[2]	-	< 0.0001
Time since last CD4 count, months	[2]	-	< 0.0001
Baseline covariates (V)			
Baseline HIV RNA, $\log_{10}$ copies/ml	1.29	0.06	< 0.001
Baseline HIV RNA not available	1.67	0.35	0.01
Baseline CD4 count, per 100 cells/mm^3	1.03	0.02	0.21
Sex, female	0.95	0.08	0.54
Age at HIV seroconversion, per 10 years	1.06	0.04	0.10
Year of HIV seroconversion	0.87	0.01	< 0.001
Route of HIV transmission, IDU	0.82	0.10	0.12
Country, versus France			< 0.0001
Germany	0.43	0.15	
Italy	0.65	0.10	
Spain	0.77	0.11	
Switzerland	0.97	0.14	
UK	0.37	0.03	
Others	0.46	0.06	
Time HIV-infected at baseline, years	0.89	0.04	0.01
Identified as HIV-infected close to seroconversion	0.85	0.12	0.23

Table 2.8: Results from the treatment model: denominator with time-dependent and baseline covariates for strategies Ia, II and III. Time modelled as a spline. [1] Not illustrated; similar to the spline illustrated previously in Figure 2.3 from the preliminary treatment model. [2] Modelled as a spline; see Figure 2.5.

Variable	Odds ratio	Standard error	p
Baseline HIV RNA, $\log_{10}$ copies/ml	1.56	0.07	< 0.001
Baseline HIV RNA not available	3.41	0.70	< 0.001
Baseline CD4 count, per 100 cells/mm^3	0.81	0.02	< 0.001
Sex, female	0.99	0.08	0.89
Age at HIV seroconversion, per 10 years	1.11	0.04	0.002
Year of HIV seroconversion	0.90	0.01	< 0.001
Route of HIV transmission, IDU	0.97	0.12	0.80
Country, versus France			< 0.0001
Germany	0.43	0.15	
Italy	0.66	0.10	
Spain	0.76	0.10	
Switzerland	0.82	0.11	
UK	0.50	0.04	
Others	0.68	0.08	
Time HIV-infected at baseline, years	0.91	0.04	0.02
Identified as HIV-infected close to seroconversion	0.77	0.10	0.06

Table 2.9: Results from the treatment model: numerator with baseline covariates only. Time modelled as a spline.

ability of treatment initiation, but there was no association between treatment initiation and either sex or route of HIV transmission. Compared to France, the odds of treatment initiation were typically lower in the other countries, particularly Germany, the UK and "Others".

In the numerator model, there were a number of other covariates which were associated with treatment initiation, namely lower baseline CD4 count, older age and not being identified as HIV-infected close to seroconversion. Further, lack of a baseline HIV RNA measurement was more strongly predictive of treatment initiation in the numerator model with a relatively large odds ratio of 3.41. We examined more closely the 711 patients with no baseline HIV RNA: a relatively large proportion were from Spain (19%) and the UK (24%). A high proportion of these patients were IDU (20% compared to 9% overall), they tended to be slightly younger (median [IQR] age at HIV seroconversion 29 [25, 34] years), seroconverted earlier (1994 [1993, 1995]) and were HIV-infected for a relatively long time before entering our study (2.1 [1.4, 3.3] years). Therefore perhaps these patients were at a later stage of disease by the time they entered our study in a way that is not entirely captured by the other covariates, and hence more likely to initiate treatment.

From the denominator model, the splines for the continuous time-dependent covariates number of previous CD4 count measurements and time since last CD4 count measurement are illustrated in Figure 2.5. A higher number of previous CD4 count measurements was associated with higher probability of treatment initiation. Either a short or long time since last CD4 count measurement was associated with a higher probability of treatment initiation, the former probably due to having had a recent clinic visit at which treatment decisions would be made, and the latter perhaps because a large gap between CD4 count measurements indicated poorer health of the patient and hence treatment was initiated or perhaps because a clinic visit did occur but we are missing a recorded CD4 measurement.

Distribution of the treatment weights

It is of interest to know from where the largest weights are arising. As discussed above, for ease we have so far considered the mean of the weights over all time intervals, but in general we expect the stabilised weights to have mean one across all time intervals. Figure 2.6 illustrates the mean and range of the weights from the six final treatment models arising from the eight strategies, plotted over time in yearly categories; Figure 2.7 is the same except after 0.1% truncation of the weights. It is clear from Figure 2.6 that there are some very large weights occurring, most noticeably under strategies Ia/II/III (which shared the same treatment model), Ib, VI, and, exceptionally, VII. While the outer 0.1 percentiles clearly depend on the distribution of the

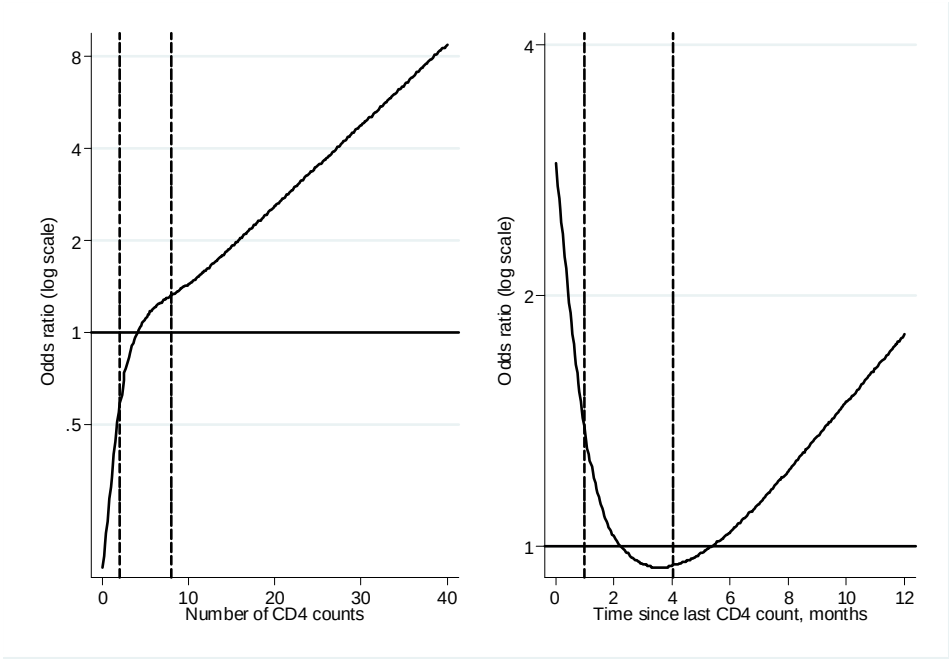


Figure 2.5: Treatment initiation by number of previous CD4 count measurements and time since last CD4 count measurement. The vertical dashed lines indicate the interquartile range.

weights for each strategy, Figure 2.7 shows that even after this relatively minimal truncation, the weights are much more well-behaved and more similar across the strategies, although still somewhat larger under strategy VII.

Across these six treatment models, the upper 0.1 percentile of the weights came from 405 patient-months in 20 patients: 9 French, 1 German, 1 Italian, 1 Spanish, 7 UK and 1 Danish. There was nothing remarkable about the baseline characteristics of these patients, except they had slightly lower median (IQR) baseline CD4 count of 601 (540, 630) cells/mm^3 compared to 641 (560, 788) cells/mm^3 across all patients and they tended to be early seroconverters with median (IQR) year of HIV seroconversion 1996 (1993, 1998) compared to 2000 (1995, 2003) across all patients.

The vast majority of the large weights were due to non-initiation at low CD4 counts (typically with high HIV RNA). Where patients were observed to eventually initiate treatment, the weights then dropped, though in two cases (both French) the weights from strategies VI and VII remained in the upper 0.1 percentile (at 127 and 222 for strategies VI and VII, respectively, for one patient; and at 28 and 321, respectively, for the second patient) and so were carried forward for the rest of follow-up (approximately 3 and 4.5 years, respectively). If the weights were large under one model, they tended to be inflated across all models, though only two patients had weights in the upper 0.1 percentile across all six treatment models (both French, due to no or delayed treatment initiation at low CD4 counts; one of these patients was the first one

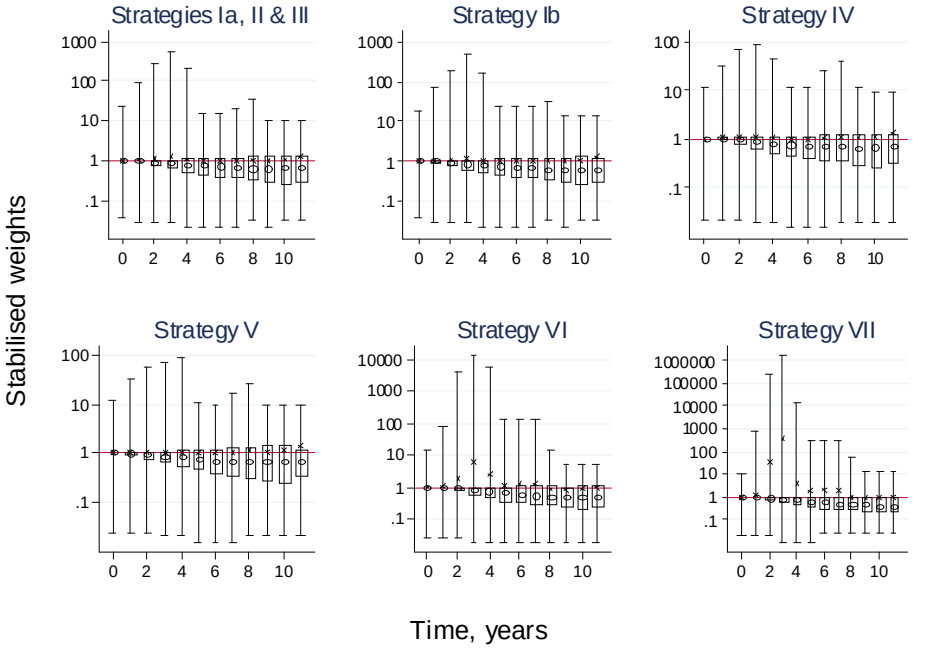


Figure 2.6: Distribution of the estimated stabilised weights for the five treatment models. Spikes = range, bars = interquartile range, o = median, x = mean. Note that the scales of the y-axes vary.

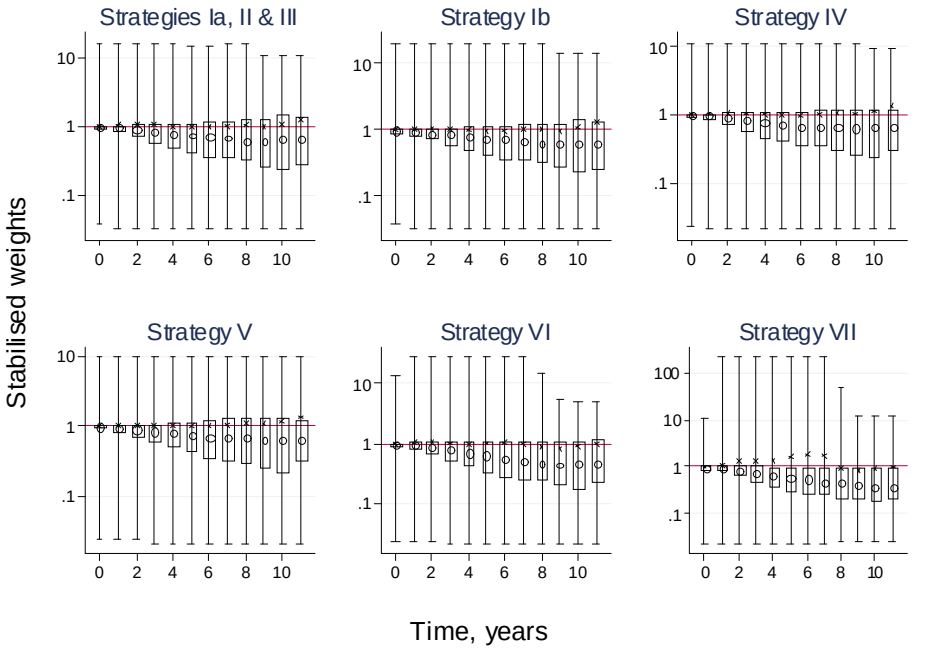


Figure 2.7: Distribution of the estimated stabilised weights for the five treatment models, after truncation of the outer 0.1 percentiles. Spikes = range, bars = interquartile range, o = median, x = mean. Note that the scales of the y-axes vary.

mentioned above who had large weights carried forward after eventually initiating treatment).

Three patients (German, Spanish and UK) had large weights due to initiation of treatment at high CD4 counts (684, 760 and 690 cells/mm^3, respectively); by definition these weights were then carried forward for the rest of follow-up (approximately 2, 5 and 3.5 years, respectively). However, the size of these weights was relatively small compared to those flagged due to non-initiation at low CD4 counts (German patient: strategy IV weight = 11, strategy V weight = 12; Spanish patient: strategy Ib weight = 24; UK patient: strategy V weight = 11).

One French patient had a somewhat odd CD4 count pattern of 820 followed by 192 (at which point they received a large weight of 41 under strategy VI for non-initiation at such a low CD4 count), then 840 and finally initiated treatment at 570 cells/mm^3 (at which point the weight under strategy VI increased to 77, and also weights from strategies Ib and IV were inflated at 23 and 16, respectively, due to initiation at a high CD4 count; these weights were then carried forward for the remaining 1.5 year follow-up). There was one further French patient whose CD4 count pattern appeared questionable: (s)he had successive CD4 counts of 624, 69 and 567 cells/mm^3 within the space of four months and received a weight of 19 under strategy Ia (in the upper 0.1 percentile) due to non-initiation at CD4 count of 69 cells/mm^3; this weight remained large for the remaining four year follow-up during which time the patient remained treatment-naïve but had quite variable CD4 counts ranging from 323 to 1006 cells/mm^3. While we could have excluded the low CD4 count of 69 cells/mm^3 on the grounds of implausibility, the resulting weights are not overly large and unlikely to affect the overall conclusions, particularly after weight truncation.

Seventeen of the 20 patients with large weights were censored before progressing to AIDS or death; the remaining three progressed to AIDS while treatment-naïve with low CD4 counts. One was the French patient mentioned above who remained treatment-naïve with exceptionally large weights across all six treatment models (562, 502, 84, 69, 10,358 and $> 1,000,000$ under strategies Ia, Ib and IV-VII, respectively), the second had large weights under strategies Ia, Ib, VI and VII (470, 249, 13,001 and $> 100,000$, respectively), and the third had a large weight under strategy V only (26).

Strategy	France	Germany	Italy	Spain	Switzerland	UK	Others
Ia, II, III	1.105 (562)	0.991 (4)	1.243 (470)	1.294 (15)	0.965 (6)	0.902 (36)	0.980 (15)
Ib	1.090 (502)	0.992 (4)	1.118 (249)	1.375 (24)	0.973 (6)	0.893 (33)	0.974 (15)
IV	1.052 (84)	1.136 (11)	0.939 (5)	1.127 (9)	1.076 (9)	0.967 (39)	0.947 (20)
V	1.043 (92)	1.125 (12)	0.941 (5)	1.168 (10)	1.101 (9)	0.969 (43)	0.958 (13)
VI	2.070 (10358)	0.973 (2)	6.215 (13001)	1.169 (8)	0.973 (6)	0.898 (40)	0.921 (14)
VII	74.215 (1727602)	0.952 (2)	76.218 (198234)	1.103 (5)	0.969 (5)	0.941 (51)	0.916 (15)

Table 2.10: Summary of mean (maximum) weights by country, across the different strategies (no truncation).

Strategy	France	Germany	Italy	Spain	Switzerland	UK	Others
Ia	1.035 (6)	0.991 (4)	0.974 (6)	1.194 (6)	0.965 (6)	0.894 (6)	0.971 (6)
Ib	1.026 (5)	0.992 (4)	0.962 (5)	1.163 (5)	0.973 (5)	0.882 (5)	0.965 (5)
II, III	1.054 (16)	0.991 (4)	0.992 (16)	1.294 (15)	0.965 (6)	0.899 (16)	0.980 (15)
IV	1.041 (11)	1.136 (11)	0.939 (5)	1.127 (9)	1.076 (9)	0.959 (11)	0.945 (11)
V	1.031 (10)	1.102 (10)	0.941 (5)	1.168 (10)	1.101 (9)	0.959 (10)	0.957 (10)
VI	1.063 (28)	0.973 (2)	0.932 (28)	1.169 (8)	0.973 (6)	0.896 (28)	0.921 (14)
VII	0.999 (8)	0.952 (2)	0.935 (8)	1.104 (5)	0.969 (5)	0.890 (8)	0.911 (8)

Table 2.11: Summary of mean (maximum) weights by country, across the different strategies with truncation as per the strategy (0.5% under strategies Ia, Ib and VII; 0.1% under the rest).

Tables 2.10 and 2.11 show the means and maxima of the weights by country, without truncation and with truncation as per the strategy, respectively. The very largest weights appeared mainly from France, and also Italy under strategies I-III, VI and VII. Where the treatment model was stratified by country, the weights were much more well-behaved. After truncation, the means of the weights were generally centred on one and the maxima much more tolerable (generally around 5-15, although >20 under strategy VI in France, Italy and the UK). However, there were some clear differences by country, with the mean of the weights always less than one for Italy, UK and Others, and always greater than one for France and Spain. It is difficult to determine why this might be, but could indicate residual confounding.

Censoring

For all three censoring types (1, LTFU; 2, irregular CD4 counts with a gap of more than 12 months; 3, administrative), strategies Ia, II and III did not add any further variables to the preliminary model incorporating just time, baseline covariates V and CD4 count. These results are shown in Table 2.12; note that this table is to illustrate the effects of incorporating censoring weights compared to the unweighted model and there is no treatment weighting. Further, no truncation of the weights has yet been performed; this will be done according to the criteria of each strategy after the treatment and censoring weights have been combined. The estimated weights from strategies Ia, II and III were centred on one with mean (SD) 1.000 (0.022), 1.003 (0.119) and 1.000 (0.012) for the censoring types 1, 2 and 3, respectively, and the estimated treatment effects were identical to 2 decimal places to the unweighted effect estimate of 0.91 (SE 0.19). This could perhaps indicate that none of the censorings were very informative. However, there were some differences in the censoring models under the other strategies.

For censoring type 1 (LTFU), strategy Ib did not include any further covariates. In contrast, strategy IV added in succession: number of HIV RNA measurements, LOCF and CD4 decrease; this yielded weights with similar mean (0.999 compared to 1.000 under strategies Ia, II and III) but with much larger maximum weights (112 compared to 1.3). The standard error of the treatment effect estimate was similar though the point estimate was somewhat smaller than the unweighted (0.86 compared to 0.91). Strategy V led to a more complex censoring model, with only time since last CD4 count and CD4 decrease removed from the "full" model, leaving nadir CD4 count, number of previous CD4 counts, number of previous HIV RNA measurements, last HIV RNA, time since last HIV RNA measurement, peak HIV RNA and LOCF. The mean of the weights was somewhat increased (1.007), as was the maximum (204), although the treatment effect estimate was similar to the unweighted estimate (OR 0.93, SE 0.20). Strategy VII (step-

wise backwards selection) removed nadir CD4 count and peak HIV RNA ($p = 0.64$ and 0.08, respectively) to leave another complex model with CD4 decrease, time since last CD4 count, number of previous CD4 counts, last HIV RNA, number of previous HIV RNA measurements, time since last HIV RNA measurement and LOCF. However, this model yielded weights with smaller mean (0.959) although once again the estimated treatment effect was similar to the unweighted estimate (OR 0.90, SE 0.19).

For censoring type 2 (irregular CD4 counts), strategy Ib added CD4 decrease (the amended variable to capture CD4 decrease at the last CD4 count which by definition was observed 12 months ago), but at the expense of weights with large mean at 1.087 (maximum 117). The point estimate for the OR of treatment effect was above one (OR 1.15, SE 0.25). Strategy IV introduced only peak HIV RNA. The mean of the weights remained centred on one (1.000) with maximum 10 and the estimated treatment effect was similar to the unweighted estimate (OR 0.94, SE 0.20). Strategy V led to the removal of (amended) CD4 decrease only, therefore yielding another complex model with nadir CD4 count, number of previous CD4 counts, number of previous HIV RNA measurements, last HIV RNA and peak HIV RNA (recall, a slightly different set L_{pot} was used for this censoring type; see section 2.4.1). The weights remained centred on one (mean 1.000) but the maximum increased hugely to 1374, perhaps raising concerns of model misspecification or non-positivity. The estimated treatment effect was close to one (OR 1.00) and poorly estimated (SE 0.24; 0.25 on the log-odds scale). Strategy VII successively removed last HIV RNA, peak HIV RNA and nadir CD4 count ($p = 0.33$, 0.63 and 0.23, respectively), leaving a model with (amended) CD4 decrease, number of previous CD4 counts and number of previous HIV RNA measurements. The mean and maximum of the weights increased considerably, to 1.260 and 1586, respectively, and the estimated treatment odds ratio was 1.29 (SE 0.34); this may raise concerns about model misspecification or non-positivity.

For censoring type 3 (administrative), strategies Ib, IV and V also led to the simple model with no additional covariates. Strategy VII successively removed number of previous CD4 counts, LOCF, last HIV RNA, number of previous HIV RNA measurements, nadir CD4 count and peak HIV RNA ($p = 0.82$, 0.80, 0.36, 0.28, 0.14 and 0.06, respectively), leaving a model with CD4 decrease, time since last CD4 count and time since last HIV RNA measurement. The estimated weights and treatment effects were broadly similar to those from the simpler model with just time, the baseline covariates V and CD4 count (mean of the weights 1.000, OR 0.90, SE 0.18).

Censoring type	Strategy	Estimated weights			Estimated treatment effect		
		Time-dependent variables included	Mean (SD)	Range	OR (SE)[1]	95% CI[1]	Log OR (SE)[1]
-	Unweighted	-	-	-	0.91 (0.19)	0.61-1.36	-0.10 (0.21)
1 (LTFU)	I-III	(Time, V and CD4 count only)	1.000 (0.022)	0.76-1.3	0.91 (0.19)	0.61-1.36	-0.09 (0.21)
	IV	Number of previous HIV RNA measurements, LOCF, CD4 decrease	0.999 (0.675)	0.14-112	0.86 (0.18)	0.57-1.30	-0.15 (0.21)
	V	Nadir CD4 count, number of previous CD4 counts, number of previous HIV RNA measurements, last HIV RNA, time since last HIV RNA measurement, peak HIV RNA, LOCF	1.007 (1.622)	0.15-204	0.93 (0.20)	0.61-1.43	-0.07 (0.22)
	VII	CD4 decrease, time since last CD4 count, number of previous CD4 counts, last HIV RNA, number of previous HIV RNA measurements, time since last HIV RNA measurement, LOCF	0.959 (0.315)	0.02-36	0.90 (0.19)	0.60-1.35	-0.10 (0.21)
2 (irregular CD4 counts)	Ia, II, III	(Time, V and CD4 count only)	1.003 (0.119)	0.44-3	0.91 (0.19)	0.61-1.36	-0.10 (0.21)
	Ib	(Amended) CD4 decrease[2]	1.087 (2.495)	0.12-117	1.15 (0.25)	0.74-1.77	0.14 (0.22)
	IV	Peak HIV RNA	1.000 (0.204)	0.29-10	0.94 (0.20)	0.62-1.42	-0.06 (0.21)
	V	Nadir CD4 count, number of previous CD4 counts, number of previous HIV RNA measurements, last HIV RNA, peak HIV RNA	1.000 (4.778)	0.02-1374	1.00 (0.24)	0.62-1.61	0.00 (0.25)
	VII	(Amended) CD4 decrease[2], number of previous CD4 counts, number of previous HIV RNA measurements	1.260 (19.162)	0.02-1586	1.29 (0.34)	0.77-2.16	0.26 (0.26)
3 (administrative)	I-V	(Time, V and CD4 count only)	1.000 (0.012)	0.58-2	0.91 (0.19)	0.61-1.36	-0.10 (0.21)
	VII	CD4 decrease, time since last CD4 count, time since last HIV RNA measurement	1.000 (0.045)	0.39-4	0.90 (0.18)	0.60-1.34	-0.11 (0.21)

Table 2.12: Results from the strategies: censoring models, weights and estimated treatment effects. SD=standard deviation. OR=odds ratio. SE=standard error. CI=confidence interval. LOCF=last (CD4) observation carried forward. [1] Robust SE calculated using clustered sandwich estimator, except for unweighted models since no weights to induce correlations. [2] Amended variable to capture CD4 decrease at the last CD4 count which by definition was 12 months ago.

We applied the same techniques as for the treatment model to investigate whether to stratify the censoring models by country. There were few Swiss patients with no baseline HIV RNA and of those all were LTFU, therefore we omitted this variable from the Swiss censoring type 1 models. There were very few patients from Italy, Switzerland or "Others" who had a change in CD4 count before being LTFU, therefore the covariate CD4 decrease was omitted from those country censoring type 1 models under strategy VI. Similarly under strategy V, there were few Swiss patients who were LTFU in each of the last HIV RNA categories, therefore that covariate was omitted from that Swiss censoring type 1 model. Lastly, all Italian or Spanish patients who were identified as HIV-infected close to seroconversion were administratively censored, therefore this covariate was removed from those country censoring type 3 models. However, none of the censoring types indicated stratifying by country under any of the strategies I-V.

Distribution of the censoring weights As outlined above, for censoring type 1 (LTFU), there were four censoring models (Table 2.12). Across these models, the upper 0.1 percentile of the weights came from 461 patient-months in 108 patients. These patients broadly matched the overall cohort demographics, although were more likely to be female (29% versus 20% overall) and infected through IDU (14% versus 9%). All these patients were eventually censored (92 LTFU, two due to irregular CD4 counts, 14 administratively). Only one patient had any weights >100; this patient was from Spain, was not observed to initiate treatment despite CD4 dropping to around 360 cells/mm^3, and received weights > 100 at month 25 under strategy V and at month 30 under strategy IV, although all the weights remained < 200 until administrative censoring at 31 months.

For censoring type 2 (irregular CD4 counts), the upper 0.1 percentile of the weights across the five censoring models came from 506 patient-months in 63 patients. As for censoring type 1, these patients were more likely to be female (32%) and infected through IDU (27%). In addition, these patients tended to be younger (median 29 years old versus the overall median of 31), seroconverted earlier (1995 versus 2000) and were less likely to have a baseline HIV RNA measurement (available for 51% versus 79%). The summary of the weights in Table 2.12 illustrates that the maxima varied considerably across the different models. There were seven patients who had any censoring type 2 weights > 100: six were French and one was from Norway (grouped under Other countries). None of these patients were observed to reach AIDS or death; five were LTFU, one was censored due to irregular CD4 counts and one was administratively censored. Six of these patients typically had high CD4 counts and were not observed to initiate treatment; all had large weights under strategies V and/or VII. Three of these patients had

Strategy	Estimated weights			Estimated treatment effect		
	Truncation	Mean (SD)	Range	OR (SE)[1]	95% CI[1]	Log OR (SE)[1]
Unweighted	-	-	-	0.91 (0.18)	0.61, 1.35	-0.10 (0.20)
Ia	0.5	1.000 (0.676)	0.04, 6	0.54 (0.14)	0.32, 0.90	-0.62 (0.26)
Ib	1	0.995 (0.702)	0.05, 5	0.63 (0.16)	0.38, 1.04	-0.46 (0.26)
II/III	0.1	1.030 (1.075)	0.03, 23	0.36 (0.12)	0.19, 0.70	-1.02 (0.34)
IV	0.1	1.031 (1.221)	0.02, 26	0.50 (0.20)	0.22, 1.11	-0.69 (0.41)
V	0.1	1.022 (2.247)	0.01, 50	0.29 (0.14)	0.11, 0.75	-1.25 (0.49)
VI	0.1	1.071 (2.583)	0.02, 66	0.35 (0.10)	0.19, 0.62	-1.06 (0.30)
VII	0.5	0.898 (1.108)	0.02, 12	0.32 (0.10)	0.17, 0.60	-1.14 (0.32)

Table 2.13: Results from the strategies: combined treatment and censoring weights and estimated treatment effects. SD=standard deviation. OR=odds ratio. SE=standard error. CI=confidence interval. [1] Robust SE calculated using clustered sandwich estimator, except for unweighted models since no weights to induce correlations.

just a single month with large weights before censoring occurred. In the other three patients, where large weights were observed over longer follow up, the weights escalated in size quite quickly. For example, one patient first received weights > 100 under both strategies V and VII at month 30, of values 135 and 117, respectively, which then increased to 1374 and 1024 by month 33 before being censored (LTFU). The one remaining patient, who was observed to initiate treatment at 5.5 years, had large weights > 100 occurring under strategy Ib from approximately 8 years onwards but these weights always remained < 120 until administrative censoring at approximately 11 years.

Across the two censoring models for censoring type 3 (administrative), the upper 0.1 percentile of the weights came from 237 patient-months in 136 patients. However, the maxima of the weights were just 2 (under the model applied to all but strategy VII) and 4 (under strategy VII). This is as we might expect for administrative censoring, which we would anticipate to be independent of patients' characteristics.

Treatment effect estimates

After obtaining the overall weights for each strategy by combining the relevant treatment and censoring weights, as previously strategy Ia led to 0.5% truncation and strategies II, IV and V indicated only the default 0.1% truncation, whereas strategy Ib indicated 1% truncation. As discussed previously, 0.1% and 0.5% truncation was applied under strategies VI and VII, respectively. The overall weights and resulting treatment effect estimates are summarised in Table 2.13 and illustrated in Figure 2.8.

Compared to the results when just incorporating treatment weights (Table 2.6), the results for strategies Ia, Ib, II/III, IV and VI were broadly similar, although the confidence interval

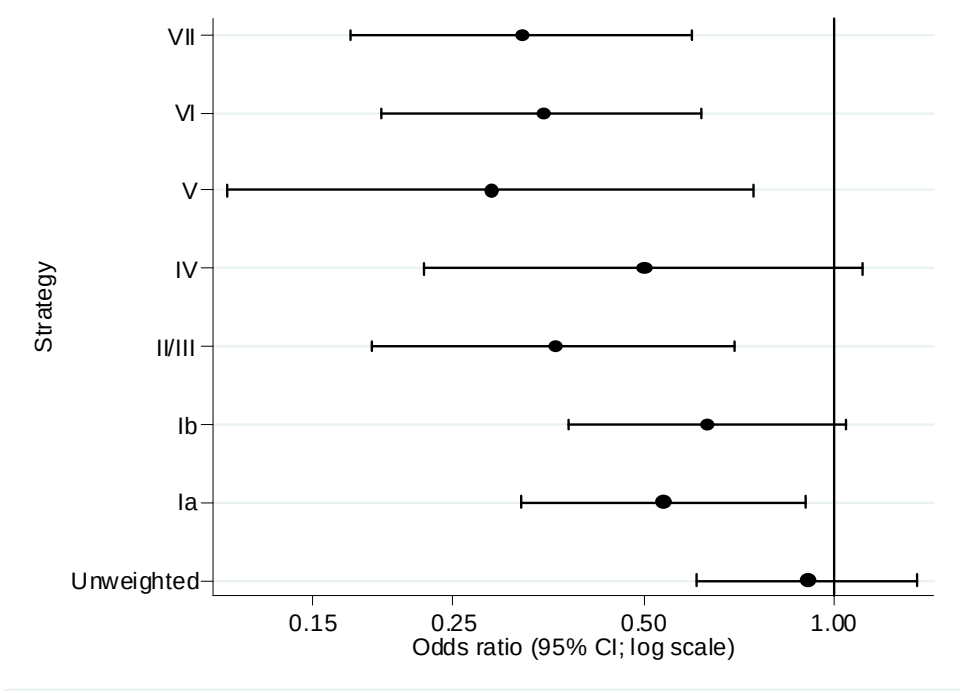


Figure 2.8: Estimated treatment effect on time to AIDS or death across the modelling strategies.

for strategy Ib after incorporating censoring weights contained one. The estimated treatment effects for strategies V and VII were somewhat more extreme, perhaps unsurprisingly since those two strategies had the most complex censoring models and therefore may have had better control for confounding due to censoring (or could perhaps be bias). This was at the expense of an increase in the standard error for strategy V (from 0.34 to 0.49 on the log-odds scale), although due to the large change in the estimated treatment effect, the confidence interval still excluded one.

All the strategies appeared to demonstrate considerable control for confounding, with the point estimates having moved away from the unweighted estimate. There was a trend towards stronger estimated treatment effects with higher strategy number (that is, those designed to be more complex), but overall the strategies led to broadly consistent results, with overlapping confidence intervals and all but strategies Ib and VI indicating a statistically significant benefit (at the 5% level) of treatment in delaying time to AIDS or death. The OR point estimates ranged from 0.29 (95% CI 0.11, 0.75) to 0.63 (0.38, 1.04), corresponding to a 37% to 71% reduction in the hazard of AIDS or death with treatment compared to no treatment. Strategies IV and V had considerably larger standard errors at 0.41 and 0.49 on the log-odds scale, compared to a maximum of 0.34 under the other strategies; this is probably due to the stratification of the treatment models by country. Strategy Ib led to a more moderate estimated treatment effect than strategy Ia; this could be related to our reason for introducing this strategy, namely that

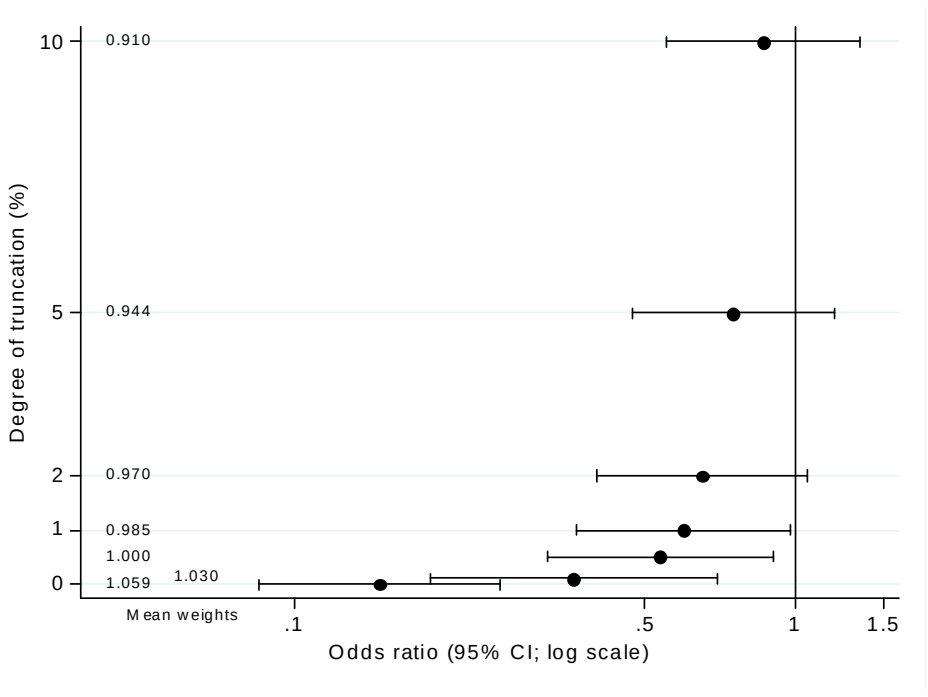


Figure 2.9: Effect of progressive truncation of the weights on the mean of the weights and the estimated treatment effect (treatment model from strategy Ia).

incorporating the direction of movement of the estimated treatment effect away from the null (as in strategy Ia) may lead to causal effect estimates that are too strong. However, strategy Ib led to more complex treatment and censoring models than strategy Ia, and therefore greater truncation was required to bring the weights under control; this was perhaps at the expense of control for confounding.

The only difference between strategies Ia and II/III was the degree of truncation of the weights. Exploring this further, Figure 2.9 illustrates the effect of progressive truncation of the weights from strategy Ia, which results in smaller mean weights and smaller estimated treatment effects, with 2% truncation yielding an estimated treatment effect which is no longer statistically significant at the 5% level, and 10% truncation yielding an estimated treatment effect which is similar to the unweighted estimate and therefore indicating little if any control for confounding. The largest jump in the odds ratio is seen between no and 0.1% truncation; the truncation may be to some extent protecting against bias due to non-positivity or model misspecification, or may be demonstrating poorer control for confounding. One might argue that 0.5% truncation may be preferred since it results in weights most closely centred on one (mean 1.000). However, as argued previously, the extreme weights capture the most information with respect to time-dependent confounding (Cole et al., 2005) and the intermediate 0.1% truncation offers perhaps a satisfactory compromise, although this is inevitably a subjective decision.

Predictors of outcome

Looking at the outcome models more closely, Table 2.14 summarises the results after applying weights from strategies Ia, Ib, II/III and IV. Under strategy Ia, being IDU and male were predictive of poor outcome, with OR 2.53 (95% CI 1.51, 4.22) for IDU versus not, and 0.58 (0.34, 1.00) for female versus male. There was a suggestion that higher baseline HIV RNA, no baseline HIV RNA, older age at HIV seroconversion, earlier year of HIV seroconversion, longer time HIV-infected at baseline and being identified as HIV-infected close to seroconversion were also all associated with poorer outcome, but not statistically significantly so (at the 5% level). There was no evidence of a difference in AIDS-free survival by country ($p = 0.73$), although the ORs ranged from 0.50 (0.09, 3.71) for Germany to 1.48 (0.85, 2.57) for the UK, compared to France. There was no association between baseline CD4 count and time to AIDS or death ($p = 0.52$), probably because the time-dependent CD4 count which is taken into account via the weighting is more important.

Except for the treatment effect estimate, the outcome results after weighting according to strategies Ib, II/III (same treatment model as for strategy Ia but with 0.1% instead of 0.5% truncation) and IV were broadly similar to those under strategy Ia (Table 2.14). The results from strategy V were also broadly similar, although the association of the lack of baseline HIV RNA (OR 9.37 [95% CI 1.88, 46.79]) and being identified as HIV-infected close to seroconversion (5.17 [1.40, 19.10]) with AIDS-free survival increased considerably though the confidence intervals were wide, and IDU was no longer associated with the outcome (1.34 [0.65, 2.74]; Table 2.15; key results to note indicated with an asterisk).

The results from strategy VI were somewhat different. There was no suggestion of an association between baseline HIV RNA or its availability (p =1.00 and 0.94, respectively; Table 2.15), probably because strategy VI incorporated time-updated HIV RNA therefore baseline HIV RNA became less important. Lower baseline CD4 was associated with faster progression (OR 0.91 [95% CI 0.83, 1.00]) and there was an indication of different AIDS-free survival by country (p =0.04) with the ORs ranging from 0.39 (0.05, 2.96) in Germany to 1.72 (0.90, 3.27) in Italy, compared to France. The results from strategy VII were broadly similar to those under strategies II/III (Table 2.15).

Covariate	Strategy Ia			Strategy Ib			Strategy II/III			Strategy IV		
	OR (SE)	95% CI	p	OR (SE)	95% CI	p	OR (SE)	95% CI	p	OR (SE)	95% CI	p
Treatment	0.54 (0.14)	0.32, 0.90	0.02	0.63 (0.16)	0.38, 1.04	0.07	0.36 (0.12)	0.19, 0.70	0.002	0.50 (0.20)	0.22, 1.11	0.09
Time, versus 0-<2 years			0.24			0.11			0.11			0.09
2-<4	1.05 (0.30)	0.60, 1.85		1.01 (0.28)	0.59, 1.74		1.54 (0.56)	0.75, 3.14		1.35 (0.58)	0.58, 3.13	
4-<6	0.75 (0.24)	0.41, 1.40		0.76 (0.25)	0.40, 1.44		0.71 (0.25)	0.36, 1.41		0.66 (0.23)	0.33, 1.33	
6-<8	1.78 (0.67)	0.86, 3.71		1.97 (0.72)	0.96, 4.05		1.64 (0.69)	0.72, 3.73		2.71 (1.19)	1.15, 6.41	
8-<10	0.48 (0.30)	0.14, 1.65		0.38 (0.24)	0.11, 1.29		0.47 (0.30)	0.13, 1.66		0.57 (0.41)	0.14, 2.34	
$\geq$10	0.70 (0.67)	0.10, 4.62		0.40 (0.38)	0.06, 2.54		0.73 (0.71)	0.11, 4.91		0.60 (0.59)	0.09, 4.09	
Baseline HIV RNA, $\log_{10}$ copies/ml	1.31 (0.22)	0.95, 1.81	0.10	1.31 (0.20)	0.96, 1.78	0.09	1.47 (0.30)	0.98, 2.19	0.06	1.30 (0.21)	0.94, 1.80	0.11
Baseline HIV RNA not available	2.63 (1.78)	0.70, 9.88	0.15	2.72 (1.80)	0.74, 9.97	0.13	3.97 (2.94)	0.93, 16.98	0.06	3.57 (2.52)	0.89, 14.26	0.07
Baseline CD4 count, per 100 cells/mm^3	1.02 (0.04)	0.95, 1.10	0.52	1.04 (0.04)	0.96, 1.11	0.35	0.98 (0.04)	0.90, 1.07	0.72	1.02 (0.04)	0.94, 1.11	0.63
Sex, female	0.58 (0.16)	0.34, 1.00	0.05	0.69 (0.21)	0.38, 1.26	0.23	0.43 (0.14)	0.23, 0.82	0.01	0.48 (0.15)	0.26, 0.87	0.02
Age at HIV seroconversion, per 10 years	1.21 (0.13)	0.98, 1.49	0.07	1.22 (0.13)	1.00, 1.50	0.05	1.10 (0.14)	0.86, 1.42	0.44	1.09 (0.14)	0.84, 1.41	0.54
Year of HIV seroconversion	0.93 (0.05)	0.83, 1.03	0.15	0.94 (0.05)	0.85, 1.04	0.23	0.89 (0.06)	0.77, 1.02	0.09	0.91 (0.06)	0.81, 1.03	0.15
Route of HIV transmission, IDU	2.53 (0.66)	1.51, 4.22	<0.001	2.77(0.71)	1.67,4.57	<0.001	2.50 (0.81)	1.32, 4.71	0.005	2.32 (0.63)	1.36, 3.94	<0.001
Country, versus France			0.73			0.38			0.69			0.52
Germany	0.50 (0.51)	0.09, 3.71		0.53 (0.55)	0.07, 3.98		0.46 (0.48)	0.06, 3.54		0.42 (0.43)	0.06, 3.17	
Italy	1.12 (0.56)	0.42, 2.97		1.21 (0.57)	0.48, 3.04		1.67 (0.85)	0.62, 4.55		0.58 (0.32)	0.19, 1.73	
Spain	0.92 (0.31)	0.48, 1.76		0.88 (0.31)	0.45, 1.74		0.63 (0.26)	0.28, 1.40		0.63 (0.24)	0.30, 1.34	
Switzerland	0.93 (0.40)	0.40, 2.16		0.93 (0.40)	0.40, 2.18		0.90 (0.40)	0.38, 2.15		0.95 (0.49)	0.34, 2.61	
UK	1.48 (0.42)	0.85, 2.57		1.65 (0.45)	0.97, 2.80		1.27 (0.50)	0.59, 2.74		1.23 (0.53)	0.52, 2.88	
Others	1.32 (0.43)	0.69, 2.51		1.55 (0.48)	0.84, 2.84		0.90 (0.40)	0.38, 2.15		0.78 (0.40)	0.28, 2.13	
Time HIV-infected at baseline, years	1.10 (0.13)	0.87, 1.39	0.42	1.09 (0.12)	0.87, 1.36	0.46	1.24 (0.15)	0.98, 1.58	0.07	1.08 (0.12)	0.87, 1.35	0.48
Identified as HIV-infected close to seroconversion	1.53 (0.73)	0.60, 3.88	0.37	1.39 (0.63)	0.57, 3.40	0.47	2.87 (1.73)	0.88, 9.37	0.08	3.35 (2.44)	0.80, 13.98	0.10

Table 2.14: Results from outcome model for time to AIDS or death, with weighting according to strategies Ia, Ib, II/III and IV.

Covariate	Strategy V			Strategy VI			Strategy VII		
	OR (SE)	95% CI	p	OR (SE)	95% CI	p	OR (SE)	95% CI	p
Treatment	0.29 (0.14)	0.11, 0.75	0.01	0.35 (0.10)	0.19, 0.62	<0.001	0.32 (0.10)	0.17, 0.60	<0.001
Time, versus 0-<2 years			0.06			<0.001			0.18
2-<4	1.29 (0.73)	0.43, 3.92		2.01 (0.58)	1.14, 3.54		1.26 (0.47)	0.61, 2.61	
4-<6	0.55 (0.28)	0.20, 1.51		0.67 (0.24)	0.33, 1.37		0.69 (0.28)	0.31, 1.54	
6-<8	1.61 (1.01)	0.47, 5.50		3.31 (1.59)	1.29, 8.50		2.51 (1.53)	0.76, 8.32	
8-<10	0.26 (0.19)	0.06, 1.09		0.66 (0.49)	0.15, 2.82		0.70 (0.55)	0.15, 3.29	
$\geq$10	2.23 (2.51)	0.25, 20.14		0.24 (0.21)	0.05, 1.30		3.64 (3.94)	0.44, 30.42	
Baseline HIV RNA, $\log_{10}$ copies/ml	2.07 (0.61)	1.16, 3.69	0.01	1.00 (0.25)*	0.61, 1.64	1.00	1.22 (0.37)	0.68, 2.20	0.51
Baseline HIV RNA not available	9.37 (7.69)*	1.88, 46.76	0.01	0.93 (0.94)*	0.13, 6.79	0.94	1.37 (1.61)	0.14, 13.77	0.79
Baseline CD4 count, per 100 cells/mm^3	1.05 (0.06)	0.94, 1.17	0.43	0.91 (0.04)*	0.83, 1.00	0.04	1.00 (0.05)	0.91, 1.10	0.98
Sex, female	0.59 (0.19)	0.31, 1.10	0.10	0.26 (0.09)	0.13, 0.50	<0.001	0.52 (0.15)	0.30, 0.91	0.02
Age at HIV seroconversion, per 10 years	0.97 (0.17)	0.69, 1.35	0.84	1.04 (0.12)	0.83, 1.30	0.75	1.14 (0.16)	0.87, 1.50	0.35
Year of HIV seroconversion	0.79 (0.09)	0.63, 1.00	0.05	0.89 (0.06)	0.78, 1.03	0.12	0.89 (0.07)	0.77, 1.04	0.14
Route of HIV transmission, IDU	1.34 (0.49)*	0.65, 2.74	0.43	3.20 (0.85)	1.90, 5.38	<0.001	2.47 (0.75)	1.37, 4.47	0.003
Country, versus France			0.41			0.04*			0.73
Germany	0.37 (0.39)	0.04, 3.02		0.39 (0.40)	0.05, 2.96		0.48 (0.50)	0.06, 3.71	
Italy	0.57 (0.33)	0.19, 1.77		1.72 (0.56)	0.90, 3.27		1.52 (0.70)	0.61, 3.76	
Spain	0.70 (0.28)	0.32, 1.51		0.41 (0.15)	0.20, 0.83		0.81 (0.29)	0.41, 1.62	
Switzerland	0.63 (0.38)	0.19, 2.03		1.14 (0.73)	0.32, 3.98		0.78 (0.43)	0.26, 2.31	
UK	1.30 (0.71)	0.44, 3.79		0.97 (0.35)	0.49, 1.95		1.37 (0.53)	0.64, 2.94	
Others	0.50 (0.32)	0.15, 1.73		0.65 (0.21)	0.34, 1.24		0.83 (0.31)	0.39, 1.73	
Time HIV-infected at baseline, years	1.36 (0.21)	1.00, 1.85	0.05	1.23 (0.14)	0.99, 1.55	0.07	1.32 (0.18)	1.02, 1.72	0.04
Identified as HIV-infected close to seroconversion	5.17 (3.45)*	1.40, 19.10	0.01	4.88 (1.34)	2.84, 8.37	<0.001	2.97 (1.47)	1.12, 7.86	0.03

Table 2.15: Results from outcome model for time to AIDS or death, with weighting according to strategies V, VI and VII. Key results to note which are discussed in the text are indicated with an asterisk.

Treatment effect modification by baseline covariates

There was evidence of nonlinearity in the outcome models for baseline HIV RNA and time HIV-infected at baseline under strategies VI and VII ($p = 0.0002$ and 0.003, respectively, under strategy VI, and $p = 0.02$ and 0.02, respectively, under strategy VII; p-values are for the test for the spline components), but not for any other baseline covariates or strategies. For consistency, to investigate the interactions between treatment and baseline covariates, we included baseline HIV RNA and time HIV-infected at baseline as splines in all strategies, and the other continuous baseline covariates as linear.

The interactions between treatment and baseline covariates as identified by the stepwise backward selection procedure are summarised in the first 3 columns of Table 2.16. There was quite a range of different interactions identified under the different strategies.

Under strategy Ia, interactions with treatment were identified for year of seroconversion and lack of a baseline HIV RNA measurement ($p = 0.03$ for both). The estimated OR (95% CI) for the effect of treatment for a patient who seroconverted in the median year 2000 and with a baseline HIV RNA measurement was 0.72 (0.37, 1.40). For a comparable patient who seroconverted one year later, the estimated treatment effect was 0.87 (0.40, 1.91). For a comparable patient who seroconverted in 2000 but without a baseline HIV RNA measurement, the estimated treatment effect was 2.73 (0.69, 10.9). The reasons for these effects are not clear, though there is a great deal of uncertainty and there may perhaps be some residual confounding.

Under strategies II/III, there was evidence of treatment effect modification by time HIV-infected at baseline, with the suggestion of a greater benefit of treatment the less time infected at baseline (Figure 2.10). This may be related to those identified closer to seroconversion generally having poorer prognosis (Tyrer et al., 2003), although these patients have been infected for at least one year before inclusion in this study. Relatedly, those infected longer at study entry have survived AIDS-free longer with high CD4 counts, therefore perhaps do not benefit as greatly from treatment as those identified closer to infection.

Under strategy V, there was initially evidence of treatment interactions with baseline HIV RNA, age at and year of seroconversion, and whether identified as HIV-infected close to seroconversion. However, if incorporating baseline HIV RNA, then it is necessary to include the indicator for availability of such a measurement; when incorporating interaction between treatment and the indicator, the interaction with baseline HIV RNA was no longer statistically significant ($p = 0.1$) therefore this was removed from the model; the rest of the interactions remained statistically significant. Therefore, for a patient with median year of seroconversion

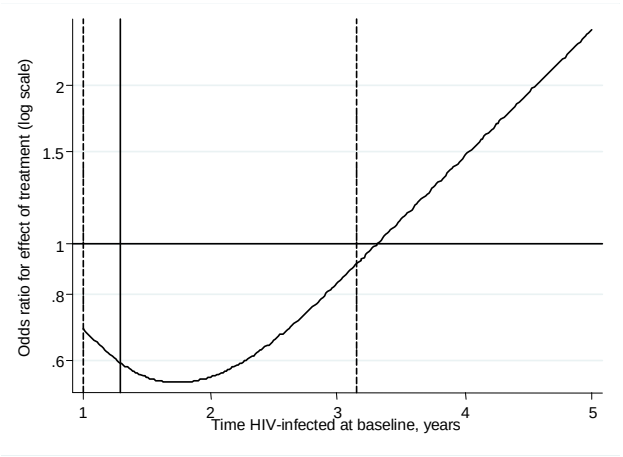


Figure 2.10: Odds ratio for estimated effect of treatment by length of time HIV-infected at baseline. Solid vertical line=median; dashed vertical lines=interquartile range.

(2000), a baseline HIV RNA measurement, median age at seroconversion (31 years) and who was not identified as HIV-infected close to seroconversion, the estimated treatment effect was 0.43 (0.19, 1.00). For a comparable patient who seroconverted one year later, the estimated treatment effect was 0.59 (0.22, 1.61), that is, in the same direction as seen under strategy Ia. For a comparable patient with no baseline HIV RNA measurement, the estimated treatment effect was 6.85 (0.58, 81.0); once again, in the same direction as seen for strategy Ia. For a comparable patient who was identified as HIV-infected close to seroconversion, the estimated treatment effect was 0.06 (0.01, 0.33). The size of this result is somewhat surprising, but patients who are identified as HIV-infected close to seroconversion are a small selected subset and this is a known surrogate for subsequent poorer prognosis (Tyrer et al., 2003), therefore we may expect a greater benefit of treatment in these patients who otherwise fare poorly. Finally, for a comparable patient who was 10 years older at seroconversion, the estimated treatment effect was 0.86 (0.34, 2.16). Since we would expect older patients to have poorer prognosis in general, this does not tie in with our argument above that those who would otherwise fare poorly benefit the most from treatment. However, our finding is in agreement with previous studies which have shown better immunological and clinical response to treatment in younger persons (Collaboration of Observational HIV Epidemiological Research Europe (COHERE) Study Group, 2008).

Strategy	Model incorporating interactions between treatment and: baseline covariates but not country		country but not baseline covariates	baseline covariates (see second column) and country	
	Baseline covariate	$p_{baseline}$	$p_{country}$	$p_{baseline}$	$p_{country}$
Unweighted	-	-	0.52	-	-
Ia	Year of seroconversion	0.03	0.29	0.03	0.36
	Lack of baseline HIV RNA measurement	0.03		0.03	
Ib	-	-	0.23	-	-
II/III	Time HIV-infected at baseline	0.02	0.28	0.01	0.37
IV	-	-	0.05	-	-
V	Year of seroconversion	0.03	0.40	0.03	0.20
	Lack of baseline HIV RNA measurement	0.01		0.01	
	Age at HIV seroconversion	0.01		0.02	
	Identified as HIV-infected close to SC	0.03		0.02	
VI	Age at HIV seroconversion	0.01	0.02	0.01	0.01
	Identified as HIV-infected close to SC	0.01		0.003	
VII	-	-	0.10	-	-

Table 2.16: Treatment effect modification by baseline covariates, across the different strategies. All continuous baseline covariates included as linear, except baseline HIV RNA and time HIV-infected at seroconversion which were included as splines. SC=seroconversion. Subscripts for the p-values indicate what interaction with treatment was being tested.

Under strategy VI, there was evidence of treatment interactions with age at HIV seroconversion and whether identified as HIV-infected close to seroconversion. For a patient with median age at seroconversion (31 years) and who was not identified as HIV-infected close to seroconversion, the estimated treatment effect was 0.45 (0.26, 0.79). For a comparable patient who was 10 years older at seroconversion, the estimated treatment effect was 0.76 (0.40, 1.45); this was 0.07 (0.02, 0.28) for a comparable patient who was identified as HIV-infected close to seroconversion. These results are similar to those seen under strategy V.

There was no evidence of treatment effect modifications under strategies Ib, IV or VII. The estimated treatment effects under strategies Ib and IV were somewhat closer to 1, therefore perhaps making the detection of interactions unlikely.

Interaction between treatment and country Finally, we considered interactions between treatment and country. As discussed in section 2.3.4, we anticipated that there should be no such interaction. However, we found evidence of such an interaction under strategies IV and VI ($p = 0.05$ and 0.02, respectively; Table 2.16), and these interactions remained even when taking into account other interactions with treatment where indicated (for example, age at seroconversion and whether identified as HIV-infected close to SC under strategy VI). Since the treatment-by-country interactions remained broadly similar regardless of whether other interactions were taken into account, and in order to compare across the different strategies by country, we proceeded with the models with interactions between treatment and country only, for illustrative purposes.

The estimated treatment effects by country for all of the strategies are illustrated in Figure 2.11 (without any other interactions included). The point estimates for treatment effect were somewhat different across the strategies albeit with wide confidence intervals. Note that across all strategies and in all countries, the weighted models yielded estimated treatment effects further from the null than the unweighted model, demonstrating control of confounding.

The only difference between strategies Ia (orange) and II/III (green) was the degree of truncation. To investigate this further, consider Figure 2.12 which illustrates the results from these strategies and in addition with no truncation. In line with previous results, and as we would anticipate, progressive truncation resulted in more moderate estimated treatment effects across all countries, except Switzerland where the results were consistent regardless of the degree of truncation. This is because, as we have seen above, Switzerland has fairly stable weights and is little affected whether 0.1 or 0.5% truncation is performed.

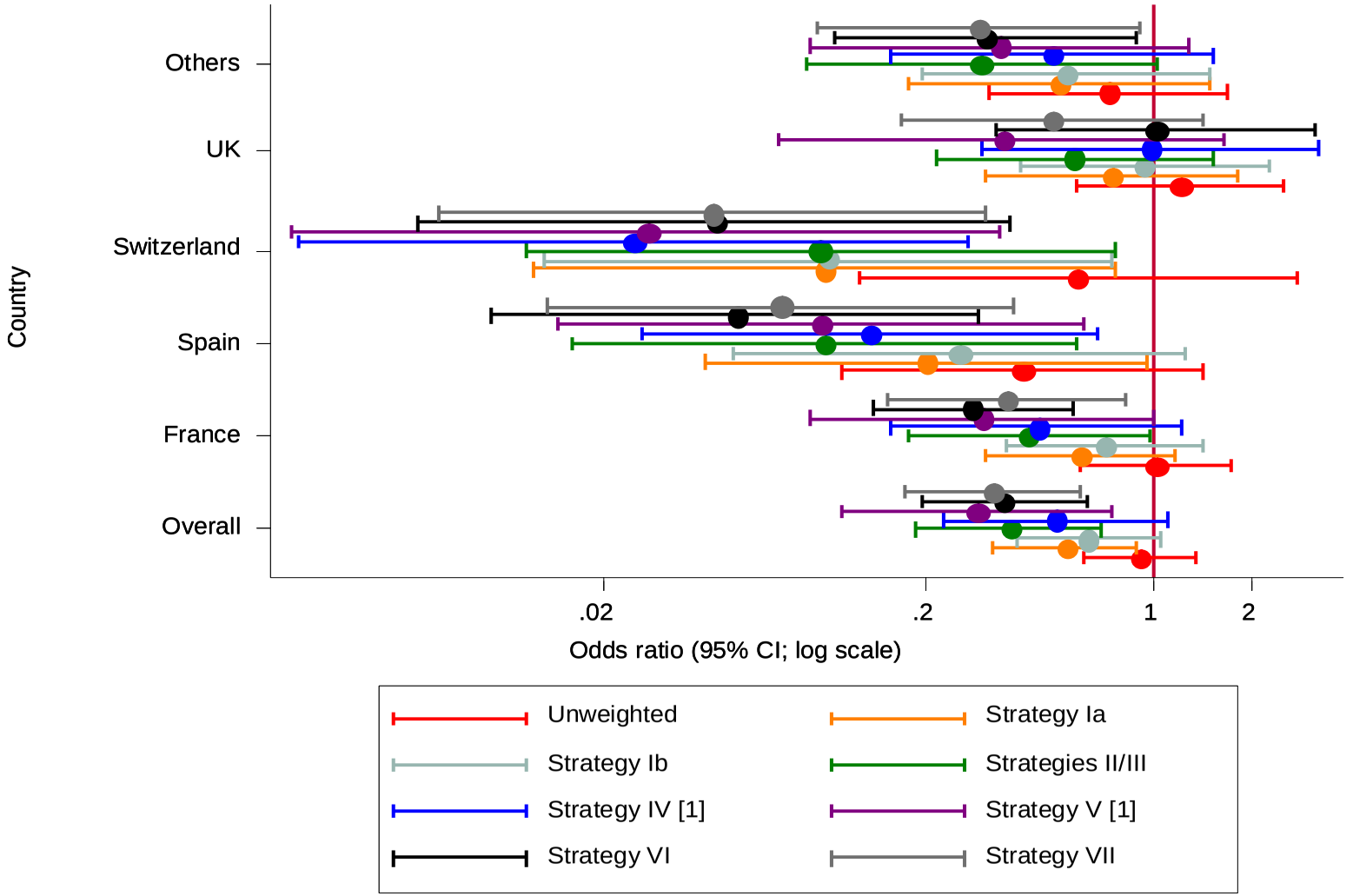


Figure 2.11: Estimated effect of treatment on time to AIDS or death by country. [1] Treatment model stratified by country.

Time, years	Strategy	No treatment	Immediate treatment	Difference
3	Unweighted	0.96 (0.95, 0.96)	0.97 (0.96, 0.98)	0.01 (0.002, 0.03)
	Ia	0.95 (0.94, 0.96)	0.97 (0.96, 0.98)	0.02 (0.009, 0.04)
	Ib	0.95 (0.94, 0.96)	0.97 (0.96, 0.98)	0.02 (0.007, 0.03)
	II/III	0.94 (0.93, 0.96)	0.98 (0.97, 0.99)	0.04 (0.02, 0.05)
	IV	0.95 (0.93, 0.97)	0.97 (0.96, 0.99)	0.02 (0.002, 0.05)
	V	0.94 (0.91, 0.96)	0.98 (0.96, 0.99)	0.04 (0.01, 0.07)
	VI	0.93 (0.91, 0.95)	0.97 (0.95, 0.99)	0.04 (0.02, 0.07)
	VII	0.94 (0.93, 0.96)	0.98 (0.97, 0.99)	0.04 (0.02, 0.05)
6	Unweighted	0.91 (0.89, 0.93)	0.93 (0.91, 0.96)	0.02 (-0.01, 0.06)
	Ia	0.91 (0.89, 0.94)	0.95 (0.93, 0.97)	0.04 (0.009, 0.07)
	Ib	0.92 (0.90, 0.94)	0.95 (0.93, 0.97)	0.03 (0.002, 0.06)
	II/III	0.90 (0.87, 0.93)	0.96 (0.94, 0.98)	0.06 (0.02, 0.09)
	IV	0.91 (0.89, 0.94)	0.95 (0.92, 0.98)	0.04 (-0.005, 0.08)
	V	0.90 (0.86, 0.94)	0.96 (0.93, 0.99)	0.06 (0.01, 0.11)
	VI	0.88 (0.85, 0.91)	0.95 (0.92, 0.98)	0.07 (0.02, 0.12)
	VII	0.90 (0.87, 0.93)	0.96 (0.94, 0.98)	0.06 (0.02, 0.10)

Table 2.17: Predicted 3 and 6 year AIDS-free survival (bootstrapped 95% confidence intervals).

Recall that strategies IV (blue) and V (purple) have treatment models stratified by country. To explore this further, consider Figure 2.13 which illustrates the results for these strategies, in addition with one overall treatment model across countries. Stratifying the treatment models by country resulted in more moderate estimated treatment effects, with the exception of Switzerland where the effect is in the opposite direction. The reasons for this are not clear; we have seen above that the weights for Switzerland are fairly stable, therefore perhaps this suggests some lack of control for confounding in the remaining countries.

AIDS-free survival

Although some interactions were detected with some weighting strategies, there was a lack of agreement across them. For illustration, the standardised survival curves for immediate versus no treatment, as described in section 2.4.1, are shown in Figure 2.14, assuming no interactions with treatment. Table 2.17 also gives the predicted AIDS-free survival at 3 and 6 years under these treatment regimes, with bootstrapped confidence intervals, and for the differences in AIDS-free survival between the two regimes at those time-points. The medians of the bootstrapped estimates were similar to the overall point estimates. Of note, two of the bootstrapped datasets only contained patients from Italy who were not observed to reach the endpoint, therefore all patients from Italy were dropped from the pooled logistic regression models. However, the results from those models were checked by eye individually and were found not to be overt outliers compared to the overall bootstrap estimates, and therefore were included for the bootstrapped confidence interval estimation.

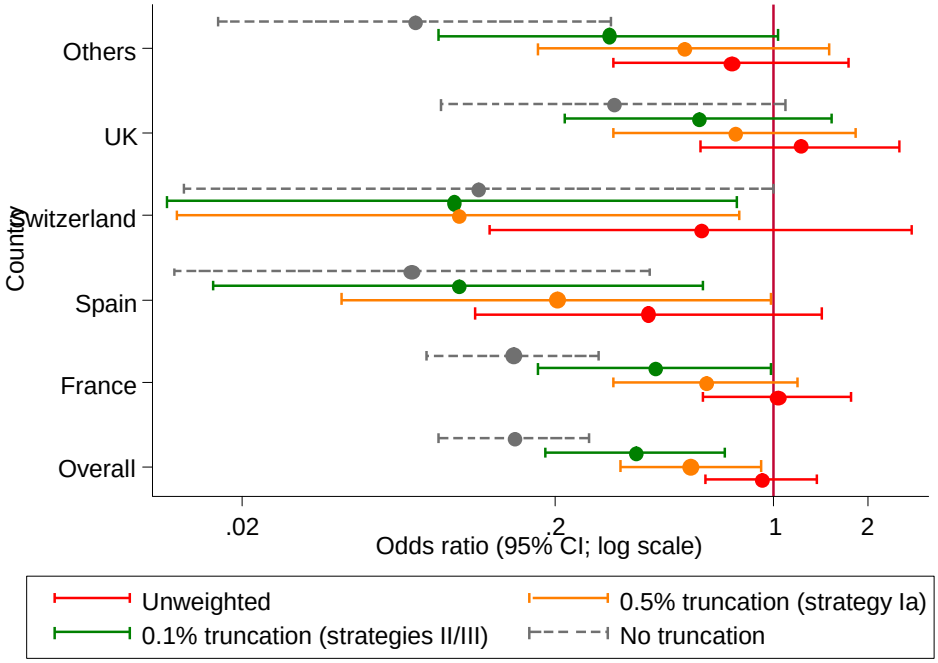


Figure 2.12: Effect of treatment by country, under weighting from treatment model of strategies I-III, with different degrees of truncation.

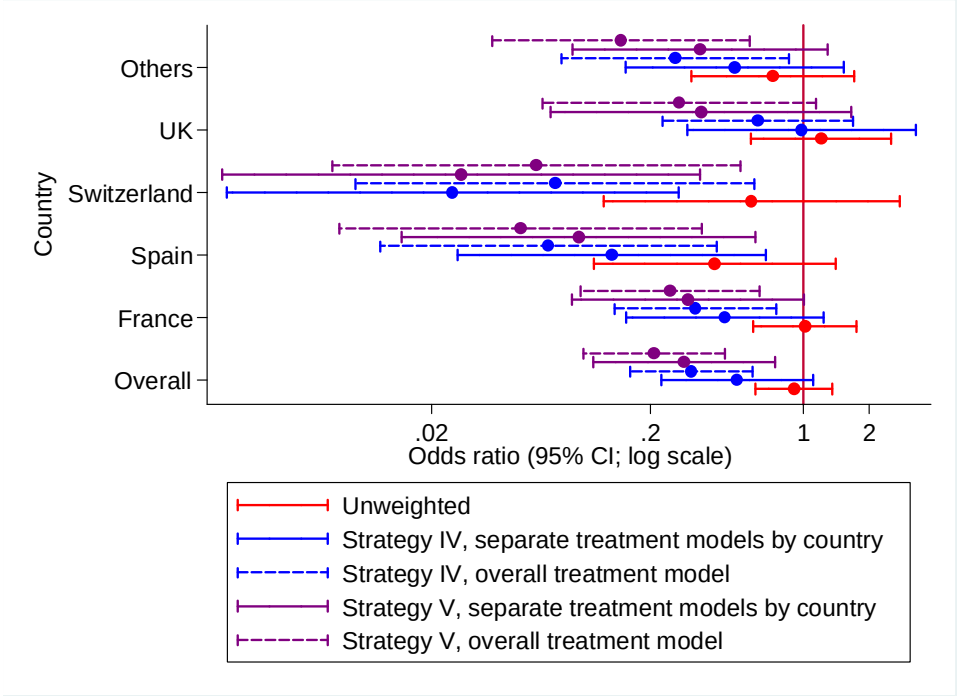


Figure 2.13: Effect of treatment by country, unweighted and under weighting from strategies IV and V, with either separate treatment models by country or overall treatment models.

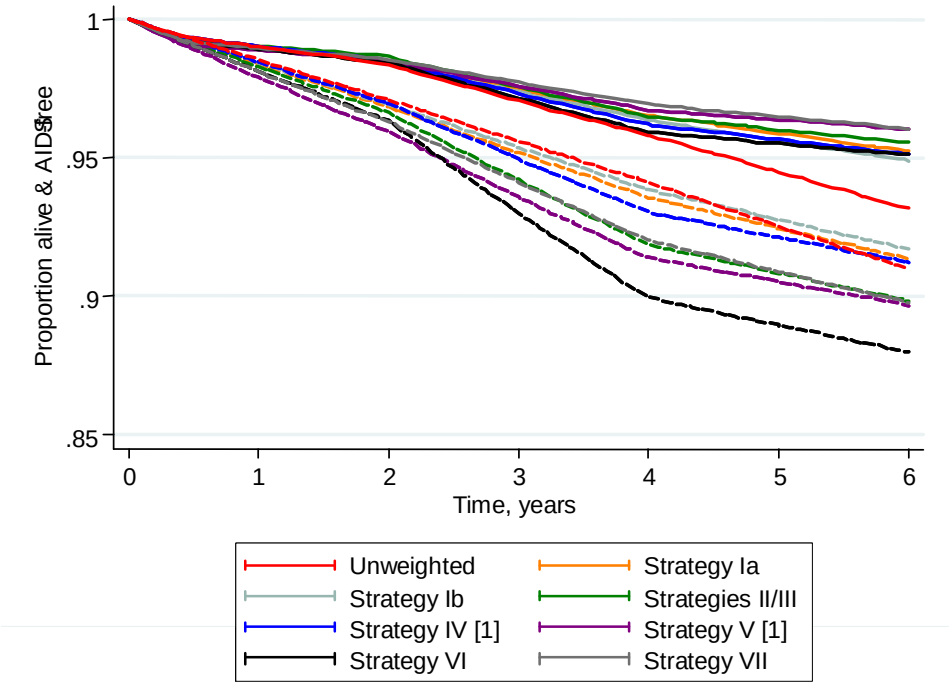


Figure 2.14: Standardised AIDS-free survival over 6 years for immediate (solid lines) versus no (dashed lines) treatment, across the different strategies and an unweighted model.

The unweighted curves for immediate and no treatment remained fairly close together over time compared to the weighted curves, which all predicted higher AIDS-free survival under immediate treatment, and typically similar or lower AIDS-free survival under no treatment, compared to the unweighted curve. While there was some departure between the weighted curves at later times, there was a great deal of uncertainty at these times, and overall they yielded fairly consistent results. At 3 years, the unweighted and weighted curves all predicted statistically significantly higher AIDS-free survival for immediate versus no treatment, although the weighted curves more so (from 2 to 4% higher AIDS-free survival compared to 1% higher survival unweighted). At 6 years, the magnitudes of the differences between the curves for immediate versus no treatment increased although the uncertainty also increased; there was no longer any evidence of a difference between the two regimes from the unweighted models. In contrast, most of the weighted models continued to predict higher AIDS-free survival under immediate versus no treatment, with the exception of strategy IV (unsurprisingly since we have already seen greater uncertainty under this strategy) and strategy V. The predicted AIDS-free survival at 6 years ranged across the weighted models from 0.88 (0.85, 0.91) to 0.92 (0.90, 0.94) under strategies VI and Ib, respectively, for a patient who had never taken treatment, and around 0.95-0.96 under all strategies for a patient who started treatment immediately. The difference in AIDS-free survival ranged from 0.03 (0.002, 0.06) to 0.07 (0.02, 0.12) under

strategies Ib and VI, respectively, compared to 0.02 (-0.01, 0.06) under the (biased) unweighted approach.

2.5 Discussion

While MSMs offer a relatively intuitive extension of standard methods to estimate causal effects, their application in practice is not straightforward. Of crucial importance is the construction of suitable inverse probability of treatment weights, which requires a number of inevitably subjective decisions. While formal methods have been proposed to develop a treatment model (Brookhart and van der Laan (2006); Mortimer et al. (2005); Petersen, Deeks, Martin, and van der Laan (2007)), these methods are not necessarily easy to implement, are somewhat opaque and still require decisions at the start regarding potential covariates and at the end to determine whether a suitable model has been achieved. We have broken down and structured the process as a set of four well-defined key decisions: the starting point (minimal model of key potential confounders versus "full" model of potential confounders), working with truncated or untruncated weights, identification of covariates to add (or remove) from the model, and the level of truncation for the final weights. We have indicated potential choices which may be preferable to different researchers, and based on these different viewpoints have constructed a set of six varied strategies (Ia, Ib, II, III, IV and V). Of course, there are many other possibilities which may yield different results; these six realistic options were chosen to explore and illustrate the potential differences that could arise, and might be expected to yield the most contrasting results.

2.5.1 The key decisions and strategies for construction of the treatment model

The six strategies led to four distinct models for treatment weights, although there was a great deal of overlap. A broad range of potential confounders were considered for inclusion in the treatment model, although no variable appeared in only one treatment model. Time since last CD4 count appeared in all models, and the number of previous CD4 counts appeared in all but one; some had in addition nadir CD4 count and the indicator for last CD4 observation carried forward. Two of the treatment models were stratified by country. Even the preliminary model with CD4 count as the only time-dependent variable demonstrated considerable control for confounding, with a reduction in the estimated OR for the effect of treatment by almost two-thirds, from 0.91 unweighted to 0.33 (after 0.1% truncation of the weights). The weights were

improved upon with the incorporation of other time-dependent covariates. All the strategies appeared to demonstrate considerable control for confounding, with the point estimates having moved away from the unweighted estimate.

Applying the same strategies to the construction of censoring weights, the strategies which were constructed to favour larger models did result in larger models, although all the ORs were fairly close to 1 indicating that the censoring processes were not very informative in this example.

Decision 1

Decision 1 regarding the starting point (minimal versus "full" model) may be most directly assessed by comparing strategies IV (starting with the minimal model) and V (starting with the "full" model). These two strategies led to similar treatment models although, perhaps conversely to expected, the forward selection procedure of strategy IV yielded a slightly larger model which included the number of previous CD4 counts in addition to those covariates which were included under strategy V.

After applying in addition the censoring weights, strategy V led to the most extreme OR seen from all the strategies (0.29) compared to 0.35 under strategy IV, albeit with larger SE (0.49 versus 0.41 on the log-odds scale).

Decision 2

Decision 2 determined the level of truncation of the weights when making decisions about which covariates to add to (or remove from) the model; working with untruncated weights led to a much more complex model, including stratification by country (comparing strategies IV versus III). This is in the direction that we might expect. It is not immediately clear whether one approach is advantageous over the other; one could argue that the untruncated weights have greater potential for capturing confounding, therefore indicating that working with truncated weights may result in important confounders being missed. Conversely, if there are positivity problems, then working with untruncated weights may result in including variables which amplify that problem.

Of note, we found that if the model building process was performed at a given degree of truncation, then this was subsequently matched by the degree of truncation selected at the final stage. This is unsurprising since the process has been directed towards that degree of truncation, but is worth noting since it implies that decisions made relating to the criteria for selecting covariates are inevitably linked to the weight truncation indicated at the end.

After applying in addition the censoring weights, strategy IV with the more complex treatment and censoring models yielded results with a more moderate estimated treatment effect (OR=0.50 versus 0.36 under strategies II/III) and less precision (SE=0.41 versus 0.34 on the log-odds scale). Therefore, it appears that working with truncated weights during the treatment model building process may be advantageous.

Decision 3

In our example, we found that decision 3 relating to different criteria for adding variables to (or removing from) the treatment model did not make a difference (comparing strategy III versus II). However, during the construction process more variables were identified as eligible under the decision criterion which favoured a larger model (strategy III) as opposed to that which favoured a smaller model (strategy II), therefore it is possible that in other applications, this decision may result in different treatment models. However, it is somewhat reassuring that somewhat different but equally reasonable strategies with respect to the incorporation of covariates are likely to lead to similar treatment models.

Decision 4

Decision 4, regarding the truncation of the final weights, will in general make the largest difference to the estimated causal effects, as illustrated by Figures 2.9 and 2.12. It is necessary for the analyst to make a reasonable judgment about whether large weights are likely due to non-positivity problems or model misspecification (and therefore truncate) or the degree of control of confounding (therefore do not truncate). This cannot typically be determined from the data, therefore it seems prudent to perform some truncation if there are extreme weights; in addition, this will most likely help increase the precision of the treatment effect estimate. In our example, we felt that some truncation was needed, and 0.1% truncation seemed sufficient to bring the weights under control.

Compared to strategy II, additional truncation was applied under strategy Ia, which favoured greater protection from non-positivity or model misspecification bias. After applying in addition the censoring weights, the impact of this decision on the estimated treatment effect was considerable, changing the estimated OR from 0.36 under strategy II/III to 0.54 under strategy Ia, with an associated reduction in the SE (0.34 to 0.26 on the log-odds scale).

Weight truncation will typically make a substantial difference to estimates of treatment effect on outcome, and perhaps researchers should be encouraged to specify a priori what level they will use or at least what criteria will be applied to determine the level of truncation at

each stage. However, we would recommend performing a range of sensitivity analyses to assess the impact of the weight truncation, which should be reported alongside the main results.

Additional non-directional strategy

We realised that incorporating the direction of change of treatment effect away from the null in the selection process may preferentially lead to an exaggerated estimate of treatment effect. We therefore introduced an additional strategy (Ib) which matched strategy Ia except it did not depend on the direction of change of the treatment effect estimate. Compared to Ia, strategy Ib incorporated one extra variable in the treatment model, namely nadir CD4 count. The estimated treatment effect was more moderate under strategy Ib than strategy Ia; this could be related to our reason for introducing this strategy, namely that incorporating the direction of movement of the estimated treatment effect away from the null (as in strategy Ia) may lead to causal effect estimates that are too strong. However, strategy Ib led to more complex treatment and censoring models than strategy Ia, and therefore greater truncation was required to bring the weights under control; this was perhaps at the expense of control for confounding.

Additional strategy with interaction between CD4 count and HIV RNA

Strategy VI, incorporating an interaction between CD4 count and HIV RNA, yielded some very large weights, most likely due to non-positivity issues (very few patients with high CD4 counts and low HIV RNA levels were observed to initiate treatment). However, after 0.1% truncation of the weights, the estimated treatment effects were very similar to those from strategy IV without stratification by country (the same model but without the CD4 count by HIV RNA interaction). Therefore our original treatment model building strategy did not appear to have missed an important confounder in HIV RNA.

Additional strategy using traditional model selection procedure

The additional strategy VII with the "traditional" stepwise procedure led to an overly complex model, incorporating a number of variables capturing HIV RNA-related data. This model yielded some very extreme weights, perhaps due to positivity issues with the large model. Greater truncation was needed (0.5% compared to 0.1% under the majority of the other strategies) in order to bring the weights under control. However, after truncation, the results were not inconsistent with those from the other strategies, therefore suggesting perhaps that even in cases of severe non-positivity, this issue may be to some extent ameliorated with simple weight truncation. After incorporating in addition the censoring weights, the results from this strategy

were not inconsistent with the results from the majority of the other strategies (OR 0.32, SE 0.32 on the log-odds scale).

Summary

In summary, strategies Ia, Ib and IV yielded somewhat more modest estimated treatment effects, with ORs of 0.54, 0.63 and 0.50, respectively. The remainder of the strategies had broadly consistent results, with the ORs ranging from 0.29 under strategy V to 0.36 under strategy II/III. All these ORs were statistically significantly different from 1. Strategies IV and V had considerably larger standard errors at 0.41 and 0.49 on the log-odds scale, compared to a maximum of 0.34 under the other strategies; this was due to the stratification of the treatment models by country. Notably, these strategies used untruncated weights during the treatment model building process. We therefore recommend strategies Ia, Ib, II and III over strategies IV and V. Analysts may be reassured that the different criteria for covariate selection did not make a difference in practice in our application to the CASCADE data. In such examples where there are concerns about violations of the positivity assumption, we may prefer to opt for the strategy which was designed to lead to a minimal model, namely strategy Ia. If there are not concerns about non-positivity, then further work simulating different scenarios may be useful to determine which strategy may be more generally preferable in different circumstances. Regardless, we recommend that a range of strategies, in particular with different degrees of weight truncation, are performed to examine the sensitivity of the results to the assumptions.

2.5.2 "Treatment refusers"

A number of the models yielded some overtly large weights; these derived mainly from a few patients with low CD4 counts who persistently delayed treatment initiation, resulting perhaps in non-positivity issues, residual confounding or model misspecification. This issue was addressed to some extent by adapting the CD4 count model to the "blunted" spline, forcing a constant probability of treatment initiation at CD4 counts <100 cells/mm^3, and also with default 0.1% truncation of the weights. This degree of truncation was somewhat arbitrary but was necessary in order to exclude unreasonably large weights (for example >1000 for the preliminary model). A further step may be to censor these persistent "treatment refusers" at those low CD4 counts, but then the dynamic element must be recognised since this censoring process is dependent on time-updated data. Therefore within the constraints of these standard MSMs, there is nothing that can easily be done to unbiasedly address this problem. However, we shall see in subsequent

chapters that it is possible to incorporate the dynamic element with history-adjusted or dynamic MSMs. In the next chapter on HAMSMs, we can simply exclude "trials" with low "baseline" CD4 count, say < 100 cells/mm^3, on the grounds that these extreme cases are irrelevant to treatment decisions at a population level. In dynamic MSMs, the lowest regime that will be considered will be to initiate when the CD4 count is first observed to drop below 200 cells/mm^3, and so these "treatment refusers" will implicitly be dealt with. That is, these patients will be progressively censored from regimes as their CD4 count drops, and finally censored from all regimes when their CD4 count dropped < 200 cells/mm^3 and they still did not initiate treatment.

2.5.3 Treatment effect modification by baseline covariates

We investigated treatment effect modification by baseline covariates. There was a lack of agreement in the interactions identified, though that may be due to lack of power, particularly in strategies IV and V where there was stratification of the treatment model by country.

Where interactions were identified, they were typically present in two strategies. These were discussed in detail in section 2.4.2. Briefly, in strategies Ia and V, later year of seroconversion and lack of a baseline HIV RNA measurement were associated with weaker treatment effect estimates, but the reasons for these associations were not clear and there may be residual confounding. The particularly strong beneficial effect of treatment in those identified close to seroconversion under strategies V and VI was somewhat surprising, but we know that such patients are a small and select subgroup, and such early presentation is a well-known predictor for worse prognosis (Tyrer et al., 2003), therefore it is plausible that treatment could be particularly beneficial among that subset of patients. Previous studies have shown better immunological and clinical response to treatment in younger persons (Collaboration of Observational HIV Epidemiological Research Europe (COHERE) Study Group, 2008), in agreement with our findings under strategies V and VI, where younger age at seroconversion was associated with stronger treatment effect. One further interaction was identified, under strategies II/III: shorter time infected at baseline was associated with stronger treatment effect. This may be related to a survivorship bias, in that those infected longer at baseline must have survived longer treatment-naïve, AIDS-free and with such high CD4 counts > 500 cells/mm^3 in order to enter the analysis.

As far as we are aware, the only previous investigation of treatment (antiretroviral therapy) effect modification by baseline covariates using MSMs in HIV-infected patients was in a series

of papers by Cole and colleagues (2007; 2005; 2003), where "baseline" was the first clinic visit in 1995 or 1996. In their 2003 paper, they found no difference in the estimated treatment effect on progression to AIDS or death by sex. However, they did find that treatment appeared to be most beneficial in those with lower baseline CD4 counts, and in fact there was no strong benefit of treatment in those with baseline CD4 counts > 350 cells/mm^3. In their 2005 paper looking at the effect of treatment on CD4 count, the authors found a larger effect of treatment in the first year among men compared to women, but there was no evidence of a difference after one year. Similarly, those with a lower baseline CD4 count experienced a greater benefit of treatment in the first year but with no difference thereafter. In 2007, the authors found evidence of a stronger effect of treatment in men compared to women on HIV RNA, but no difference by baseline CD4 count. In contrast to those studies, our population includes patients with high baseline CD4 counts, and subsequent CD4 decline, rather than the starting value, is likely to be more important, therefore it is no surprise that we did not see any treatment effect modification by baseline CD4 count. This may be different when we progress to HAMSMs where we can look at treatment effect modification by trial "baseline", that is time-dependent, CD4 count. We did not find any evidence of an interaction of sex with treatment across any of the strategies.

2.5.4 Model checking using country

As suggested in section 2.3.4, we were able to exploit the existence of different countries to test for an interaction with treatment, as a model checking procedure.

Recall that a number of countries with few patients were combined, namely Australia, Canada, Denmark, the Netherlands and Norway. There is no reason that treatment or outcome in these countries would necessarily be similar in any way; greater numbers of patients would enable analyses split by these countries too and perhaps add to our understanding. Those countries which were included separately are known to contain different populations; in particular, the populations from the UK and Germany were predominantly men infected through having sex with men, while the populations from Italy and Spain had high proportions of patients infected through IDU. The frequency of CD4 count measurements varied by country, with the medians ranging from 3.0 to 5.6 months.

There was evidence of differential treatment effects by country under strategies IV and VI. Of note, strategy VI was based on strategy IV, but with the incorporation of an interaction between CD4 count and HIV RNA in the treatment model. Strategy IV led to the largest treatment model of the original strategies, but aside from this it was not clear how this strategy differed

to indicate treatment effect modification by country where the others did not. However, Figure 2.11 illustrates that, despite not reaching conventional statistical significance, there appeared to be some differential effects of treatment by country across all the strategies. In particular, the treatment effect appeared to be strongest in Switzerland. Notably, the estimated treatment effects for Switzerland were least affected by truncation of the weights, perhaps suggesting that the strong treatment effect estimates seen for Switzerland better captured the truth and we may be missing residual confounders in the other countries. However, given that treatment guidelines across these countries are broadly consistent, and based mainly on CD4 count, it is difficult to imagine what those confounders might be.

Regardless of the result, we found this process to be helpful in examining and understanding the data and would encourage others to consider applying such an approach. We did not see any great advantage in stratifying the treatment models by country. While other examples may be different, this offers some reassurance to other studies where stratification may not be possible, for example by subpopulations of clinical centres which may not be recorded.

2.5.5 Limitations

All results presented here rely on a number of assumptions (section 2.2.2), in particular consistency, no unmeasured confounders between treatment and outcome, no misspecification of the treatment or outcome models and positivity. The consistency assumption is likely to be a reasonable one in general, but the others are perhaps more debatable. There was some empirical evidence of non-positivity. We have considered a range of different models, but we cannot explicitly test whether the models, in terms of specification and incorporation of all confounders, are correct. Truncation of the weights should provide some protection against violation of these assumptions, at the potential expense of bias.

Patients with less than one month of follow-up were excluded, therefore the probability of AIDS-free survival in the first month was artificially equal to 1 in our analysis, although there were only 46 such patients.

The median follow-up was only 2.3 years, restricted in part by the follow-up time starting at least one year after seroconversion and by the large proportions of patients being censored due to lack of availability of CD4 counts. With longer follow-up with complete CD4 count data, and hence greater power, some of the differences arising under the different strategies, for example identification of different interactions of treatment with baseline covariates, may have been resolved.

2.5.6 Application to other disease areas

Our approach of defining the treatment model building process as a series of (subjective) decisions helps to ensure the decision-making process is transparent and may facilitate greater involvement of collaborators such as clinicians. Our range of strategies helped to illustrate the potential differences in results that may arise from different modelling approaches. While no approach is more "correct" than another, this may help researchers understand potential differences seen in literature published previously or in the future, and may for example help inform any systematic reviews by having indicated likely sources of any heterogeneity between results, such as the degree of weight truncation. Further, we suggest that authors may wish to consider more than one of the strategies outlined here, in order to explore the potential problems of positivity, model misspecification or residual confounding. Of course, there are other possible strategies which may also be considered. These recommendations apply to the field of HIV research and more widely.

2.5.7 Summary

In this chapter, we have proposed a transparent process for the construction of treatment models in terms of a series of decisions, and illustrated how these may be combined to form different modelling strategies. These may be adapted for the estimation of causal effects in any setting. We have applied these strategies to our population of patients from CASCADE, and illustrated the need for weighting to appropriately adjust for time-dependent confounders which are used in the treatment decision process. Across all strategies, we demonstrated a beneficial effect of treatment in terms of reduction in the risk of AIDS or death. There were some differences in the point estimates obtained, but overall the results were broadly consistent. In addition, we have estimated survival according to the non-dynamic treatment regimes of immediate versus no treatment, adjusting for baseline covariates only. We will compare these estimates to those obtained in subsequent chapters under different approaches.

We have explored treatment effect modification by baseline covariates, but, as previously described, it is not possible to investigate such effect modification by time-dependent covariates with standard MSMs. A natural follow-on question is when to initiate treatment; perhaps an intermediate time would still provide the benefits afforded by immediate treatment initiation, but reduce the time spent on treatment over a patient's lifetime, thus potentially reducing the risks associated with long-term treatment such as side effects and development of drug resistance. In the following chapter, we will incorporate interactions between treatment and

CD4 count using HAMSMs to address the question of whether to initiate or defer treatment with respect to current CD4 count, which is the situation faced by clinicians and patients at each clinic visit. In chapter 4, we then proceed to consider pre-specified, well-defined dynamic treatment regimes in terms of CD4 count, whose effects are estimated using dynamic MSMs. The construction of inverse probability of treatment weights for unbiased estimation of treatment effects via history-adjusted and dynamic MSMs follow the same principles as those for standard MSMs, therefore we will make use of the weights constructed in this chapter.

Chapter 3

History-adjusted marginal structural models

3.1 Introduction

As discussed in section 2.5.7, a limitation of standard MSMs is that they cannot directly incorporate interactions between treatment and time-dependent covariates. For example, we found that baseline CD4 count was not a treatment effect modifier in our population of HIV-infected persons from CASCADE, perhaps unsurprisingly given that, by design, all patients had a high CD4 count at the time of study entry. However, one might hypothesise that treatment is most beneficial at subsequent low CD4 counts; such treatment effect modifications cannot be addressed using standard MSMs.

We introduced the concept of history-adjusted static treatment regimes and their estimation using history-adjusted MSMs (HAMSMs) in section 1.4.3, with the idea of a series of "trials" and a common standard MSM assumed at each time-point (Petersen, Deeks, Martin, and van der Laan, 2007). In chapter 2, we outlined a range of potential strategies for estimation of the inverse probability weights, applied these to the CASCADE data to obtain an array of estimated weights, and demonstrated a treatment effect on the time to AIDS or death in our population of CASCADE patients.

In this chapter, we build on that work, firstly introducing the theory of HAMSMs and then applying these methods to the CASCADE data using the different sets of estimated weights from chapter 2, and in particular exploring treatment effect modification by time-updated CD4 count. As outlined in section 1.5, previous researchers have looked at estimating the causal effects of immediate versus deferred treatment, given current (or past) CD4 count, using CASCADE

data (although with all initial CD4 counts, not restricted to those with a first CD4 count ≥ 500 cells/mm^3 as in our population; Writing Committee for the CASCADE Collaboration (2011)). As we shall see, this question may be extended to consider the effects of treatment initiation immediately versus never (that is, no subsequent treatment), using HAMSMs (Hernán et al., 2008). We consider and compare both of these approaches, and obtain causal estimates of treatment given current CD4 count. It will be of interest to return to these results in chapter 5 to compare them with those obtained from the optimisation of dynamic treatment regimes, in chapter 4. While the application of history-adjusted and dynamic MSMs typically answer different questions, we might anticipate some consistency across the two approaches, and the application of both may offer additional insights to the inference of interest.

3.2 Methodology

As in section 1.4.3, in their most basic form HAMSMs can be used to estimate the effects of initiating treatment sequentially at each given time-point, given treatment and covariate history, ignoring whether treatment is subsequently initiated by those patients who initially deferred treatment, that is considering only the effect of starting treatment now versus not starting now and assuming behaviour in those deferring treatment is generalisable. This approach may be extended to estimate "adherence-adjusted" effects (Hernán et al., 2008), where appropriate adjustment is made for those patients who initially deferred but subsequently initiated treatment, in order to estimate the effects of immediate versus no treatment. We now describe these two scenarios further and the appropriate methods for estimation.

3.2.1 Treatment regimes Immediate versus Deferred treatment

Notation

We use exactly the same set-up and notation for time-dependent covariates $L(k)$, treatment $A(k)$ and outcome $Y(t)$ as introduced in section 2.2.1, with overbars representing history to that time. The key concept behind HAMSMs is to consider each small time interval, for example month, of patient follow-up as the start of a new "trial", and then investigate treatment effects by performing estimation across the pooled trials. Therefore, in practice, the first step is to expand the data, treating each month in which a patient remains alive, event-free, in follow-up and previously treatment-naïve as the start of a new trial. For the first trial, given by $k = 1$, we use the first month of follow-up, that is time $[0, 1)$, to determine $A(1)$ and hence the treatment status for that first trial, where $A(1) = 0$ means the patient is following the regime Deferred

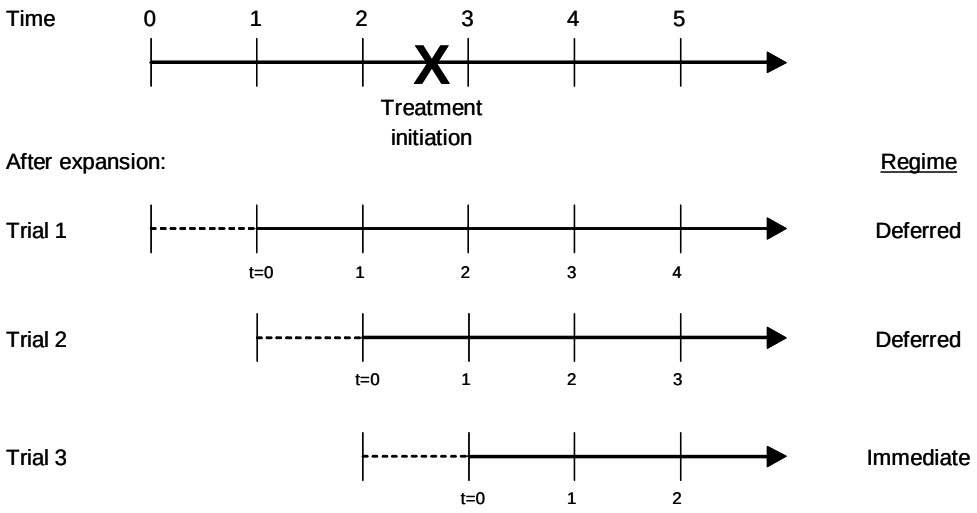


Figure 3.1: Illustration of expansion of data for an example patient who was treatment-naïve up to time-point 2 but had initiated treatment by time-point 3. We create a series of trials, starting at each time-point the patient remains alive, event-free, in follow-up and previously treatment-naïve. The solid lines indicate the follow-up time included for that trial; the dashed lines indicate the time during which the treatment status is determined for that trial (Immediate or Deferred).

treatment and $A(1) = 1$ means the patient is following the regime Immediate treatment. Follow-up of that first trial then begins at time 1. In general, for a trial k, the treatment status of the patient is determined by $A(k)$ and follow-up begins at time k. We use $t = 0, 1, 2, ...$ to indicate follow-up time within a given trial k. The "baseline" covariates for trial k are those given by $L(k-1)$; this ensures temporality in that the "baseline" covariates are always measured before determination of treatment status. In order to clearly distinguish between different variables, we refer to those at overall study entry as "true-baseline" covariates (constant for each patient across all trials) and those at the start of each trial as "trial-baseline" covariates (constant within each trial but different for each patient-trial). Expansion is performed for all patient-months for which the patient remains in follow-up with $\overline{A}(k-1) = 0$ and $\overline{Y}(k) = 0$. Of note, this requires for the first trial that $Y(1) = 0$, on the grounds of temporality; if treatment was initiated in the first month then we cannot be sure that it was not in response to the event.

Consider the example patient illustrated in Figure 3.1. In the first two months, the patient did not initiate treatment, therefore $A(1) = A(2) = 0$, and so for the first two trials, given by $k = 1, 2$, we consider the patient to be following the regime Deferred treatment. In the third month, the patient did initiate treatment, and so $A(3) = 1$ and, in this third trial, the patient is considered to be following the regime Immediate treatment. The patient does not contribute to any further trials.

The equivalent RCT would involve the randomisation of patients who are currently treatment-

naïve at each time-point to one of the two treatment regimes of Immediate or Deferred treatment. Note that the treatment regime Deferred permits treatment initiation at any subsequent time; see further discussion relating to this in section 3.2.2.

Cox proportional hazards model

When comparing the regimes of Immediate versus Deferred treatment, since the covariates $L(k-1)$ can be considered as baseline covariates for trial k, and we do not take into account subsequent treatment, this means that we do not have the problem of time-dependent confounding for the comparison of these two regimes. This means that straightforward adjustment of the trial-baseline covariates will provide an unbiased estimate for the effect of initiating treatment immediately versus deferral on the outcome of interest, under standard assumptions (section 1.2.4). In this case, weighting is not required and it is possible to easily use standard models, such as Cox proportional hazards models for our time-to-event outcome, since there are no time-varying weights to incorporate. Compared to pooled logistic regression models, this reduces computational time since only one record per patient per trial is required, rather than one record per patient per trial per month, so less data expansion is required. In addition, standard model building approaches can be employed.

Consider trial k; we could fit the following Cox proportional hazards model in patients who are in follow-up, previously treatment-naïve and event-free:

$$\lambda_T^{(k)}\left\{t|(A(k),\overline{A}(k-1)=0,\overline{L}(k-1),\right\}=\lambda_0^{(k)}(t)\exp\left\{\alpha A(k)+\beta L(k-1)\right\} \tag{3.1}$$

where T is the time to event of interest (now measured from the start of trial k). $A(k)$, $\overline{A}(k-1)$ and $\overline{L}(k-1)$ are as explained above. Equation 3.1 therefore consists of the baseline (perhaps trial-dependent) hazard $\lambda_0^{(k)}(t)$, and parameters α and β which we seek to estimate. In particular, $\widehat{\alpha}$ will provide our estimate of the effect of the regimes Immediate versus Deferred treatment. To clarify for the equivalent RCT analysed by ITT, $A(k)$ represents the randomised regime (Immediate or Deferred treatment), rather than whether the patient actually initiated treatment or not, in order to preserve the randomisation balance.

Assuming homogeneity in the treatment effect across trials, we can pool across all k, although then we may choose to include k in the model, preferably as a smooth function. This homogeneity assumption can be tested by incorporating an interaction between treatment and k (Hernán et al., 2008). The treatment effect modification by a trial-baseline covariate can then be explored using interactions between that covariate and treatment regime.

This approach was used by Writing Committee for the CASCADE Collaboration (2011) to estimate the effect of treatment on time to AIDS or death by CD4 count in CASCADE participants (not restricting to those with a high initial CD4 count); the authors found that treatment initiation, compared to deferral, was beneficial at CD4 counts < 350 cells/mm^3, and more greatly so at lower CD4 counts. There was some suggestion that initiation of treatment at CD4 counts of $350 - 500$ cells/mm^3 may be beneficial, but the authors noted that the event rate was low in this range of CD4 counts. The authors also considered an alternative approach, whereby instead of direct adjustment of the baseline covariates for each trial in the model via $L(k-1)$ in equation 3.1, they estimated inverse probability of treatment weights and used these in a Cox proportional hazards model to account for the non-random differences between the patients who were and were not observed to initiate treatment at the start of each trial. They noted that the results were very similar, as one would expect.

3.2.2 Treatment regimes Immediate versus No treatment

The approach described in section 3.2.1 comparing the effect of immediate initiation of treatment versus deferral does not attempt to take into account the subsequent treatment (or not) in those patients who initially deferred treatment. This in itself is useful clinically, where clinicians and patients are typically faced with the decision to immediately initiate or defer treatment at successive clinic visits. However, "deferral" in this instance encompasses a broad range of subsequent treatment options. If instead interest lies in the effect of initiating treatment immediately versus never, then this approach would in general yield a conservative estimate.

In a study investigating the effect of postmenopausal hormone therapy on the risk of coronary heart disease using observational data, Hernán et al. (2008) began with the direct-adjustment method of Writing Committee for the CASCADE Collaboration (2011), but then progressed to an "adherence-adjusted" approach. This involved censoring patients when they discontinued their trial-baseline treatment regime; that is, those who initiated treatment immediately at the start of a given trial were censored if they subsequently stopped treatment, and those who initially deferred treatment were censored if they subsequently initiated treatment. They used inverse probability weighting of pooled logistic regression models to account for this potentially informative censoring, thus up-weighting those patients who remained on their trial-baseline regime to allow for those censored from their trial-baseline regime due to non-adherence. Of note, while Hernán et al. (2008) censored both those patients who initially initiated treatment but later stopped and those who initially deferred treatment but later started, as outlined

in section 1.6.1, we will only be concerned with the latter, under the assumption that once treatment is initiated, it is continued for life in HIV-infected persons. This is similar to other studies in HIV infection (Gran et al., 2010; Writing Committee for the CASCADE Collaboration, 2011).

More recently, Gran et al. (2010) employed sequential Cox proportional hazards models to estimate the direct causal effect of treatment in HIV-infected persons. As in the adherence-adjusted approach of Hernán et al. (2008), patients were censored if they initially deferred but subsequently initiated treatment, and weights were applied accordingly. They estimated the parameters of these models using composite (pseudo) likelihood, stratifying the Cox models by trial start time, and estimated the standard errors using a jackknife approach. These models are the same as those approximated by the weighted pooled logistic regression models of Hernán et al. (2008), which are easier to implement with time-varying weights in standard software. As discussed in section 2.2.5, the odds ratios obtained from pooled logistic regression models can be interpreted as hazard ratios providing the probability of an event in each time interval is small (D'Agostino et al., 1990).

Petersen, Deeks, Martin, and van der Laan (2007) used a similar approach in a different field of HIV: when to switch from a failing treatment regime. They included patients with virological failure and estimated the effect of each additional month delay until switching on the CD4 count eight months later. As in the approach used by Writing Committee for the CASCADE Collaboration (2011), they constructed a number of trials, starting at each month that a patient remained in the study. Inverse probability weighting was used to adjust for treatment switches after the baseline time. Petersen, Deeks, Martin, and van der Laan (2007) refer to their models as HAMSMs, since they employed standard MSMs with multiple baseline times. In particular, they used a range of baseline times, and for each there was a single fixed time after the start of that trial at which the outcome was evaluated. That is, for each trial, there was a trial-specific outcome, namely CD4 count eight months later. However, in a subsequent commentary, Robins et al. (2007) suggest that the term HAMSM should be reserved for the scenario where there are a number of baseline times mapping to at least one overarching outcome time across the study, rather than the one-to-one relationship between start and end times in Petersen, Deeks, Martin, and van der Laan (2007). Robins et al. (2007) argue that the danger of such HAMSMs is that if one allows realistically flexible models, then there may be a risk of model incompatibilities. However, such inconsistencies will not arise provided the assumptions of correct model specification and no unmeasured confounding are met.

At each time-point k, the estimates from the Immediate versus No treatment regimes determine the optimal history-adjusted static treatment regime from that time k onwards. Petersen, Deeks, Martin, and van der Laan (2007) demonstrated that following the optimal history-adjusted static treatment regime determined at time k will in general yield a poorer outcome compared to sequentially following the optimal history-adjusted static treatment regime determined at time k, followed by the optimal history-adjusted static treatment regime determined at time $k+1$, and so on. Consider a simple example, where the optimal history-adjusted treatment regime is to initiate if the current CD4 count is < 350 cells/mm^3. If a patient, who was previously treatment-naïve, had a CD4 count > 350 cells/mm^3 at time k, then following the optimal history-adjusted treatment regime determined at time k from that time onwards would mean not initiating treatment for the remainder of follow-up. However, sequentially following the optimal history-adjusted static treatment regime determined at time k, followed by that at time $k+1$, and so on, would mean that the patient would initiate treatment if they have a subsequent CD4 count < 350 cells/mm^3, and this will in general yield a better outcome than remaining off treatment.

These sequential optimal history-adjusted static treatment regimes can be considered to map to a dynamic treatment regime. Petersen, Deeks, Martin, and van der Laan (2007) show that in a simple scenario with just one time-point, their optimal history-adjusted static treatment regime yields the optimal dynamic treatment regime (see chapter 4). However, with more time-points, their statically-optimal dynamic treatment regime may be inferior to the optimal dynamic treatment regime. In addition, they note that if the outcome is for example CD4 count m months later, rather than for example CD4 count at a specific time K at the end of the study, then their HAMSMs and the dynamic MSMs of Robins et al. (2008) are optimising different quantities since the outcome is different.

As outlined above, estimation of the effects of the Immediate versus No treatment regimes firstly requires censoring patients who deviate from their initial treatment regime. As we assume that treatment is continued once initiated, we are only concerned with the censoring of patients who initially deferred but subsequently initiated treatment. This censoring may be informative and we address this using inverse probability weighting.

Inverse probability weighting

We firstly assume that there is no "usual" censoring, for example due to loss to follow-up; this is addressed below. Since the artificial censoring from the regime Deferred treatment is directly related to treatment initiation, we use an analogous approach as for the inverse probability of

treatment weighting of standard MSMs. We fit the same model as in 3.1, except that patients who initially deferred but later initiated treatment are censored from the time they initiated, and inverse probability weights are incorporated to account for this potentially informative censoring.

As in section 2.2.3, we define:

$$p_A(u) := \Pr\left\{A(u) = 0 | \overline{A}(u-1) = 0, Y(u) = 0, \overline{L}(u-1)\right\}$$

for $u = 1, 2, ...$ Estimation of $p_A(u)$ follows as previously, using pooled logistic regression on the *unexpanded* data (hence we have used u to denote time, rather than t, to avoid confusion with follow-up time within a trial k). However, the remainder of the estimation differs.

As above, the data are expanded into one record per patient per trial k per month of follow-up, while patients remain in follow-up, treatment-naïve and event-free. All patients receive weight 1 at time $t = 0$ in each trial, because the trial-baseline covariates may be adjusted for directly in the outcome model. In addition, since we assume that treatment is continued once initiated, those patients with $A(k) = 1$, who are considered to be following the regime Immediate treatment, receive weight 1 for all follow-up in that trial k. For the patients who initially deferred treatment in trial k, we use the estimates $\widehat{p}_A(u)$ to obtain the (cumulative) probability of remaining off treatment. That is, for a given patient in trial $k = 1, 2, ...,$ the weight at times $t = 0, 1, 2, ...$ is estimated by:

$$\widehat{q}_A^{(k)}(t) = \begin{cases} 1 & \text{if } t = 0 \text{ or } A(k) = 1 \\ \prod_{u=k+1}^{k+t} \widehat{p}_A(u) & \text{if } t \geq 1 \text{ and } A(k) = 0 \end{cases} \tag{3.2}$$

As in section 2.2.3, the (non-stabilised) weights then are estimated as:

$$\widehat{W}_A^{(k)}(t) = \frac{1}{\widehat{q}_A^{(k)}(t)}$$

Stabilisation The weights can be stabilised as previously using the true-baseline covariates (section 2.2.3). Alternatively, we can now incorporate the trial-baseline covariates with the aim of increasing the efficiency. Define:

$$p_A^{(k)\dagger}(t) := \Pr\left\{A(t) = 0 | \overline{A}(t-1) = 0, Y(t) = 0, \overline{L}(k-1)\right\}$$

which is the same as $p_A^*(t)$ defined in section 2.2.3, except replacing V with $\overline{L}(k-1)$, and using the superscript (k) to indicate the trial-dependence. Estimation of $p_A^{(k)\dagger}(t)$ follows from pooled logistic regression models for the probability of treatment initiation, estimated on the *expanded* data over patient-months with $t \geq 1$ in trials where treatment was initially deferred, that is $A(k) = 0$. The numerator for the stabilised weights is estimated similarly to $q_A^{(k)}(t)$:

$$\widehat{q}_A^{(k)\dagger}(t) = \begin{cases} 1 & \text{if } t = 0 \text{ or } A(k) = 1 \\ \prod_{s=1}^{t} \widehat{p}_A^{(k)\dagger}(s) & \text{if } t \geq 1 \text{ and } A(k) = 0 \end{cases}$$

and the stabilised weights are then simply given by:

$$\widehat{SW}_A^{(k)}(t) = \frac{\widehat{q}_A^{(k)\dagger}(t)}{\widehat{q}_A^{(k)}(t)}$$

Pooled logistic regression model

Still assuming no "usual" censoring, we estimate for $t = 1, 2, ...$:

$$p(t) = \Pr\left\{Y(t+1) = 1 | Y(t) = 0, k, A(k), \overline{A}(k-1) = 0, \overline{L}(k-1)\right\}$$

using for example a pooled logistic regression model, in patients event-free and previously treatment-naïve, with estimated weights $\widehat{SW}_A^{(k)}(t)$:

$$\text{logit}\left\{p(t)\right\} = \alpha A(k) + \beta L(k-1) + \gamma f(k) + \delta f^{\dagger}(t)$$

where $f(k)$ and $f^{\dagger}(t)$ are functions of the trial k and follow-up time t within that trial, respectively. As indicated above, this assumes heterogeneity across the trials k; this can be tested by incorporating an interaction between $A(k)$ and $f(k)$.

Comparison with standard MSMs

As we have seen in chapter 2, standard MSMs may be susceptible to large inverse probability of treatment weights, resulting in unstable treatment effect estimates. These may arise in particular at treatment initiations when the estimated probability of treatment is small. In contrast, with these history-adjusted models, patients who initially deferred treatment are censored at the time of subsequent treatment initiation, therefore such large weights at treatment initiations will be censored (Gran et al., 2010). Further, under standard MSMs, the estimated weights at treatment initiation are carried forward for the remaining follow-up; again this is not the case

under history-adjusted estimation. Of note, the weights are, strictly-speaking, inverse probability of (artificial) censoring weights, but since the (artificial) censoring is determined based on treatment history, and in order to distinguish from "usual" censoring, we will refer to the inverse probability of (artificial) censoring weights as inverse probability of treatment weights henceforth.

As discussed in chapter 2, large weights may arise when the data are close to non-positivity. In CASCADE, we have seen that there is a small subset of patients who continued to defer treatment initiation despite having low CD4 counts. With the standard MSMs, we were unable to do anything further, other than truncate the weights, unless we had taken the rather drastic and potentially biased approach of excluding those patients completely. However, with history-adjusted models, it is easy to restrict the trials, with respect to the trial-baseline covariates, to the population of interest. We suspected that these patients who persistently deferred treatment initiation despite low CD4 counts are not part of the population in which we wished to estimate the effects of treatment, and therefore it was possible to simply restrict our analyses to those trials in which the trial-baseline CD4 count is above a certain threshold, that is for when the question of Immediate versus Deferred treatment initiation is a clinically relevant choice.

While the treatment effect parameters from the standard MSMs and the HAMSMs discussed above are not the same and therefore not directly analogous, both approaches are an attempt to understand the effects of treatment and we would anticipate that the results would be broadly compatible. Gran et al. (2010) found their results to be very similar to previous treatment effect estimates from applying standard MSMs to the same data; they argued that this supports the validity of each approach. However, as mentioned, the treatment parameters in these models are not identical. In particular, the adherence-adjusted treatment effect estimate (looking at immediate versus no treatment) from the HAMSM is adjusted for the trial-baseline covariates, whereas that from the standard MSM is not.

Comparison with effect of regimes Immediate versus Deferred treatment

Any differences observed in the estimated effects of the treatment regimes Immediate versus No treatment, compared to Immediate versus Deferred, will depend on the treatment initiation patterns in relation to the time-dependent covariates. We might anticipate that treatment is less likely to be delayed for long periods of time at low current CD4 counts, therefore at such CD4 counts the actual differences between the Immediate and Deferred treatment regimes may be relatively small due to those initially deferring treatment subsequently initiating soon after. In contrast, the Immediate versus No treatment estimation will censor those patients, and

upweight accordingly comparable patients who remain off treatment, therefore we may expect to see stronger treatment effects under this approach, and particularly at low CD4 counts, compared to the treatment regimes Immediate versus Deferred treatment.

3.2.3 Censoring

In the presence of "usual" censoring, for example due to LTFU, weights may be applied in a similar way to those as in section 2.2.4. As for the treatment weights described above, the denominator of the "usual" censoring weights is typically estimated using the unexpanded data, and the numerator is typically estimated based on the expanded data to include the trial-baseline covariates. For adherence-adjusted estimation, the overall weights are obtained as in chapter 2. The outcome models must also then condition on being uncensored due to "usual" censoring.

3.2.4 Standard error estimation

As for standard MSMs, we use robust variance estimators in the outcome models. For the comparison of the regimes Immediate versus Deferred treatment, this is necessary since patients may contribute to more than one trial. This reasoning also applies to the adherence-adjusted estimation, but also to allow for correlated observations induced by the use of time-dependent weights estimated from the data.

3.3 Application to CASCADE

Our ultimate aim is to apply dynamic MSMs to the CASCADE data to investigate the question of when to initiate treatment in HIV-infected persons. Using standard MSMs in chapter 2, we have begun by demonstrating a treatment effect on the time to AIDS or death in the population of patients from CASCADE in which we will apply the dynamic MSMs, and have considered effect modification by true-baseline covariates. We now propose the following analyses:

1. use the direct-adjustment approach of Writing Committee for the CASCADE Collaboration (2011) to estimate the effect on time to AIDS or death of the regimes Immediate versus Deferred treatment, ignoring subsequent treatment in those patients who initially deferred treatment, in our subset of CASCADE participants with CD4 $\geq$ 500 cells/mm^3 at study entry; and

2. extend these analyses to obtain the adherence-adjusted estimates of Hernán et al. (2008), by accounting for subsequent treatment initiations in the patients who initially deferred

with the use of censoring and inverse probability weighting, to give estimates of the effect of the regimes Immediate versus No treatment.

For both, we will investigate treatment effect modification by time-dependent (trial-baseline) CD4 count. This will help inform our subsequent work with dynamic MSMs and enable us to make comparisons between the different methods.

3.3.1 Methods

We firstly outline the methods for estimating the effects of the regimes Immediate versus Deferred treatment, ignoring subsequent treatment initiations in those patients who initially deferred treatment. Recall that, since the regime is determined at the start of the trial, adjustment for the trial-baseline covariates in the outcome model is sufficient, and this can be done straightforwardly using Cox proportional hazards models. That is, no weighting is required. Secondly, we detail the methods for the adherence-adjusted approach, which censors patients who initially deferred but subsequently initiated treatment. Inverse probability weights are required to account for this potential informative censoring, and we use (weighted) pooled logistic regression models due to limitations of current software.

Treatment regimes Immediate versus Deferred treatment

We began by using Cox proportional hazards models adjusted for the true-baseline covariates only (including country), then incorporated firstly just trial-baseline CD4 count and then the other time-dependent trial-baseline covariates. We considered a "full" model with all the covariates a priori identified as potential confounders (see Table 1.3 of chapter 2, with the categorisations as given there for the categorical variables and splines for the continuous variables, with five knots at the 5, 25, 50, 75 and 95^{th} percentiles) and then used a stepwise backwards selection procedure for the trial-baseline covariates (except CD4 count which was kept in the model) to identify a more parsimonious model (remove if $p > 0.05$, re-enter if $p < 0.01$). Trial start time k, measured from overall entry into the study as detailed in section 1.6.1, was included as a spline (five knots at the 5, 25, 50, 75 and 95^{th} percentiles). To test for heterogeneity in the treatment effect across trials, we considered including an interaction between treatment regime and trial k.

To investigate treatment effect modification by trial-baseline CD4 count, we included an interaction between CD4 count and treatment, with underlying CD4 count modelled as a five knot spline and the interaction with treatment based on categorical CD4 count. We began by

categorising CD4 count as Writing Committee for the CASCADE Collaboration (2011), but had limited data in the lowest CD4 count category of < 50 cells/mm^3 (5 patients contributing to 29 trials) therefore combined the two lowest categories and considered < 200, $200 - 349$, $350 - 499$ and ≥ 500 cells/mm^3.

We performed a range of sensitivity analyses, as follows:

1. Stratified by trial-baseline CD4 count, therefore permitting different effects of the other confounders on the time to AIDS or death by the trial-baseline CD4 count.

2. Excluded the trials beginning in the first month, since a high number of treatment initiations occurred in the first month (161 (5%) patients initiated in the first month following entry into the study as detailed in section 1.6.1).

3. Excluded trials with no previous HIV RNA information, since we suspected that HIV RNA might be important and the lack thereof indicative of a different prognosis.

4. Excluded trials with trial-baseline CD4 count < 100 cells/mm^3. As discussed previously, we were concerned that such "treatment refusers" might be different in some way to our population of interest.

5. Relaxed the LTFU and regular CD4 count requirements (as defined in section 2.4.1). This was required in the standard MSM approaches since we needed regular CD4 counts in order to reliably estimate the inverse probability of treatment weights. While we may be concerned about the implications of LTFU or irregular CD4 counts, we can be reassured that this censoring did not appear to be very informative in the estimation of the standard MSMs, and we were able with the history-adjusted models to relax that requirement. That is, patients were no longer censored during the course of a trial if they had irregular CD4 counts or met the criteria for LTFU (no CD4 count measured for > 12 months). However, patients did not contribute to new trials once they were considered censored under these criteria, since trial-baseline data would not have been available.

6. Used pooled logistic regression instead of Cox proportional hazards models, to check that the approximation of the pooled logistic regressions (which were used for the adherence-adjusted estimation with time-dependent weights, presented next) was reasonable. In this model, time since trial start t was included as a spline (five knots at the 5, 25, 50, 75 and 95^{th} percentiles).

Treatment regimes Immediate versus No treatment

We proceeded to apply the adherence-adjusted approach to estimate the effects of immediate versus no treatment, by censoring patients who initially deferred treatment at the start of a trial but subsequently initiated, and using inverse probability weighting to account for this potentially informative censoring. Since it is not straightforward to apply time-dependent weights with Cox proportional hazards models, we used pooled logistic regression models. We also considered the effects of applying the censoring only, without the upweighting, labelled the "unweighted" approach. Note that this will in general be biased for the causal estimates of interest, since it fails to account for the potential informative censoring of patients who initially deferred but subsequently initiated treatment. Further note that this is different to the approach above comparing Immediate versus Deferred treatment regimes, which addresses a different question and where weighting is not required.

We used the treatment models derived in the different strategies of chapter 2 to estimate a range of weights, with the numerator determined using the trial-baseline values of the time-dependent covariates in the respective treatment model. However, we did not include strategy VI since it was not possible to reliably estimate the model with interactions between CD4 count and both HIV RNA and treatment. We estimated an additional set of weights based on the model selected by the stepwise backwards procedure from the estimation of the effect of the regimes Immediate versus Deferred treatment (labelled strategy VIII). Time was included in the numerator models as outlined above (k and t as splines). Truncations were applied as indicated in chapter 2 (prior to incorporating censoring weights); that is, 0.1% truncation for all strategies, except for strategies Ia, Ib and VII where 0.5% truncation was applied. Since the weights from strategy VIII were somewhat unstable, 0.5% truncation was applied. All weight summaries presented were based on the trials in which the patient initially deferred treatment, and did not include the first month of each trial. That is, the summaries were only over the patient-months in which the weights were estimated, and not those in which the weights were set to 1 (see equation 3.2). This is to avoid including a lot of patient-months in which the weight is known to be 1, which does not help inform the performance of the weight estimation.

In the outcome models, time was included as above (k and t as splines). Although the treatment parameters from the standard MSMs and the adherence-adjusted HAMSMs are not directly comparable, it is reasonable to expect that they might be consistent, therefore we first estimated average hazard ratios across all trial-baseline CD4 counts. We then proceeded to incorporate an interaction between treatment and CD4 count as above to look at the effect of

treatment by trial-baseline CD4 count.

Standard error estimation

As outlined in the methods, robust variance estimators were used throughout. However these may be conservative, therefore for the main analyses we also bootstrapped (1000 repetitions) with resampling stratified by country (though as for the standard MSMs, we had to group Italy with Others since there were few patients in Italy). We assumed fixed weights since re-estimating the weights is extremely time- and computer-intensive, and previous work with the standard MSMs indicated that the additional uncertainty associated with re-estimating the weights on each bootstrap sample is likely to be relatively small.

Censoring

As outlined in section 3.2.3, we proceeded to incorporate censoring weights. We used the censoring models as determined previously, with the numerator estimated based on the trial-baseline covariates for each respective model, as for the treatment weights. The simplest censoring weights (from strategy Ia) were combined with the treatment weights from the new strategy VIII to create the overall strategy VIII weights.

Model checking using country

As under the standard MSMs, we considered incorporating an interaction between country and treatment, as a form of model checking.

AIDS-free survival

As for standard MSMs, these HAMSMs assume no effect of the length of time spent on treatment. To assess effect modification by CD4 count with a time-varying or cumulative effect of treatment, it would be necessary to include interactions between time on treatment and trial-baseline CD4 count. However, due to limited numbers of patients and events within these categories, this is not possible without substantially collapsing the CD4 categories, therefore this has not been addressed here.

Trial	N patients	N initiated treatment immediately	N events	N events in those who initiated immediately
1	3356	161	157	10
2	3156	32	144	1
3	3071	22	142	0
4	3015	19	140	3
5	2938	21	134	0
6	2871	25	129	0
7	2803	24	128	1
8	2737	19	123	0
9	2681	33	121	2
10	2603	29	115	2

Table 3.1: Illustration of the expansion of the CASCADE data to create a new trial for each month that a patient remains alive, AIDS-free, in follow up and treatment-naive (for first 10 trials).

3.3.2 Results

Data

Our initial dataset was the same as that used for the application of standard MSMs in chapter 2. Of note, patients with less than one month of follow-up (including due to AIDS or death) were excluded from that dataset, therefore meeting the requirement of $Y(1) = 0$ as indicated in section 3.2.1. Of the 3382 patients, 26 had less than two months follow-up before being censored, therefore, although they contributed to the estimation of censoring weights, they did not contribute to the outcome model. The remaining 3356 patients contributed cumulatively to a total of 84,029 patient-trials, with a median of 18 trials per patient (IQR 11, 33). The maximum number of trials per patient was 147; 10 patients contributed to at least 130 trials each. Table 3.1 illustrates, for the first 10 trials, the number of patients contributing to each trial, the number of patients initiating treatment immediately, and the number of subsequent events. Of note, a large number of patients initiated treatment in the first month despite having high CD4 counts; this may in part be related to all patients by definition having a clinic visit at that time (a CD4 count was recorded); these were excluded under the second sensitivity analysis, as detailed above.

	< 200 cells/mm^3 $n = 360$		$200 - 349$ cells/mm^3 $n = 5683$		$350 - 499$ cells/mm^3 $n = 18103$		≥ 500 cells/mm^3 $n = 59883$	
	Deferred 272 (76%)	Immediate 88 (24%)	Deferred 5395 (95%)	Immediate 288 (5%)	Deferred 17864 (99%)	Immediate 239 (1%)	Deferred 59436 (99%)	Immediate 447 (1%)
Sex, female	72 (26%)	11 (13%)	980 (18%)	45 (16%)	3094 (17%)	45 (19%)	12647 (21%)	96 (21%)
Age at SC, years	30 (26, 36)	31 (27, 38)	31 (26, 38)	36 (29, 41)	31 (26, 38)	31 (24, 38)	31 (26, 37)	30 (26, 36)
Year of SC	1996 (1993, 2000)	1998 (1995, 2000)	1998 (1995, 2001)	1999 (1996, 2001)	1999 (1995, 2002)	1997 (1994, 2001)	1998 (1995, 2001)	1997 (1995, 2000)
Route of HIV transmission, IDU	78 (29%)	7 (8%)	486 (9%)	9 (3%)	1356 (8%)	22 (9%)	5571 (9%)	46 (10%)
Identified close to SC[1]	9 (3%)	6 (7%)	399 (7%)	19 (7%)	1257 (7%)	14 (6%)	4760 (8%)	18 (4%)
Time HIV-infected at study entry, years	1.5 (1.1, 3.4)	1.4 (1.1, 2.0)	1.4 (1.2, 1.9)	1.3 (1.1, 1.8)	1.4 (1.1, 1.9)	1.4 (1.2, 2.0)	1.3 (1.1, 2.0)	1.4 (1.1, 2.2)
Country								
France	105 (39%)	26 (30%)	1695 (31%)	139 (48%)	7981 (45%)	143 (60%)	29991 (50%)	303 (68%)
Germany	0 (0%)	1 (1%)	90 (2%)	2 (1%)	265 (1%)	2 (1%)	1234 (2%)	4 (1%)
Italy	26 (10%)	2 (2%)	267 (5%)	12 (4%)	869 (5%)	9 (4%)	3577 (6%)	28 (6%)
Spain	8 (3%)	3 (3%)	410 (8%)	13 (5%)	1076 (6%)	13 (5%)	4981 (8%)	34 (8%)
Switzerland	5 (2%)	2 (2%)	301 (6%)	23 (8%)	935 (5%)	17 (7%)	3116 (5%)	17 (4%)
UK	66 (24%)	41 (47%)	1934 (36%)	79 (27%)	5023 (28%)	33 (14%)	12228 (21%)	33 (7%)
Others	62 (23%)	13 (15%)	698 (13%)	20 (7%)	1715 (10%)	22 (9%)	4309 (7%)	28 (6%)

Continued overleaf...

	< 200 cells/mm^3		$200 - 349$ cells/mm^3		$350 - 499$ cells/mm^3		≥ 500 cells/mm^3	
	Deferred	Immediate	Deferred	Immediate	Deferred	Immediate	Deferred	Immediate
CD4 count, cells/mm^3	162 (119, 185)	155 (119, 174)	310 (275, 332)	279 (241, 311)	433 (397, 467)	406 (375, 450)	670 (575, 821)	597 (539, 713)
LOCF	176 (65%)	23 (26%)	3736 (69%)	149 (52%)	13299 (74%)	143 (60%)	47025 (79%)	208 (47%)
CD4 count decrease								
Large increase[2]	0 (0%)	0 (0%)	11 (<1%)	0 (0%)	168 (1%)	4 (2%)	2280 (4%)	18 (4%)
Small increase[3]	6 (2%)	4 (5%)	275 (5%)	18 (6%)	1166 (7%)	22 (9%)	2376 (4%)	29 (6%)
No change	178 (65%)	23 (26%)	3770 (70%)	151 (52%)	13395 (75%)	145 (61%)	50469 (85%)	371 (83%)
Small decrease[3]	41 (15%)	31 (35%)	731 (14%)	79 (27%)	1705 (10%)	32 (13%)	2263 (4%)	15 (3%)
Large decrease[2]	47 (17%)	30 (34%)	608 (11%)	40 (14%)	1430 (8%)	36 (15%)	2048 (3%)	14 (3%)
Time since last CD4 count, months	1.5 (0.6, 3.2)	0.5 (0.2, 1.1)	1.8 (0.8, 3.1)	1.0 (0.5, 2.0)	2.1 (1.0, 3.5)	1.5 (0.6, 3.1)	2.3 (1.0, 4.0)	0.8 (0.0, 2.4)
Nadir CD4 count, cells/mm^3	156 (118, 180)	153 (118, 171)	290 (252, 323)	269 (230, 309)	400 (356, 443)	380 (352, 425)	600 (520, 726)	568 (515, 665)
Number of previous CD4 counts	9 (6, 13)	12 (9, 16)	8 (5, 13)	9 (5, 14)	6 (3, 10)	6 (4, 9)	3 (1, 6)	2 (1, 4)
Number with HIV RNA data	266 (98%)	87 (99%)	5262 (98%)	284 (99%)	17301 (97%)	234 (98%)	52174 (88%)	413 (92%)
Number of previous HIV RNAs	9 (5, 13)	12 (8, 15)	8 (5, 12)	8 (5, 14)	6 (4, 10)	6 (3, 9)	3 (2, 6)	2 (1, 3)
Last HIV RNA, $\log_{10}$ copies/ml[4,5]	4.7 (3.8, 5.2)	5.1 (4.7, 5.5)	4.4 (3.9, 4.9)	4.8 (4.3, 5.3)	4.2 (3.7, 4.7)	4.8 (4.2, 5.1)	3.9 (3.3, 4.5)	4.4 (3.6, 5.0)
Time since last HIV RNA, months[4]	1.8 (0.7, 3.6)	0.7 (0.3, 1.6)	1.9 (0.9, 3.5)	1.1 (0.5, 2.3)	2.2 (1.0, 3.8)	1.8 (0.7, 3.3)	2.4 (1.0, 4.2)	0.7 (0.0, 2.0)
Peak HIV RNA, $\log_{10}$ copies/ml[4,5]	5.0 (4.4, 5.4)	5.3 (5.0, 5.7)	4.7 (4.3, 5.2)	5.0 (4.6, 5.5)	4.6 (4.1, 5.0)	4.9 (4.5, 5.3)	4.2 (3.6, 4.7)	4.5 (3.8, 5.0)

Table 3.2: Characteristics by trial-baseline CD4 count and "randomisation" of Immediate versus Deferred treatment. Patients may contribute more than once to multiple trials within each CD4 stratum. Values are n (column %, except in header which is row %) for categorical variables and median (IQR) for continuous variables. IDU=injecting drug use. LOCF=last (CD4) observation carried forward. [1] Defined as last negative and first positive HIV tests within 30 days, or laboratory evidence of seroconversion. [2] > 100 cells/mm^3. [3] ≤ 100 cells/mm^3. [4] Of the trials with prior HIV RNA data available. [5] If no subsequent measurements available, last HIV RNA measurement carried forward regardless of the length of time.

	Trial-baseline CD4 count stratum, cells/mm^3			
	< 200	$200 - 349$	$350 - 499$	≥ 500
Number of trials	360	5683	18103	59883
Number of patients[1]	115	698	1458	3356
Follow-up, person-years[1]	331	2164	4638	10974
Number of events[1]	8	38	78	157
Treatment regime				
Defer	272	5395	17864	59436
Immediate	88	288	239	447
Subsequently initiated after Deferred[2]	157 (58%)	3219 (60%)	7156 (40%)	14365 (24%)

Table 3.3: Summary of trials, patients, follow-up, subsequent treatment initiations and events, by trial-baseline CD4 count. [1] Follow-up time and the numbers of patients and events are unique within but not across the strata. [2] Percentage of those who initially deferred treatment at the trial start.

Demographics and follow-up

The characteristics of patients who immediately initiated versus deferred treatment, by trial-baseline CD4 count, are shown in Table 3.2. As we would expect, the probability of treatment initiation was higher at lower CD4 counts, and within each CD4 stratum those who initiated treatment had lower CD4 counts and higher HIV RNA levels. UK (and to some extent Other) patients made up a higher proportion of those who initiated at lower versus higher CD4 counts; conversely, the French made up a smaller proportion. At lower CD4 counts, females and IDUs were more likely to defer treatment; these factors are likely to be correlated and this higher rate of deferral amongst IDU was observed by Writing Committee for the CASCADE Collaboration (2011). Also at lower CD4 counts, patients with decreases in CD4 count were more likely to initiate treatment immediately, as we might expect. Further, no change in CD4 count, LOCF and longer time since last CD4 count were associated with deferral of treatment at lower CD4 counts; these are likely to be proxy measures for no recent clinic visit and hence no CD4 measured. At higher CD4 counts, the nadir CD4 count tended to be slightly lower amongst those who initiated treatment immediately, but this was not seen at the lower CD4 counts, perhaps because at that stage of infection, the current CD4 count is a more influential factor in the treatment decision than past values. At lower CD4 counts, the median number of previous CD4 counts was lower in those who deferred treatment; again, this may be a proxy measure for no recent CD4, but also it may be that the type of patient who attends fewer visits is less likely to begin treatment.

Subsequent follow-up, including treatment initiations in patients who initially deferred and events, is summarised in Table 3.3. There was a great deal more follow-up at higher CD4 counts, since patients all began with CD4 counts ≥ 500 cells/mm^3. As we might expect, a large

Trial-dependent covariates included in the model	Estimated treatment effect		
	HR (SE)[4]	95% CI[4]	p
-	0.83 (0.12)	0.63, 1.10	0.20
CD4 count[1]	0.71 (0.13)	0.50, 1.00	0.05
CD4 count plus all other trial-dependent covariates[2]	0.58 (0.10)	0.41, 0.82	0.002
CD4 count and peak HIV RNA[3]	0.58 (0.11)	0.40, 0.83	0.003

Table 3.4: Estimated effect of treatment, with different time-dependent covariates included in the Cox proportional hazards model. All models included time of trial start as a five knot spline, plus the true-baseline covariates. HR=hazard ratio. SE=standard error. CI=confidence interval. [1] Spline with five knots at the 5, 25, 50, 75 and 95^{th} percentiles. [2] See text for further details of the trial-dependent covariates. [3] As determined by a refined stepwise backwards selection procedure (see text for more details). [4] Robust standard errors.

percentage of patients who initially deferred treatment in any given trial went on to initiate subsequently, particularly at lower trial-baseline CD4 counts.

Effect on time to AIDS or death of regimes Immediate versus Deferred treatment (no adjustment for subsequent treatment initiation if initially deferred)

There was no evidence of a significant difference in the regimes Immediate versus Deferred treatment on the time to AIDS or death in the model adjusting for true-baseline covariates only (HR=0.83 [95% CI 0.63, 1.10], $p = 0.20$; Table 3.4). However, after adjusting for trial-baseline CD4 count, Immediate treatment was associated with a 29% reduction in the hazard of AIDS or death compared to Deferred treatment (0.71 [0.50, 1.00], $p = 0.05$). After further adjustment for all trial-baseline covariates, a stronger benefit of treatment was apparent (0.58 [0.41, 0.82], $p = 0.002$).

The stepwise backwards selection procedure successively dropped time since last RNA measurement, LOCF, last HIV RNA, number of previous CD4 counts and nadir CD4 count (p = 0.87, 0.85, 0.60 and 0.35, respectively), leaving a model with CD4 decrease, time since last CD4 count, number of previous HIV RNA measurements and peak HIV RNA (along with CD4 count, the true-baseline covariates and time). The estimated treatment regime effect was similar to the full model, but with slightly more precision (0.56 [0.40, 0.79], $p = 0.001$).

However, there were concerns about over-fitting in this model due to effect estimates of certain covariates being in the opposite direction to that expected based on HIV epidemiology, and colinearity was suspected between some covariates. For example, both shorter and longer time since last CD4 count at trial-baseline were associated with lower risk of AIDS or death, as was no change in CD4 count compared with any change (whether increase or decrease), but the CD4 decrease category of no change captures to some extent the lack of a recent

CD4 measurement. Therefore we undertook the following refinements to the model: firstly, we refitted the model without time since last CD4 count. This resulted in no change to 2 decimal places in the estimated effect of the regime Immediate versus Deferred treatment, nor in much difference to the effects of the other covariates, except that it rendered CD4 decrease non-significant. Therefore, in the interests of a parsimonious model, it was decided to omit both (trial-baseline) time since last CD4 count and CD4 decrease, yielding a similar HR for the regimes Immediate versus Deferred treatment of 0.55 (0.39, 0.77).

There were a number of HIV RNA-related variables in the model which gave rise to further concerns about colinearity. In particular, both the absence of a true-baseline HIV RNA measurement and higher true-baseline HIV RNA, if available, were associated with lower risk of progression to AIDS or death, contrary to expectations. This effect was reversed if (trial-baseline) peak HIV RNA was omitted from the model, supporting our concern of colinearity. However, this resulted in a weaker treatment effect estimate (HR=0.72), suggesting that peak HIV RNA may be an important confounder and should be included in the model. Therefore, true-baseline HIV RNA and the indicator for its availability were removed from the model. In this revised model, the (trial-baseline) number of previous HIV RNA measurements was no longer significant ($p = 0.14$), therefore this was also dropped from the model (although an indicator for availability of any previous HIV RNA measurement was included).

In conclusion, the final model, via this refined stepwise backwards selection procedure, included the same true-baseline covariates as previously except for baseline HIV RNA, and included only the trial-baseline covariates CD4 count and peak HIV RNA (and its availability). The resulting estimated HR for the regimes Immediate versus Deferred treatment was 0.58 (0.40, 0.83). Including an interaction between treatment regime and time of trial start yielded a p-value of 0.53, indicating homogeneity across trials, therefore we proceeded with the model pooled across trials.

The results from this final model are given in Table 3.5. There was no evidence of a difference in outcome by trial-baseline CD4 count ($p = 0.45$). Higher peak HIV RNA was associated with faster time to AIDS or death, as we would expect (Figure 3.2). There was a strong trend towards the lack of any previous HIV RNA measurement being predictive of AIDS or death, but the confidence interval was extremely wide. Being HIV-infected via IDU was associated with faster time to AIDS or death, as was shorter time HIV-infected at study entry. This may be because those entering further from seroconversion were a different type of patient in that they had to have survived that long with a high CD4 count in order to enter the study. There

Covariate	HR (95% CI)[1]	p
Immediate, versus Deferred treatment	0.58 (0.40, 0.83)	0.003
Trial-baseline covariates		
CD4 count, cells/mm^3	[2]	0.45
Peak HIV RNA, $\log_{10}$ copies/ml	[2]	< 0.001
No previous HIV RNA measurement available	3.13 (0.10, 98.2)	0.52
True-baseline covariates		
CD4 count, per 100 cells/mm^3	1.02 (0.94, 1.12)	0.60
Sex, female	0.69 (0.37, 1.28)	0.24
Age at seroconversion, per 10 years	1.17 (0.96, 1.43)	0.13
Year of seroconversion	0.94 (0.87, 1.03)	0.18
Route of HIV transmission, IDU	1.96 (1.10, 3.49)	0.02
Country, versus France		0.94
Germany	0.50 (0.07, 3.79)	
Italy	0.74 (0.26, 2.12)	
Spain	0.94 (0.39, 2.28)	
Switzerland	1.06 (0.42, 2.66)	
UK	0.75 (0.43, 1.32)	
Others	0.94 (0.48, 1.85)	
Time HIV-infected at entry, years	0.79 (0.62, 1.01)	0.06
Identified as HIV-infected close to seroconversion	1.71 (0.80, 3.64)	0.16

Table 3.5: Predictors of time to AIDS or death. Time included as a five knot spline. HR=hazard ratio. SE=standard error. CI=confidence interval. [1] Robust standard errors. [2] Splines used for continuous variables.

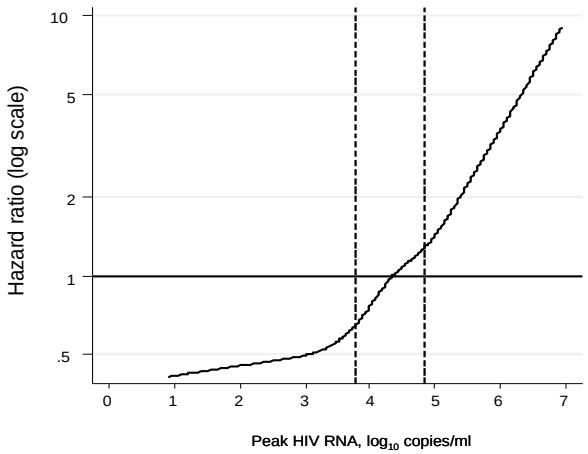


Figure 3.2: Hazard ratios for time to AIDS or death for peak HIV RNA (five knot spline). Centred on HR=1 for the median. The vertical dashed lines indicate the interquartile range.

was no association between time to AIDS or death and true-baseline CD4 count, sex, age at or year of seroconversion, country or whether identified as HIV-infected close to seroconversion.

Sensitivity analyses Table 3.6 shows the overall results for this model and the sensitivity analyses (outlined in section 3.3.1); the results are all broadly similar.

Treatment effect modification by CD4 count We proceeded to incorporate an interaction between treatment regime and CD4 count in the original model to look at effect modification by trial-baseline CD4 count. While the interaction was not significant ($p = 0.27$ and 0.26 with CD4 categorical or continuous, respectively), there was a trend towards greater benefit of treatment at lower CD4 counts (Table 3.6). At trial-baseline CD4 counts < 350 cells/mm^3, there was clear evidence of a benefit of immediate compared to deferred treatment, although the confidence intervals were wide, probably due to the more limited data in these strata (HR 0.20 [0.05, 0.77] and 0.44 [0.22, 0.88] for CD4 counts < 200 and $200 - 349$ cells/mm^3, respectively). There was a suggestion of a benefit of immediate compared to deferred treatment at trial-baseline CD4 counts ≥ 350 cells/mm^3 but the confidence intervals included one and overlapped considerably (0.79 [0.46, 1.37] and 0.70 [0.44, 1.09] for CD4 counts $350 - 499$ and ≥ 500 cells/mm^3, respectively). The effects of the other true- and trial-baseline covariates were similar to those from the model without the CD4 count by treatment interaction.

Sensitivity analyses The results from the original model are presented along with those by CD4 stratum from the sensitivity analyses in Table 3.6 and Figure 3.3. The results were all fairly consistent across the different sensitivity analyses. Restricting to those trials with trial-baseline CD4 count ≥ 100 cells/mm^3, the HR for the lowest CD4 category (now $100 - 199$ compared to $0 - 199$ cells/mm^3 previously) was closer to one at 0.31 (0.08, 1.14), as we would expect since by definition the trials with the very lowest CD4 counts were omitted.

Sensitivity analysis	Number of trials	Overall	Trial-baseline CD4 count stratum, cells/mm^3				$p^{[1]}$
			< 200	$200 - 349$	$350 - 499$	≥ 500	
Original	84029	0.58 (0.40, 0.83)	0.20 (0.05, 0.77)	0.44 (0.22, 0.88)	0.79 (0.46, 1.37)	0.70 (0.44, 1.09)	0.27
1	84029	0.57 (0.40, 0.82)	0.16 (0.04, 0.64)	0.49 (0.25, 0.98)	0.79 (0.47, 1.34)	0.68 (0.43, 1.07)	0.19
2	80673	0.52 (0.35, 0.78)	0.20 (0.05, 0.77)	0.44 (0.22, 0.87)	0.79 (0.46, 1.36)	0.59 (0.34, 1.02)	0.34
3	76021	0.55 (0.38, 0.81)	0.22 (0.06, 0.83)	0.45 (0.22, 0.91)	0.80 (0.46, 1.40)	0.62 (0.38, 1.01)	0.39
4	83966	0.62 (0.44, 0.88)	0.31 (0.08, 1.14)	0.49 (0.24, 0.98)	0.79 (0.46, 1.37)	0.70 (0.45, 1.10)	0.62
5	84029	0.60 (0.44, 0.80)	0.19 (0.05, 0.66)	0.41 (0.21, 0.78)	0.73 (0.46, 1.17)	0.75 (0.52, 1.06)	0.20
6	84029	0.58 (0.40, 0.83)	0.20 (0.05, 0.76)	0.43 (0.22, 0.87)	0.79 (0.46, 1.37)	0.70 (0.44, 1.09)	0.34

Table 3.6: Results from the original and sensitivity analyses: estimated effect of the regimes Immediate versus Deferred treatment, overall and by trial-baseline CD4 count. Results given are hazard ratio (95% confidence interval, with robust standard errors). See section 3.3.1 for further details of the different sensitivity analyses; briefly, 1=stratified by trial-baseline CD4 count. 2=excluded first trials, 3=excluded trials without HIV RNA, 4=excluded trials with trial-baseline CD4 count <100 cells/mm^3, 5=relaxed LTFU and regular CD4 count restrictions, 6=pooled logistic regression. [1] p-value for interaction between treatment and trial-baseline CD4 count category.

Regimes Immediate versus No treatment (adherence-adjusted estimation, adjusting for those trials where treatment was initially deferred but subsequently initiated)

As detailed in the methods, to obtain adherence-adjusted estimates, the patients who initially deferred treatment at the start of a trial but then subsequently initiated were censored at the time of treatment initiation, and weighting was required to adjust for this potentially informative censoring. We proceeded with the original analysis above only (that is, we did not repeat any of the sensitivity analyses detailed in the previous section), since the treatment effect estimates did not appear to be sensitive to these assumptions. However, the weighting was applied using the range of treatment and censoring models developed under the different strategies of chapter 2. Of note, trial-baseline CD4 count was always included in the models for the denominator and numerator of the weights.

Distribution of the inverse probability weights The estimated inverse probability weights, for the artificial censoring of the patients who initially deferred but subsequently initiated treatment, are illustrated over time in Figures 3.4 and 3.5 (the former with no truncation and the latter with 0.1% truncation for illustration). Once again, there were some very large weights occurring, but in contrast to the weights employed for the standard MSMs, there were also some very small weights. The weights were much more centrally located on one, with narrower interquartile ranges, compared to those for the standard MSM. After 0.1% truncation, the weights were again much more well-behaved, although there were still somewhat large weights under strategies VII and VIII (neither of which were the main strategies recommended in chapter 2), although still < 100.

As outlined in section 3.3.1, the treatment model building strategies were used to determine what degree of truncation should be applied, which was 0.1% truncation across all strategies, except strategies Ia, Ib and VII where 0.5% truncation was applied, and 0.5% truncation was applied to the new strategy VIII. After applying these strategy-specific truncations, the means of the weights were all slightly less than one, ranging from 0.969 under strategy VIII to 0.996 under strategy V (Table 3.7). All sets of weights were fairly stable with smaller standard deviations and ranges compared to the weights under the standard MSMs (maximum weight 11, under strategies II/III).

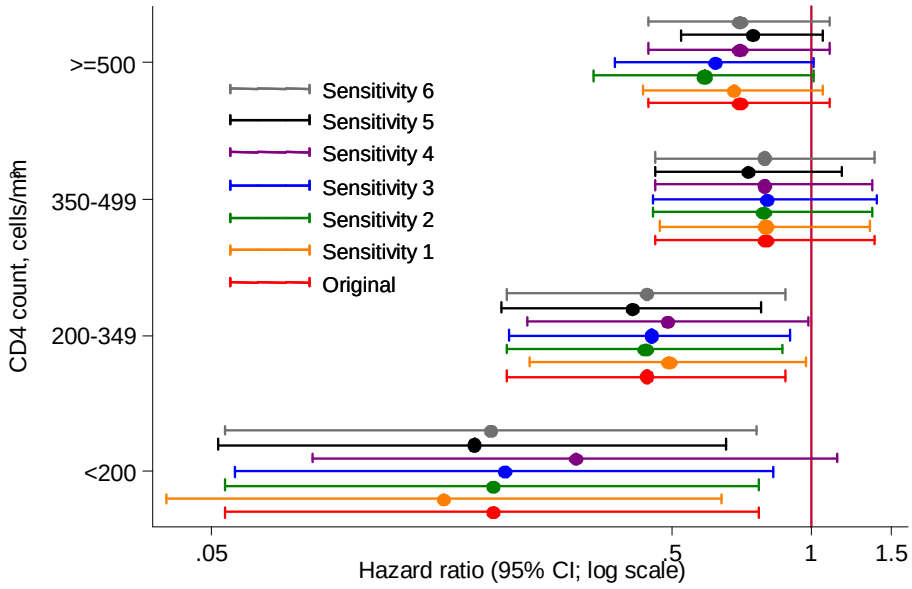


Figure 3.3: Results from the original and sensitivity analyses: estimated effect of regimes Immediate versus Deferred treatment by trial-baseline CD4 count. See section 3.3.1 for further details on the different sensitivity analyses; briefly, 1=stratified by trial-baseline CD4 count, 2=excluded first trials, 3=excluded trials without HIV RNA, 4=excluded trials with trial-baseline CD4 count <100 cells/mm^3, 5=relaxed LTFU and regular CD4 count restrictions, 6=pooled logistic regression.

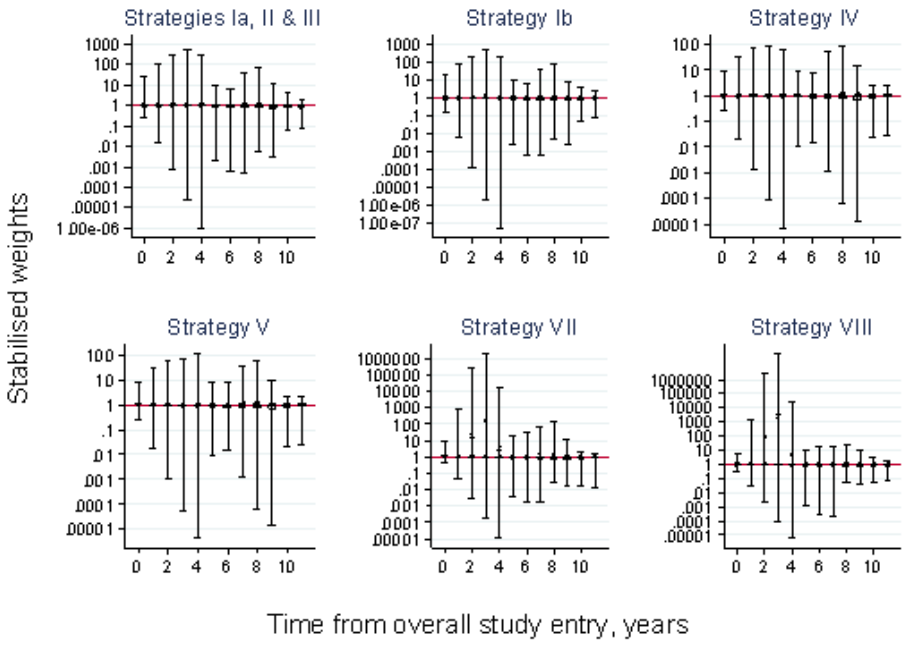


Figure 3.4: Distribution of the estimated stabilised weights for the different treatment models. Spikes = range, bars = interquartile range, o = median, x = mean. Note that the scales of the y-axes vary.

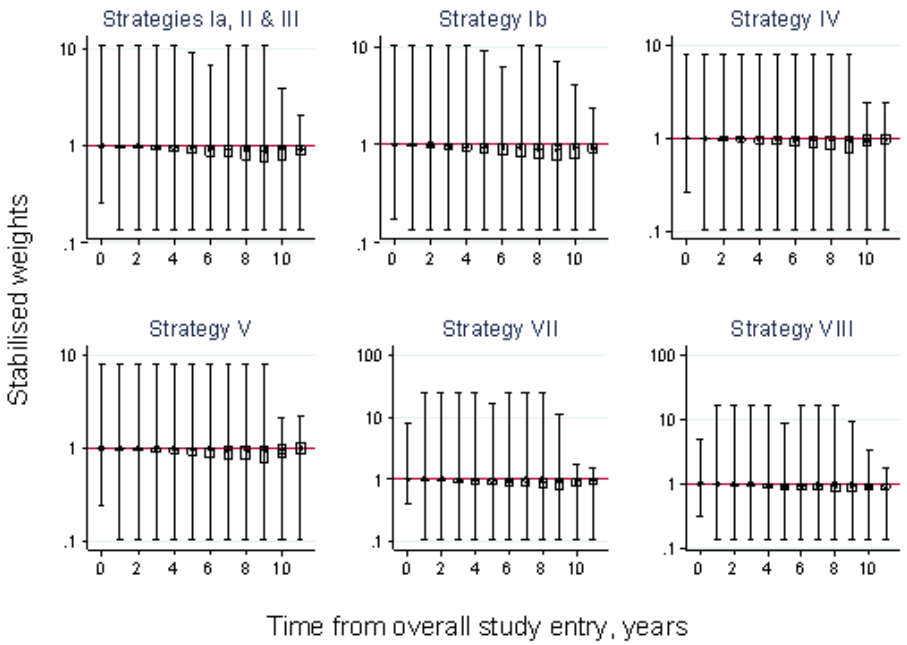


Figure 3.5: Distribution of the estimated stabilised weights for the different treatment models, after truncation of the outer 0.1 percentiles, for illustration. Spikes = range, bars = interquartile range, o = median, x = mean. Note that the scales of the y-axes vary.

Strategy	Estimated weights				Estimated effect of regimes Immediate versus No treatment				Comparison with standard MSM		
	Time-dependent variables included	Truncation[1]	Mean (SD)	Range	OR	95% CI		SE (log scale)[2]	OR	95% CI[2]	SE (log scale)[2]
						Robust[2]	BS[3]				
Unweighted[4]	-	-	-	-	0.46	0.25,0.86	-	0.32	0.91	0.61, 1.35	0.20
Ia	Time since last CD4 count, number of previous CD4 counts	0.5	0.974 (0.257)	0.32,3	0.31	0.15,0.62	0.15,0.65	0.36	0.52	0.31, 0.89	0.27
Ib	Time since last CD4 count, number of previous CD4 counts, nadir CD4 count	0.5	0.977 (0.255)	0.32,3	0.30	0.15,0.61	0.14,0.65	0.36	0.57	0.34, 0.95	0.26
II/III	As I	0.1	0.990 (0.475)	0.14,11	0.25	0.13,0.48	0.11,0.56	0.34	0.39	0.21, 0.75	0.33
IV	Time since last CD4 count, number of previous CD4 counts, nadir CD4 count, LOCF; stratified by country	0.1	0.994 (0.400)	0.10,8	0.24	0.09,0.67	0.08,0.70	0.52	0.60	0.30, 1.17	0.34
V	Time since last CD4 count, nadir CD4 count, LOCF; stratified by country	0.1	0.996 (0.403)	0.10,8	0.21	0.07,0.59	0.07,0.61	0.54	0.54	0.28, 1.05	0.34
VII	Time since last CD4 count, number of previous CD4 counts, LOCF, HIV RNA, peak HIV RNA, time since last HIV RNA, number of HIV RNA measurements	0.5	0.974 (0.308)	0.27,4	0.22	0.12,0.40	0.11,0.43	0.31	0.43	0.23, 0.78	0.31
VIII	Peak HIV RNA	0.5	0.968 (0.267)	0.30,3	0.20	0.11,0.39	0.10,0.43	0.33	-	-	-

Table 3.7: Estimated effect of regimes Immediate versus No treatment, across the different treatment model building strategies of chapter 2 (treatment (artificial censoring) weights only), compared with results from standard MSMs. All models included CD4 count and time. OR=odds ratio. CI=confidence interval. SE=standard error. [1] Truncation of x means that weights $< x^{th}$ or $> (100-x)^{th}$ percentile were replaced with the value of the x^{th} and $(100-x)^{th}$ percentile, respectively. [2] Robust SE calculated using clustered sandwich estimator, except for unweighted models since no weights to induce correlations. [3] Nonparametric bootstrap, 1000 replications. [4] Censoring applied (therefore different to the previous results which were estimating the effects of Immediate versus Deferred treatment with no censoring performed) but no weighting applied (therefore in general biased for the causal effects of the regimes Immediate versus No treatment).

Average estimated treatment effects The results for the effect of immediate versus no treatment, across all CD4 strata, arising from each of the different strategies are shown in Table 3.7. All strategies demonstrated considerable control for confounding when compared to the unweighted estimate of 0.46 (0.25, 0.86). Recall that this unweighted approach involved applying the censoring of patients who initially deferred but subsequently initiated treatment, but not the weighting, and therefore is biased for the causal effect of the regimes Immediate versus No treatment; it is included to demonstrate the effects of weighting. Strategies Ia and Ib yielded odds ratios closest to one (0.31 [0.15, 0.62] and 0.30 [0.15, 0.61], respectively), as we might have expected since these had the greatest truncation of 0.5% applied. However, strategy VIII (the model obtained by the adapted stepwise backwards selection procedure) also had 0.5% truncation applied yet yielded the odds ratio furthest from one (0.20 [0.11, 0.39]). Of note, the standard errors arising from strategies IV and V, where the treatment models were stratified by country, were considerably larger than those from the other strategies (0.52 and 0.54 on the log-odds scale, respectively, compared to 0.31-0.36 under the other strategies).

Comparing the estimated treatment effects with those from the standard MSMs and those of the regimes Immediate versus Deferred treatment As discussed above, although the treatment parameters of the standard MSMs and adherence-adjusted HAMSMs are not the same, we would expect effects in the same direction, as observed. With respect to the magnitude of effect, the adherence-adjusted odds ratios were all consistently lower (further from one) than those from the standard MSMs (Table 3.7). In fact, the results from the standard MSMs more closely matched the estimated effect of the regimes Immediate versus Deferred treatment (0.56 [0.40, 0.79], Table 3.4). Therefore, by adjusting for the subsequent treatment initiations in those who initially deferred treatment, a greater benefit of treatment is apparent, as we might expect.

The standard errors from the HAMSMs were all somewhat larger than those from the standard MSMs, perhaps contrary to what we might expect. The bootstrapped CIs were comparable to the robust CIs, and if anything were a little wider (Table 3.7). We incorporated a further 500 bootstraps for strategies Ia and III, but the resulting CIs were very similar. The medians were similar to the point estimates.

Treatment effect modification by CD4 count Incorporating an interaction between treatment and trial-baseline CD4 count, the odds ratios for Immediate versus No treatment across all strategies and CD4 count strata were lower than the unweighted estimates (that is, with cen-

Strat-egy	CD4 count, cells/mm^3				p[2]
	< 200	$200 - 349$	$350 - 499$	≥ 500	
Unweighted[1]	0.08 (0.02, 0.35)	0.28 (0.12, 0.65)	0.92 (0.46, 1.82)	0.83 (0.48, 1.42)	0.007
Ia	0.04 (0.01, 0.17)	0.19 (0.08, 0.46)	0.73 (0.35, 1.52)	0.73 (0.42, 1.27)	< 0.001
Ib	0.03 (0.01, 0.15)	0.21 (0.09, 0.50)	0.78 (0.37, 1.66)	0.70 (0.40, 1.22)	< 0.001
II/III	0.03 (0.01, 0.14)	0.16 (0.07, 0.37)	0.57 (0.26, 1.24)	0.63 (0.35, 1.11)	< 0.001
IV	0.02 (0.004, 0.15	0.17 (0.06, 0.50)	0.71 (0.31, 1.61)	0.63 (0.34, 1.20)	< 0.001
V	0.02 (0.003, 0.12)	0.15 (0.05, 0.45)	0.70 (0.30, 1.63)	0.57 (0.29, 1.12)	< 0.001
VII	0.03 (0.01, 0.13)	0.15 (0.07, 0.36)	0.39 (0.19, 0.84)	0.44 (0.24, 0.81)	0.004
VIII	0.03 (0.006, 0.11)	0.13 (0.06, 0.32)	0.47 (0.22, 1.02)	0.44 (0.24, 0.82)	< 0.001

Table 3.8: Estimated effect of regimes Immediate versus No treatment, by trial-baseline CD4 count, across the different treatment model building strategies of chapter 2 (treatment (artificial censoring) weights only). Results are odds ratio (95% confidence interval, with robust standard errors). Note that the overall results are shown in Table 3.7. [1] Censoring applied (therefore different to the previous results which were estimating the effects of Immediate versus Deferred treatment with no censoring performed) but no weighting applied (therefore in general biased for the causal effects of the regimes Immediate versus No treatment). [2] p-value for the interaction between treatment and trial-baseline CD4 count category.

soring but not weighting applied), as expected, since these estimates account for the patients who initially deferred but subsequently initiated treatment (Table 3.8 and Figure 3.6). We saw a similar pattern to the results from the regimes Immediate versus Deferred treatment, but the evidence for a greater benefit of treatment at lower CD4 count strata was much stronger. Compared to the Immediate versus Deferred treatment results, the estimated odds ratios were broadly similar for CD4 counts ≥ 350 cells/mm^3, but in contrast were now somewhat lower (further from one) for CD4 counts < 350 cells/mm^3, at least for strategies II/III, VII and VIII. This is in line with our observation that, at lower trial-baseline CD4 counts, higher percentages of patients were subsequently observed to initiate treatment. The confidence intervals after the artificial censoring and weighting for non-adherence to trial-baseline regime were wider than under the regimes Immediate versus Deferred treatment, as we might expect by the nature of weighted estimates.

As previously, while there was a suggestion of a benefit of treatment at CD4 counts ≥ 350 cells/mm^3, the confidence intervals spanned one for the majority of the strategies. However, strategies VII and VIII indicated that there may be a benefit of treatment at CD4 counts $350 - 499$ cells/mm^3 (albeit borderline for strategy VIII) and even ≥ 500 cells/mm^3 (0.39 [0.19, 0.84] and 0.44 [0.24, 0.81], respectively, for strategy VII; 0.47 [0.22, 1.02] and 0.44 [0.24, 0.82], respectively, for strategy VIII). However, given the issues raised earlier in this section regarding potential collinearity between a number of HIV RNA variables, we might be concerned about the results from strategy VII, which incorporated a number of covariates based on HIV RNA data.

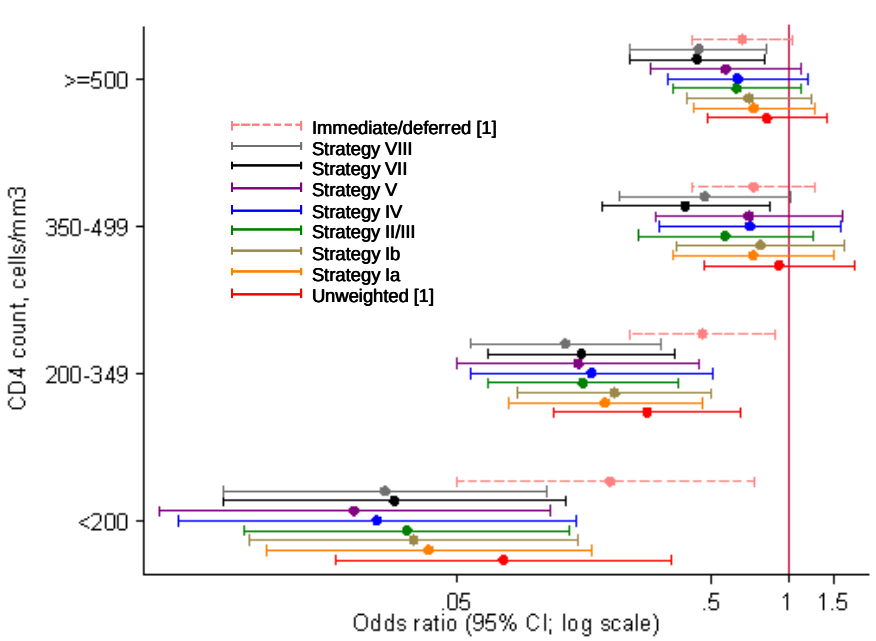


Figure 3.6: Effect of regimes Immediate versus No treatment on time to AIDS or death by trial-baseline CD4 count, and by the different strategies (treatment (artificial censoring) weights only). [1] "Unweighted" approach censored patients if they initially deferred but subsequently initiated treatment (biased for the causal effect of Immediate versus No treatment); "immediate/deferred" approach applied neither censoring nor weighting of such patients (different question).

Censoring The results after incorporating the weights for the "usual" censoring are shown in Table 3.9. The means, standard deviations and maxima of the estimated weights were all slightly larger than with the treatment weights only, most noticeably for strategy V where the standard deviation and maximum more than doubled (from 0.403 to 1.027 and from 8 to 24, respectively); this strategy had rather complex censoring models. These changes in the estimated weights were reflected in the larger standard errors of the average (across CD4 count strata) estimated treatment effects, except for strategies IV and V where the standard errors were smaller. These two strategies also yielded more extreme average estimated treatment effects once censoring was taken into account (average ORs reduced from 0.24 to 0.17 and from 0.21 to 0.17, respectively). The average estimated treatment effects under the other strategies were all broadly similar; those for strategies VII and VIII were slightly larger than previously but the confidence intervals were widely overlapping.

Strategy	Estimated weights			Estimated effect of regimes Immediate versus No treatment					
	Truncation	Mean (SD)	Range	Overall	CD4 count, cells/mm^3				$p^{[1]}$
					< 200	$200 - 349$	$350 - 499$	≥ 500	
Ia	0.5	1.017 (0.302)	0.002-3	0.32 (0.15, 0.66) [0.37]	0.06 (0.01, 0.33)	0.38 (0.16, 0.92)	0.98 (0.41, 2.31)	0.72 (0.32, 1.63)	0.02
Ib	1	1.019 (0.325)	0.06-3	0.32 (0.15, 0.67) [0.37]	0.06 (0.01, 0.30)	0.41 (0.17, 1.00)	1.03 (0.44, 2.42)	0.73 (0.35, 1.53)	0.02
II/III	0.1	1.032 (0.491)	< 0.001-11	0.28 (0.13, 0.57) [0.37]	0.06 (0.01, 0.29)	0.34 (0.14, 0.82)	0.82 (0.34, 1.97)	0.63 (0.27, 1.49)	0.03
IV	0.1	1.027 (0.523)	< 0.001-10	0.17 (0.06, 0.46) [0.50]	0.02 (0.003, 0.16)	0.32 (0.13, 0.79)	0.70 (0.27, 1.78)	0.49 (0.21, 1.15)	0.03
V	0.1	1.028 (1.027)	< 0.001-23	0.17 (0.07, 0.43) [0.46]	0.02 (0.004, 0.16)	0.35 (0.14, 0.90)	0.71 (0.28, 1.81)	0.38 (0.14, 0.99)	0.02
VII	0.5	0.965 (0.477)	0.003-5	0.26 (0.13, 0.53) [0.37]	0.05 (0.01, 0.28)	0.36 (0.15, 0.89)	0.55 (0.21, 1.42)	0.49 (0.21, 1.17)	0.09
VIII	0.5	1.007 (0.284)	0.002-3	0.24 (0.12, 0.46) [0.34]	0.05 (0.01, 0.24)	0.30 (0.13, 0.70)	0.70 (0.29, 1.66)	0.49 (0.21, 1.09)	0.04

Table 3.9: Estimated effect of Immediate versus No treatment, overall and by trial-baseline CD4 count, across the different model building strategies of chapter 2 (treatment (artificial censoring) and "usual" censoring weights). Treatment effect estimates are odds ratio (95% confidence interval, with robust standard errors) [robust standard error on the log-odds scale, where applicable]. [1] p-value for the interaction between treatment and trial-baseline CD4 cout category.

Treatment effect modification by CD4 count after censoring applied Considering the results by CD4 strata (Table 3.9 and Figure 3.7), the estimated ORs tended to be slightly larger (closer to one) in the < 200 cells/mm^3 stratum compared to previously, but the confidence intervals were somewhat larger. The most noticeable differences were seen in the $200 - 349$ cells/mm^3 stratum, where the ORs roughly doubled across all strategies when using censoring weights as well as treatment weights (for example from 0.19 to 0.38 under strategy Ia). The estimated treatment effects in the $350 - 499$ cells/mm^3 stratum also increased, but not quite so dramatically. In contrast, there was no clear pattern in the ≥ 500 cells/mm^3 stratum; under strategy Ia there was little change in the estimated treatment effect (0.73 compared with 0.72 previously), under strategy IV the estimated odds ratio dropped from 0.60 to 0.49, and under strategy VIII the estimated odds ratio increased from 0.44 to 0.49.

Model checking using country Figure 3.8 illustrates the estimated effects of the regimes Immediate versus No treatment by country for each of the different strategies (all CD4 count strata combined; estimation performed by incorporating an interaction between treatment and country). Visually, there appears to be some difference in the estimated treatment effects by country, although not statistically significant under any of the strategies (p-values shown in brackets in the Figure). As under the standard MSMs, the treatment effect estimates appeared to be strongest for Switzerland and Spain, while weakest for France and the UK. Overall, we may be reassured that there is no strong evidence of a difference in the treatment effect estimates by country.

3.4 Discussion

We have estimated that the effect of immediate treatment initiation compared to deferral (ignoring any subsequent treatment initiation), with straightforward adjustment for time-dependent covariate history, is associated with a 42% (17, 60) reduction in the risk of AIDS or death in our CASCADE population. Although not statistically significant, we observed a trend towards a greater benefit of immediate treatment initiation compared to deferral at lower current CD4 counts, with treatment associated with a 80% (23, 95) reduction in the risk of AIDS or death in those with current CD4 counts < 200 cells/mm^3 compared to 30% reduction (9% increase to 46% reduction) in those with current CD4 counts ≥ 500 cells/mm^3. These results were robust to a broad range of sensitivity analyses. One of these sensitivity analyses relaxed the LTFU and regular CD4 requirements, by not censoring patients within a trial if they had irregular

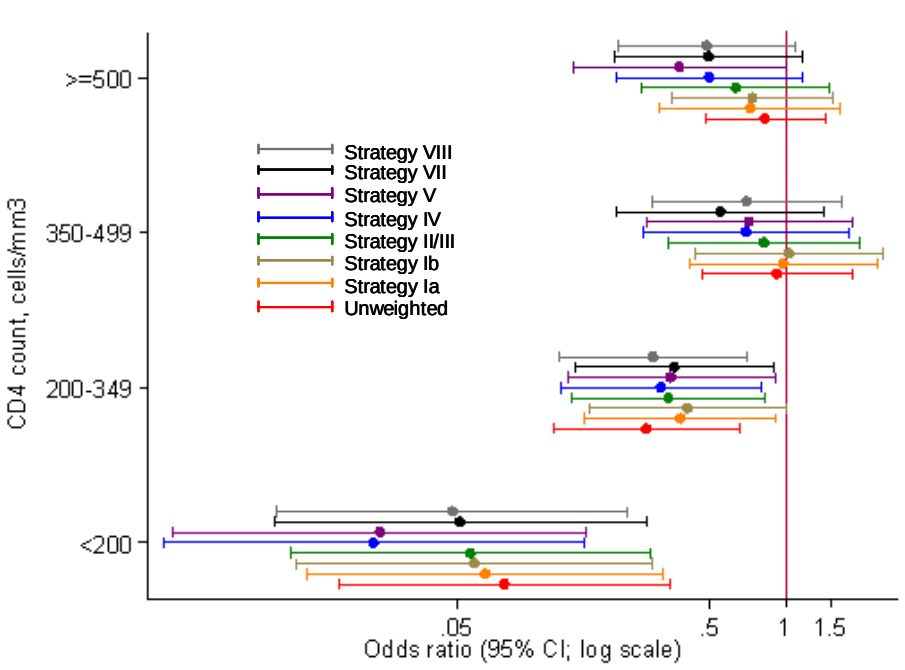


Figure 3.7: Effect of regimes Immediate versus No treatment on time to AIDS or death by trial-baseline CD4 count, and by the different strategies (treatment (artificial censoring) and "usual" censoring weights). "Unweighted" approach censored patients if they initially deferred but subsequently initiated treatment (biased for the causal effect of Immediate versus No treatment).

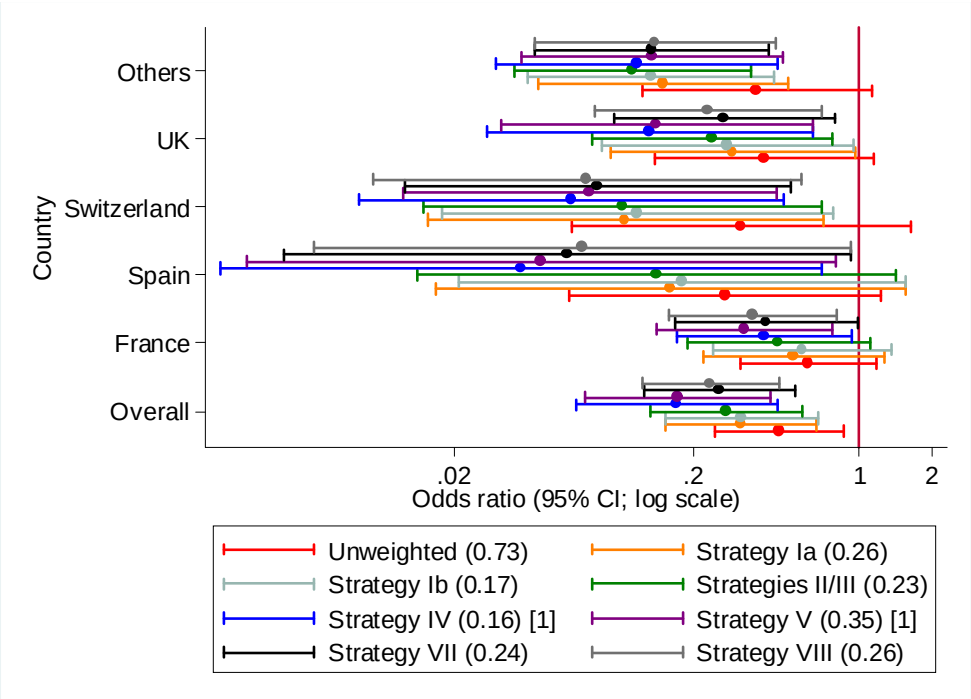


Figure 3.8: Estimated effect of regimes Immediate versus No treatment on time to AIDS or death by country. Values in brackets are the p-values for the interaction between regime and country. "Unweighted" approach censored patients if they initially deferred but subsequently initiated treatment (biased for the causal effect of Immediate versus No treatment). [1] Treatment model stratified by country.

CD4 counts or met the LTFU criteria (no CD4 count measured for > 12 months). However, patients did not contribute to new trials once considered censored under these criteria. This could have been further extended to include patients again if new CD4 counts were available, although there may be concern about what happened to those patients in the interim, which it may not be possible to adjust for using observed data.

We have outlined why "adherence-adjusted" estimates may be desirable, shown how these may be obtained using inverse probability weighting of HAMSMs, and applied these methods to our CASCADE population using a range of weight-estimation strategies. By "adherence-adjusted" estimates, we mean the estimation of the effect of the regimes Immediate versus No treatment. Although the principles of weight estimation are exactly the same as for standard MSMs, the weights are applied slightly differently. In particular, for a patient who initially deferred but subsequently initiated treatment, their follow-up is censored at the time of treatment initiation and therefore large weights, which may occur under standard MSMs due to a low probability of treatment initiation, may be avoided (Gran et al., 2010). This was apparent in the more stable weights observed here, with means close to one and small range, compared to those for the standard MSM.

After applying the appropriate censoring and weighting for those patients who initially deferred but subsequently initiated treatment, the estimated odds ratios were much further from one compared to those from the Immediate versus Deferred treatment analysis. This is as we would expect since the regime Deferred treatment encompasses a broad range of subsequent treatment paths and implicitly assumes that treatment will be started at some later time-point, rather than being artificially withheld forever. The results of this adherence-adjusted analysis of Immediate versus No treatment were broadly consistent across the different weighting strategies, as expected, with odds ratios ranging from 0.20 (0.11, 0.39) to 0.31 (0.15, 0.62). If weighting was not applied to account for the artificial censoring process, we saw a somewhat larger odds ratio (0.46 [0.25, 0.86]), which nicely illustrates the important role of weighting for unbiased estimation in this scenario.

We observed a stronger trend towards greater benefit of treatment at lower current CD4 counts under the regimes Immediate versus No treatment, compared to the regimes of Immediate versus Deferred treatment, due to the larger proportions of patients who initially deferred treatment at lower CD4 counts subsequently initiating treatment, as we might expect. Although the confidence intervals were wide, there were differences in the treatment effect estimates for current CD4 counts < 200, $200 - 349$ and ≥ 350 cells/mm^3, with strong evidence of a benefit of

treatment when current CD4 counts were < 200 and $200 - 349$ cells/mm^3 (0.06 [0.01, 0.33] and 0.38 [0.16, 0.92], respectively) but no clear benefit of treatment at when current CD4 counts of $350 - 499$ or ≥ 500 cells/mm^3 (0.98 [0.41, 1.63] and 0.72 [0.32, 1.63], respectively).

The overall effect estimate from the Immediate versus Deferred treatment regimes (odds ratio 0.58 [0.40, 0.83]) broadly resembled the treatment effect estimates from the standard MSMs (for example, 0.54 [0.32, 0.90] under strategy Ia), while the estimates from the Immediate versus No treatment regimes were considerably smaller (odds ratios of 0.20-0.31 as given above). As discussed in section 3.2.2, Gran et al. (2010) suggested that the treatment effects of interest from the standard MSMs and adherence-adjusted HAMSMs should be similar, and they found this to be the case. However, they also obtained similar estimates with and without the adherence-adjustments, that is, with and without taking into account subsequent treatment initiations in those patients who initially deferred treatment. The reason for this is not clear, but will be dependent on the subsequent treatment initiation patterns of those patients who initially deferred treatment in relation to their CD4 count trajectories. While the treatment effect estimates of Gran et al. (2010) were considerably further from one than our estimates (their overall HR was 0.17 [0.08, 0.34]), their results were not dissimilar to the estimates we obtained in the Swiss data (visible in Figure 3.8; their study used data from the Swiss HIV Cohort, which feeds in to CASCADE). Also, in contrast to their study, our treatment effect estimates were somewhat different with and without the weighting to adjust for adherence. Fundamentally, the standard MSMs and adherence-adjusted HAMSMs are estimating different quantities (Gran et al., 2010), so it is possible that the fact that their estimates were similar was coincidental.

The approach used by Writing Committee for the CASCADE Collaboration (2011) is most comparable to our analysis of the regimes of Immediate versus Deferred treatment regimes, since the authors permitted patients to follow any treatment path following initial deferral. However, they included all patients regardless of initial CD4 count, rather than focus on the group with high CD4 counts shortly after seroconversion. Overall, our results were broadly consistent with their findings, albeit with wider confidence intervals because of the smaller sample size due to our stringent eligibility criteria (our estimates: 0.20 [0.05, 0.77], 0.44 [0.22, 0.88], 0.79 [0.46, 1.37] and 0.70 [0.44, 1.09] for CD4 counts < 200, $200 - 349$, $350 - 499$ and ≥ 500 cells/mm^3, respectively; their results: 0.32 [0.17, 0.59], 0.48 [0.31, 0.74], 0.59 [0.43, 0.81], 0.75 [0.49, 1.14] and 1.10 [0.67, 1.79] for CD4 counts < 50, $50-199$, $200-349$, $350-499$ and $500-799$ cells/mm^3, respectively, with sample size 9455 patients).

As discussed in chapter 2, our population is unlike many others in that all patients enter

the risk set with a high CD4 count. In particular, this meant we had limited data on trials with low trial-baseline CD4 count. Longer follow-up may address this issue, however if all patients initiated treatment according to current guidelines then there would be no patients with CD4 counts < 350 cells/mm^3 remaining off treatment.

Of note, Hernán et al. (2008) performed the treatment model fitting with the time-dependent covariates (which then contributes to the denominator of the inverse probability weights) on the *expanded* data, with one treatment model per trial. This was possible because they had a limited number of trials, only 8. In contrast, we fitted the treatment models with the time-dependent covariates on the *unexpanded* data, following Petersen, Deeks, Martin, and van der Laan (2007), because we had a large number of trials (median 18 per patient, and one patient contributed to 147). No heterogeneity was detected between the trials, therefore it was possible to use a model pooled across the trials.

We have demonstrated treatment effect modification by time-dependent CD4 count, with treatment having a greater effect in those with lower current CD4 count. In chapter 5, we will return to the results presented here to compare with those obtained from the optimisation of dynamic treatment regimes, which are explored in the next chapter. As previously highlighted, while the application of history-adjusted and dynamic MSMs typically answer different questions, we might anticipate some consistency across the two approaches, and such a comparison may offer additional insights to the inference of interest.

Chapter 4

Dynamic marginal structural models

4.1 Introduction

In previous chapters, we have explored the estimation of causal effects using MSMs. We have proposed and applied a range of plausible strategies for the estimation of the inverse probability weights, and have considered treatment effect modification by baseline covariates. We progressed to HAMSMs, to allow the effects of treatment to depend on time-dependent covariates. We now move to an approach which will allow us to look directly at the estimation of pre-defined dynamic treatment regimes, that is a set of regimes which are defined in advance in terms of a patient's time-dependent covariates (see section 1.3). Our motivating clinical example is when to initiate treatment in HIV-infected persons, with respect to their CD4 count.

In this chapter, we begin by outlining the methodology of dynamic MSMs, which are a relatively recent approach. These methods have recently been extended to incorporate permitted delays in treatment initiation (grace periods; see section 4.2.2; Cain et al. (2010)). However, these extensions have been rarely applied in practice (Cain et al., 2010; HIV-CAUSAL collaboration, 2011; Shepherd et al., 2010), and their implications have not previously been investigated. We discuss and explore some of the issues surrounding these methods, firstly via simulation studies and then applied to the CASCADE data.

4.1.1 A hypothetical randomised trial

A recommended approach to defining dynamic causal questions is to consider the hypothetical randomised trial we would ideally conduct (Cain et al., 2010; Hernán et al., 2008). To address our question of when treatment should be initiated with respect to CD4 count in HIV-infected persons, we could imagine a trial which enrols treatment-naïve patients with CD4 counts ≥ 500 cells/mm^3 and randomises them to start treatment when their CD4 count

first falls below x cells/mm^3, with the range of x to be considered perhaps given by $x \in \{200, 210, 220, ..., 490, 500\}$. Alternatives to this set are discussed in section 4.6.3. As in any randomised trial, we may specify certain requirements in the protocol, such as that patients should have their CD4 count measured every month and start treatment within a month of their CD4 count reaching the value defined by their randomised regime x. It would then be possible to compare these regimes by looking at the AIDS-free survival at say 10 years, and selecting the optimal x as that which maximises 10-year AIDS-free survival. However, in practice it would not be trivial to conduct such a trial, since very large numbers of patients would be required with very long follow-up. We wish to mimic this randomised trial using causal methods with observational data; this could help inform a more limited set of potential optimal regimes for consideration in future trials.

4.2 Methodology

4.2.1 Dynamic marginal structural Cox model

As in the hypothetical randomised trial (section 4.1.1), HIV-infected treatment-naïve persons are included from the time of first observed CD4 count ≥ 500 cells/mm^3 and regimes are defined by:

"initiate treatment when observed CD4 count first falls below x cells/mm^3"

where $x \in \{200, 210, ..., 500\}$. For brevity, we refer to these regimes by their index x, for example regime $x = 350$ means to initiate treatment when observed CD4 count first drops below 350 cells/mm^3. Let T_x be the time to AIDS or death for a given patient under a regime x. If we could observe T_x for all patients and regimes x, or indeed if a sufficiently large number of patients were randomised to each x as in section 4.1.1, then the optimal regime x would simply be that which minimises the risk of AIDS or death across all patients, assuming constant treatment regime effects across patients. However, even in the absence of any other censoring, it is clearly not possible to observe T_x for all patients and regimes; in practice, for each patient we observe only a subset of regimes (which may be empty, or have one or more elements), and the regime(s) that any given patient is observed to follow may be confounded by their prognosis. In particular, any patient who initiated treatment at a CD4 count above their nadir (lowest value to date) no longer contributes to any regimes. Assuming for now that there is no censoring, for each patient we observe the time to event T and, under the assumption of consistency, T under observed x is T_x (section 1.2.4). For those regimes x which a patient is not compliant with throughout

follow-up, T_x remains counterfactual.

We define a dynamic marginal structural Cox model for the time to AIDS or death by:

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

where V is a vector of baseline covariates and $g(x)$ is some function of the regimes, which could for example simply be categorical or linear, or more complex such as a spline or fractional polynomial (Royston and Sauerbrei, 2008). This is an extension of the standard MSM (equation 2.2). Since T_x remains counterfactual for some patients and regimes x, we cannot fit this model directly. However, we can estimate the causal parameter of interest α of this MSM using inverse probability weighting methods, similar to those of section 2.2.3. There are three main steps to the method.

Step 1. All patients initially follow all regimes

As previously, we split time into suitable intervals, given by $t = 1, 2, ...$ We use exactly the same set-up and notation for time-dependent covariates $L(t)$, treatment $A(t)$ and outcome $Y(t)$ as introduced in section 2.2.1, with overbars representing history to that time. Of note, as in previous chapters, since we model outcome $Y(t+1)$ given $A(t)$, this means that we assume treatment in $[t, t+1)$ is independent of the outcome in that interval. We consider all patients to be compliant with all regimes initially; patients are then artificially censored from regimes when they first become noncompliant with that regime due to their covariate and treatment history. Define $C_x(t)$ to be an indicator for "artificial" censoring, taking value 0 if the patient's observed data is still consistent with regime x prior to time t, and value 1 otherwise.

Consider the example patient shown in Figure 4.1. This patient had monthly CD4 counts, where, as per the notation introduced in section 2.2.1, $CD4(t)$ refers to the latest CD4 count measured in $[t-1, t)$ (that is, the latest CD4 count assumed to be available to inform the treatment decision at time t). Treatment was initiated in the interval $[5, 6)$, therefore $A(t) = 0$ for $t \leq 5$ and $A(t) = 1$ for $t \geq 6$. The patient may then have experienced AIDS or death at some later time $t > 6$.

Firstly, imagine the simple case of just three regimes, defined by $x = 200, 350, 500$. The patient is compliant with all regimes to start with. Under the regime given by $x = 500$, we have $C_{500}(t) = 0$ for $t \leq 2$ and $C_{500}(t) = 1$ for $t \geq 3$, since the patient did not initiate treatment in the interval $[2, 3)$ in response to the first observed CD4 count below that threshold. It is important to note that although the censoring indicator takes the value 1 in this interval, any

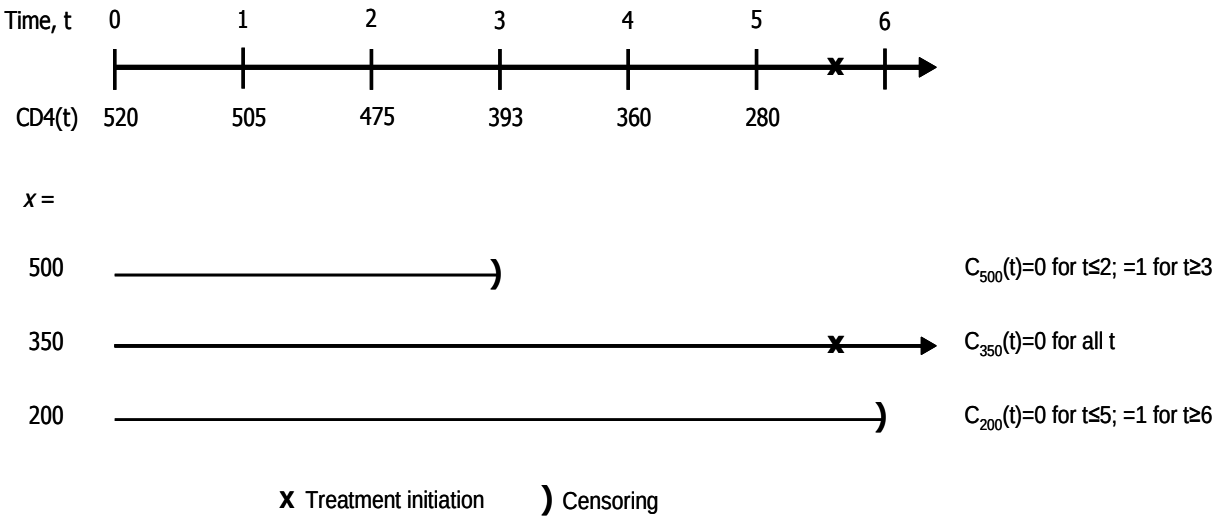


Figure 4.1: Illustration of compliance over time for an example patient with CD4 counts observed monthly and with regimes given by $x = 200, 350$ and 500 (no grace period). Recall that $CD4(t)$ is the latest CD4 count measured in the time interval $[t-1, t)$. $C_x(t)$ is an artificial censoring indicator; see text for more details.

AIDS or death events occurring in this interval would be included in the analysis; only AIDS or death events occurring from time 3 onwards are censored. The reason for this is to ensure that events occurring in patients who did and did not initiate treatment in a given interval are handled in the same way, to avoid introducing bias. Henceforth, when we refer to a patient being censored from (or non-compliant with) a regime from time t onwards, we mean that AIDS or death events occurring from time t onwards are no longer included in the analysis, and we have $C_x(s) = 0$ for $s \leq t-1$ and $C_x(s) = 1$ for $s \geq t$.

Considering the regime given by $x = 350$, since the patient did initiate treatment in response to his first observed CD4 count below that threshold, he is considered to be compliant with that regime for all time. Lastly, under the $x = 200$ regime, the patient is censored from time 6 onwards, since he initiated treatment in the interval $[5, 6)$ when his last CD4 count was still > 200 cells/mm^3.

Expanding this example to all regimes of interest given by $x = 200, 210, ..., 500$, Table 4.1 (no grace period) and panel A of Figure 4.2 illustrate the regimes the same example patient is considered to be compliant with over time (see section 4.2.2 for discussion on the grace period). Prior to time 3, this patient is compliant with all regimes since $CD4(t-1) > 500$ cells/mm^3 for $t < 3$. However, when he does not initiate treatment in response to $CD4(2) = 475$ cells/mm^3, he is censored from time 3 onwards from all higher regimes given by $x > 475$ and therefore is still compliant only with the 28 regimes given by $x = 200, 210, ..., 470$. While the patient remains off treatment, the number of regimes with which he remains compliant drops with his observed

t	$CD4(t-1)$ (cells/mm^3)	$A(t)$	Regimes from which uncensored from t onwards			
			No grace period ($m=1$)		Grace period ($m=2$)	
			$x=$	N	$x=$	N
1	520	0	200-500	31	200-500	31
2	505	0	200-500	31	200-500	31
3	475	0	200-470	28	200-500	31
4	393	0	200-390	20	200-470	28
5	360	0	200-360	17	200-390	20
6 onwards	280	1	290-360	8	290-390	11

Table 4.1: Compliance over time of the example patient of Figure 4.1 with multiple regimes given by $x = 200, 210, ..., 500$ with no grace period ($m = 1$) and a grace period ($m = 2$; see section 4.2.2). Recall that $CD4(t-1)$ is the latest CD4 count measured in the time interval $[t-2, t-1)$.

CD4 count. Once treatment was initiated following the observed CD4 count of 280 cells/mm^3, the patient is thereafter compliant with just the 8 regimes given by $x = 290, 300, ..., 360$. Of note, once a patient is observed to initiate treatment, they will never be censored off the regimes with which they were compliant at treatment initiation.

Cain et al. (2010) indicate that an alternative to this expansion method would be to randomly allocate each patient to one of the multiple regimes with which they are compliant, although this would be statistically inefficient compared to the approach applied here of including all patients on all regimes which with they remain compliant, and adjusting the variance estimates accordingly for multiple observations per patient (either approach requires the weighting as detailed in the next section for unbiased estimation).

Formalising our notation, let $Q_x(t)$ be an indicator for whether a patient's CD4 count has dropped $< x$ cells/mm^3 prior to time t. Then the censoring indicator $C_x(t)$ is a deterministic function of A, Y and x, given for $t = 1, 2, ...$ by:

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

and $C_x(t)$ is missing if $Y(t) = 1$. As noted above, if $A(t) = 1$ and $C_x(t) = 0$ then $C_x(s) = 0$ for all $s > t$, since if treatment was initiated in compliance with a given regime then the patient will be compliant with that regime for the remainder of their follow-up. This broadly follows Cain et al. (2010) but we have used discrete indicators. Of course, this artificial censoring process is likely to be informative; we take account of this using inverse probability weighting.

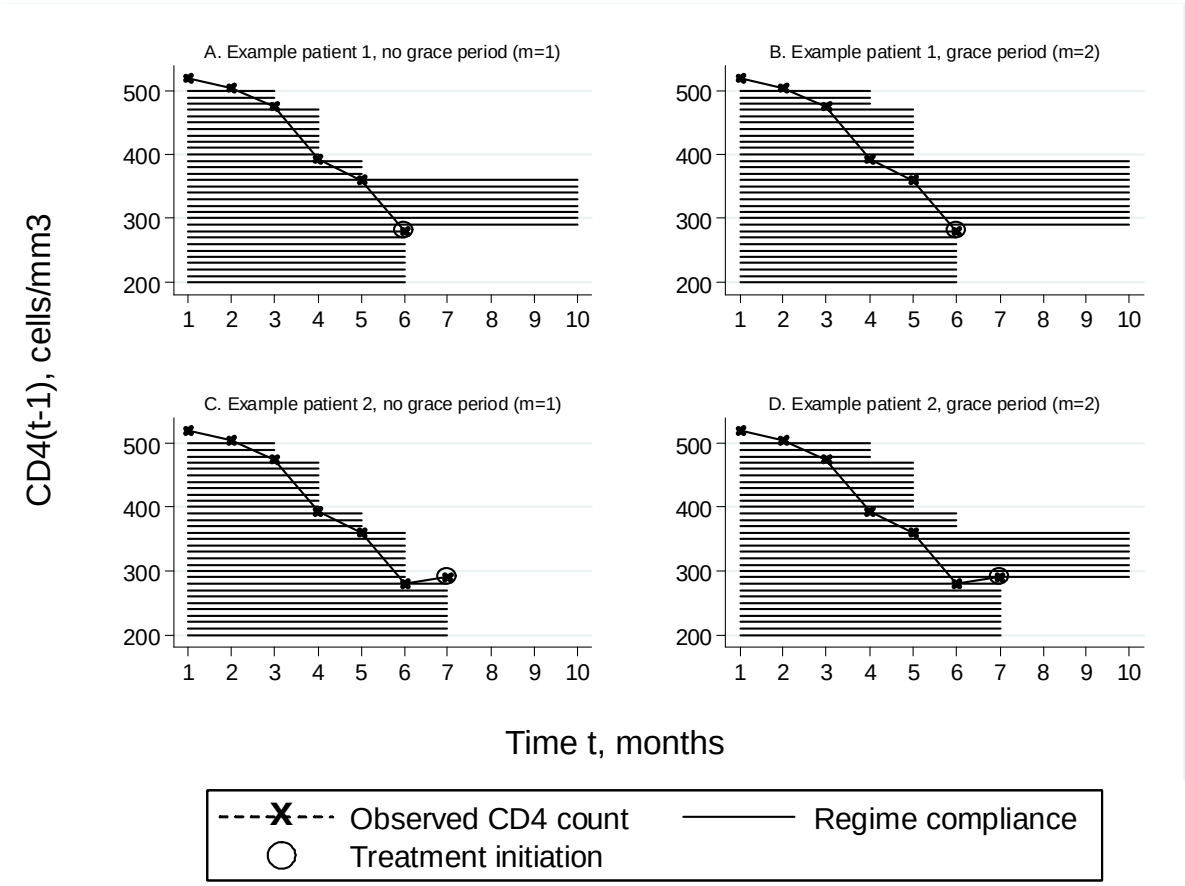


Figure 4.2: Illustration of compliance over time with regimes given by $x = 200, 210, ..., 500$ of: the example patient of Figure 4.1, (A) under no grace period ($m = 1$) and (B) with a grace period of $m = 2$ months, and a second patient who has the same CD4 trajectory as the first patient, but delays treatment initiation for one month, by which time his CD4 count increased above the nadir, (C) under no grace period ($m = 1$) and (D) with a grace period of $m = 2$ months. Recall that $CD4(t-1)$ is the latest CD4 count measured in the time interval $[t-2, t-1)$.

Step 2. Estimate inverse probability weights

Under the assumption of no unmeasured confounders for censoring and outcome (see section 2.2.2), and given baseline and time-updated covariates and treatment history, the weight for a patient on regime x at time t is the inverse probability of remaining uncensored to t:

$$W_x(t) = \frac{I\left[C_x(t) = 0\right]}{\prod_{j=1}^{t} \Pr\left\{C_x(j) = 0 | \overline{C_x}(j-1) = 0, Y(j) = 0, \overline{L}(j-1)\right\}}$$

where $I\left[\cdot\right]$ is an indicator equal to 1 if $\cdot$ is true, and 0 otherwise.

As with HAMSMs, the probability of remaining uncensored for any given regime to a given time is the same as the probability of the observed treatment history to that time, conditional on baseline and time-updated covariates and treatment history (Hernán et al., 2006; Robins et al., 2008). Therefore the weights can equivalently be given by:

$$\begin{aligned} W_x(t) &= \frac{I\left[C_x(t) = 0\right]}{\prod_{j=1}^{t} \Pr\left\{A(j) | \bar{A}(j-1), Y(j) = 0, \bar{L}(j-1)\right\}} \\ &= \frac{I\left[C_x(t) = 0\right]}{\prod_{j=1}^{t} p_A(j)^{I[A(j)=0]} \left\{1 - p_A(j)\right\}^{I[A(j-1)=0, A(j)=1]}} \end{aligned}$$

where

$$p_A(j) := \Pr\left\{A(j) = 0 | \overline{A}(j-1) = 0, Y(j) = 0, \overline{L}(j-1)\right\}$$

is the probability of not initiating treatment given covariate history as in equation 2.3, noting that after treatment initiation the probability of treatment is 1. As in section 2.2.3, we can estimate $p_A(j)$ from the data using a pooled logistic regression model. The treatment probabilities are independent of treatment regime x, so we fit this model on a dataset with one observation per patient (per month), and when we expand to one observation per patient per regime (per month), at any given time the weights are constant for each patient across all regimes with which they are still compliant.

Stabilisation of the weights As in section 2.2.3, we may wish to stabilise the weights. However, we cannot use the same approach as for standard MSMs; not only must the numerator of the stabilised weights not depend on time-updated covariates, but it also cannot depend on past treatment. That is, the numerator can depend on x, V and Y, but not on L nor A. Cain et al. (2010) suggest a natural choice is the cumulative product over time of

$\Pr\left\{C_x(j) = 0 | \overline{C_x}(j-1) = 0, Y(j) = 0, x, V\right\}$. Therefore while the patient remains uncensored, the stabilised weights may be given by:

$$SW_x(t) = \frac{\prod_{j=1}^{t} \Pr\left\{C_x(j) = 0 | \overline{C_x}(j-1) = 0, Y(j) = 0, x, V\right\}}{\prod_{j=1}^{t} p_A(j)^{I[A(j)=0]} \left\{1 - p_A(j)\right\}^{I[A(j-1)=0, A(j)=1]}}$$

We can similarly fit a pooled logistic regression model for the numerator, but, because x is a factor in the linear predictor, the model for the numerator must be estimated on the expanded data (that is, with all patients following all regimes until censored, as in Step 1, so one observation per patient per regime [per month]) and over all time intervals in which patients remain uncensored, regardless of whether they are on treatment on not. Separate numerator models could be used for each treatment regime, or if there are many regimes then it may be more efficient to use just one model incorporating regime, perhaps as a smooth function.

If a patient has initiated treatment in accordance with a given regime, then we know they cannot be censored from that regime, but they will continue to contribute to the numerator censoring model. This means that, while the non-stabilised weights will remain constant after treatment initiation, the stabilised weights will not. Non-stabilised weights essentially weight the data such that all patients follow all regimes for all time (with patient numbers at later times declining only because of patients dropping out due to events, or "usual" censoring). Conversely, the stabilised weights depend on the "artificial" censoring process, therefore the weighted follow-up will reflect that of the (artificially) censored but unweighted follow-up over time.

Cain et al. (2010) state that these stabilised weights may not necessarily reduce the variance. It has also not been described how to stabilise the weights under a scenario with uniform initiation across a grace period (see section 4.2.2), which is not trivial. Therefore, we will only use non-stabilised weights hereafter.

Step 3. Weighted discrete-time survival regression with a smooth function for regime x

Once we have estimated the weights, a simple approach would be to estimate the (weighted) survival for each regime. In realistic scenarios, this may be somewhat unstable, since few patients will be following any one regime at a given time. Instead, Cain et al. (2010) suggest applying a pooled logistic regression model to the weighted data with one observation per patient per regime (per month) and using a smooth function for regime x. That is, they suggest employing

a model such as:

$$\text{logit}\Pr\{Y(t+1)=0|Y(t)=0,C_x(t)=0,x,V\}=\exp\{\alpha g(x)+\beta V+\gamma f(t)+\delta g(x)f(t)\} \tag{4.1}$$

where $f(t)$ is some function of time, as in equation 2.4. The parameters of this model may be estimated using weighted maximum pseudo-likelihood, with robust standard errors, since there are multiple non-independent observations per patient. Of note, as highlighted above, we model $Y(t+1)$ conditional on $C_x(t)=0$, therefore even if $C_x(t+1)=1$, we include AIDS or death events which occur in that interval $[t,t+1)$.

The assumption of proportional hazards is highly likely to be implausible when looking at dynamic treatment regimes. For example, consider the regimes given by $x=200$ and 350; these regimes are identical until the patient's CD4 count drops <350 cells/mm^3. Therefore it is important to allow for time-dependent effects of the regimes ($g(x)f(t)$ in the above pooled logistic regression model) and so we will consider survival curves rather than hazard ratios. We will plot survival curves, estimated from the pooled logistic regression parameters, over time and focus on 10-year AIDS-free survival rates. The optimal regime x is therefore determined as that which minimises the risk of AIDS or death by 10 years.

Alternative approaches There are alternatives to the pooled logistic regression model as outlined above. For example, we could first obtain the weighted Kaplan-Meier estimates for the event of interest for each regime, and then perhaps perform some smoothing over these estimates. A global procedure would require modelling on the bounded $[0,1]$ scale of the survivor function which is unlikely to be appropriate. Rather, a local smoothing procedure may be preferable, although if there is a great deal of uncertainty in the estimates, then relatively heavy local smoothing may be required. The estimation of standard errors would not be straightforward, but bootstrapping could be applied.

4.2.2 Grace periods

In clinical practice, there may be a delay between the taking of bloods for CD4 count measurement, performing the analysis, informing the patient of the results and the patient being prescribed and finally initiating treatment. Further delay may result if the patient or clinician requests a second confirmatory CD4 count before initiating treatment, although the treatment initiation may still be considered in response to the first CD4 measurement. For these reasons, allowing delayed initiation may better reflect the processes that led to the observed data. Cain

et al. (2010) refer to this permitted delayed action as a "grace period" indexed by m. Formally, the regimes are defined as:

"initiate treatment within m months after observed CD4 count first falls below x cells/mm^3"

and such regimes, with a permitted delay, may be more typical of those implemented via the protocol of an RCT. Note that Cain et al. (2010) consider "immediate" initiation of treatment to be given by $m = 0$, but since "immediate" initiation refers to within the first month we prefer to consider this as $m = 1$ and so true grace periods here are given by $m \geq 2$. As in the scenario where there is no grace period, patients are censored if they initiate treatment *before* becoming eligible for a given regime (that is, observed CD4 count $> x$ cells/mm^3). However, for all other patients, since we are allowing $m > 1$ months for initiation after having observed CD4 count dropping below the given threshold, none will be censored during the grace period, but will be censored after the m^{th} interval of the grace period if they have not initiated treatment by that time. As in the $m = 1$ situation, patients who are not censored at that point will remain uncensored for the rest of their follow-up. Our definition of $C_x(t)$ can be extended to allow for a grace period of m months as follows (Cain et al., 2010):

$$\begin{aligned} C_x(t) = 0 \quad \text{if and only if, for all } j \leq t, \quad & A(j) = 0 \text{ when } Q_x(j-1) = 0, Y(t) = 0 \\ \text{and} \quad & A(j+m-1) = 1 \text{ when } Q_x(j-1) = 1, Y(t) = 0 \end{aligned}$$

and again $C_x(t)$ is missing if $Y(t) = 1$. Again, note that we model $Y(t+1)$ conditional on $C_x(t) = 0$.

Example patients

Consider our example patient (Figure 4.1), whose regime compliance permitting a grace period of $m = 2$ months is given in the last two columns of Table 4.1 and illustrated in panel B of Figure 4.2. Compared to the scenario with no grace period, the patient is compliant with at least as many regimes within each time interval, and often more.

A benefit of a grace period is that if the patient's observed CD4 count rises slightly from the nadir before treatment initiation, then the patient may still be considered to be compliant with some regimes with which they would not have been considered compliant if no grace period were permitted. For example, consider a second patient with the observed CD4 counts as given in Table 4.2. The covariate and treatment history of this patient is the same as that of the first example patient, except that this patient delayed treatment for one month, by which time

t	$CD4(t-1)$ (cells/mm^3)	$A(t)$	Regimes from which uncensored from t onwards			
			No grace period ($m=1$)		Grace period ($m=2$)	
			$x=$	N	$x=$	N
1	520	No	200-500	31	200-500	31
2	505	No	200-500	31	200-500	31
3	475	No	200-470	28	200-500	31
4	393	No	200-390	20	200-470	28
5	360	No	200-360	17	200-390	20
6	280	No	200-280	9	200-360	17
7 onwards	290	Yes	-	0	290-360	8

Table 4.2: Compliance of a second example patient with multiple regimes over time give by $x = 200, 210, ..., 500$ with no grace period ($m = 1$) and a grace period ($m = 2$). Recall that $CD4(t-1)$ is the latest CD4 count measured in the time interval $[t-2, t-1)$.

his observed CD4 count had risen slightly from 280 to 290 cells/mm^3. Under no grace period, this patient would be censored from all regimes from time 7 onwards (illustrated in panel C of Figure 4.2). However, under a grace period of $m = 2$ months, this patient is considered to be compliant with the eight regimes given by $x = 290, 300, ..., 360$ from time 7 onwards, a scenario which is perhaps more clinically realistic given known measurement error and natural fluctuations in CD4 count (Table 4.2 and panel D of Figure 4.2).

In observational data, since such fluctuations in observed CD4 count and treatment initiation patterns may be common, permitting a grace period may result in fewer of the observed treatment initiations being censored, therefore perhaps leading to more efficient estimation. It is important to note that to allow a grace period is to ask a different question, that is the effect of the regimes permitting a maximum delay in treatment initiation, compared to the original question which considers the effects of the regimes assuming no delay in treatment initiation. However, one may be prepared to accept the potential bias associated with an interpretation assuming no grace period, although one was permitted for analysis, in order to exploit the potential gain in efficiency. Below, we seek to evaluate these trade-offs through simulation.

Regimes are not fully identified

In the presence of a grace period, the regimes are not fully identified, since for each x there is more than one possible treatment path which is consistent with the definition, namely those in which treatment is initiated in any of the m intervals of the grace period. Of note, Young et al. (2011) refer to regimes without grace periods as deterministic, and to those with grace periods as random, since there may be a random element to the time at which treatment is initiated during the grace period. Cain et al. (2010) considered two examples. The first can be more

precisely specified as "do not initiate treatment before the CD4 count is $<x$ cells/mm^3, and do initiate exactly m months after the CD4 count first drops below x cells/mm^3 if treatment has not already been initiated in the first $m-1$ months of the grace period". The authors describe their second example as "initiate treatment within m months after the CD4 count first drops below x cells/mm^3, such that there is a uniform probability of starting in each of the months $1, 2, ..., m$", though note this is still not fully specified since the treatment probabilities could be conditional on covariates such as CD4 count yet still achieve uniform marginal probabilities of treatment initiation across the grace period. These choices have implications for the weight estimation; we now describe how the weights may be estimated in each of these two scenarios.

Weight estimation

Note that since we model $Y(t+1)$ (equation 4.1), the weights $W_x(t)$ are used to upweight outcome in the next month. Patients who reach the end of the grace period without initiating treatment are censored at the end of the grace period; there is no censoring during the grace period.

First approach The weight estimation is simplest under the first approach of Cain et al. (2010), where only the patients who initiated treatment in the m^{th} interval of the grace period are weighted up (in the next month) to account for those censored at the end of the grace period due to non-initiation of treatment. Let the time q_x be such that $Q_x(q_x - 1) = 0$ and $Q_x(q_x) = 1$. Then the (non-stabilised) weights are estimated as follows:

$$W_x(t) = \frac{I\left[C_x(t) = 0\right]}{\left\{\prod_{j=1}^{t} p_A(j)^{I[j<q_x]}\right\} \times \left\{1 - p_A(q_x + m)\right\}^{I[t \geq q_x + m]}}$$

where $p_A(j)$ is estimated from the data as in the case where $m = 1$, that is, on the dataset before expansion. The first component of the denominator of these weights is the probability of remaining uncensored while CD4 count is $\geq x$ cells/mm^3, that is, off treatment. The second component of the denominator is the probability of remaining uncensored after the m^{th} interval of the grace period (which is m months after treatment indicated by the regime and CD4 count history); this probability is given by the probability of treatment initiation at that time. Therefore, those patients who initiated treatment in the m^{th} interval of the grace period are upweighted to account for those censored at the end of the grace period due to non-initiation of treatment.

Second approach Under the second approach of Cain et al. (2010), we assume that the probability of treatment initiation is uniform across the grace period, and the patients who initiated treatment at any point during the grace period are weighted up to account for those censored at the end of the grace period because they did not initiate by that time. The (non-stabilised) weights are estimated as follows:

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

The first component of the weights is identical to that of the first approach (the probability of remaining uncensored, that is, off treatment, while CD4 count is $\geq x$ cells/mm^3). The second part of the weights spans the grace period, $l = 1, ..., m$, that is, covering the m months after treatment is indicated by the regime and CD4 count history. The denominator is based on the probabilities of observed treatment, that is, the probability of remaining off treatment while treatment-naïve during the grace period, multiplied by the probability of initiating treatment when (if) it is initiated during the grace period. The numerator of the second part of these weights is to form the uniform distribution of treatment initiation over the grace period. In the l^{th} interval of the grace period, the numerator takes value $1/(m+1-l)$ for patients who initiated in that interval (that is $1/m, 1/(m-1), ..., 1/2, 1$ for intervals $l = 1, ..., m$, respectively) and value $1 - 1/(m+1-l)$ for those who did not initiate in that interval (that is values $1-1/m, 1-1/(m-1), ..., 1/2, 0$ for intervals $l = 1, ..., m$, respectively). Given this adjustment, it is difficult to express these weights in terms of the probability of remaining censored at a given time t, but they serve to upweight those patients who initiated during the grace period to account for those who are censored at the end of the grace period due to non-initiation of treatment.

Comparison of these approaches

The first approach upweights the patients who initiated treatment in the m^{th} interval of the grace period to account for those who did not initiate by the end of the grace period; this may potentially be a small subset of patients who may not be comparable to those who initiated earlier in the grace period. The second approach assumes uniform treatment initiation across the grace period which also may not be plausible; for example, if m is large then perhaps patients may be more likely to initiate earlier in the grace period, with few patients delaying treatment initiation for m months.

Of course, other choices are possible. For example, suppose it was anticipated that if the regimes were implemented in practice then the majority of patients would initiate in the first interval of the grace period. Then one could assume for example 80% of the patients would initiate in the first interval, and uniform initiation across the remainder of the grace period. Further, the treatment initiation pattern could be data-driven, that is based on what was observed in the real data. However, the correct weights would need to be determined (adjustments made to the numerator) and, strictly-speaking, the results should then be interpreted in the same vein. This approach would only be advantageous if it was thought that clinicians would employ the same treatment initiation patterns when implementing the results of the study, which is perhaps unlikely, since if they are changing practice then the treatment initiation patterns are likely to also change.

4.2.3 Other censoring

Other types of "usual" censoring, such as LTFU or administrative, may be incorporated in a similar way as for standard MSMs (section 2.2.4).

4.2.4 Interactions between treatment effect and baseline characteristics

We have so far assumed a constant regime effect across all patients; that is, the optimal regime(s) is the same for all patients. In reality, the optimal regime x may vary by baseline patient characteristics such as age or sex. These can be addressed using interactions, for example by replacing the function $g(x)$ of regime x in the pooled logistic regression model for the outcome (equation 4.1, in the components $\alpha g(x) + \delta g(x) h(t)$) with some function $g(x, V)$ of the regime x and baseline covariates V, for example:

$$g(x, V) = \sum_j (1 + \delta_j V) x^{p_j}$$

Robins et al. (2008) described the use of such interactions, but to our knowledge this has not been done in practice in the context of optimal dynamic treatment regimes.

4.2.5 Gaps in the methodological literature

While dynamic MSMs have been applied in practice a number of times previously (Cain et al., 2010; HIV-CAUSAL collaboration, 2011; Shepherd et al., 2010), and we know that asymptotically the methods are unbiased for causal estimation of the effects of dynamic treatment regimes (Robins et al., 2008), their performance in realistically-sized datasets has not been sys-

tematically explored. In addition, the impact on the optimal regime of factors such as the rate of decline and measurement frequency of the biomarker which is used to define the dynamic regimes, and also length of the grace period, has not been investigated. Such knowledge could, for example, enable comparison between different studies which have been conducted under different conditions to help understand any differences in the results.

Further, while we may be interested in inferences under the assumption of no grace period, which may be easier to interpret and implement in practice, there could perhaps be potential gains in efficiency by permitting a grace period for the purposes of analysis, since fewer treatment initiations will be censored. This may be at the risk of bias for the inference of interest (assuming no grace period); this bias-variance trade-off has not previously been studied.

We investigated these issues via simulation studies.

4.3 Simulation study 1

4.3.1 Motivation

There were two overarching aims for our first simulation study. The first was concerned with the effects of different CD4 observation frequencies and grace periods on the optimal regime. Since we are defining the optimal regimes in terms of maximising a time-to-event outcome, it is not possible to easily determine the optimal regimes directly. Therefore we simulated large randomised trials for this purpose. The second aim was related to the performance of these methods in realistic situations, therefore we simulated observational studies.

First aim (via randomised trials)

Our first aim was to explore the effects of different observation frequencies of CD4 count (for the purposes of treatment initiation) and different length grace periods on the optimal regime. As mentioned above, this has not previously been systematically investigated. These two issues are clearly closely related. For example, individuals monitored less frequently or permitted a delay in treatment initiation may need to be directed to initiate earlier at higher CD4 counts to prevent long periods of time before treatment initiation and hence CD4 counts dropping to levels associated with increased risk of AIDS or death. We also considered populations with different average treatment-naïve CD4 declines. Of note, these scenarios with different CD4 declines, CD4 count observation frequencies and grace periods are expected to lead to different results since they are addressing different questions. As mentioned above, initially we were interested in the true effects of these factors on the optimal regime, therefore we simulated large

randomised trials to address these issues.

Second aim (via observational studies)

Our second aim was to explore the performance of these methods using realistically-sized observational datasets, in terms of bias and precision. We know that asymptotically the methods will be unbiased, but in practice limited data will be available.

A key motivation for incorporating grace periods is to attempt to minimise the number of censored treatment initiations in (likely limited) observational data, thereby aiming to increase efficiency. Since permitting a grace period is to ask a different question than a scenario without a grace period, we were also interested in the potential bias arising from interpreting results from a study with a grace period as if there was no grace period. If the gain in efficiency outweighed the potential bias, then even in scenarios where inference was desired in the absence of a grace period, it may be beneficial to allow a grace period anyway for the purposes of analysis. This bias-variance trade-off has not previously been studied, and was therefore part of our second aim.

To investigate these issues, we simulated realistically-sized observational studies. In particular, the questions we wished to address were:

1. With realistically-sized datasets, are the methods unbiased (compared to the results from the RCT simulations for the same population in terms of treatment-naïve CD4 decline and CD4 count observation frequency, and also the same grace period)?

2. What is the precision of a single analysis of this size?

3. What is the bias-variance trade-off in allowing grace periods of $m > 1$ months, when in fact the question of interest is under the scenario of no grace period ($m = 1$)? That is, assuming that we want to interpret the results under no grace period, we compared the results from the RCT with a given population (in terms of treatment-naïve CD4 decline and CD4 count observation frequency) and no grace period with the results from the observational studies with the same population but permitting a grace period. As discussed above, by increasing the grace period, we may gain efficiency but potentially at the expense of bias for the inference of interest.

4.3.2 Methods

Data generation

Simulated patients were included with observed baseline CD4 counts uniformly in $[500, 550]$ cells/mm^3; this narrow range of baseline CD4 counts was chosen to avoid lengthy amounts of time spent with CD4 count ≥ 500 cells/mm^3, which would not contribute to the comparison between the defined regimes. Follow-up over 10 years was divided into monthly intervals.

Modelling CD4 count trajectory Our models were based on previous work modelling CD4 count using CASCADE data by A Babiker (personal communication, 10 September 2010). This previous work suggested a piecewise linear mixed effects model for square-root CD4 count, with a change-point at treatment initiation and one year after initiation, and incorporating Brownian motion (this was superior over the standard mixed effects model). We now describe these models in more detail, and give the parameters as estimated by that previous work. Note that all the following parameter estimates are those which were used for the population with regular treatment-naïve CD4 decline; the changes made to consider populations with different CD4 declines are described below.

Let $CD4_i^T(t)$ and $CD4_i^O(t)$ represent the true and observed CD4 count, respectively, for patient i at time t. Measurement error was incorporated as follows:

$$\sqrt{CD4_i^O(t)} = \sqrt{CD4_i^T(t)} + E_i(t) \tag{4.2}$$

where $E_i(t)$ are independent random measurement errors with distribution $N(0, \sigma_E^2)$, with the variance σ_E^2 dependent on whether treatment had been initiated or not. We used the following model for the treatment-naïve CD4 trajectory:

$$\sqrt{CD4_i^T(t)} = B_i^T + S_{0,i}(t/12) + W_{0,i}(t)$$

where time t is measured in months and, for patient i, $(B_i^T)^2$ is the true baseline CD4 count and $S_{0,i}$ is the random slope drawn from $N(\mu_{S_0}, \sigma_{S_0}^2)$. Note that B_i^T is random, determined from the uniformly-simulated observed baseline CD4 count $CD4_i^O(0)$ in $[500, 550]$ cells/mm^3 (see above) and equation 4.2. No correlation between the baseline CD4 count and subsequent slope was incorporated, since time was from trial entry and patients are captured within a narrow range of observed baseline CD4 counts (of note, the range of true baseline CD4 counts was somewhat wider than the observed due to the measurement error, though not large; see

results in Table 4.3). Outside a simulation model, we would be unlikely to capture patients within such a narrow range of CD4 counts, but here the baseline CD4 count is not for example representative of the CD4 count at seroconversion, so we would not necessarily expect a correlation between the baseline CD4 count and subsequent slope. $W_{0,i}(t)$ represents the Brownian motion process, which is independent of the baseline CD4 count or slope, and has $W_0(0) = 0$, distribution $N(0, \delta_0 t/12)$ and $corr\left[W_0(t_1), W_0(t_2)\right] = min(t_1, t_2)/\sqrt{t_1 t_2}$. Appendix A describes how $W_{0,i}(t_2)$ is simulated in practice, given $W_{0,i}(t_1)$. The parameters were previously estimated from CASCADE data to be $\mu_{S_0} = -1.10, \sigma_{S_0} = 0.50, \delta_0 = 6.89$ and $\sigma^2_{E_1} = 2.26$.

After treatment initiation, the CD4 trajectory was modelled as follows, with time t' in months from treatment initiation:

$$\sqrt{CD4_i^T(t')} = \begin{cases} R_i + S_{1,i}(t'/12) + W_{1,i}(t') & \text{if } t' < 12 \text{ months} \\ R_i + S_{1,i} + S_{2,i}(t'/12 - 1) + W_{1,i}(t') & \text{if } t' \geq 12 \text{ months} \end{cases}$$

where, for patient i, R_i^2 is the true CD4 count at treatment initiation, and $S_{1,i}$ and $S_{2,i}$ are the slopes during the first year and from one year after treatment initiation, respectively, on the square-root scale. Therefore, R_i^2 is known (the true CD4 count at treatment initiation) and $S_{1,i}$ and $S_{2,i}$ were simulated conditional on R_i (see appendix A). Overall, these three random variables followed a trivariate Normal distribution, with mean vector and variance-covariance matrix given by:

$$\begin{aligned} \mu &= \begin{pmatrix} \mu_R \\ \mu_{S_1} \\ \mu_{S_2} \end{pmatrix} = \begin{pmatrix} 19.69 \\ 2.93 \\ 0.10 \end{pmatrix} \\ \text{and } \Sigma &= \begin{pmatrix} \sigma^2_R & \sigma_{R,S_1} & \sigma_{R,S_2} \\ \sigma_{R,S_1} & \sigma^2_{S_1} & \sigma_{S_1,S_2} \\ \sigma_{R,S_2} & \sigma_{S_1,S_2} & \sigma^2_{S_2} \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2_R & r_{R,S_1}\sigma_R\sigma_{S_1} & r_{R,S_2}\sigma_R\sigma_{S_2} \\ r_{R,S_1}\sigma_R\sigma_{S_1} & \sigma^2_{S_1} & r_{S_1,S_2}\sigma_{S_1}\sigma_{S_2} \\ r_{R,S_2}\sigma_R\sigma_{S_2} & r_{S_1,S_2}\sigma_{S_1}\sigma_{S_2} & \sigma^2_{S_2} \end{pmatrix} \end{aligned}$$

where $r_{x,y} = corr(x, y) = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$. The variance-covariance parameters were previously estimated from CASCADE data to be $\sigma_R = 5.71, \sigma_{S_1} = 2.06, \sigma_{S_2} = 0.57$ and $r_{R,S_1} = -0.44$; r_{R,S_2} was found to be -1, that is the slope after one year after treatment initiation was a linear function of R_i (the square-root of the true CD4 count at treatment initiation), and therefore

$r_{S_1,S_2} = -r_{R,S_1}$. Once again, $W_{1,i}(t)$ represents a Brownian motion process, independent of R, S_1 and S_2, with $W_1(0) = 0$, distribution $N(0, \delta_1 t/12)$ and correlation as given above (simulated as described in appendix A). The remaining parameters were given by $\delta_1 = 7.83$ and $\sigma^2_{E_2} = 2.19$.

In practice, successive CD4 counts were determined as follows:

$$\sqrt{CD4_i^T(t)} = \sqrt{CD4_i^T(t-1)} + S_{j,i}/12 - W_{j,i}(t-1) + W_{j,i}(t)$$

That is, the random effects and Brownian motion components were additive. In our simulation study, CD4 counts were truncated at 0 if estimated as < 0 cells/mm^3, and values > 1000 cells/mm^3 were truncated at 1000 cells/mm^3, due to the high biological variation at such high CD4 counts and little difference in the probability of reaching AIDS/death (or initiating treatment, in the observational study) at those levels. The numbers of observations truncated at each of these limits are shown in the results (Tables 4.3 and 4.8).

Modelling event rates Let $p_i(t)$ represent the probability of AIDS or death at a given time t for patient i; this was dependent on true CD4 count and treatment, as follows:

$$\begin{aligned}
\log \frac{p_i(t)}{1-p_i(t)} &= \begin{cases} \lambda_0 - \nu_0\sqrt{CD4_i^T(t-1)} & \text{if } A_i(t-1) = 0 \\ \lambda_1 - \nu_1\sqrt{CD4_i^T(t-1)} & \text{if } A_i(t-1) = 1 \end{cases} \\
&= \begin{cases} 0.582 - 0.266\sqrt{CD4_i^T(t-1)} & \text{if } A_i(t-1) = 0 \\ 0.763 - 0.415\sqrt{CD4_i^T(t-1)} & \text{if } A_i(t-1) = 1 \end{cases}
\end{aligned}$$

where the parameters $\lambda_0, \nu_0, \lambda_1, \nu_1$ were chosen to equate to the probability of the event being 0.01 and 0.0005 while off treatment for CD4 counts of 200 and 500 cells/mm^3, respectively, and 0.006 and 0.0002 while on treatment for the same CD4 counts, respectively. These values were based on previous work estimating event rates using CASCADE data (A Babiker, personal communication, 23 August 2010). The event probability curves are illustrated in Figure 4.3.

Determining the optimal regime The optimal regime was that with the highest AIDS-free survival at 10 years (see below for how this was estimated in the RCTs and observational studies).

The randomised trials

A total of 31 million simulated patients were randomised equally across the 31 regimes given by "initiate treatment within one month of observed CD4 count first dropping $< x$ cells/mm^3"

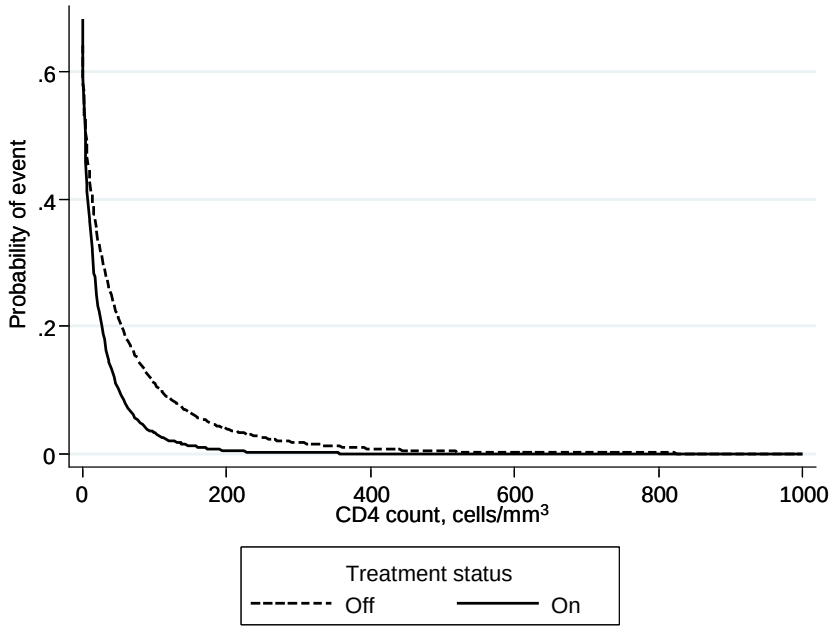


Figure 4.3: Model for probability of AIDS or death given true CD4 count and treatment status.

where $x = 200, 210, ..., 500$ (1 million patients on each regime). AIDS-free survival was estimated by Kaplan-Meier. The sample size was chosen to be large enough to obtain sufficiently stable results; further exploration of the impact of the sample size on precision is discussed below.

Exploration of the sample size Due to the large measurement errors and low event rates, a large number of patients were required. By using Kaplan-Meier estimation, we were able to see any residual uncertainty in estimating the optimal regime by plotting the 10-year AIDS-free survival against regime. In theory, with sufficient patient numbers and high enough event rate, by the construction of the models, the curve should be smooth. However, even with such a large sample size of 1 million patients per regime, it was still possible to detect a small amount of uncertainty. Whilst we could have used a global smoothing method, such as modelling regime with a spline or fractional polynomials, there was concern that this may only serve to hide the uncertainty and perhaps lead to incorrect inferences. In particular, the uncertainty was greatest at higher regimes due to the lower event rates at high CD4 counts; these extremities of the data may unduly influence such models. It is important to note that the differences in the 10-year AIDS-free survival between neighbouring regimes (that is, differing by 10 cells/mm^3) close to the optimal regime were very small, and typically they were identical to 3 decimal places. Such differences are not of great interest clinically but we wanted to be certain that we were correctly determining the optimal regime. Therefore, we considered a number of different sensitivity analyses:

1. We applied a least squares local smoothing technique ("lowess" in Stata 11.1; StataCorp (2009)). This procedure performs a series of weighted linear regressions of the dependent variable of interest y (here, the estimated 10-year AIDS-free survival) on the independent variable of interest z (here, regime x) to obtain smoothed estimates, with one regression centred on each (y_i, z_i). We used a bandwidth of 0.2, meaning that 20% of the data were used for each regression; this relatively small bandwidth was chosen to ensure only very local smoothing. The regressions were weighted with the greatest weight going to the central data pair; we used "tricube" weighting, which means that for each of the observations (y_j, z_j) contributing to the regression centred on (y_i, z_i), the following weight was applied:

$$\omega_j = \left[1 - \left(\frac{|z_j - z_i|}{\Delta}\right)^3\right]^3$$

where $\Delta = 1.0001max\left(z_{i_+} - z_i, z_i - z_{i_-}\right)$, and z_{i_+} and z_{i_-} are the maximum and minimum values of z contributing to the (y_i, z_i) regression, respectively.

2. We used an ad-hoc local smoothing approach by weighting the 10-year AIDS-free survival estimate for each regime x (given by $\widehat{u}_x$) as follows: $(\widehat{u}_{x-10} + 2\widehat{u}_x + \widehat{u}_{x+10})/4$, with no change for the most extreme datapoints.

3. We considered the "minimum acceptable regime", defined as that given by the lowest x with no worse than 0.5% poorer AIDS-free survival at 10 years than that of the optimal regime. The reasoning behind this is that the 10-year AIDS-free survival estimates are very similar close to the optimal regime, therefore this lower bound of the minimum acceptable regime may be more stable. Note that the minimum acceptable regime is not the same as the optimal regime, and answers a different question.

Variations We explored a variety of scenarios via the RCT simulations, including populations with different mean treatment-naïve CD4 count declines, less-frequently observed CD4 counts and permitting grace periods.

Treatment-naïve CD4 decline The above models assumed a mean absolute treatment-naïve CD4 decline of 1.10 per year on the square-root scale (referred to as the "regular-decline" population), based on the previous work mentioned above using CASCADE data. We considered the effect of different populations, with slower or faster average decline, to look at the impact on the optimal regime. That is, we changed the mean decline per year on the square-root scale to either 0.76 or 1.44 (based on the lower and upper quartiles, respectively, of the regular decline

distribution; labelled the "slow" and "fast" decline populations, respectively). For a patient with a CD4 count of 500 cells/mm^3 and mean decline at the population level, their CD4 count one year later would be 467, 452 or 438 cells/mm^3 in populations with slow, regular or fast decline, respectively. Similarly, a patient with a CD4 count of 350 cells/mm^3 and mean decline at the population level would have a CD4 count one year later of 322, 310 or 298 cells/mm^3, respectively, in those three populations. We would anticipate estimated optimal regimes given by higher x in populations with faster treatment-naïve CD4 decline.

Frequency of observed CD4 count If CD4 counts are observed less frequently than monthly for treatment initiation, then these methods may be applied in exactly the same way. However, this is likely to have an impact on the results and the interpretation. For example, if CD4 counts are only observed every $p > 1$ months, then initiating treatment when CD4 count is first observed to drop below a given threshold will tend to be later, in terms of CD4 count at treatment initiation and hence cumulative event risk, than if CD4 counts had been observed monthly ($p = 1$), due to the time lag. Therefore, under schedules where CD4 count is observed less frequently, we might expect optimal regimes to be given by higher x compared to scenarios where CD4 count is observed more frequently, in order to attempt to address that time lag. We considered the impact on the optimal regime if CD4 counts were observed every $p = 3$, 6 or 12 months. Of note, for the RCTs, the frequency of observed CD4 counts relates to those observed for the purposes of treatment initiation only; that is, true CD4 counts were still estimated monthly for the purposes of applying the event rates and the outcome estimation was still applied with time split into monthly intervals.

Grace periods No grace period (that is, $m = 1$, "immediate" treatment initiation) has so far been assumed. We considered allowing grace periods of $m = 3$, 6 or 12 months; we would anticipate higher estimated optimal regimes with longer grace periods. As discussed above, regimes in the presence of grace periods are not fully identified; we chose to apply uniform initiation across the grace period (second approach of Cain et al. (2010)). For the RCTs, this meant that all patients identified for treatment initiation at a given time, based on their randomised regime and CD4 count history, were treated as if they had been randomly allocated to initiate in one of the following intervals of the grace period, with probability of initiation in each interval given by $1/m$. Note that, after this allocation, patients may have been removed from the risk set before treatment initiation during the grace period due to reaching the event (that is, while waiting for their allocated treatment initiation time during the grace period), but

since the event rates were low this would typically be a very small number of patients. The use of this approach means that, in the equivalent observational study, we avoid estimating weights based on a small and potentially unrepresentative group of subjects initiating treatment in the last interval of the grace period.

The observational studies

We firstly simulated a large observational study with $n = 100,000$ patients, with regular treatment-naïve CD4 decline, monthly observed CD4 counts and no grace period ($m = 1$), and only considered the 3 regimes given by $x = 200$, 350 and 500, to check that we obtained similar results to the equivalent RCT (the number of patients and regimes were limited by the computational power required at the data expansion step). There was some uncertainty remaining despite the large sample size, therefore we also considered different scenarios (3-monthly observed CD4 counts and 3-month grace periods, all with regular treatment-naïve CD4 decline) and then repeated all of these large observational studies (different starting seed) to look at the variation in the 10-year AIDS-free survival estimates.

We then simulated 1000 datasets each with $n = 3000$ patients, from a population with regular treatment-naïve CD4 decline and CD4 counts measured monthly. We repeated these simulations with CD4 counts observed only every 3 months, which is the median frequency observed in our CASCADE data. As mentioned above, the CD4 observation frequency is relevant for the purposes of treatment initiation; the event rates were applied to the true CD4 counts which were always calculated monthly. However, for the observational studies, the observed CD4 counts were in addition used for performing the weight estimation. These datasets were then used with different grace periods for the estimation of the optimal regime (with regimes $x = 200, 210, ..., 500$). Of note, we considered the grace periods to be a step in the data analysis, not in the data generation.

Modelling treatment initiation The curve for the probability of treatment initiation given CD4 count was chosen to resemble that of the model from chapter 2 (page 72), with the parameters of the curve determined by the probability of treatment initiation at CD4 count 200 and 500 cells/mm^3 being 0.23 and 0.01, respectively (approximately based on the results of the chapter 2 model). The model used was:

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

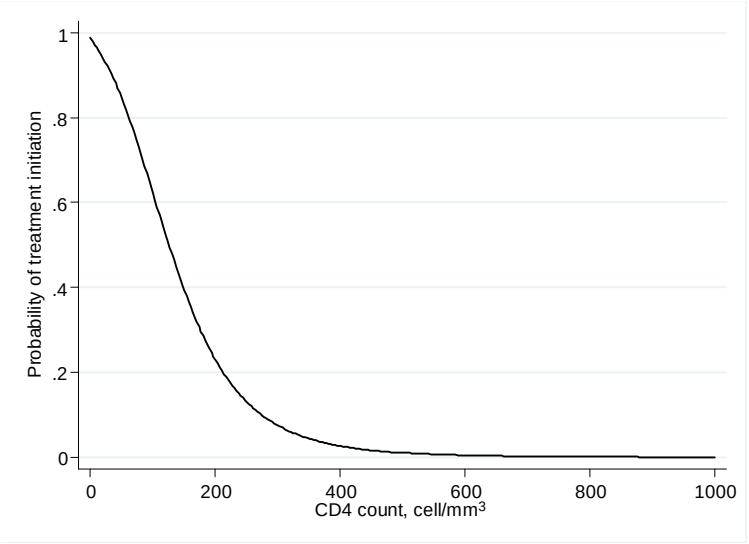


Figure 4.4: Model for probability of treatment initiation given current observed CD4 count.

where $p(t)$ represents the probability of treatment initiation at time t, and this is illustrated in Figure 4.4. If CD4 counts were observed less frequently than monthly, then the last observed CD4 count was carried forward and the treatment probabilities applied to that. Of note, since $A(t)$ was generated based on $CD4^O(t)$, this was reflected in the treatment model, rather than modelling $A(t)$ conditional on $CD4^O(t-1)$ as indicated in section 4.2. The latter approach is in order to be conservative with real data, where CD4 counts in the same month as treatment initiations may not actually have been available at the time of treatment initiation and therefore not contributed to the treatment decision, or indeed may have been measured after treatment initiation; this concern does not apply to these simulated data, and the interpretation and generalisability of the results is not affected.

Weight estimation We used a pooled logistic regression model for treatment initiation in patients previously treatment-naïve, given current observed CD4 count (square-root transformed and as a continuous variable; this mimics the data generation). From this, we estimated the (non-stabilised) weights under the second approach of Cain et al. (2010) as described above in section 4.2.2. The weights were truncated at maximum 20.

Outcome model In such realistically-sized datasets, there will undoubtedly be a great deal of uncertainty in the (weighted) Kaplan-Meier estimates of the AIDS-free survival. One approach is to model the outcome of AIDS or death using the pooled logistic regression models of Cain et al. (2010), as outlined in section 4.2. However, this global procedure may be heavily influenced by the extremes of the data. Therefore we also considered different approaches, as indicated in

section 4.2. That is, we estimated the optimal regimes (that with the highest 10-year AIDS-free survival) under the following approaches:

1. Based on the raw Kaplan-Meier estimates.

2. Applying local smoothing to the Kaplan-Meier estimates. Due to the much smaller sample size and hence greater uncertainty than the randomised trials, much heavier local smoothing was required. We used the command "smooth" of Stata 11.1 (StataCorp, 2009), which applies robust non-linear smoothing. A smoother of span r produces smoothed values of the variable y of interest by taking the median of each y_i and the $r-1$ values around y_i (with linear interpolation if r is even). We applied multiple smoothers in sequence to ensure relatively heavy smoothing, along with the Hanning smoother (Velleman and Hoaglin, 1981), which applies a smoother of span 3 with binomial weights (specified by "H"), and some further refinements: firstly, special treatment of the ends of the data (specified by "E"); secondly, "splitting" repeated values with a smoother of span 3 to avoid flat-topped peaks and troughs (specified by "S"); lastly, repeating an odd-spanned smoother until the smoothed variable did not change anymore (specified by "R"). The complete command we used was: 753SR8642EH.

3. Using a spline in a weighted pooled logistic regression model, as Cain et al. (2010). Regime and time were modelled as four-knot splines (with knots at the 5, 35, 65 and 95^{th} centiles), and interactions between regime (as a spline) and time (in two-yearly categories) were incorporated.

The questions of interest Addressing the questions of interest as outlined above:

1. To investigate bias in these realistically-sized datasets, we compared the mean and median of the optimal regimes from the 1000 datasets to the optimal regime from the equivalent randomised trial (that is, with the same treatment-naïve CD4 decline, frequency of observed CD4 and grace period). In addition, we looked at the proportion of estimates which were less than the minimum acceptable regime from the equivalent RCT.

2. To look at the precision of a single analysis of this size, we estimated the standard error using the standard deviation of the estimates from the 1000 simulated datasets.

3. To investigate the bias-variance trade-off in allowing grace periods of $m > 1$ months, when the inference of interest is under no grace period ($m = 1$), we compared the results

from the observational study simulations, with grace periods of $m = 1$, 3 and 6 months, with that from the equivalent randomised trial except with no grace period (that is, the same population treatment-naïve CD4 decline and frequency of observed CD4 counts). We assessed the bias-variance trade-off by examining the mean square error, calculated as the square of the estimated standard error from (2) plus the square of the difference in the estimated optimal regimes from the observational study and RCT (Burton et al., 2006). In addition, we considered the relative efficiency, which for a given CD4 count observation frequency was calculated for each approach and choice of grace period as the square of the estimated standard error from (2) divided by the square of the standard error under the pooled logistic regression approach with no grace period ($m = 1$; Lebanon (2006)); this was chosen as the reference group since this method is commonly used in the literature.

4.3.3 Results: the randomised trials

We firstly present detailed results for the population with regular treatment-naïve CD4 decline, where CD4 counts were observed monthly for the purpose of treatment initiation, and with no grace period ($m = 1$).

For illustration, Figure 4.5 shows the path of CD4 count over time for an example patient, who was randomised to initiate treatment when their CD4 count was first observed to drop < 200 cells/mm^3. The black lines indicate the underlying CD4 slopes, with decline while treatment-naïve, relatively steep increase after treatment initiation at 57 months, and more gradual increase from one year after treatment initiation onwards. The blue lines show the path of the true CD4 count over time, that is after incorporating the Brownian motion. The red lines illustrate the observed CD4 count, that is after allowing for measurement error. Of note, this patient initiated treatment at a low observed CD4 count of 169 cells/mm^3, while their true CD4 count was at 333 cells/mm^3. This behaviour was common across the population, by the nature of CD4 decline and the definitions of the regimes. This issue is discussed further in section 4.6.

Summary of baseline characteristics and treatment

Summary statistics of the baseline characteristics and treatment for the $1,000,000$ patients on each of the three regimes given by $x = 200$, 350 and 500 are given in Table 4.3. As anticipated for a large randomised trial, the baseline CD4 counts (observed or true) and treatment-naïve slopes were very similar across all regimes. While the observed baseline CD4 counts were all by definition in $[500, 550]$ cells/mm^3, this did not necessarily hold for the true baseline CD4

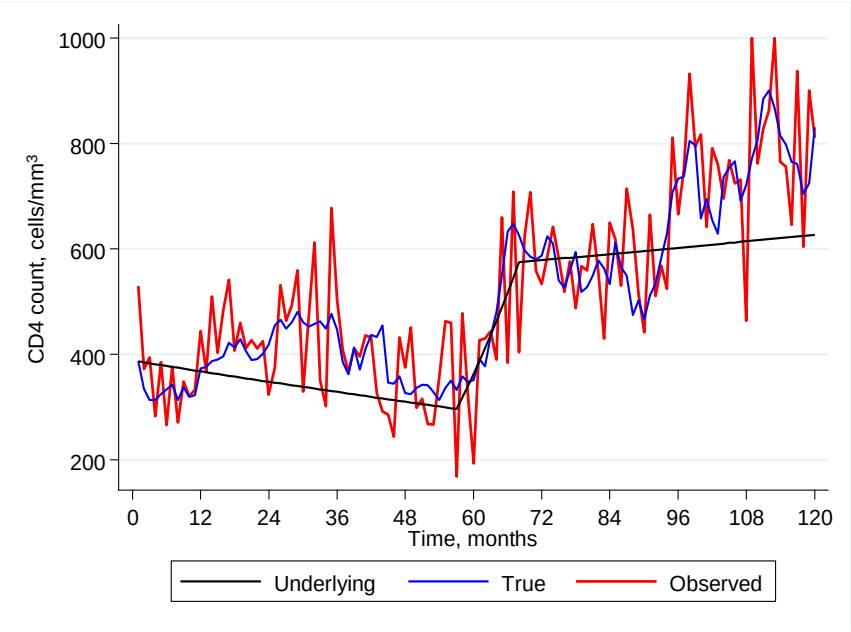


Figure 4.5: Illustration of underlying, true and observed CD4 count over time for an example patient randomised to initiate when CD4 count was first observed to be < 200 cells/mm^3 (from the scenario with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period). This patient initiated treatment at 57 months, when their observed and true CD4 counts were 169 and 333 cells/mm^3, respectively.

counts, as shown by the interquartile ranges of 457 to 598 cells/mm^3 across the regimes.

The summary results relating to treatment initiation reflect what we might expect by definition of the regimes. At regimes defined by higher x, a greater proportion of patients were observed to initiate treatment ($> 99\%$ versus 78% for the $x = 500$ and 200 regimes, respectively) and sooner (median 3 versus 37 months, respectively). The percentage of follow-up time spent on treatment was 94% for regime $x = 500$ compared with just 49% for regime $x = 200$.

By definition, the observed CD4 counts at treatment initiation were all $< x$ cells/mm^3 for each of the regimes, with median 432 versus 175 cells/mm^3 for regimes $x = 500$ and 200, respectively. The true CD4 counts at treatment initiation tended to be higher than that defined by the regime, influenced by unusually low observed CD4 counts resulting in treatment initiation despite higher true CD4 count, as exemplified by the CD4 count paths of the example patient shown in Figure 4.5. This was more noticeable at regimes defined by lower x, with median true CD4 counts at treatment initiation of 503 versus 275 cells/mm^3 for regimes $x = 500$ and 200, respectively.

As a consequence of the random error structure of the simulated data, treatment initiation at lower CD4 counts was associated with faster initial and subsequent CD4 count increase on the square-root scale. The median increase in square-root CD4 count over the first year

	Regimes given by x		
	200	350	500
Baseline			
Observed CD4 count, cells/mm^3	525 (513, 538)	525 (512, 537)	525 (512, 538)
True CD4 count, cells/mm^3	525 (457, 598)	525 (457, 598)	525 (457, 598)
Annual slope, square-root scale	1.10 (0.76, 1.44)	1.10 (0.76, 1.44)	1.10 (0.76, 1.44)
Treatment			
N patients observed to initiate treatment	783,766 (78%)	952,144 (95%)	995,522 (>99%)
Time to initiation, months[1]	37 (21, 62)	10 (4, 25)	3 (2, 5)
Observed CD4 count at initiation, cells/mm^3[1]	175 (154, 189)	313 (282, 333)	432 (379, 469)
True CD4 count at initiation, cells/mm^3[1]	275 (240, 313)	425 (381, 472)	503 (441, 565)
Initial annual slope after initiation, square-root scale[1,2]	3.42 (2.16, 4.68)	2.78 (1.52, 4.04)	2.51 (1.24, 3.77)
Annual slope one year after initiation, square-root scale[1,2]	0.41 (0.30, 0.52)	0.01 (-0.10, 0.12)	-0.17 (-0.31, -0.03)
Percentage of follow-up time spent on treatment	49%	79%	94%

Table 4.3: Simulation study 1 (RCT): summary of baseline characteristics and treatment for $n = 1,000,000$ patients on each of the three regimes given by $x = 200$, 350 and 500 (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period). Unless otherwise stated, values are n (%) for categorical variables and median (interquartile range) for continuous variables. Of note, $< 1\%$ of true CD4 counts were truncated at 0 cells/mm^3, and approximately 1, 2 and 3% of true CD4 counts were truncated at 1000 cells/mm^3 on the regimes given by $x = 200$, 350 and 500 cells/mm^3 respectively. [1] In those patients who were observed to initiate treatment. [2] As assigned at treatment initiation.

after treatment initiation was 2.51 and 3.42 for regimes $x = 500$ and 200, respectively. In the absence of Brownian motion, and for patients who initiated when their true CD4 counts were 500 and 200 cells/mm^3, these median increases translated to a CD4 count one year later of 618 and 308 cells/mm^3, respectively (a broadly similar increase on the absolute scale). The median increase in square-root CD4 count after the first year on treatment was -0.17 and 0.41 per year thereafter for regimes $x = 500$ and 200, respectively. Similarly, these translated to a CD4 count of 577 and 384 cells/mm^3, respectively, five years later. Of note, for regimes defined by higher x, the slope beyond one year after treatment initiation tended to be negative, thereby introducing a penalty for early treatment initiation. Figures 4.6 and 4.7 illustrate the distribution of true CD4 count over time for the regimes $x = 200$, 350 and 500 (in a random subset of $n = 100,000$ patients per regime due to computational limitations); in Figure 4.6, it is possible to see the slightly negative slope over the longer term for the $x = 500$ regime.

Outcome results

Overall, 17%, 13% and 14% of patients were observed to progress to AIDS/death on regimes given by $x = 200$, 350 and 500, respectively, during the 10 year follow up. Figure 4.8 shows the estimated AIDS-free survival curves for these three regimes; the estimated 10-year AIDS-free survival was 0.8278, 0.8657 and 0.8587, respectively, for these regimes.

Figure 4.9 illustrates the estimated AIDS-free survival at 10 years by regime; the peak of the curve at $x = 350$ is the optimal regime. As mentioned above, the probability of surviving AIDS-free to 10 years under this regime was 0.8657. As discussed in the methods, there was some residual uncertainty apparent in the plot, particularly at regimes given by higher x where the event rate is much lower. Applying the local smoothing by either least squares or weighting, the optimal regime was given by $x = 360$ (with 10-year AIDS-free survival of 0.8656 in both cases). The smoothed plots are illustrated in Figure 4.10. Consideration of the minimum acceptable regime, as illustrated in Figure 4.9, yielded $x = 290$ regardless of whether local smoothing was applied (with 10-year AIDS-free survival rates of 0.8621, 0.8619 and 0.8620 for no smoothing, least squares smoothing or weighted smoothing, respectively).

Frequency of observed CD4 counts

In the population with regular treatment-naïve CD4 decline and with no grace period, reducing the frequency of observed CD4 counts from monthly to every 3, 6 or 12 months increased the optimal treatment regime from $x = 350$ to 410, 460 and 490, respectively, and the 10-year AIDS-free survival on those optimal regimes decreased from 0.8657 to 0.8650, 0.8634 and

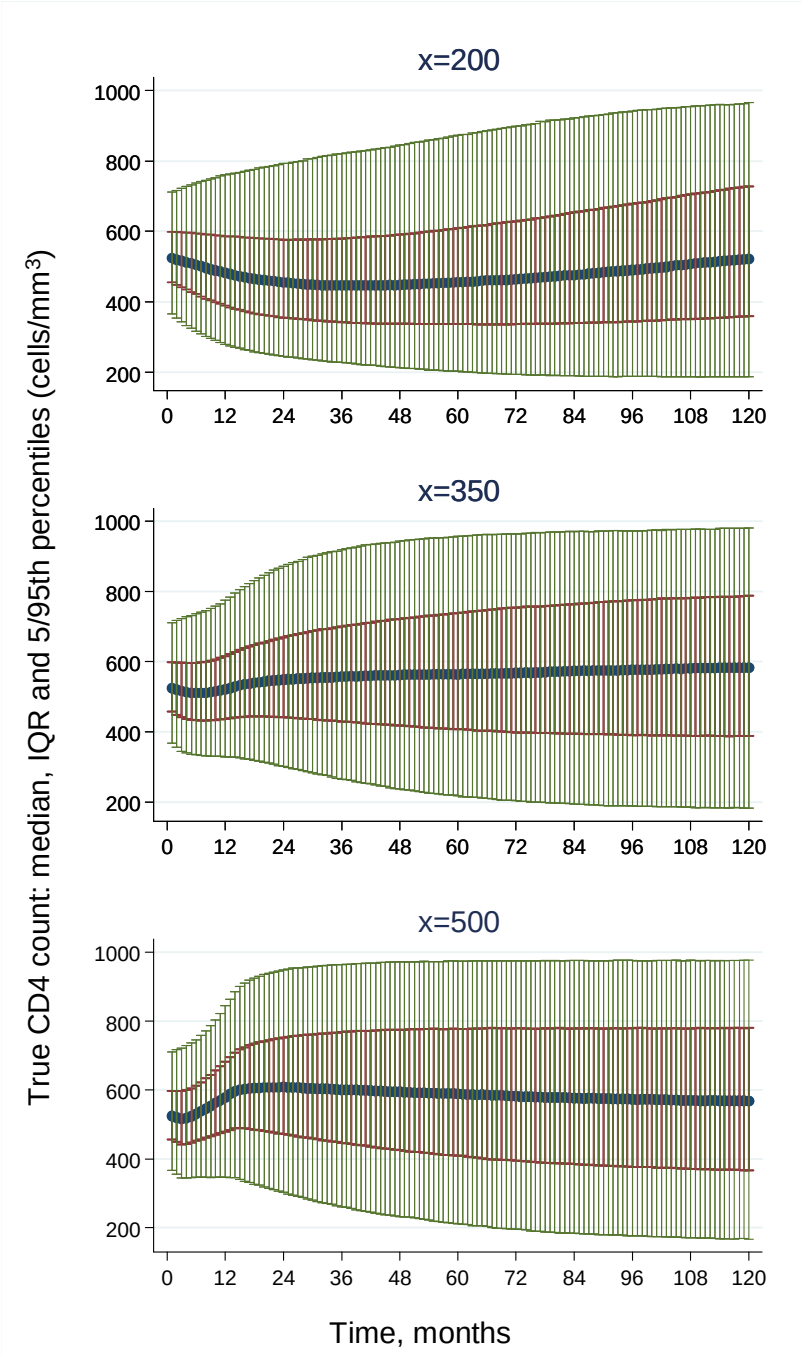


Figure 4.6: Simulation study 1 (RCT): true CD4 count over time (median, interquartile range and $5/95^{th}$ percentiles) for a subset of $n = 100,000$ patients in the RCT on each of the three regimes given by $x = 200$, 350 and 500 (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period).

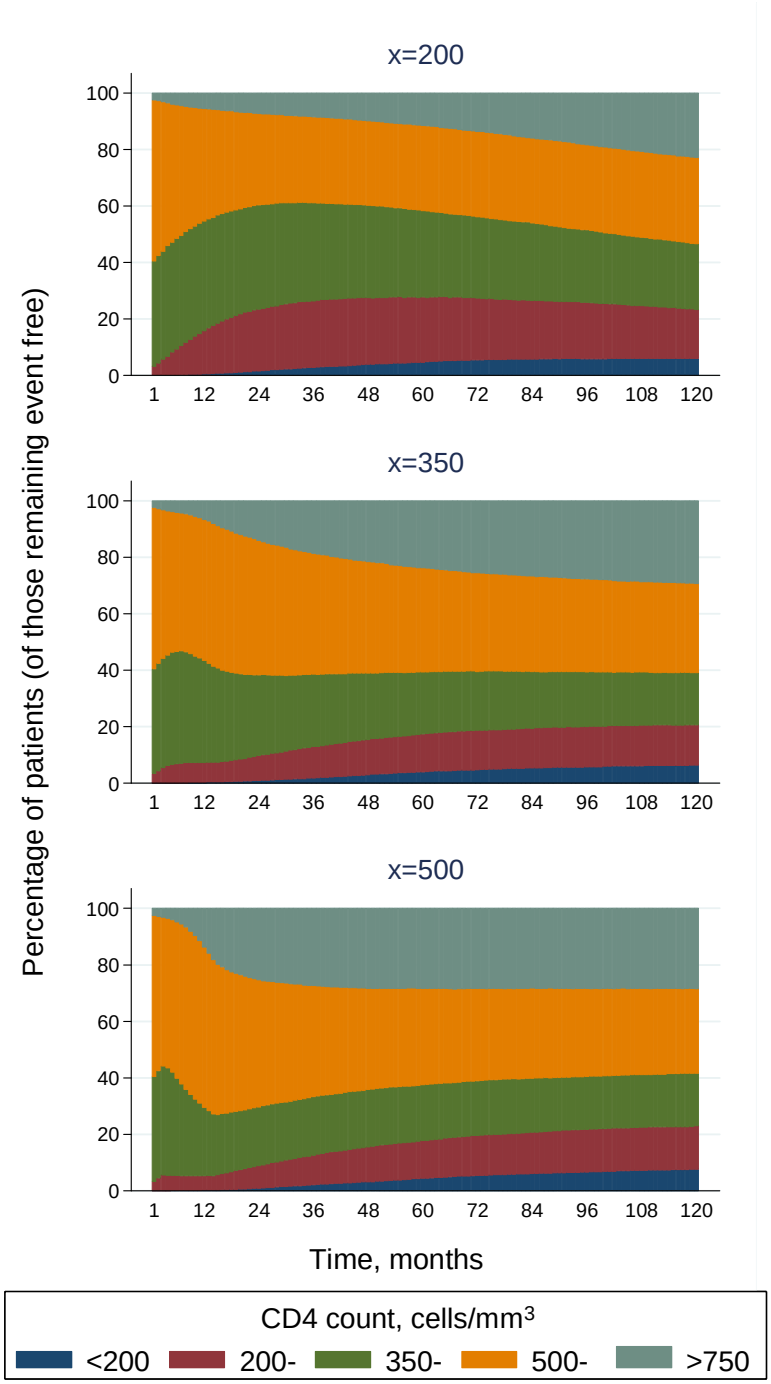


Figure 4.7: Simulation study 1 (RCT): true CD4 count categorised over time from trial start for a subset of $n = 100,000$ patients on each of the three regimes given by $x = 200$, 350 and 500 (population with regular treatment-naïve CD4 decline CD4 decline, CD4 counts observed monthly and no grace period).

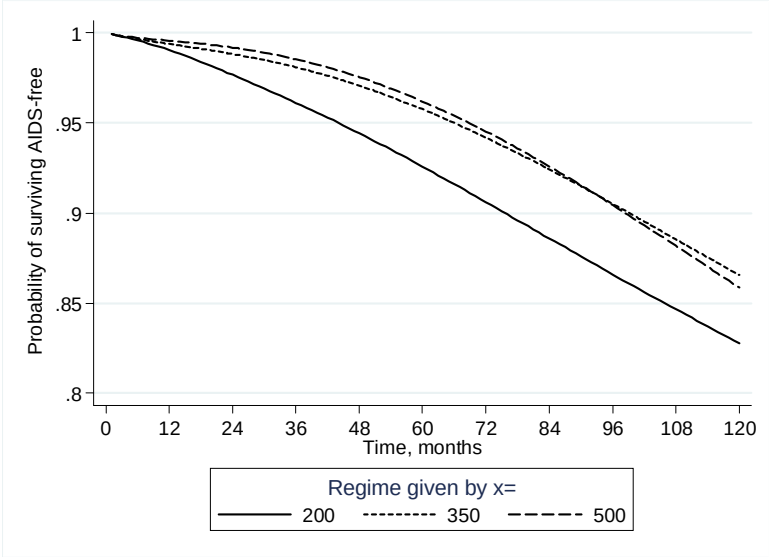


Figure 4.8: Simulation study 1 (RCT): AIDS-free survival curves over 10 years for the three regimes given by $x = 200$, 350 and 500 (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period).

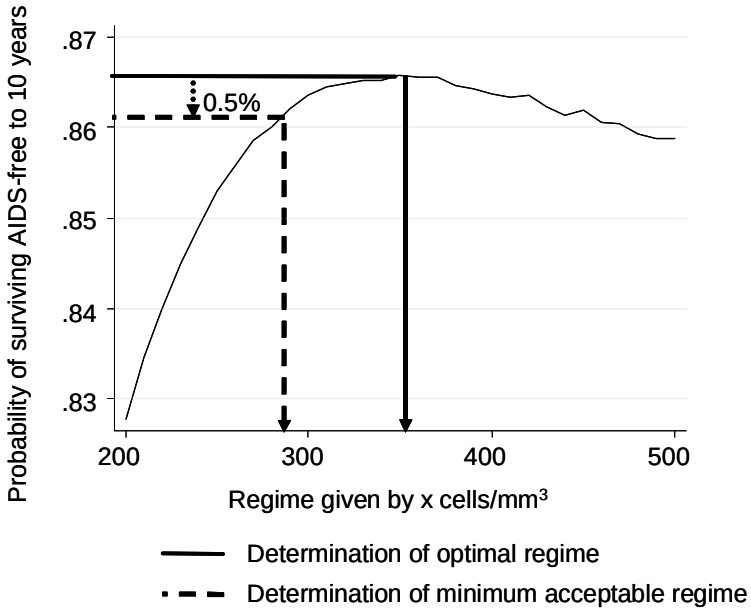


Figure 4.9: Simulation study 1 (RCT): probability of surviving AIDS-free to 10 years by regime (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period). The optimal regime is determined by that with maximum 10-year AIDS-free survival (solid lines). The minimum acceptable regime is defined as the lowest with no worse than 0.5% poorer 10-year AIDS-free survival than under the optimal regime (dashed lines).

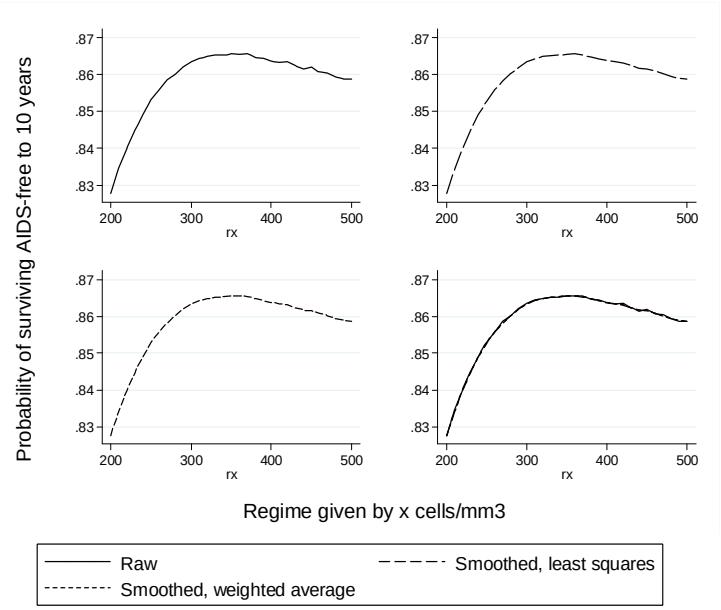


Figure 4.10: Simulation study 1 (RCT): probability of surviving AIDS-free to 10 years by regime, with no smoothing, least squares smoothing or weighted average smoothing (and all three compared in the bottom right plot; they overlap considerably; population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period).

0.8564, respectively (Table 4.4 and Figure 4.11). Figure 4.12 illustrates that observing CD4 counts only annually results in 10-year AIDS-free survival under the optimal regime more than 0.5 percentage points lower than if CD4 counts were observed monthly.

However, applying the optimal regime from the population with regular treatment-naïve CD4 decline and CD4 counts observed monthly (namely, $x = 350$; no grace period) to the comparable scenario but with CD4 counts observed 3-, 6- or 12-monthly, the 10-year AIDS-free survival would be 0.8616, 0.8528 and 0.8304, respectively. Similarly, if the optimal regime from the population with regular treatment-naïve CD4 decline and 3-monthly observed CD4 counts was applied to the comparable scenario but with 6- or 12-monthly observed CD4 counts, then the 10-year AIDS-free survival would be 0.8615 and 0.8484, respectively.

Similar patterns were observed in the populations with slower or faster treatment-naïve CD4 decline (Table 4.4 and Figure 4.11). As we might expect, in populations with faster CD4 decline, the optimal regime tended to be given by higher x, and the 10-year AIDS-free survival was lower. If CD4 counts were observed only 6- or 12-monthly, the impact of the different population CD4 declines was less apparent, but the optimal regime for the population with fast CD4 decline when CD4 counts were observed so infrequently was estimated at the maximum of the permitted range ($x = 500$) and so may be higher in reality.

CD4 decline[1]	Frequency of observed CD4 counts, months[2]			
	1	3	6	12
Slow	310 (0.8731)	380 (0.8720)	460 (0.8705)	490 (0.8671)
	320 (0.8727)	-	420 (0.8703)	480 (0.8668)
	-	-	420 (0.8703)	-
Regular	350 (0.8657)	410 (0.8650)	460 (0.8634)	490 (0.8564)
	360 (0.8656)	-	-	-
	360 (0.8656)	-	-	-
Fast	360 (0.8601)	460 (0.8592)	460 (0.8569)	500 (0.8471)
	370 (0.8600)	450 (0.8591)	500 (0.8569)	-
	370 (0.8600)	-	500 (0.8569)	-

Table 4.4: Simulation study 1 (RCTs): **optimal regimes** in populations with different treatment-naïve CD4 declines and frequencies of observed CD4 count (no grace period, $m = 1$). For each population, the first line gives the results with no local smoothing, and the second and third line shows the results under local smoothing using least squares and weighting, respectively (the smoothed results are only shown if the optimal differs from that under no smoothing). Values in brackets are the estimated probabilities of surviving AIDS-free to 10 years under that regime. [1] Population CD4 decline while treatment-naïve; see text for details regarding the rates. [2] Frequency with which CD4 count was observed for treatment initiation.

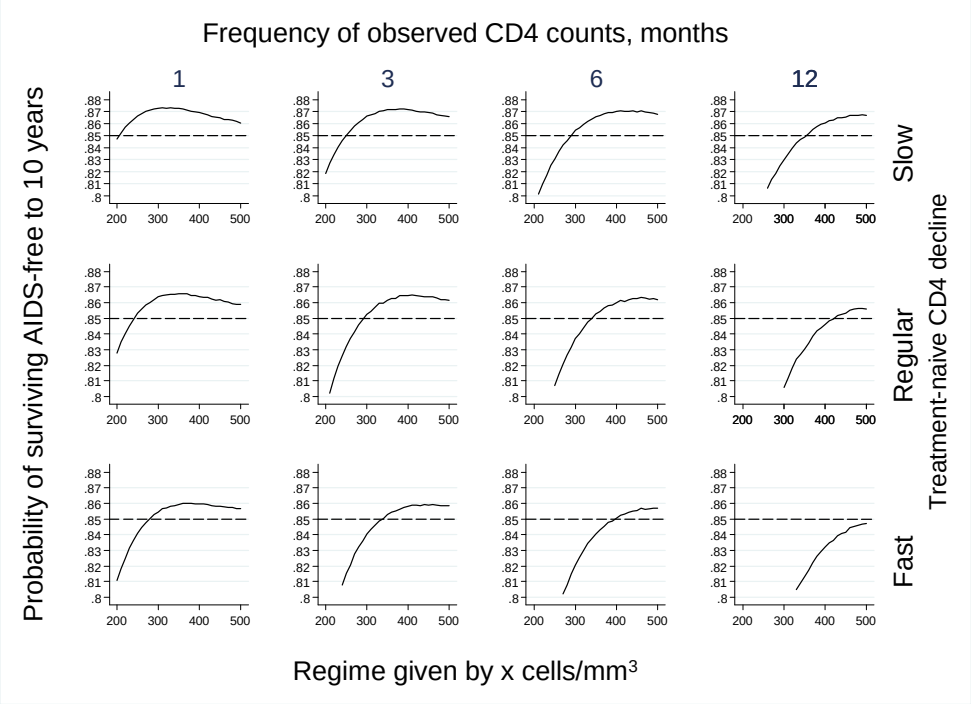


Figure 4.11: Simulation study 1 (RCTs): probability of surviving AIDS-free to 10 years by regime, across different treatment-naïve CD4 declines and frequencies of observed CD4 counts (no grace period, $m = 1$). Note that probabilities were only plotted if ≥ 0.80 to preserve a common scale. Horizontal lines drawn at 0.85 to aid comparison between plots.

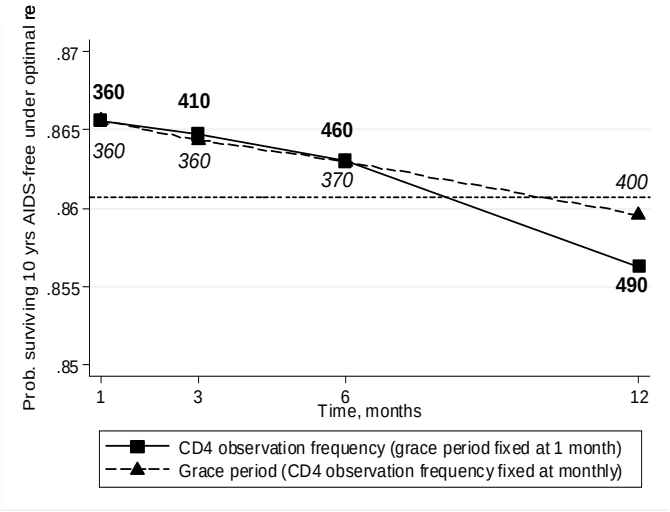


Figure 4.12: Simulation study 1 (RCTs): probability of surviving AIDS-free to 10 years under the optimal regime, for the population with regular treatment-naïve CD4 decline and different CD4 observation frequencies (with grace period fixed at 1 month) and grace periods (with CD4 observation frequency fixed at monthly). The horizontal dashed line is set at 0.5 percentage points lower AIDS-free survival than that under the scenario when CD4 counts were observed monthly and with no grace period. Numbers on the plot show the optimal regime for each scenario after local smoothing applied (in bold where the CD4 observation frequency was varied, and in italic where the grace period was varied).

Smoothing Local smoothing of the results resulted in some small changes to the estimated optimal regime (Table 4.4), but typically by no more than 10 cells/mm^3. There were two exceptions, both where CD4 counts were observed every 6 months, and in both cases the two methods of local smoothing yielded the same optimal regimes. Firstly, in the population with slow declining treatment-naïve CD4 count, the optimal regime estimated after smoothing was 420 compared to 460 cells/mm^3 without smoothing. Secondly, in the population with fast declining CD4 count, the optimal with smoothing was 500 compared to 460 cells/mm^3 without. However, the estimated 10-year AIDS-free survival probabilities on the optimal regimes from smoothing and not smoothing were very similar, and it is clear from Figure 4.11 that the curves were quite flat in these regions.

Minimum acceptable regime The same patterns were observed when considering the minimum acceptable regime (Table 4.5). As anticipated, these minimum acceptable regimes were somewhat more stable than the optimal regimes, with only three instances of the smoothed and non-smoothed approaches leading to different estimated optimal regimes. In a population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period, delaying treatment initiation until CD4 count was first observed to drop below 290 cells/mm^3 was associated with 10-year AIDS-free survival no worse than 0.5 percentage points lower than the

CD4 decline[1]	Frequency of observed CD4 counts, months[2]			
	1	3	6	12
Slow	260 (0.8682)	310 (0.8673)	350 (0.8657)	410 (0.8622)
	-	-	-	-
	-	-	-	-
Regular	290 (0.8621)	350 (0.8616)	390 (0.8587)	430 (0.8517)
	-	340 (0.8602)	-	-
	-	340 (0.8601)	-	-
Fast	310 (0.8564)	370 (0.8552)	410 (0.8526)	460 (0.8445)
	-	360 (0.8541)	-	450 (0.8422)
	-	360 (0.8541)	-	450 (0.8421)

Table 4.5: Simulation study 1 (RCTs): **minimum acceptable regimes** in populations with different treatment-naïve CD4 declines and frequencies of observed CD4 count (no grace period, $m = 1$). For each population, the first line gives the results with no local smoothing, and the second and third line shows the results under local smoothing using least squares and weighting, respectively (the smoothed results are only shown if the optimal differs from that under no smoothing). Values in brackets are the estimated probabilities of surviving AIDS-free to 10 years under that regime. [1] CD4 decline while treatment-naïve; see text for details regarding the rates. [2] Frequency with which CD4 count is observed for treatment initiation.

optimal (under the regime given by $x = 350$). For a patient with the median treatment-naïve CD4 decline, this translates to a delay in treatment initiation of approximately 18 months, in the absence of Brownian motion or measurement error (time for CD4 count to drop 60 cells/mm^3 from 350 to 290 cells/mm^3).

Grace periods

Fixing the frequency with which CD4 counts were observed as monthly, and with regular treatment-naïve CD4 decline, increasing the grace period from $m = 1$ to 3, 6 or 12 months resulted in an increase in the optimal regime from $x = 350$ to 360, 370 and 380, respectively, and the 10-year AIDS-free survival on those optimal regimes decreased from 0.8657 to 0.8644, 0.8631 and 0.8598, respectively (Table 4.6 and Figure 4.13). Therefore, the effect of permitting a grace period of 12 months had much less of an impact on the optimal regime than reducing the observation frequency to 12 monthly. This is as we might anticipate, for at least two reasons. Firstly, with only yearly observed CD4 counts, patients were only able to initiate treatment at yearly time-points, whereas under the 12-month grace period, only 1/12 patients eligible to initiate treatment delayed for the full 12 months. Secondly, and perhaps more importantly, there is an asymmetry due to the regimes being defined by CD4 counts dropping *below* a threshold. For example, when the CD4 counts were observed monthly, the same patients 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 ac-

CD4 decline[1]	Grace period (m), months			
	1	3	6	12
Slow	310 (0.8731)	310 (0.8722)	340 (0.8708)	350 (0.8679)
	320 (0.8727)	320 (0.8721)	350 (0.8707)	360 (0.8674)
	-	320 (0.8721)	350 (0.8707)	-
Regular	350 (0.8657)	360 (0.8644)	370 (0.8631)	380 (0.8598)
	360 (0.8656)	-	-	400 (0.8596)
	360 (0.8656)	-	-	400 (0.8596)
Fast	360 (0.8601)	400 (0.8589)	420 (0.8575)	450 (0.8538)
	370 (0.8600)	390 (0.8588)	-	-
	370 (0.8600)	-	-	-

Table 4.6: Simulation study 1 (RCTs): **optimal regimes** in populations with different treatment-naïve CD4 declines and grace periods (CD4 counts observed monthly). For each population, the first line gives the results with no local smoothing, and the second and third line shows the results under local smoothing using least squares and weighting, respectively (the smoothed results are only shown if the optimal differs from that under no smoothing). Values in brackets are the estimated probabilities of surviving AIDS-free to 10 years under that regime. Of note, the first column of results is the same as that presented in Table 4.4. [1] CD4 decline while treatment-naïve; see text for details regarding the rates.

cording to a given regime, but that CD4 count was not observed on the 12-monthly CD4 count schedule, then that patient would not have been identified for treatment initiation under the less-frequent CD4 observation until some time later. In particular, if that CD4 count observed on the monthly but not 12-monthly schedule was a random low value, and the following observed CD4 counts were higher (closer to the underlying trend), then that patient may not have been identified for treatment initiation until much later on the 12-monthly schedule. Therefore the regime would need to be higher in order to identify such patients for treatment initiation, hence the optimal regime is higher. However, it is still important to note that allowing grace periods of 12 months resulted in 10-year AIDS-free survival more than 0.5 percentage points lower than immediate treatment initiation ($m = 1$) on the optimal regime (Figure 4.12). Therefore if grace periods are to be used to potentially increase power, then grace periods of this length will be associated with substantial bias. Broadly similar patterns were observed across the populations with different treatment-naïve CD4 decline.

Smoothing Again, local smoothing of the estimates resulted in some small changes in the estimated optimal regime, though typically no more than 10 cells/mm^3. The one exception was in the scenario with regular treatment-naïve CD4 decline and a 12 month grace period, where the smoothed optimal regimes were 400 compared to 380 cells/mm^3 in the absence of smoothing. Once again, the estimated 10-year AIDS-free survival probabilities were very similar, and the curves were fairly flat in this region (Figure 4.13).

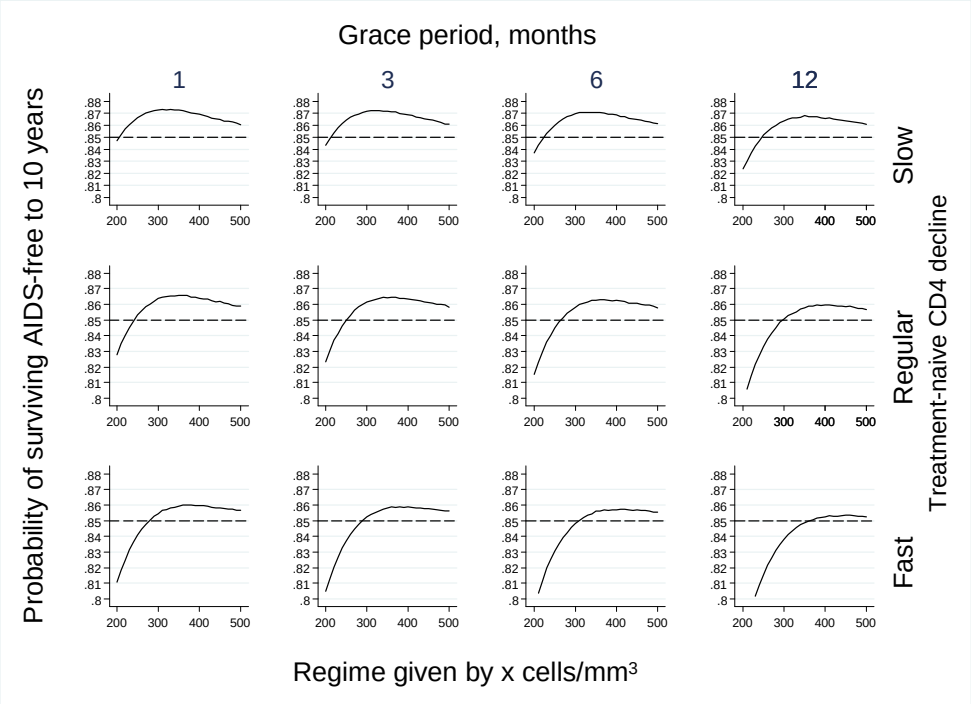


Figure 4.13: Simulation study 1 (RCTs): probability of surviving AIDS-free to 10 years by regime, across different treatment-naïve CD4 declines and grace periods (CD4 counts observed monthly). Note that probabilities were only plotted if ≥ 0.80 to preserve a common scale. Horizontal lines drawn at 0.85 to aid comparison between plots.

Minimum acceptable regime The same patterns were observed when considering the minimum acceptable regime (Table 4.7), but with no differences between the smoothed and non-smoothed optimal regimes, indicating the greater stability of the minimum acceptable regimes.

Different combinations of CD4 count observation frequency and grace periods

All the results above either hold grace period at $m = 1$ months and vary the frequency of CD4 measurements, or vice versa. Combinations of these are likely to be of interest.

For a population with regular treatment-naïve CD4 decline, observing CD4 counts every 3 months and permitting a 3- or 6-month grace period, the optimal regime was given by $x =$ 410 and 460, respectively (with 10-year AIDS-free survival of 0.8638 and 0.8625, respectively, which is 0.0019 and 0.0032 less, respectively, than the optimal under the scenario where CD4 counts were observed monthly and with no grace period, when the optimal regime was given by $x = 350$). The local smoothing methods led to the same optimal regime in the presence of the 3-month grace period, but were somewhat lower at 420 with a grace period of 6 months; the curve was very flat at high x (10-year AIDS free survival under this optimal regime after local smoothing using least squares and weighting was 0.8621 and 0.8662, respectively). The minimum acceptable regimes were given by $x = 350$ (0.8600) and 360 (0.8584) under 3- and

CD4 decline[1]	Grace period (m), months			
	1	3	6	12
Slow	260 (0.8682)	270 (0.8682)	280 (0.8670)	300 (0.8638)
	-	-	-	-
	-	-	-	-
Regular	290 (0.8621)	290 (0.8600)	300 (0.8581)	330 (0.8553)
	-	-	-	-
	-	-	-	-
Fast	310 (0.8564)	310 (0.8542)	330 (0.8535)	350 (0.8489)
	-	-	-	-
	-	-	-	-

Table 4.7: Simulation study 1 (RCTs): **minimum acceptable regimes** in populations with different treatment-naïve CD4 declines and grace periods (CD4 counts observed monthly). For each population, the first line gives the results with no local smoothing, and the second and third line shows the results under local smoothing using least squares and weighting, respectively (the smoothed results are only shown if the optimal differs from that under no smoothing). Values in brackets are the estimated probabilities of surviving AIDS-free to 10 years under that regime. Of note, the first column of results is the same as that presented in Table 4.4. [1] CD4 decline while treatment-naïve; see text for details regarding the rates.

6-month grace periods, respectively, with no change under local smoothing.

If instead CD4 counts were measured every 6 months and with a 6-month grace period, then the optimal regime was given by $x = 460$ (with 10-year AIDS-free survival of 0.8603, which is 0.0054 less than under the optimal regime if CD4 counts were observed monthly and with no grace period). Local smoothing under either method led to a slightly different optimal regime of $x = 470$, with corresponding 10-year AIDS-free survival of 0.8601. The minimum acceptable regime was $x = 410$ (0.8569), again with no change under local smoothing.

Summary

The simulation of these large RCTs has highlighted some important results. We have seen that the measurement error in CD4 counts may be large, and that large numbers of patients are required for precise estimation. In addition, it is clear that sufficient follow-up time is required in order to see differences between the regimes. In these data, the AIDS-free survival rates are broadly similar at high CD4 counts. In a population with regular treatment-naïve CD4 decline, monthly observed CD4 counts and no grace, the optimal regime is given by $x = 360$ (after smoothing). As discussed above, decreasing the frequency of observed CD4 counts substantially raised the optimal regimes, whereas increasing the grace period had less of an effect.

4.3.4 Results: single large observational study

Summary of baseline characteristics and treatment

Summaries of the baseline characteristics and treatment for the $n = 100,000$ patients in the single large observational study, after expansion to the three regimes given by $x = 200, 350, 500$, are shown in Table 4.8, both unweighted and after applying weights (truncated at maximum 20; note this is for a population with regular treatment-naïve decline, monthly observed CD4 counts and no grace period). The baseline results were similar to those from the RCT (see Table 4.3). The median follow-up time (censoring when no longer compliant with a given regime) was longer after weighting, at 55 versus 27 months for the $x = 200$ regime, and 4 versus 2 months for the $x = 500$ regime, as we would anticipate since we are upweighting those patients who remain uncensored to account for those who have been censored.

The median time of treatment initiation was typically longer after weighting and more comparable with that from the RCT, for example at 33 months after weighting versus 24 without weighting for the $x = 200$ regime (compared to 37 months in the RCT; Table 4.3), although was not noticeably different for the $x = 500$ regime (2 months with or without weighting, compared to 3 months in the RCT). Similarly, the observed and true CD4 counts at treatment initiation were higher under all three regimes after weighting, compared to no weighting, making them more comparable to those in the RCT, although still somewhat lower for the $x = 500$ regime. The post-treatment slopes moved in different directions after weighting, but in all cases moved closer to those seen under the RCT.

Looking at the distribution of true CD4 counts over time, in the absence of weighting (Figure 4.14) and comparing to that from the RCT (Figure 4.6), we do not see the initial decline in CD4 under the 200 regime, we see an initial increase under the 350 regime and for the 500 regime we see a big initial increase followed by a sharper decline. In contrast, after the application of weights (truncated at maximum 20; Figure 4.15), the plots much more closely resemble those from the RCT.

Of the 16,773 patients who were observed to initiate treatment in compliance with at least one regime, 1534 (9%) and 153 (1%) initiated treatment in compliance with 2 and 3 regimes, respectively. Those patients who initiated in compliance with all three regimes tended to have low true baseline CD4 counts, with median 335 (IQR 295, 392) cells/mm^3, and the observed CD4 count was by definition > 500 cells/mm^3 at baseline but then plummeted soon after, with 93% of those patients initiating in the next month.

Figure 4.16 shows compliance over time, by whether on or off treatment, for the three regimes

	Regime given by x		
	200	350	500
Baseline			
Observed CD4 count, cells/mm^3	525 (513, 537)	525 (513, 537)	525 (513, 537)
True CD4 count, cells/mm^3	525 (457, 598)	525 (457, 598)	524 (456, 597)
Annual slope, square-root scale	1.10 (0.77, 1.44)	1.10 (0.77, 1.44)	1.10 (0.77, 1.44)
Follow-up in compliance with regime			
Follow-up time, months	27 (12,64)	10 (4,31)	2 (1,5)
	55 (24,120)	*60 (9,120)*	*4 (1,120)*
Treatment			
N patients observed to initiate treatment	8887 (9%)	6786 (7%)	2938 (3%)
Time to initiation, months[1]	24 (13,43)	8 (4,20)	2 (2,4)
	33 (19,54)	*10 (4,23)*	*2 (2,4)*
Observed CD4 count at initiation, cells/mm^3[1]	169 (146,186)	289 (248,321)	361 (288,423)
	174 (153,188)	*309 (278,330)*	*385 (335,437)*
True CD4 count at initiation, cells/mm^3[1]	284 (248,323)	413 (369,461)	458 (388,526)
	278 (243,315)	*424 (381,471)*	*478 (414,541)*
Initial annual slope after initiation, square-root scale[1,2]	3.39 (2.12,4.66)	2.83 (1.56,4.08)	2.62 (1.30,3.85)
	3.38 (2.09,4.68)	*2.78 (1.51,4.06)*	*2.50 (1.24,3.78)*
Annual slope one year after initiation, square-root scale[1,2]	0.38 (0.27,0.49)	0.04 (-0.08,0.15)	-0.07 (-0.22,0.10)
	0.40 (0.29,0.51)	*0.01 (-0.10,0.12)*	*-0.12 (-0.26,0.03)*
Percentage of follow-up time spent on treatment	17%	25%	37%
	47%	*79%*	*89%*

Table 4.8: Simulation study 1 (large observational study): summary of baseline characteristics and treatment for $n = 100,000$ patients, after expansion to the three regimes given by $x =$ 200, 350 and 500 (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period). However, patients who initiated treatment in the first month while by definition observed CD4 count was > 500 cells/mm^3 were immediately censored from all regimes, and therefore these summary statistics are based on the 99,108 patients who were not immediately censored. Results in regular text are based on unweighted data; those in italics are after weighting (with truncation at maximum 20). Unless otherwise stated, values are n (%) for categorical variables and median (interquartile range) for continuous variables. Of note, no true CD4 counts were truncated at 0 cells/mm^3, and approximately 1, 2 and 2% of true CD4 counts were truncated at 1000 cells/mm^3 on the regimes given by $x = 200$, 350 and 500 respectively. [1] In those patients who were observed to initiate treatment. [2] As assigned at treatment initiation.

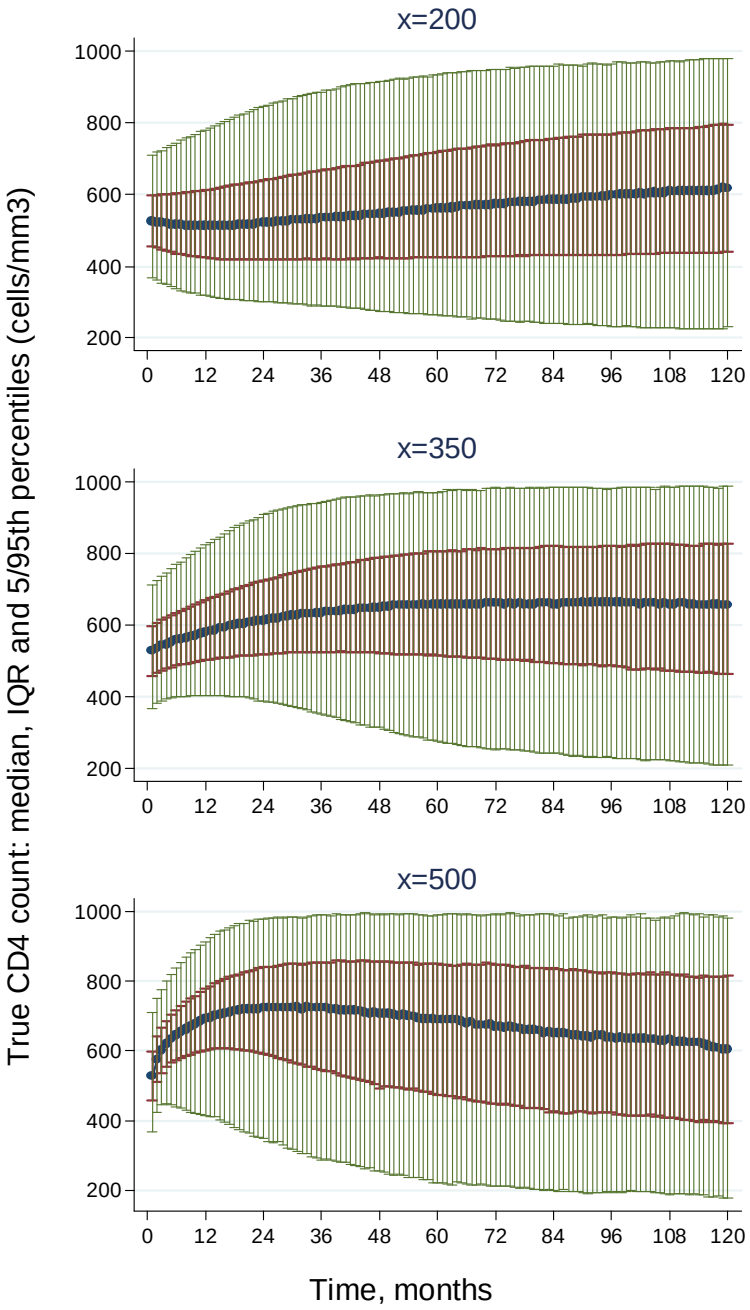


Figure 4.14: Simulation study 1 (large observational study): true CD4 count over time (median, interquartile range and $5/95^{th}$ percentiles) for the $n = 100,000$ patients, after expansion to each of the three regimes given by $x = 200$, 350 and 500, with **no weighting** (population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period).

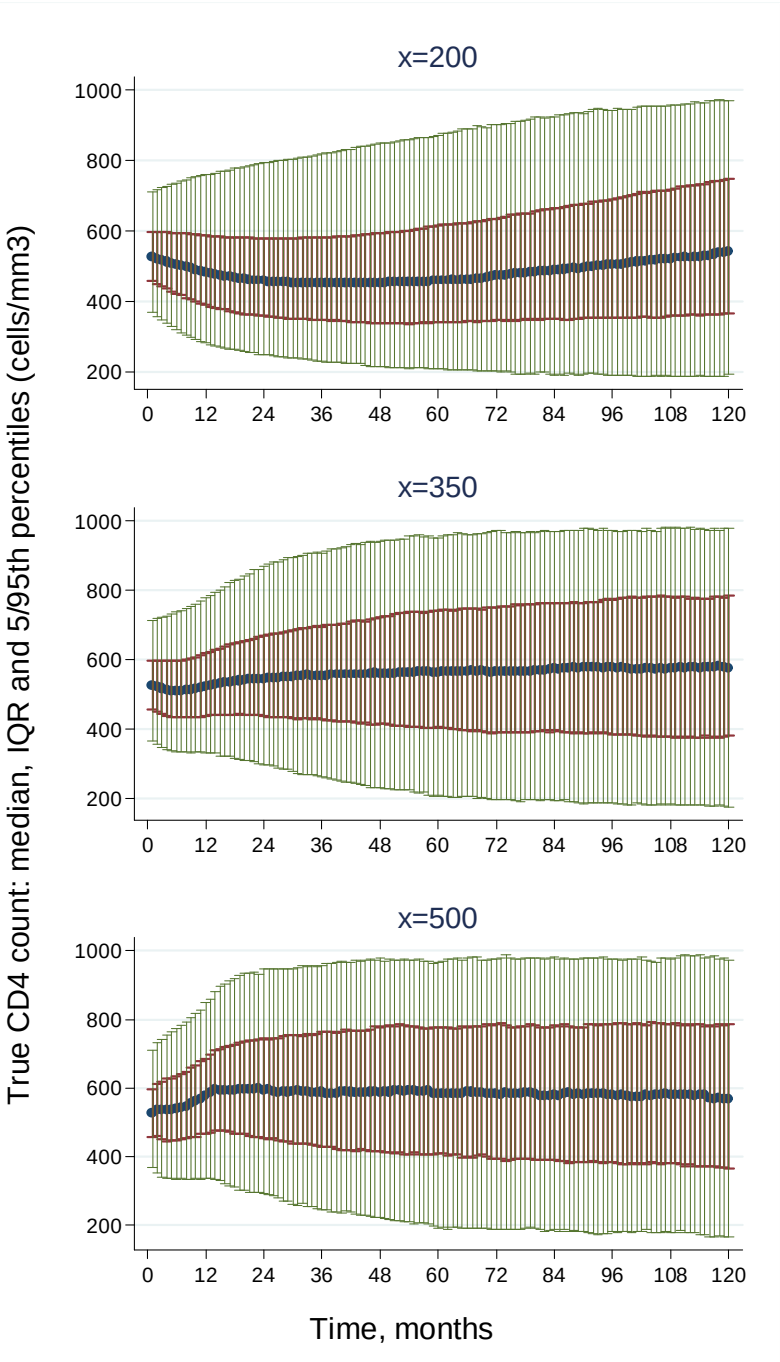
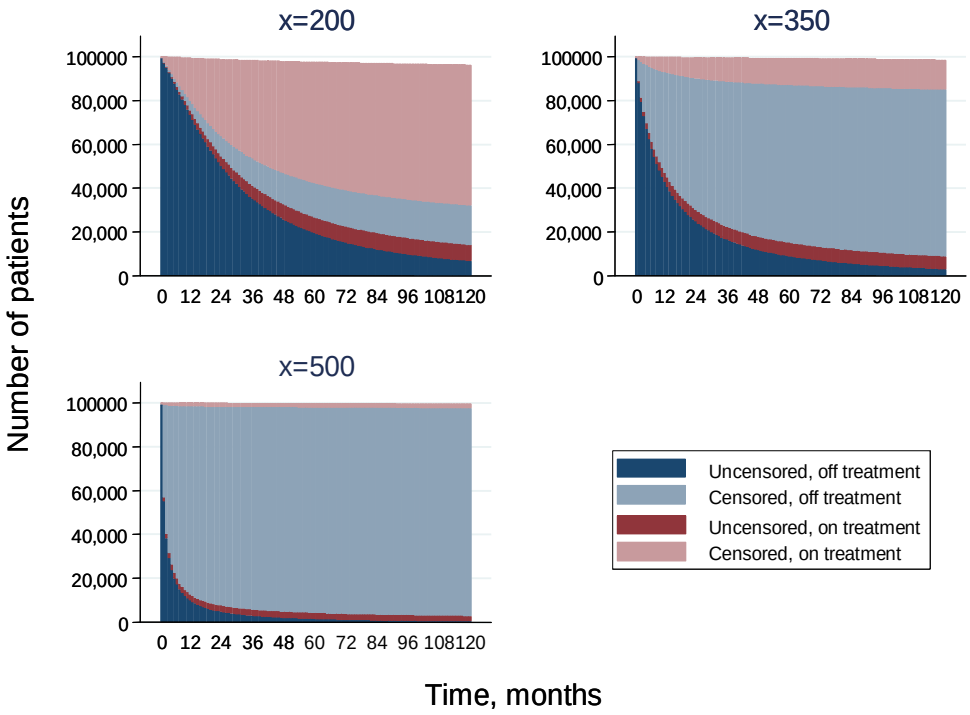


Figure 4.15: Simulation study 1 (large observational study): true CD4 count over time (median, interquartile range and $5/95^{th}$ percentiles) for the $n = 100,000$ patients, after expansion to each of the three regimes given by $x = 200$, 350 and 500, **after application of weights** (truncation at maximum 20; population with regular treatment-naïve CD4 decline, CD4 counts observed monthly and no grace period).



Regime	Time	0-	12-	24-	36-	48-	60-	72-	84-	96-	108-120
	200	684 (92%)	651 (83%)	511 (73%)	397 (59%)	384 (46%)	314 (35%)	273 (27%)	283 (22%)	252 (19%)	236 (15%)
	350	370 (93%)	196 (80%)	110 (58%)	119 (38%)	123 (26%)	137 (20%)	169 (7%)	159 (7%)	144 (4%)	153 (4%)
	500	160 (91%)	27 (44%)	29 (14%)	39 (5%)	54 (0%)	61 (0%)	62 (2%)	63 (2%)	60 (0%)	55 (2%)

Figure 4.16: Simulation study 1 (large observational study): compliance over time of $n = 100,000$ patients with the three regimes given by $x = 200$, 350 and 500, by whether on or off treatment. The table shows the numbers of observed AIDS or death events in each 12 month period, and the percentage of those events that occurred while the patient was still off treatment. If patients reached AIDS or death while uncensored, then they were removed from the risk set, but those censored were carried forward for all time to illustrate the cumulative impact of censoring.

given by $x = 200$, 350 and 500 (if patients reached AIDS or death then they were removed from the risk set, but those censored were carried forward to illustrate the cumulative impact of the censoring; note that no weighting has been applied here). As expected, the predominant censoring on the 200 regime was due to early initiation of treatment, before CD4 count was observed to drop below 200 cells/mm^3, whereas on the 500 regime, the vast majority of patients were censored due to remaining off treatment when their CD4 count was first observed to drop < 500 cells/mm^3. The higher number of events in the regimes defined by lower x is clear. Across all regimes, the proportion of events happening on treatment increased over time, simply due to more patients initiating treatment.

Frequency of observed CD4 count, months	Grace period, months	Approach	Regime given by x		
			200	350	500
1	1	RCT	0.8278	0.8657	0.8587
		Obs 1	0.8298	0.8653	0.8581
		Obs 2	0.8285	0.8647	0.8460
	3	RCT	0.8232	0.8642	0.8581
		Obs 1	0.8282	0.8640	0.8584
		Obs 2	0.8282	0.8638	0.8508
3	1	RCT	0.7926	0.8616	0.8614
		Obs 1	0.8051	0.8631	0.8559
		Obs 2	0.8120	0.8643	0.8589
	3	RCT	0.7861	0.8600	0.8612
		Obs 1	0.7975	0.8583	0.8630
		Obs 2	0.7989	0.8622	0.8635

Table 4.9: Simulation study 1: comparison of the 10-year AIDS-free survival from the RCT with $n = 1,000,000$ patients per regime and as estimated by two large observational studies with $n = 100,000$ patients per regime (different starting seeds; population with regular treatment-naïve CD4 decline and CD4 counts observed every 1 or 3 months, and with grace periods of 1 or 3 months).

Outcome results

The 10-year AIDS-free survival, as estimated by weighted Kaplan-Meier, was 0.8298, 0.8653 and 0.8581 on the regimes given by $x = 200$, 350 and 500, respectively, matching to two decimal places the results obtained from the equivalent RCT with regular treatment-naïve CD4 decline, monthly observed CD4 counts and no grace period.

Precision of results In order to look at the variability in the results, we also considered similar large observational studies with CD4 counts observed every 3 months and 3-month grace periods, and then we repeated each of these (different starting seed). The results illustrate the variability which remains in the estimates despite the large sample size (Table 4.9), reassuring us that any differences between the results from the large observational studies and the RCTs are consistent with sampling variability and do not show any evidence of bias.

4.3.5 Results: 1000 realistically-sized observational studies

We simulated 1000 observational studies each with 3000 patients, considering the regimes $x = 200, 210, ..., 500$. For illustration, the Kaplan-Meier estimates for the 10-year AIDS-free survival from the first 12 simulations for the population with regular treatment-naïve CD4 decline, monthly observed CD4 counts and no grace period ($m = 1$) are shown in Figure 4.17, with the locally-smoothed estimates overlaid. Of note is the variability in the estimates within and between plots, and that the optimal regime is quite frequently at the highest value of x, namely

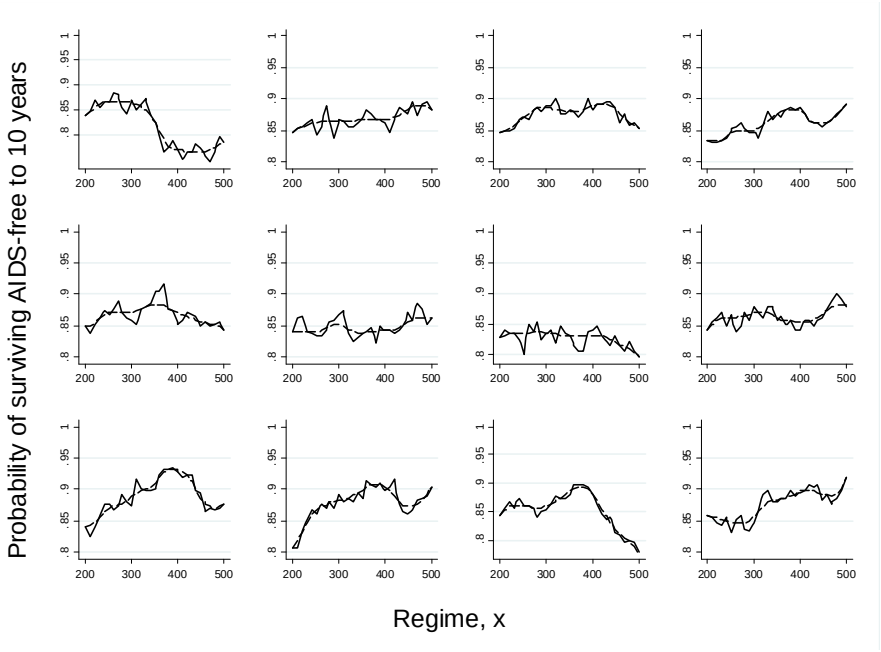


Figure 4.17: Simulation study 1 (small observational studies): estimated probability of surviving AIDS-free for 10 years by regime, for the first 12 of the simulated datasets, as estimated by Kaplan-Meier and locally smoothed (population with regular treatment-naïve CD4 decline and CD4 counts observed monthly, and with no grace period).

500.

When CD4 counts were observed monthly, the raw Kaplan-Meier estimates yielded means and medians which were fairly close to, though slightly higher than, the optimal regime determined from the equivalent RCT (for example, with no grace period, the mean was 374 compared to the optimal regime of 360 from the RCT; Table 4.10). Of note, a peak in the histograms of the raw Kaplan-Meier estimates was visible at $x = 500$, probably due to the optimal regime under some of those simulations being given by $x > 500$ (Figure 4.18). The standard deviations were large, at 71-77 across the different grace periods. In particular, extending the grace period from 1 to 6 months reduced the SD by only 9%. Across all grace periods, the percentage of optimal regime estimates which were lower than the minimum acceptable regime from the equivalent RCT was 13%.

The local smoothing and pooled logistic regression approaches yielded fairly similar results to each other, but with means, medians and standard deviations typically higher than under the raw Kaplan-Meier approach (for example, with no grace period, the standard deviations were 85 and 89, respectively; Table 4.10). It is apparent from the histograms that this is due to the optimal regime frequently being estimated at the maximum range of x, namely 500 (Figure 4.18).

CD4 freq., months	Approach	Summary	Grace period, months					
			1		3		6	
1	RCT	Optimal regime (MA)	360	(290)	360	(290)	370	(300)
	Raw KM	Mean (SD)	374	(77)	375	(73)	379	(71)
		Median (%<RCT MA)	370	(13%)	370	(13%)	375	(13%)
	Smoothed KM[1]	Mean (SD)	383	(85)	381	(81)	390	(76)
		Median (%<RCT MA)	380	(14%)	380	(13%)	390	(11%)
	Pooled logistic[2]	Mean (SD)	374	(89)	378	(86)	387	(80)
		Median (%<RCT MA)	370	(21%)	370	(14%)	375	(11%)
3	RCT	Optimal regime (MA)	410	(340)	410	(350)	420	(360)
	Raw KM	Mean (SD)	395	(70)	418	(57)	420	(60)
		Median (%<RCT MA)	400	(23%)	420	(12%)	430	(16%)
	Smoothed KM[1]	Mean (SD)	405	(73)	426	(62)	428	(62)
		Median (%<RCT MA)	410	(21%)	430	(12%)	430	(15%)
	Pooled logistic[2]	Mean (SD)	402	(76)	424	(66)	430	(63)
		Median (%<RCT MA)	400	(24%)	420	(15%)	420	(14%)

Table 4.10: Simulation study 1: results from the 1000 simulated observational studies. Population with regular treatment-naïve CD4 decline, with CD4 counts observed every 1 or 3 months, and grace periods of 1, 3 or 6 months. RCT results shown are the optimal and minimum acceptable regimes (after local smoothing applied). Results shown from the observational studies are the mean, standard deviation and median of the estimated optimal regimes, and the percentage of estimated optimal regimes that were less than the minimum acceptable regime from the equivalent RCT. KM=Kaplan-Meier. MA=minimum acceptable. SD=standard deviation. [1] Local smoothing procedure applied to the Kaplan-Meier estimates. [2] Pooled logistic regression applied to the raw data. See text for further details on all these methods.

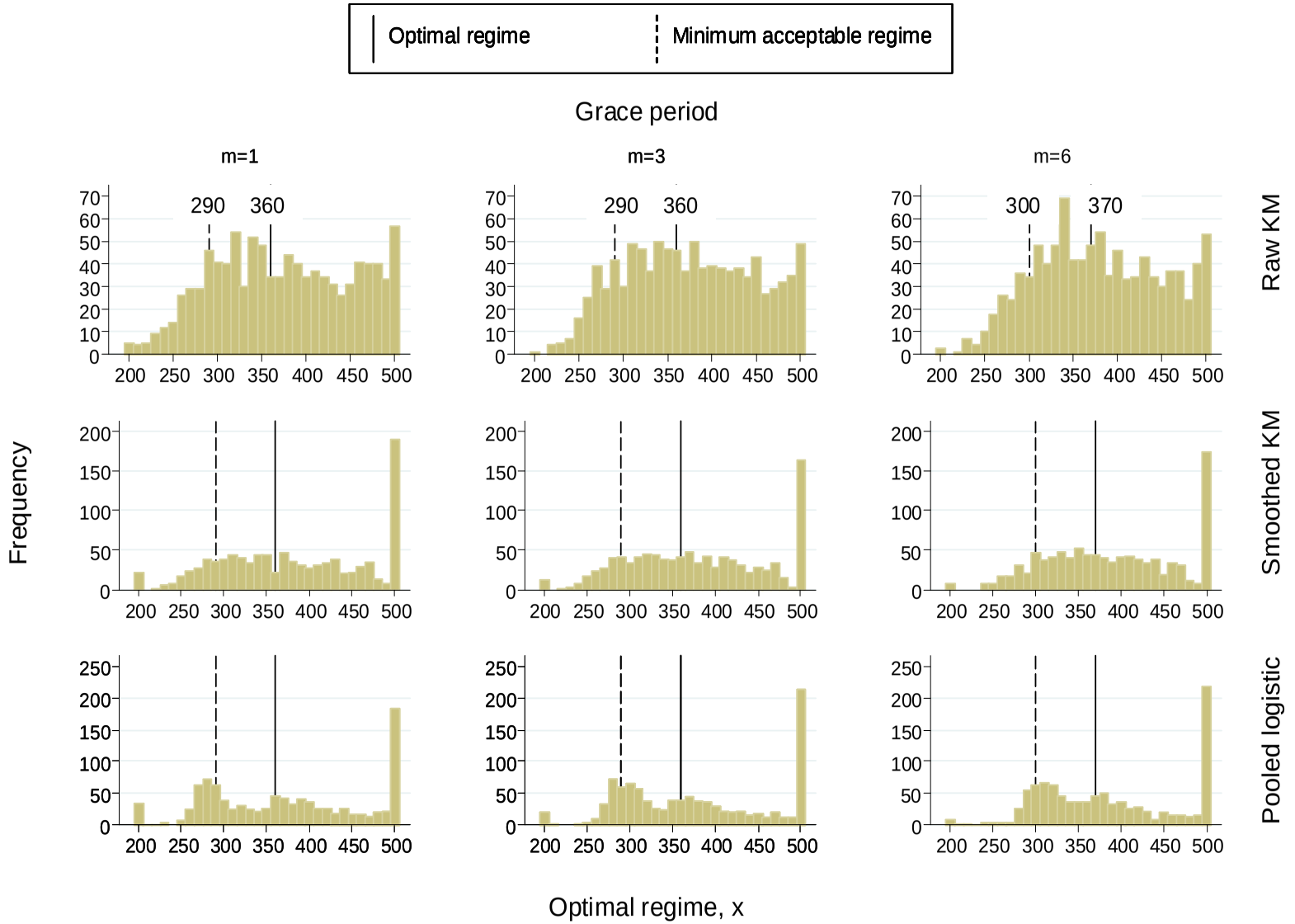


Figure 4.18: Simulation study 1: estimated optimal regimes from the 1000 simulated observational studies, with regular treatment-naïve CD4 decline and **CD4 counts observed monthly**. The solid and dashed vertial lines indicate the optimal and minimum acceptable regimes respectively from the equivalent RCT (after smoothing; values shown on the top plots). Note the different scales on the y-axis for the different methods.

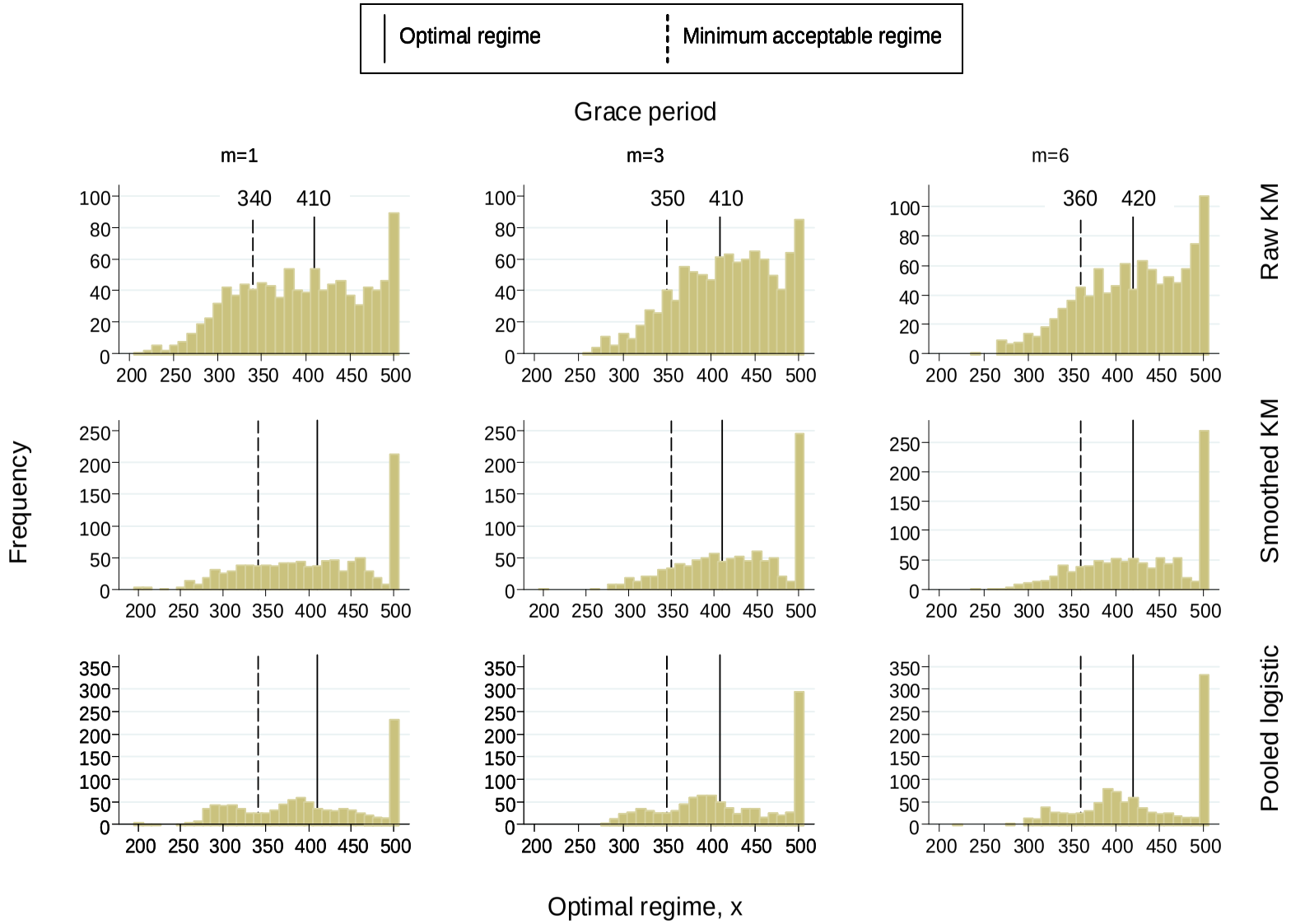


Figure 4.19: Simulation study 1: estimated optimal regimes from the 1000 simulated observational studies, with regular treatment-naïve CD4 decline and **CD4 counts observed every 3 months**. The solid and dashed vertial lines indicate the optimal and minimum acceptable regimes respectively from the equivalent RCT (after smoothing; values shown on the top plots). Note the different scales on the y-axis for the different methods.

When the CD4 counts were observed every 3 months, we know that the optimal regimes as determined by the RCTs are higher, and this was reflected in the observational study simulation results (Table 4.10 and Figure 4.19). Similar patterns with respect to the results of the equivalent RCTs were seen as when the CD4 counts were observed monthly. However, the results from the local smoothing and pooled logistic regression approaches were slightly improved, with means and medians closer to the optimal regime from the equivalent RCT and smaller standard deviations, although this was probably due to the range of x over which the AIDS-free survival rates were broadly constant being smaller (higher minimum acceptable regimes, and the maximum of the range at 500).

Of note, under both CD4 observation frequencies, longer grace periods were associated with slightly higher mean and median optimal regimes and slightly lower standard deviations, as we would anticipate.

Inference under assumption of no grace period

For each CD4 count observation frequency, Table 4.11 shows the performance (bias, mean square error (MSE), and relative efficiency (RE) with reference to the pooled logistic regression approach with no grace period) of the different approaches and grace periods, assuming that the inference of interest is under no grace period ($m = 1$). Under monthly observed CD4 counts, the biases under all approaches were all > 0, indicating overestimation of the optimal regime. As anticipated, there was a trend towards greater bias, but smaller MSE and RE, with longer grace periods. The raw Kaplan-Meier approach consistently performed better than the other two approaches, related to the smaller variances (see the standard deviations in Table 4.10).

When CD4 counts were observed every 3 months, we saw broadly similar patterns, except for two key differences. Firstly, the bias, MSE and RE all tended to be smaller, compared to when CD4 counts were observed monthly. Under all three approaches, with no grace period the bias was negative (indicating underestimation of the optimal regime). Secondly, there was no clear benefit of permitting a 6-month compared to 3-month grace period, since the bias increased under all approaches and the MSE and RE were either broadly similar or larger.

Summary

In this example, due to the large measurement error in CD4 count and the broadly constant 10-year AIDS-free survival rates at higher CD4 counts, a single analysis with a realistic sample size may yield an estimate quite "far" from the optimal regime. In particular, the estimates tended to be biased towards higher regimes. Although lacking precision, the raw Kaplan-

CD4 count frequency, months	Approach		Grace period, months		
			1	3	6
1	Raw KM	Bias	14	15	19
		MSE	6149	5543	5386
		RE	0.75	0.67	0.64
	Smoothed KM[1]	Bias	23	21	30
		MSE	7733	6932	6711
		RE	0.91	0.82	0.74
	Pooled logistic[2]	Bias	14	18	27
		MSE	8084	7692	7112
		RE	1 (ref)	0.93	0.81
3	Raw KM	Bias	-15	8	10
		MSE	5099	3273	3644
		RE	0.84	0.55	0.61
	Smoothed KM[1]	Bias	-5	16	18
		MSE	5388	4066	4135
		RE	0.92	0.66	0.65
	Pooled logistic[2]	Bias	-8	14	20
		MSE	5874	4537	4333
		RE	1 (ref)	0.75	0.68

Table 4.11: Simulation study 1: bias, mean square error (MSE) and relative efficiency (RE) when comparing the results from the observational studies with grace periods of 1, 3 or 6 months, compared to the equivalent RCT but with no grace period ($m = 1$). Note that the variance under each scenerio is given by the square of the standard deviation in Table 4.10. Population with regular treatment-naïve CD4 decline and CD4 counts observed every 1 or 3 months. RCT results after least square smoothing applied. KM=Kaplan-Meier. [1] Local smoothing procedure applied to the Kaplan-Meier estimates. [2] Pooled logistic regression applied to the raw data. See text for further details on all these methods.

Meier approach performed best overall, since the smoothed Kaplan-Meier and pooled logistic regression approaches frequently estimated the optimal regime at the upper bound of the set of regimes under consideration (namely, at $x = 500$).

Of note for the CASCADE analyses, when CD4 counts were observed every 3 months, permitting a grace period of 3 months may not result in a large bias for the estimation of the optimal regime in the absence of a grace period, and may offer benefits in terms of greater precision. Extension to a 6-month grace period may not offer any additional advantages. If a grace period is permitted for the purposes of potentially increase precision, then under the inference of no grace period, there will naturally be bias towards higher regimes.

4.4 Simulation study 2

4.4.1 Motivation

The results of the first simulation study reported above naturally raised the question of whether in a scenario where the optimal regime is more distinct (that is, the outcome-by-regime curve is less flat and has a clearer peak) and with a greater number of patients, the application of dynamic MSMs to the observational data would yield results closer to those of the equivalent RCT. We therefore repeated the simulation study above, but with a larger number of patients and artificially enforcing a greater penalty for early and late treatment initiation, with respect to CD4 count, to create an outcome-by-regime curve with a more distinct peak for the optimal regime.

4.4.2 Methods

As in the first simulation study, we simulated large RCTs and a large number of realistically-sized observational studies. The study was conducted in exactly the same way as the first, except for three differences. Firstly, we increased the number of patients in the observational studies to $n = 7000$, which, within computational limitations, is closer to the size of other observational studies investigating applying dynamic MSMs to look at the effects of HIV treatment (Cain et al. (2010); Young et al. (2011); these used prevalent rather than incident cohorts like CASCADE and therefore had access to greater numbers of patients). Secondly, we applied a penalty for early or late treatment initiation, with respect to CD4 count (see below). Lastly, we performed 500 rather than 1000 observational study simulations, to reduce computational time and because the interpretations were fairly clear even with this smaller number of simulations.

Penalty for early or late treatment initiation

We reduced the CD4 slope from one year after treatment initiation, if treatment was initiated when true CD4 count was < 300 or > 400 cells/mm^3. This penalty was a fixed linear function of true CD4 count at treatment initiation, with the new slope $S_2^{'}$ on the square-root scale given by:

$$S_2^{'} = \begin{cases} S_2 - 1.2 + 0.004R^2 & \text{if } R^2 < 300 \\ S_2 & \text{if } R^2 \geq 300 \text{ and } \leq 400 \\ S_2 + 1.2 - 0.003R^2 & \text{if } R^2 > 400 \end{cases}$$

where, as previously, R^2 is the true CD4 count at treatment initiation and S_2 is the CD4 slope on the square-root scale from one year after treatment initiation. This equates to a reduction in S_2 by 0.4 and 0.3 if the true CD4 count at treatment initiation was 200 and 500 cells/mm^3, respectively. Note that this function is continuous at all values of CD4 count (with change-points at 300 and 400 cells/mm^3).

4.4.3 Results: the randomised trials

Optimal regimes

Figure 4.20 clearly shows that the 10-year AIDS-free survival by regime curves were less flat and had clearer peaks, as intended. Under the scenario where CD4 counts were observed monthly, the optimal regimes under grace periods of 1, 3 and 6 months were 290, 300 and 310, respectively (with 10-year AIDS-free survival probabilities on those optimal regimes of 0.8559, 0.8533 and 0.8490, respectively; Table 4.12). There were no changes under smoothing, except for the grace period of 1 month, where the optimal regime was 300 under both local smoothing methods (with very similar 10-year AIDS-free survival probabilities).

When CD4 counts were observed every 3 months, the optimal regime was given by $x = 350$, regardless of the length of the grace period, although with poorer 10-year AIDS-free survival with longer grace periods (Table 4.12). However, there were some differences under local smoothing, with the optimal regimes being given by $x = 340$ and 350 with no grace period ($m = 1$) when smoothing by least squares and weighting, respectively, and $x = 360$ under both local smoothing methods for grace periods of both 3 and 6 months. The reason for few if any differences in the optimal regimes across the different grace periods is probably due to the AIDS-free survival by regime curves having clearer peaks, resulting from the penalties imposed for early or late treatment initiation with respect to CD4 count (Figure 4.20).

CD4 count frequency, months	Grace period, months	Optimal regime	Minimum acceptable regime
1	1	290 (0.8559)	270 (0.8530)
		300 (0.8557)	-
		300 (0.8557)	-
	3	300 (0.8533)	270 (0.8498)
		-	-
		-	-
	6	310 (0.8490)	280 (0.8454)
		-	-
		-	-
3	1	350 (0.8521)	310 (0.8476)
		340 (0.8517)	-
		350 (0.8518)	-
	3	350 (0.8493)	320 (0.8455)
		360 (0.8489)	-
		360 (0.8489)	-
	6	350 (0.8453)	330 (0.8418)
		360 (0.8448)	-
		360 (0.8448)	-

Table 4.12: Simulation study 2 (RCTs): optimal and minimum acceptable regimes in populations different CD4 count observation frequencies and grace periods (regular treatment-naïve CD4 decline). For each scenario, the first line gives the results with no local smoothing, and the second and third lines show the results under local smoothing using least squares and weighting, respectively (the smoothed results are only shown if the optimal differs from that under no smoothing). Values in brackets are the estimated probabilities of surviving AIDS-free to 10 years under that regime.

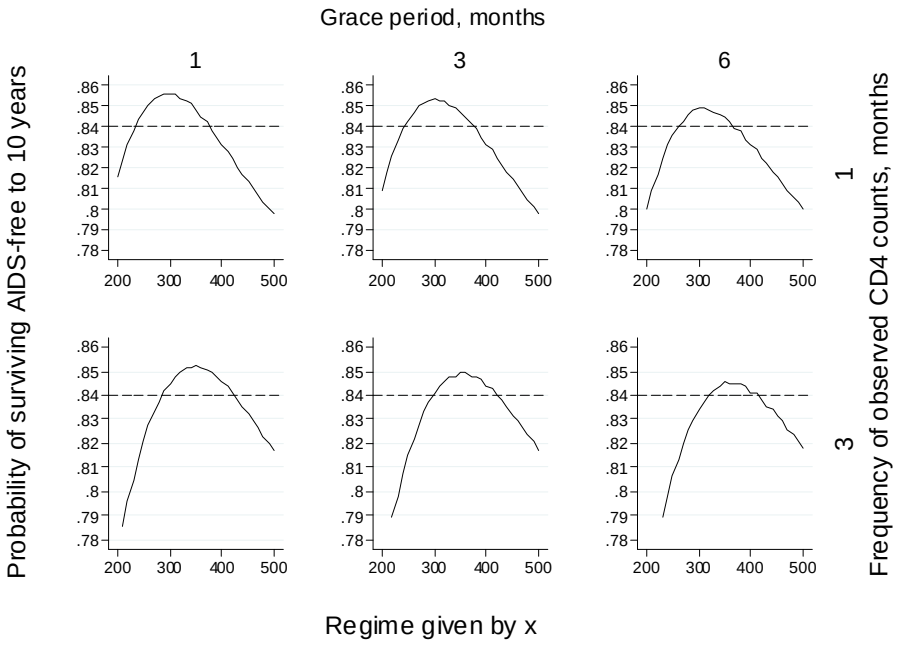


Figure 4.20: Simulation study 2 (RCTs): probability of surviving AIDS-free to 10 years by regime, across different CD4 count observation frequencies and grace periods (population with regular treatment-naïve CD4 decline). Note that probabilities were only plotted if ≥ 0.78 to preserve a common scale. Horizontal line drawn at 0.84 to aid comparison between plots.

Minimum acceptable regimes

The minimum acceptable regimes were given by $x = 270$, 270 and 280 under grace periods of 1, 3 and 6 months, respectively, when CD4 counts were observed monthly (Table 4.12). The corresponding figures were 310, 320 and 330 when CD4 counts were observed every 3 months. There were no changes under either local smoothing method.

4.4.4 Results: the observational studies

Under both CD4 count observation frequencies, all approaches performed somewhat better compared to the first simulation study, in particular with smaller standard deviations and smaller percentages of estimated optimal regimes being less than the minimum acceptable regime from the equivalent RCT (Table 4.13). This was most noticeable for the pooled logistic regression approach. This is illustrated in the histograms, which were much more centred on the optimal regime from the equivalent RCT, as we would expect in this scenario where there is a more defined optimal curve (Figures 4.21 and 4.22).

When looking at inference under the assumption of no grace period, the most noticeable difference compared to the first simulation study was the reduction in MSE across all scenarios, related to the reductions in variances (Table 4.14; see also standard deviations in Table

CD4 freq., months	Approach	Summary	Grace period, months					
			1		3		6	
1	RCT	Optimal regime (MA)	300	(270)	300	(270)	310	(280)
	Raw KM	Mean (SD)	314	(52)	316	(42)	321	(38)
		Median (%<RCT MA)	300	(15%)	310	(11%)	320	(10%)
	Smoothed KM[1]	Mean (SD)	312	(49)	315	(40)	321	(37)
		Median (%<RCT MA)	300	(10%)	310	(6%)	320	(7%)
	Pooled logistic[2]	Mean (SD)	303	(44)	308	(33)	315	(27)
		Median (%<RCT MA)	290	(6%)	300	(2%)	310	(1%)
3	RCT	Optimal regime (MA)	340	(310)	360	(320)	360	(330)
	Raw KM	Mean (SD)	363	(53)	363	(41)	372	(39)
		Median (%<RCT MA)	350	(12%)	360	(12%)	370	(9%)
	Smoothed KM[1]	Mean (SD)	365	(54)	361	(38)	372	(40)
		Median (%<RCT MA)	350	(10%)	350	(8%)	370	(8%)
	Pooled logistic[2]	Mean (SD)	361	(52)	357	(36)	369	(33)
		Median (%<RCT MA)	360	(11%)	350	(6%)	370	(6%)

Table 4.13: Simulation study 2: results from the 1000 simulated observational studies. Population with regular treatment-naïve CD4 decline, with CD4 counts observed every 1 or 3 months, and grace periods of 1, 3 or 6 months. RCT results shown are the optimal and minimum acceptable regimes (after local smoothing applied), Results shown from the observational studies are the mean, standard deviation and median of the estimated optimal regimes, and the percentage of estimated optimal regimes that were less than the minimum acceptable regime from the equivalent RCT. KM=Kaplan-Meier. MA=minimum acceptable. SD=standard deviation. [1] Local smoothing procedure applied to the Kaplan-Meier estimates. [2] Pooled logistic regression applied to the raw data. See text for further details on all these methods.

CD4 count frequency, months	Approach		Grace period, months		
			1	3	6
1	Raw KM	Bias	14	16	21
		MSE	2919	2045	1890
		RE	1.38	0.90	0.74
	Locally-smoothed KM[2]	Bias	12	15	21
		MSE	2594	1832	1808
		RE	1.24	0.81	0.70
	Pooled logistic[3]	Bias	3	8	15
		MSE	1982	1123	955
		RE	1 (ref)	0.54	0.38
3	Raw KM	Bias	23	23	32
		MSE	3380	2202	2573
		RE	1.06	0.62	0.57
	Locally-smoothed KM[2]	Bias	25	21	32
		MSE	3604	1873	2578
		RE	1.10	0.53	0.58
	Pooled logistic[3]	Bias	21	17	29
		MSE	3124	1561	1908
		RE	1 (ref)	0.48	0.40

Table 4.14: Simulation study 2: bias, mean square error (MSE) and relative efficiency (RE) when comparing the results from the observational studies with grace periods of 1, 3 or 6 months, compared to the equivalent RCT but with no grace period ($m = 1$). Note that the variance under each scenerio is shown in Table 4.10. Population with regular treatment-naïve CD4 decline and CD4 counts observed every 1 or 3 months. RCT results after least square smoothing applied. KM=Kaplan-Meier. [1] Local smoothing procedure applied to the Kaplan-Meier estimates. [2] Pooled logistic regression applied to the raw data. See text for further details on all these methods.

4.13). The pooled logistic regression approach performed consistently better than the other approaches, with smaller bias, MSE and RE. As before, permitting a 3-month grace period led to improvements in the MSE and RE, compared to no grace period ($m = 1$). When CD4 counts were observed 3-monthly, there was no clear benefit in allowing a 6-month grace period, due to increases in the MSE across all approaches (the larger biases outweighed the gains in efficiency).

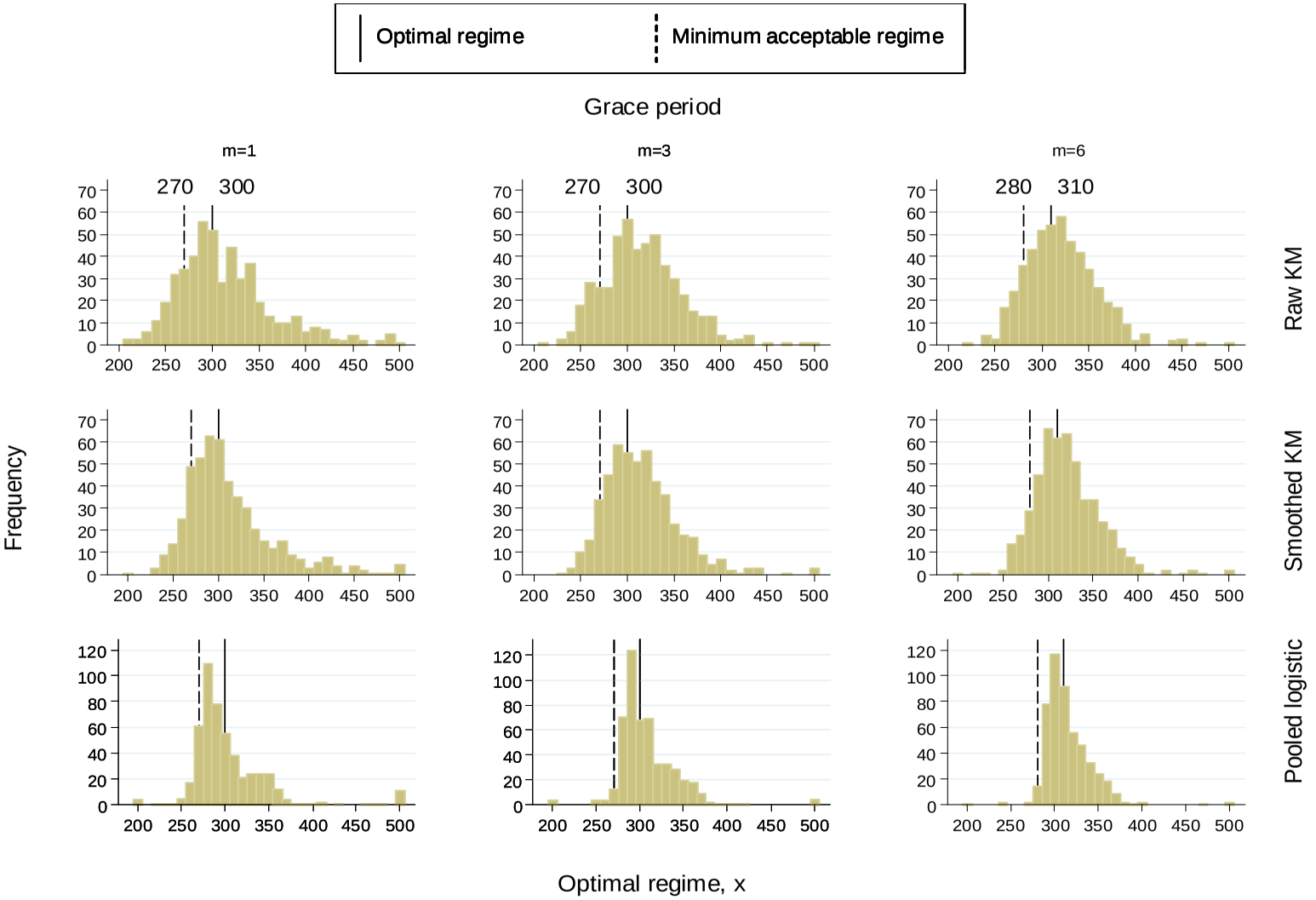


Figure 4.21: Simulation study 2: estimated optimal regimes from the 500 simulated observational studies, with artifical penalty for early or late treatment initiation (otherwise regular treatment-naïve CD4 decline; **CD4 counts observed every month**). The solid and dashed vertial lines indicate the optimal and minimum acceptable regimes respectively from the equivalent RCT (after smoothing; values shown on the top plots). Note the different scales on the y-axis for the different methods.

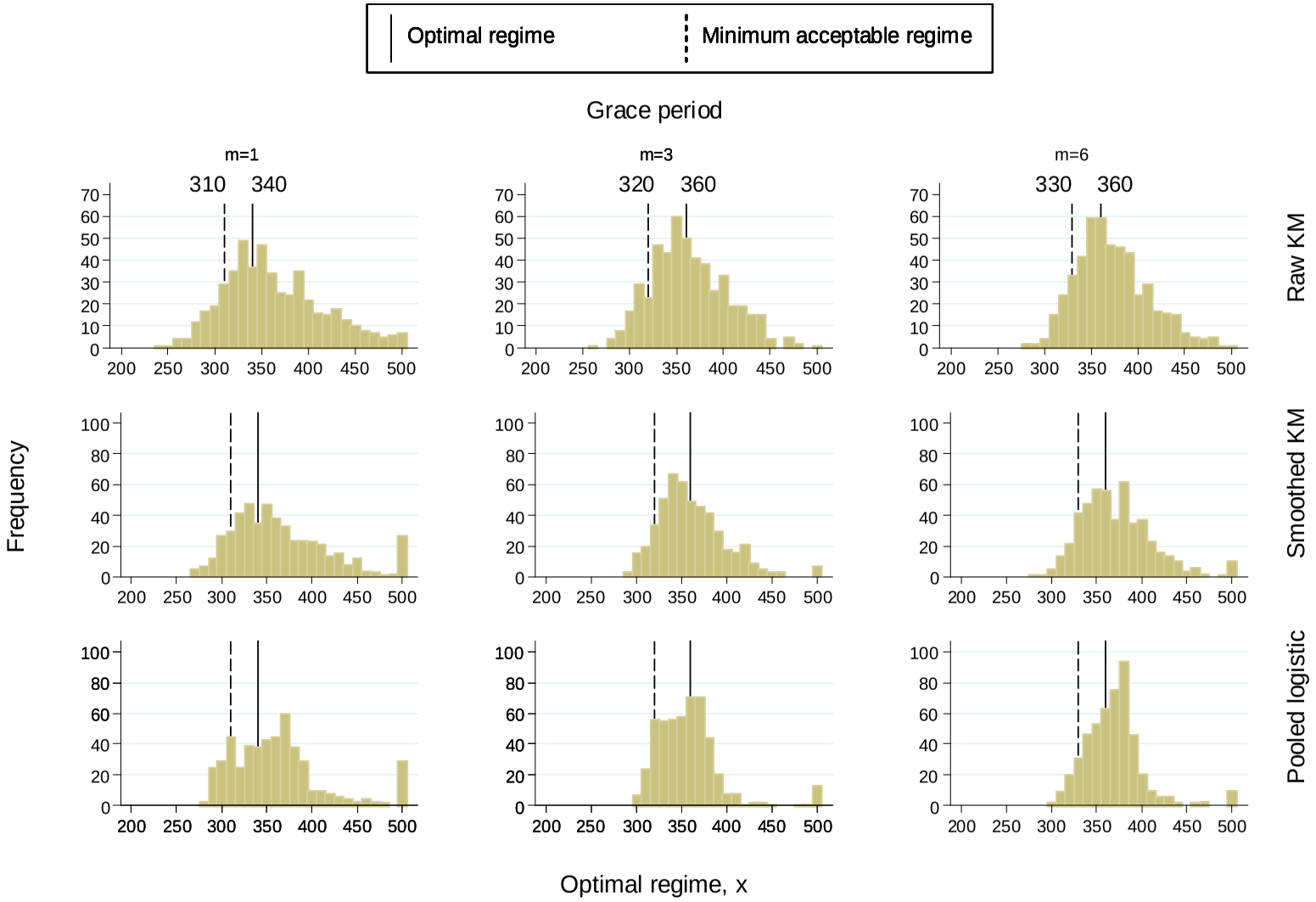


Figure 4.22: Simulation study 2: estimated optimal regimes from the 500 simulated observational studies, with artifical penalty for early or late treatment initiation (otherwise regular treatment-naïve CD4 decline; **CD4 counts observed every three months**). The solid and dashed vertial lines indicate the optimal and minimum acceptable regimes respectively from the equivalent RCT (after smoothing; values shown on the top plots). Note the different scales on the y-axis for the different methods.

4.5 Application to CASCADE

A number of previous researchers have attempted to estimate the optimal time to initiate treatment in HIV-infected persons with respect to CD4 count, as outlined in section 1.5. While these studies have typically benefited from a greater sample size than we have available in CASCADE, we have the advantage of a seroconverter, as opposed to seroprevalent, cohort. The pros and cons of our approach are discussed further in chapter 5.

The work of our first simulation study with realistic CD4 count trajectories (section 4.3) indicated that in populations where CD4 counts are observed every 3 months, permitting a grace period of 3 months may not result in a large bias for the estimation of the optimal regime in the absence of a grace period, and may offer benefits in terms of greater precision. Extension to a 6-month grace period may not offer any additional advantages. Therefore we used the CASCADE data to estimate the optimal regime in terms of when to initiate treatment with respect to CD4 count, in those with CD4 counts $\geq$ 500 cells/mm^3, allowing a 3-month grace period for treatment initiation. We used the second approach of Cain et al. (2010) where uniform treatment initiation across the grace period is assumed, as was applied in the simulation studies.

4.5.1 Methods

Treatment regimes

We used the treatment regimes considered in the simulation studies, that is, defined by $x = 200, 210, ..., 500$. However, as we have seen in previous chapters, 15% of the 1082 treatment initiations observed in our population of 3382 patients occurred in the first month following study entry. Since by definition all patients had CD4 count $\geq$500 cells/mm^3 at that time, none of these treatment initiations would be compliant with any of the treatment regimes given by $x = 200, 210, ..., 500$ and so would all be censored, resulting in a substantial loss of information. Therefore, we considered incorporating an additional regime defined as "initiate treatment immediately following study entry", since it has some clinical meaning with respect to our entry criteria (namely, first CD4 count $\geq$ 500 cells/mm^3 within 1-5 years after seroconversion). While all our results are based on this large number of regimes, for clarity we will sometimes present summaries of the data for just the key regimes given by $x = 200, 350, 500$ and "initiate immediately". Note that our immediate treatment initiation regime has a different meaning to immediate initiation at first observed CD4 count, regardless of the CD4 count level; our results refer only to the subpopulation who have a first observed CD4 count $\geq$ 500 cells/mm^3.

Weight estimation

We used the treatment models as determined under the different strategies of chapter 2 to estimate the weights. As indicated in section 4.2, it is not trivial to stabilise the weights while permitting a grace period with $m > 1$, and the stabilised weights are not guaranteed to increase the precision, therefore we used non-stabilised weights throughout. When using grace periods, the "probabilities" in the numerator of the non-stabilised weights may be < 1 (including while the denominator is equal to 1), therefore the non-stabilised weights may be < 1, in contrast to non-stabilised weights under standard MSMs. However, the value of the numerator is a simple function of the interval of the grace period, and therefore will not suffer from extreme values; with a grace period of 3, it will have a lower bound of 1/3 (see section 4.2.2). Therefore, rather than truncating the outer percentiles, we truncated the upper 1%, which was typically close to the value of 20 used in the simulation studies. Of note, the only difference between strategies Ia and II/III of chapter 2 was the degree of truncation, therefore with this blanket truncation across all strategies there was no longer any difference between these strategies and they are presented here as one.

AIDS-free survival

As in the simulation studies, we estimated survival using Kaplan-Meier methods, with and without local smoothing, and also using pooled logistic regression models. We obtained both weighted and unweighted estimates, to look at the impact of the weighting.

To smooth the Kaplan-Meier estimates, we used the same approach as in the simulation studies, but only across the range $x = 200$ to 500; the immediate treatment initiation regime estimates were left unchanged.

In the pooled logistic regression models, time was included as a 5 knot spline as in chapter 2, and categorised as 0-<0.5, 0.5-<1, 1-<2 and ≥2 years for the interaction with regime (Cain et al., 2010). Regime was included as a 4 knot spline for values $x = 200$ to 500, with knots at the 5, 35, 65 and 95^{th} percentiles (Harrell, 2001) which translated to x given by 210, 290, 380 and 490, and a separate indicator was used for the immediate treatment initiation regime. Of note, the non-stabilised weights adjust for the time-dependent as well as the time-independent covariates, therefore it was not necessary to include the baseline covariates in the outcome model. We directly predicted survival from the pooled logistic regression models; the resulting survival curves are analogous to the standardised survival curves of chapter 2.

As we have seen in the simulation studies, it is important to allow long follow-up when

seeking to optimise dynamic treatment regimes. However, of course we were limited by the observed follow-up in our population of CASCADE patients (median 2.3 years), therefore we focussed on the AIDS-free survival at 3 and 6 years.

Interval estimation

95% confidence intervals were estimated by bootstrap with resampling stratified by country (500 repetitions).

Censoring

"Usual" censoring may be incorporated using weights as in previous chapters, but based on the results from those chapters we would expect this to make little difference in practice, and so was not incorporated here.

4.5.2 Results

We used the same dataset of 3382 patients as throughout the thesis, but, as in chapter 3, 26 patients were censored in the second month of follow-up, and therefore those patients contributed to the weight estimation only and not the outcome estimations.

Compliance with treatment regimes

In the absence of a grace period, 2325, 2072 and 1438 patients remained compliant throughout their follow-up with the regimes given by $x = 200$, 350 and 500, respectively. Incorporating a 3-month grace period, 2356, 2166 and 1538 patients were compliant with those three regimes, respectively. Overall, permitting a 3-month grace period, 35% of the 1082 observed treatment initiations were compliant with at least one regime given by $x = 200, 210, ..., 500$, and 20% initiated treatment immediately. Therefore, incorporating the regime of immediate treatment initiation meant that 55% of the observed treatment initiations were in compliance with at least one regime. Of note, 20% of the treatment initiations were at a CD4 count which had been carried forward for more than 3 months, therefore were censored due to the treatment initiation being beyond the permitted grace period.

Treatment initiations across the grace period

Table 4.15 illustrates the treatment initiation patterns across the grace period for those observed treatment initiations which were compliant with each of the four regimes given by $x = 200$,

Interval of the grace period	Regime			
	x =200	350	500	Imm
1	46 (63%)	51 (40%)	42 (41%)	161 (75%)
2	18 (25%)	46 (37%)	30 (29%)	33 (15%)
3	9 (12%)	29 (23%)	30 (29%)	22 (10%)
Total	73 (100%)	126 (100%)	102 (100%)	216 (100%)

Table 4.15: Application to CASCADE: pattern of treatment initiation across the grace period, for those observed treatment initiations which were in compliance with the regimes given by x = 200, 350, 500, and immediate treatment initiation. Values are the number of treatment initiations in a given interval of the grace period which were in compliance with the given regime (% of total number of treatment initiations across the grace period which were in compliance with the given regime). Imm=immediate treatment initiation regime.

350, 500 and immediate treatment initiation. A higher percentage of patients who initiated in compliance with regime $x = 200$ initiated in the first interval of the grace period (63%) compared to those who initiated in compliance with regimes $x = 350$ or 500 (40 and 41%, respectively). This pattern is as we may expect, since at lower CD4 counts clinicians and patients may be keen to initiate treatment sooner, whereas at higher CD4 counts they may not be concerned about a small delay. However, a large percentage of patients who initiated in compliance with the immediate treatment regime did so in the first interval of the grace period; this is probably related to the definition of that regime. Of note, the first approach of Cain et al. (2010) would only upweight those patients who waited until the last interval of the grace period after their CD4 count had first dropped below the given threshold to initiate. For the regime given by $x = 200$ in particular, this subset of patients is unlikely to be representative of the remainder of the patients who initiated earlier in the grace period.

Weights

Summaries of the weights are presented in Table 4.16; of note, these are non-stabilised weights therefore we no longer expect the mean to be close to 1. The maxima of the untruncated weights were between 135 and 361 for the strategies Ia to V, and were much larger under strategies VI and VII (1044 and 42034, respectively). After truncation of the upper 1% of the weights, the maxima were between 14 and 17 across all strategies. The means of the truncated weights ranged from 1.542 under strategy V to 1.606 under strategy Ib. The standard deviations were around 2.

AIDS-free survival

Overall, 103, 89, 55 and 15 AIDS or death events were observed in patients remaining compliant with the regimes given by $x = 200$, 350, 500 and immediate treatment initiation. Of these, only

Strategy	No truncation		Truncation of upper 1%	
	Mean (SD)	Range	Mean (SD)	Range
Ia/II/III	1.745 (3.557)	0.33, 135	1.604 (2.011)	0.33, 15
Ib	1.743 (3.569)	0.33, 172	1.606 (2.028)	0.33, 16
IV	1.767 (4.595)	0.33, 342	1.587 (2.102)	0.33, 17
V	1.742 (4.591)	0.33, 361	1.542 (1.785)	0.33, 14
VI	1.779 (4.349)	0.33, 1044	1.601 (2.082)	0.33, 16
VII	1.810 (45.30)	0.33, 42034	1.570 (2.023)	0.33, 16

Table 4.16: Application to CASCADE: summary of the estimated weights from each of the different strategies. SD=standard deviation.

2, 7, 4 and 11, respectively, occurred following treatment initiation. Of note, the 4 events which occurred while treatment-naïve and in compliance with the immediate treatment initiation regime all occurred during the 3-month grace period (none initiated treatment before the event; all patients still had CD4 counts over 500 cells/mm^3 therefore these events were also included in the events while treatment-naïve for the regimes given by $x = 200$, 350 and 500).

Figure 4.23 illustrates the probability of surviving AIDS-free to 6 years from study entry for the regimes given by $x = 200$, 350, 500 and immediate initiation, unweighted and weighted based on the different weight estimation strategies (under the raw Kaplan-Meier approach). Overall, there was little difference between the weight estimation strategies; all showed some greater separation between the AIDS-free survival curves compared to the unweighted estimation. There appeared to be little difference between the treatment regimes given by $x = 200$ and 350, except a suggestion of poorer AIDS-free survival on the regime $x = 350$ at later times, perhaps contrary to what we might expect. Immediate treatment initiation appeared to be preferable when considering AIDS-free survival to 6 years, compared to delaying treatment to any of the three CD4 count thresholds considered (except perhaps under strategy VI).

Looking instead at the AIDS-free survival curves predicted from the pooled logistic regression models (Figure 4.24), there was again greater separation between the regimes for the weighted compared to the unweighted curves. Similarly to the raw Kaplan-Meier AIDS-free survival curves, there was little difference between the regimes given by $x = 200$ and 350. In contrast to the raw Kaplan-Meier curves, there was a suggestion that waiting until the CD4 count is first observed to drop < 500 cells/mm^3 may be preferable in terms of 6-year AIDS-free survival compared to immediate treatment initiation, at least under strategies IV, V, VI and VII.

For illustration and focussing on the time-points of 3 and 6 years, Tables 4.17 and 4.18 show the estimated AIDS-free survival at those times, respectively, under the different weighting strategies and estimation approaches, for the four regimes given by $x = 200$, 350, 500 and immediate treatment initiation.

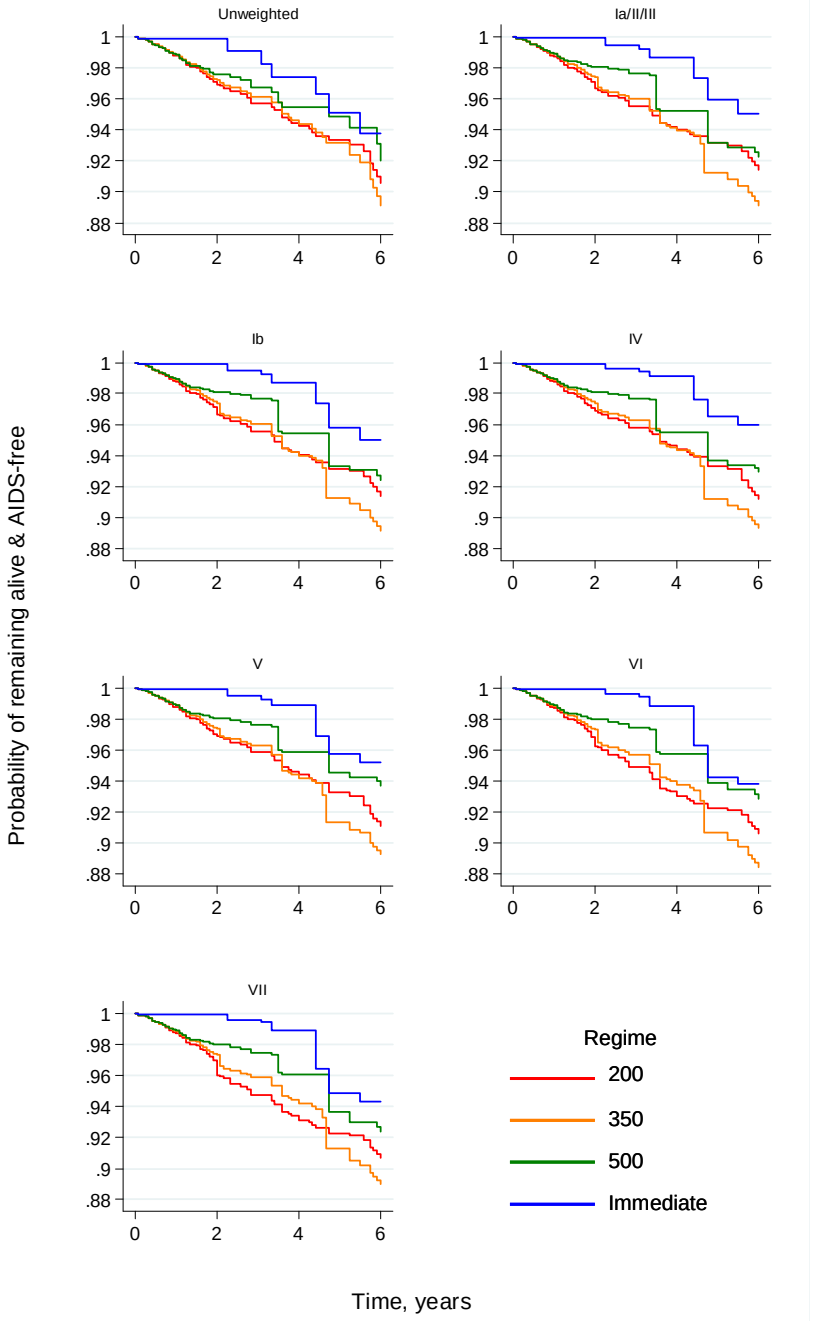


Figure 4.23: Application to CASCADE: probability of remaining alive & AIDS-free to 6 years, estimated using the raw Kaplan-Meier approach, under the different weight estimation strategies (and with no weighting), for the four regimes given by $x = 200$, 350 and 500, and immediate treatment initiation.

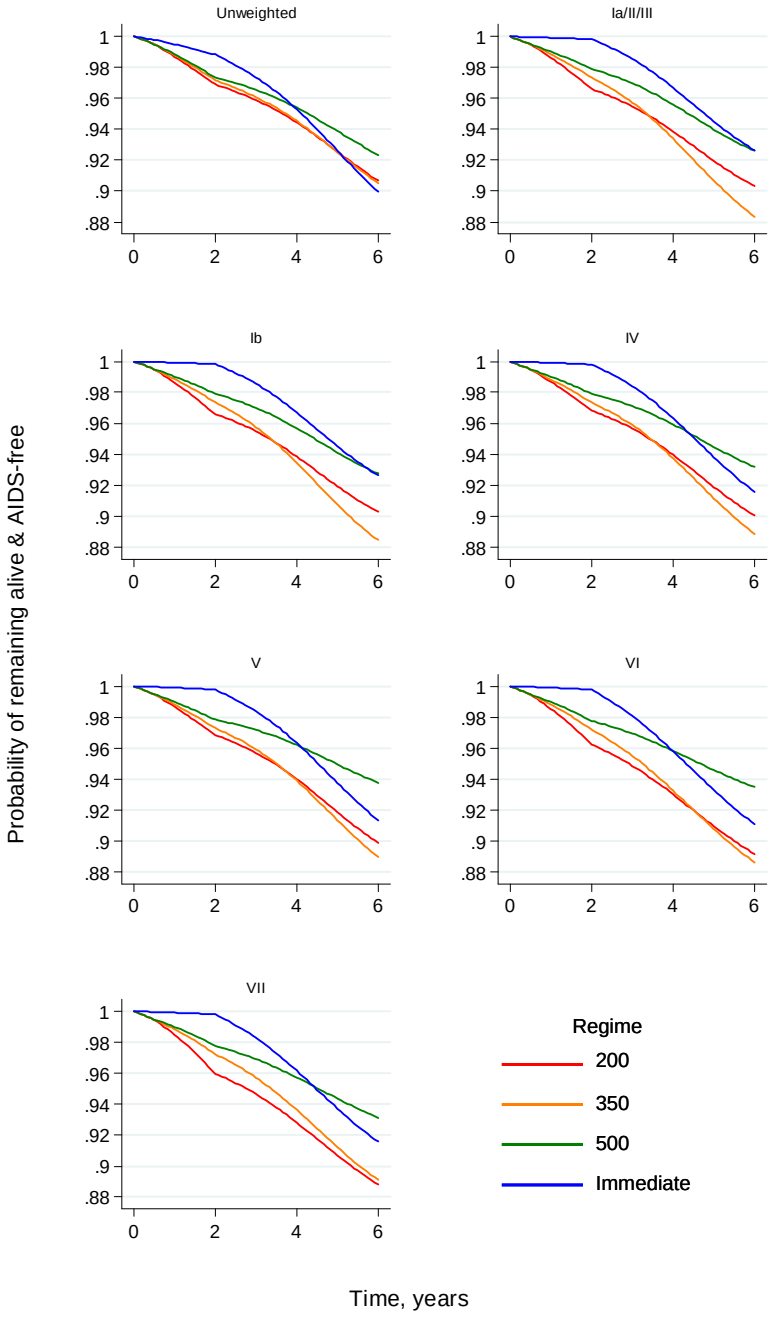


Figure 4.24: Application to CASCADE: probability of remaining alive & AIDS-free to 6 years, estimated using the pooled logistic regression model approach, under the different weight estimation strategies (and with no weighting), for the four regimes given by $x = 200$, 350 and 500, and immediate treatment initiation.

Approach	Regime			
Strategy	200	350	500	Imm
Raw KM				
Unweighted	0.957 (0.946,0.967)	0.961 (0.949,0.971)	0.968 (0.955,0.978)	0.991 (0.973,1.000)
Ia/II/III	0.956 (0.942,0.966)	0.960 (0.942,0.973)	0.977 (0.968,0.983)	0.995 (0.984,1.000)
Ib	0.955 (0.942,0.966)	0.960 (0.942,0.973)	0.977 (0.968,0.984)	0.995 (0.984,1.000)
IV	0.958 (0.948,0.967)	0.963 (0.951,0.974)	0.977 (0.969,0.984)	0.996 (0.989,1.000)
V	0.959 (0.949,0.968)	0.963 (0.950,0.973)	0.976 (0.967,0.983)	0.995 (0.985,1.000)
VI	0.949 (0.931,0.963)	0.957 (0.934,0.972)	0.975 (0.965,0.982)	0.997 (0.990,1.000)
VII	0.947 (0.923,0.964)	0.959 (0.939,0.973)	0.975 (0.965,0.982)	0.996 (0.989,1.000)
Smoothed KM				
Unweighted	0.957 (0.946,0.967)	0.961 (0.950,0.971)	0.968 (0.955,0.978)	0.991 (0.973,1.000)
Ia/II/III	0.956 (0.942,0.966)	0.960 (0.942,0.973)	0.977 (0.968,0.983)	0.995 (0.984,1.000)
Ib	0.955 (0.942,0.966)	0.960 (0.942,0.973)	0.977 (0.968,0.983)	0.995 (0.984,1.000)
IV	0.958 (0.948,0.967)	0.963 (0.951,0.973)	0.977 (0.969,0.984)	0.996 (0.989,1.000)
V	0.959 (0.949,0.968)	0.963 (0.950,0.973)	0.976 (0.967,0.983)	0.995 (0.985,1.000)
VI	0.949 (0.931,0.963)	0.957 (0.933,0.972)	0.975 (0.965,0.982)	0.997 (0.990,1.000)
VII	0.947 (0.923,0.964)	0.959 (0.938,0.973)	0.975 (0.965,0.982)	0.996 (0.989,1.000)
Pooled logistic				
Unweighted	0.959 (0.949,0.967)	0.961 (0.950,0.970)	0.966 (0.954,0.974)	0.974 (0.957,0.988)
Ia/II/III	0.955 (0.942,0.965)	0.957 (0.942,0.968)	0.970 (0.957,0.980)	0.985 (0.974,0.994)
Ib	0.955 (0.941,0.965)	0.958 (0.943,0.969)	0.970 (0.958,0.980)	0.986 (0.974,0.994)
IV	0.957 (0.946,0.965)	0.959 (0.946,0.968)	0.971 (0.957,0.981)	0.984 (0.973,0.994)
V	0.957 (0.946,0.966)	0.959 (0.946,0.968)	0.972 (0.961,0.980)	0.984 (0.972,0.993)
VI	0.949 (0.931,0.962)	0.955 (0.940,0.967)	0.970 (0.958,0.979)	0.981 (0.964,0.992)
VII	0.947 (0.921,0.963)	0.957 (0.943,0.968)	0.969 (0.957,0.978)	0.983 (0.967,0.993)

Table 4.17: Application to CASCADE: AIDS-free survival at **3 years**, estimated by the three approaches of raw Kaplan-Meier, smoothed Kaplan-Meier or pooled logistic regression, under the different weight estimation strategies (and with no weighting), for the four regimes given by $x = 200$, 350 and 500, and immediate treatment initiation. Values in brackets are 95% bootstrap confidence intervals. Imm=immediate treatment initiation regime.

Approach	Regime			
Strategy	200	350	500	Imm
Raw KM				
Unweighted	0.906 (0.880,0.931)	0.891 (0.858,0.921)	0.920 (0.881,0.957)	0.938 (0.882,0.980)
Ia/II/III	0.914 (0.889,0.936)	0.891 (0.842,0.932)	0.923 (0.871,0.966)	0.951 (0.896,0.989)
Ib	0.914 (0.888,0.936)	0.891 (0.842,0.932)	0.924 (0.875,0.966)	0.950 (0.894,0.989)
IV	0.912 (0.885,0.937)	0.893 (0.842,0.934)	0.929 (0.878,0.971)	0.960 (0.909,0.992)
V	0.911 (0.884,0.936)	0.893 (0.845,0.931)	0.937 (0.899,0.969)	0.952 (0.891,0.990)
VI	0.906 (0.881,0.931)	0.885 (0.835,0.926)	0.929 (0.882,0.967)	0.938 (0.858,0.991)
VII	0.907 (0.877,0.930)	0.890 (0.839,0.930)	0.924 (0.869,0.967)	0.943 (0.870,0.991)
Smoothed KM				
Unweighted	0.906 (0.880,0.931)	0.890 (0.857,0.921)	0.920 (0.881,0.957)	0.938 (0.882,0.980)
Ia/II/III	0.914 (0.889,0.936)	0.885 (0.837,0.924)	0.923 (0.870,0.966)	0.951 (0.896,0.989)
Ib	0.914 (0.888,0.936)	0.886 (0.837,0.924)	0.924 (0.873,0.966)	0.950 (0.894,0.989)
IV	0.912 (0.885,0.937)	0.883 (0.839,0.926)	0.929 (0.878,0.971)	0.960 (0.909,0.992)
V	0.911 (0.884,0.936)	0.886 (0.844,0.924)	0.937 (0.898,0.969)	0.952 (0.891,0.990)
VI	0.906 (0.881,0.931)	0.881 (0.833,0.921)	0.929 (0.882,0.967)	0.938 (0.858,0.991)
VII	0.907 (0.877,0.930)	0.886 (0.837,0.923)	0.924 (0.867,0.967)	0.943 (0.870,0.991)
Pooled logistic				
Unweighted	0.907 (0.883,0.926)	0.905 (0.878,0.927)	0.924 (0.893,0.947)	0.900 (0.846,0.946)
Ia/II/III	0.903 (0.873,0.927)	0.884 (0.842,0.919)	0.926 (0.872,0.962)	0.926 (0.877,0.967)
Ib	0.903 (0.872,0.927)	0.885 (0.843,0.919)	0.928 (0.875,0.963)	0.927 (0.876,0.967)
IV	0.900 (0.870,0.925)	0.889 (0.845,0.922)	0.932 (0.877,0.967)	0.916 (0.860,0.970)
V	0.899 (0.866,0.924)	0.890 (0.851,0.923)	0.938 (0.902,0.964)	0.914 (0.854,0.967)
VI	0.892 (0.852,0.925)	0.886 (0.847,0.921)	0.935 (0.892,0.964)	0.911 (0.835,0.965)
VII	0.888 (0.845,0.922)	0.891 (0.850,0.925)	0.931 (0.882,0.964)	0.916 (0.845,0.967)

Table 4.18: Application to CASCADE: AIDS-free survival at **6 years**, estimated by the three approaches of raw Kaplan-Meier, smoothed Kaplan-Meier or pooled logistic regression, under the different weight estimation strategies (and with no weighting), for the four regimes given by $x = 200$, 350 and 500, and immediate treatment initiation. Values in brackets are 95% bootstrap confidence intervals. Imm=immediate treatment initiation regime.

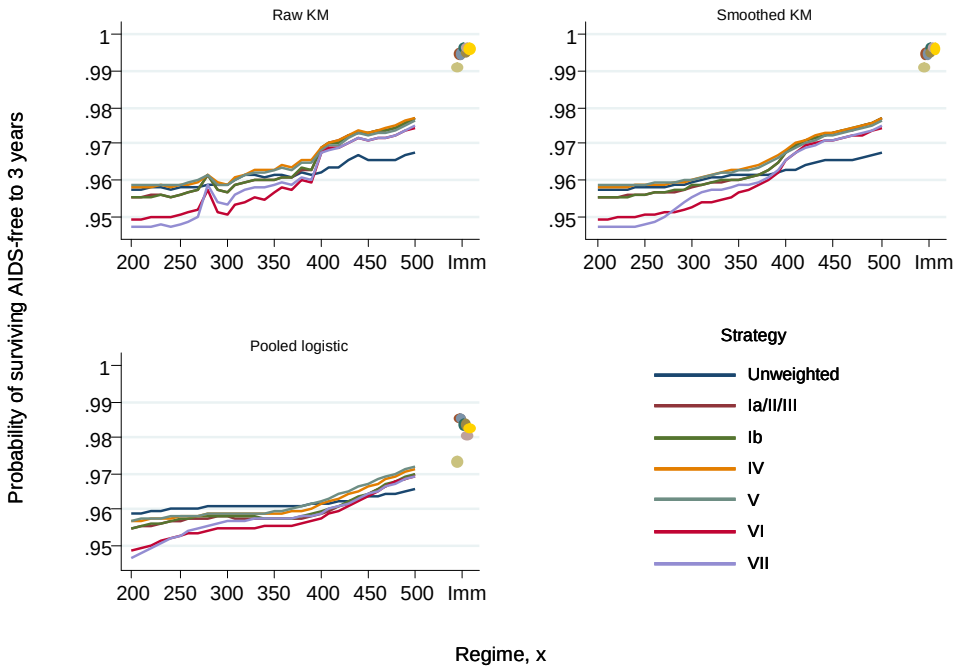


Figure 4.25: Application to CASCADE: AIDS-free survival at **3 years** by regime, estimated by the three approaches of raw Kaplan-Meier, smoothed Kaplan-Meier or pooled logistic regression, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime. The estimates for the immediate treatment initiation regime are staggered to aid clarity.

Optimal regimes at 3 and 6 years

We now consider the whole range of the regimes, to determine the optimal regime as defined by 3- and 6-year AIDS-free survival.

Comparing all the treatment weighting strategies and estimation approaches, it is clear that the optimal regime with respect to 3-year AIDS-free survival is immediate treatment initiation (Figure 4.25, and with 95% confidence intervals in Figures 4.26, 4.27 and 4.28 for the raw Kaplan-Meier, smoothed Kaplan-Meier and pooled logistic regression model approaches, respectively). The curves were broadly similar across all the different strategies and approaches, although the unweighted curves tended to underestimate somewhat the AIDS-free survival at regimes given by higher x.

Considering the 6-year AIDS-free survival probabilities, we saw quite different shapes in the AIDS-free survival by regime curves (Figure 4.29). Contrary to what we might expect based on the realistic simulation study (section 4.3) and the known benefits of treatment at CD4 counts ≤ 350 cells/mm^3, there appeared to be a trough at regimes given by intermediate x (300 to 400). Similarly to the 3-year AIDS-free survival estimation, immediate treatment initiation appeared

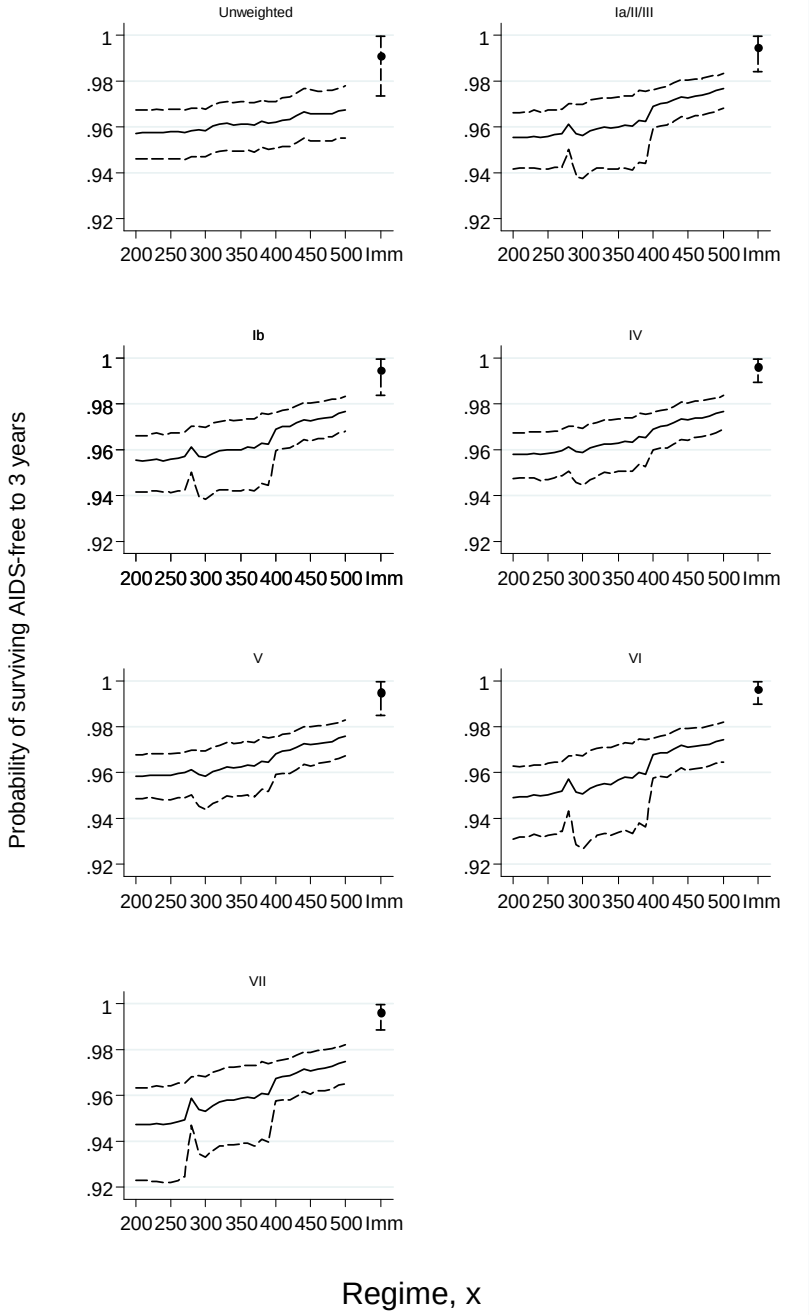


Figure 4.26: Application to CASCADE: AIDS-free survival at **3 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **raw Kaplan-Meier approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

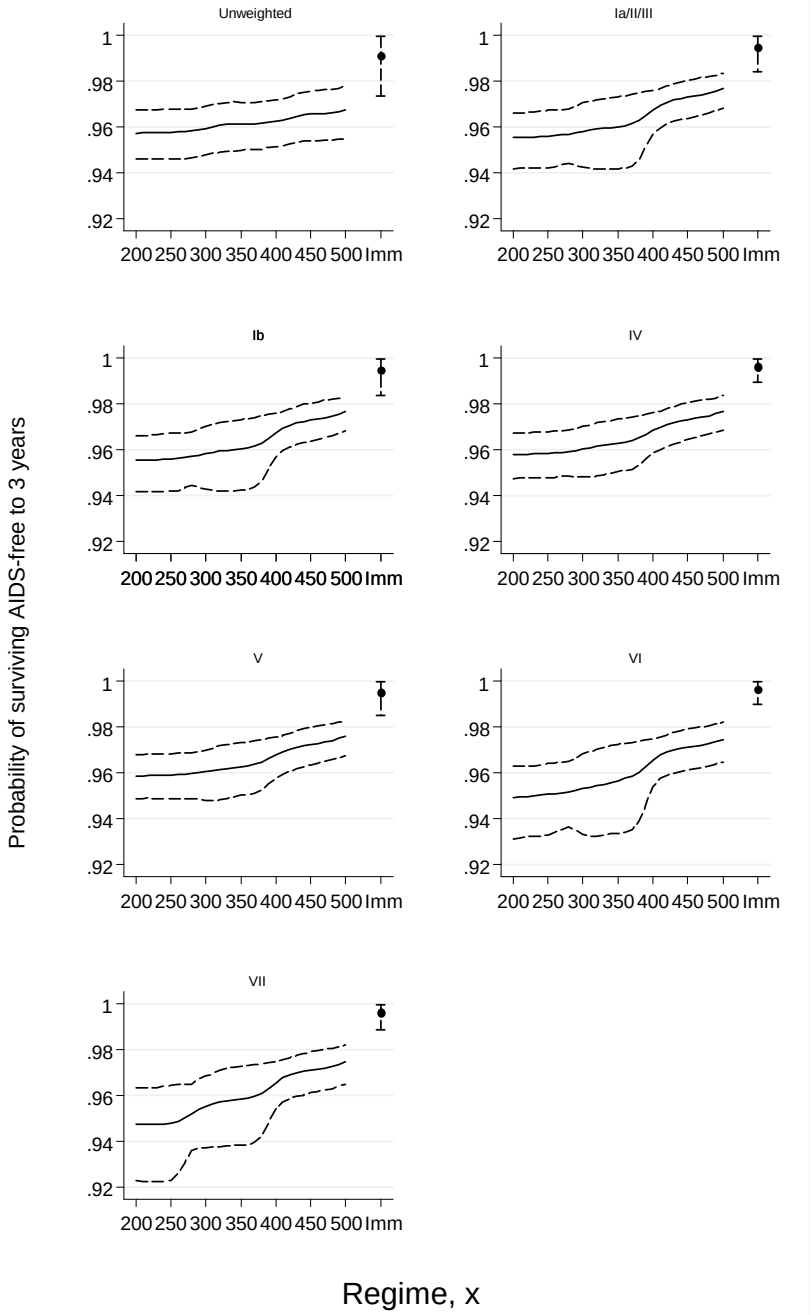


Figure 4.27: Application to CASCADE: AIDS-free survival at **3 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **smoothed Kaplan-Meier approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

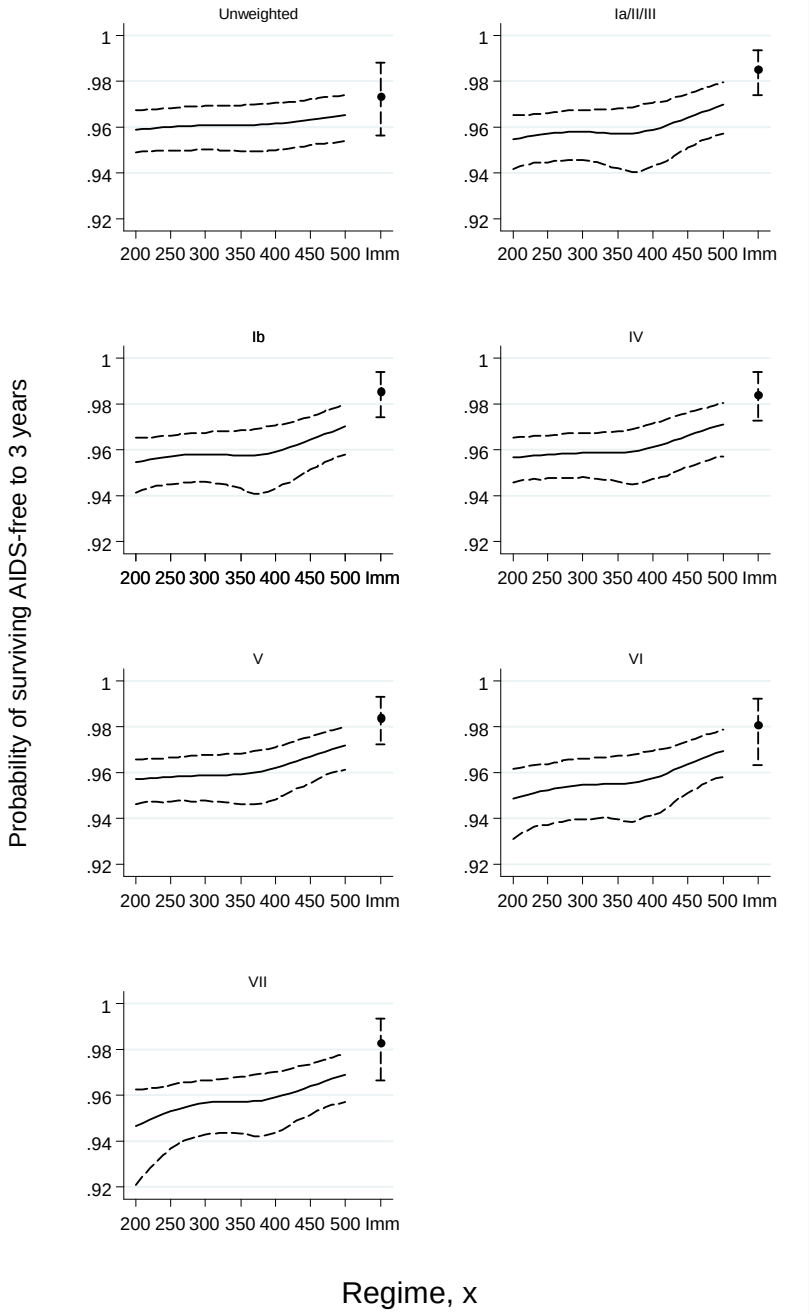


Figure 4.28: Application to CASCADE: AIDS-free survival at **3 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **pooled logistic regression model approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

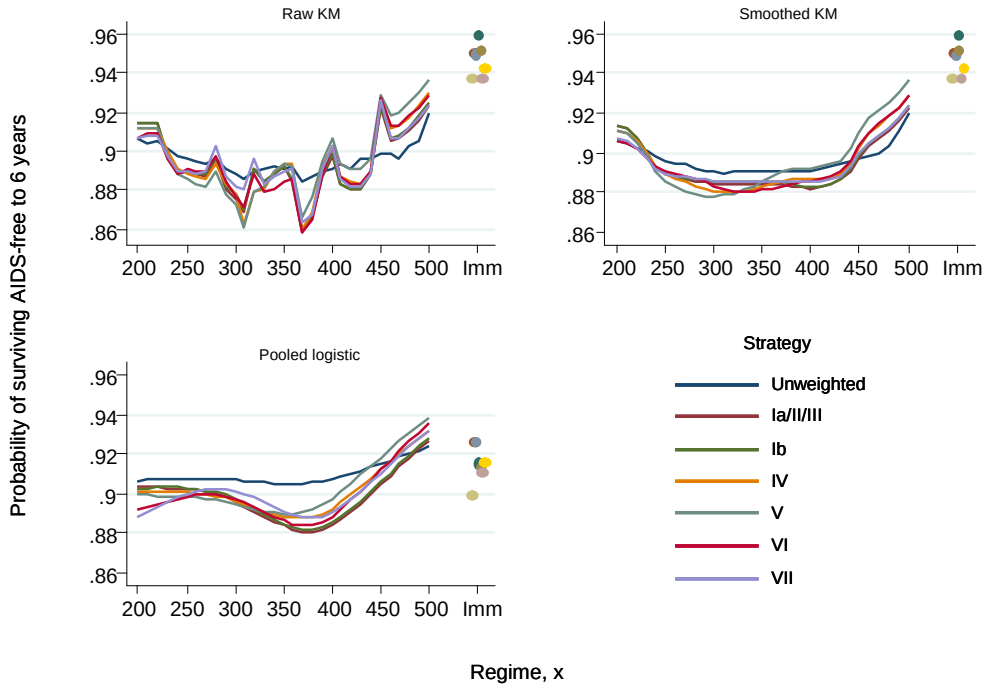


Figure 4.29: Application to CASCADE: AIDS-free survival at **6 years** by regime, estimated by the three approaches of raw Kaplan-Meier, smoothed Kaplan-Meier or pooled logistic regression, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime. The estimates for the immediate treatment initiation regime are staggered to aid clarity.

to be the optimal choice under the raw and smoothed Kaplan-Meier approaches, although the confidence intervals overlapped considerably with those of the other regimes (Figures 4.30 and 4.31). However, under the pooled logistic regression approach, the point estimates for the 6-year AIDS-free survival tended to be lower on the immediate treatment initiation regime, compared to for example the regime given by $x = 500$; although, once again the confidence intervals overlapped considerably (Figure 4.32). The greater uncertainty is due to less uncensored follow-up to 6 years.

The optimal regimes based on these results are shown in Table 4.19, along with the minimum acceptable regimes. Across most scenarios, the optimal regime was immediate treatment initiation, although, when considering 6-year AIDS-free survival, the pooled logistic regression model approach yielded optimal regimes of $x = 500$ across all but one weighting strategy. Of note, the raw and smoothed Kaplan-Meier estimated optimal regimes were unlikely be different when the optimal regime for the former was immediate treatment initiation, since the smoothing was only performed over the range $x = 200$ to 500, and the immediate treatment initiation estimates were left unchanged.

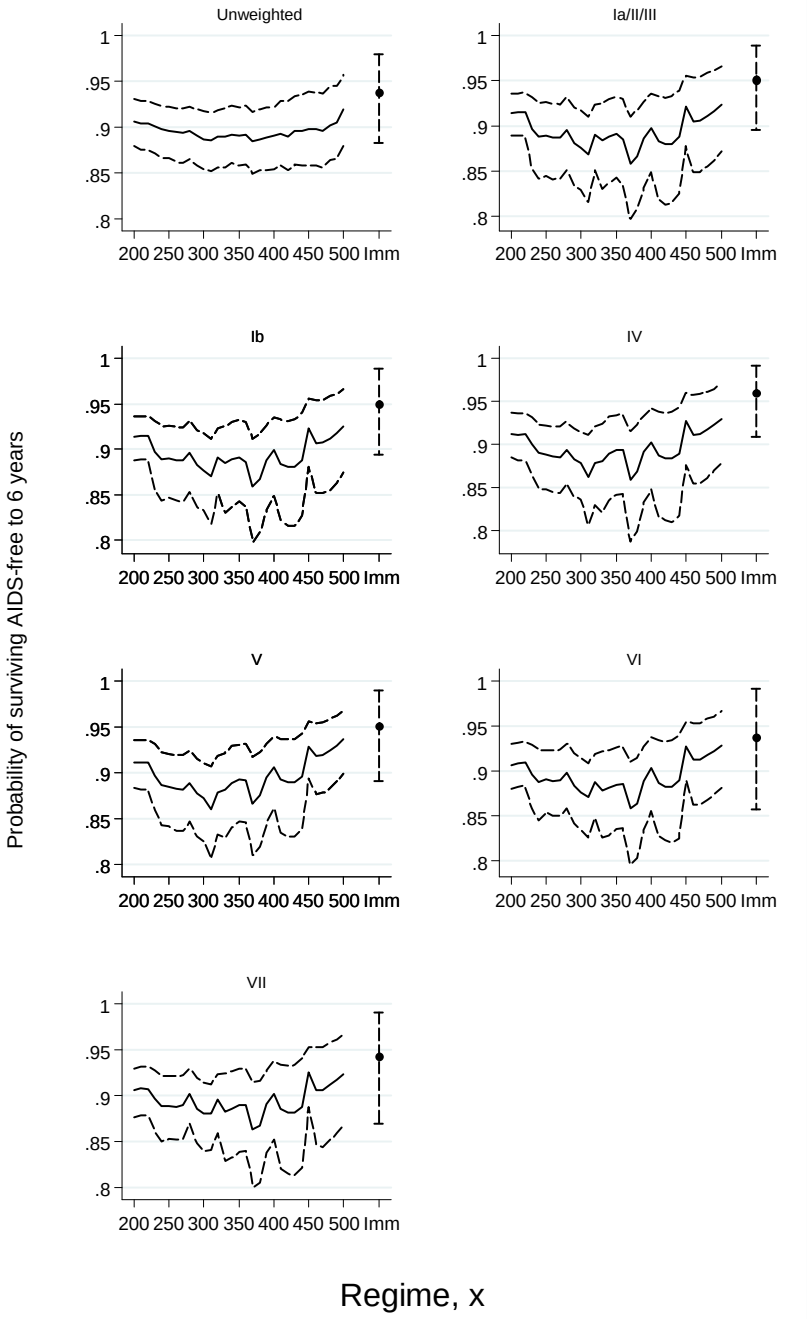


Figure 4.30: Application to CASCADE: AIDS-free survival at **6 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **raw Kaplan-Meier approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

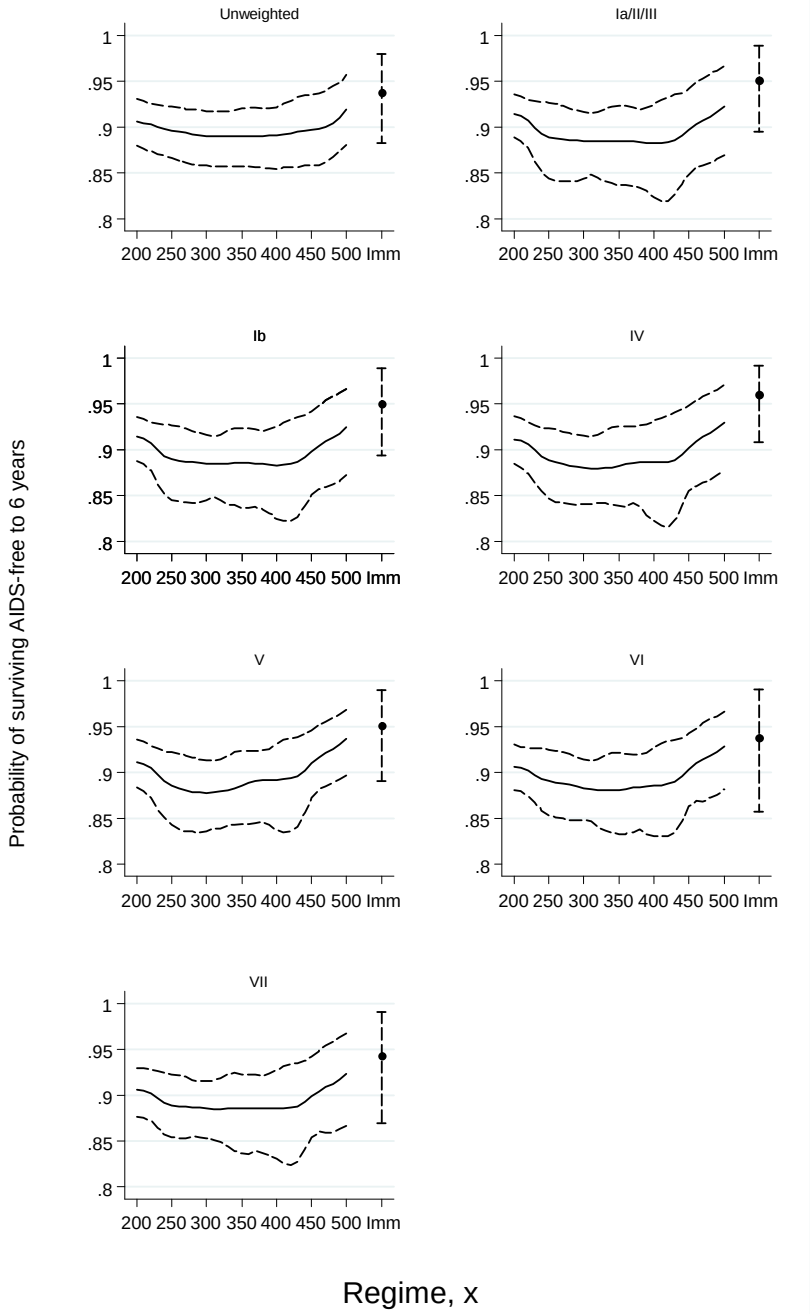


Figure 4.31: Application to CASCADE: AIDS-free survival at **6 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **smoothed Kaplan-Meier approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

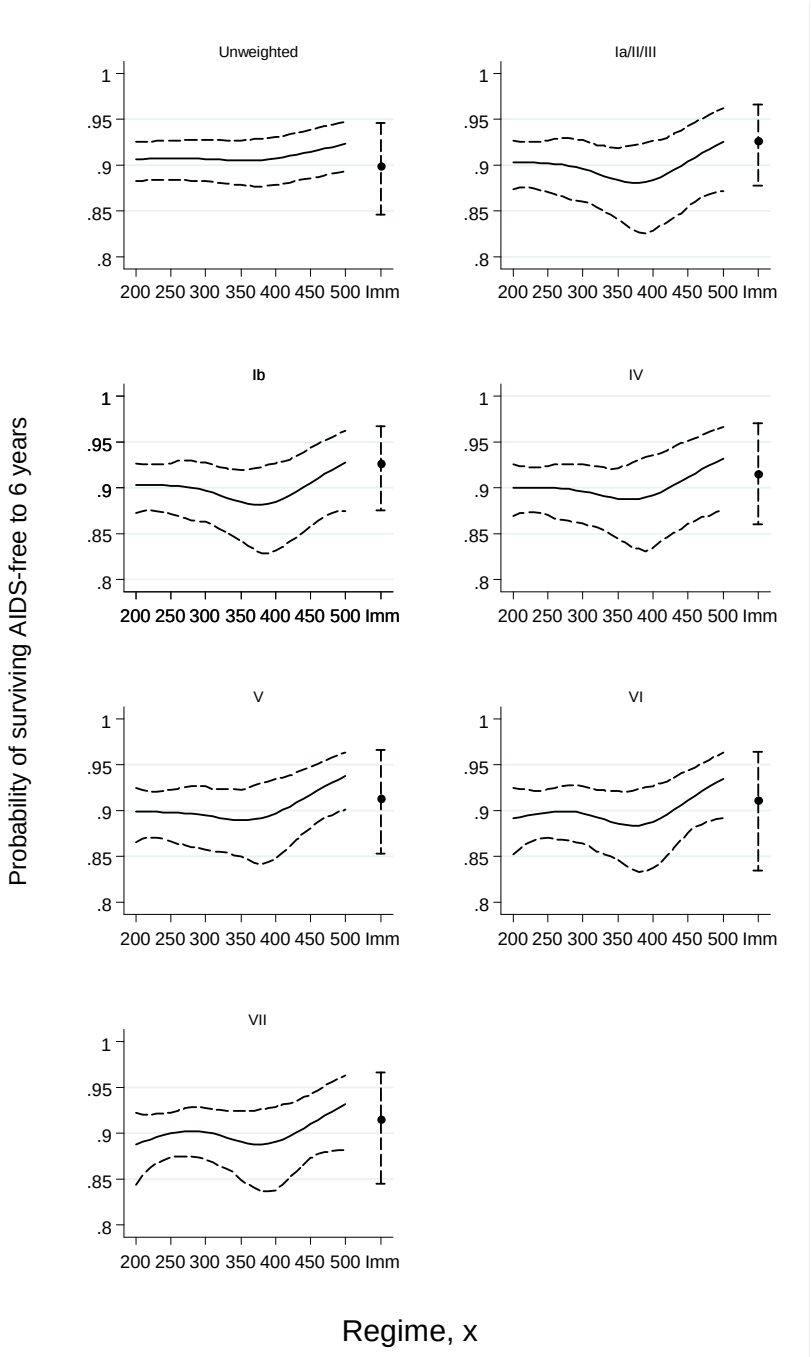


Figure 4.32: Application to CASCADE: AIDS-free survival at **6 years** by regime, with 95% bootstrap confidence intervals, as estimated by the **pooled logistic regression approach**, across the different weight estimation strategies (and under no weighting). Imm=immediate treatment initation regime.

Time-point / Strategy	Raw KM	Smoothed KM	Pooled logistic
3 years			
Unweighted	Imm (0.991)	Imm (0.991)	Imm (0.974)
	-	-	-
Ia/II/III	Imm (0.995)	Imm (0.995)	Imm (0.985)
	-	-	-
Ib	Imm (0.995)	Imm (0.995)	Imm (0.986)
	-	-	-
IV	Imm (0.996)	Imm (0.996)	Imm (0.984)
	-	-	-
V	Imm (0.995)	Imm (0.995)	Imm (0.984)
	-	-	-
VI	Imm (0.997)	Imm (0.997)	Imm (0.981)
	-	-	-
VII	Imm (0.996)	Imm (0.996)	Imm (0.983)
	-	-	-
6 years			
Unweighted	Imm (0.938)	Imm (0.938)	500 (0.924)
	-	-	480 (0.920)
Ia/II/III	Imm (0.951)	Imm (0.951)	Imm (0.926)
	-	-	490 (0.922)
Ib	Imm (0.950)	Imm (0.950)	500 (0.928)
	-	-	490 (0.924)
IV	Imm (0.960)	Imm (0.960)	500 (0.932)
	-	-	490 (0.928)
V	Imm (0.952)	Imm (0.952)	500 (0.938)
	-	-	490 (0.934)
VI	Imm (0.938)	Imm (0.938)	500 (0.935)
	-	-	490 (0.931)
VII	Imm (0.943)	Imm (0.943)	500 (0.931)
	-	-	490 (0.928)

Table 4.19: Application to CASCADE: optimal and minimum acceptable regimes with respect to 3- and 6-year AIDs free survival. The first row for each time-point/strategy shows the optimal regime, and the second shows the minimum acceptable regime (if different to the optimal regime). Recall, the minimum acceptable regime is defined as that given by lowest x which has < 0.005 poorer AIDS-free survival compared to under the optimal regime. Values in brackets are the estimated probabilities of surviving AIDS-free to that time-point under that regime. Imm=immediate treatment initiation regime. KM=Kaplan-Meier.

Minimum acceptable regimes at 3 and 6 years

Due to the relatively large observed higher 3- and 6-year AIDS-free survival under the immediate treatment initiation regime, there were typically no other regimes which met the stringent criterion for acceptability (no worse than 0.5% poorer AIDS-free survival). However, where the estimated optimal regime was given by $x = 500$ (when considering 6-year AIDS-free survival under the pooled logistic regression approach), the regime given by $x = 490$ met this criterion for acceptability (Table 4.19).

4.6 Discussion

4.6.1 Methodological findings

In this chapter, we have explored the optimisation of pre-defined treatment regimes using dynamic MSMs, via the clinical question of when to initiate treatment with respect to CD4 count in HIV-infected persons. As outlined in section 4.1.1, these methods are best approached via the concept of the RCT which one would ideally conduct (Cain et al., 2010; Hernán et al., 2008). This enables the correct framing of the question to be addressed using the observational data. We have demonstrated via simulations of large RCTs and observational cohorts that, with sufficient data and under the standard assumptions (section 1.2.4), these methods yield the correct answers. However, in our clinical example where there are large natural fluctuations and measurement error in the biomarker CD4 count which defines the dynamic treatment regimes, and where the event (AIDS or death) rates are low, a great deal of uncertainty is present. This was evident in our large simulated observational studies of 100,000 participants, and even to some extent in the simulated RCTs of 31 million individuals. We have reinforced the current view that large collaborative clinical cohorts are required to answer such causal questions.

A related issue encountered in our simulation study based on a realistic scenario (simulation study 1) was that the outcome used to determine the optimum regime, that is, 10-year AIDS-free survival, was broadly constant at high values of CD4 count, the time-dependent covariate used to define the regimes x. This is encouraging in terms of support for current HIV treatment guidelines, which typically recommend treatment initiation at CD4 counts of around 350 cells/mm^3, and also reassuring for patients and clinicians, in that the optimal timing of treatment initiation may not be critical before the CD4 count drops to around that threshold. However, in terms of the application of these methods, the lack of a clear "peak" and hence optimal regime means that, under any single analysis, the method may yield an estimate quite

"far" from the optimal regime, as illustrated by the individual simulations (Figures 4.18 and 4.19). Consideration of the shape of the AIDS-free survival curve by regime as estimated by the raw Kaplan-Meier approach helped our understanding of the data, and indeed when the curve was so "flat", we found that this approach outperformed the pooled logistic regression approach. Our second simulation study illustrated that with a clearer "peak" (and a greater number of patients), the methods perform better, and in this case the pooled logistic regression approach outperformed the others. Therefore, in general, we would recommend that both the raw Kaplan-Meier and pooled logistic regression approaches are applied, and urge caution in the interpretation of optimal regimes, which should be done with regard to the shape of the optimal criterion-by-regime curve and recognising that the precision may be low.

If broadly constant AIDS-free survival rates at regimes defined by higher x were observed in real data, then clinically this "flatness" could be interpreted in different ways: at the expense of a small increase in AIDS or death, treatment could be delayed a little past the optimal regime, perhaps to the minimum acceptable regime, to preserve resources which could perhaps be more beneficially used in other areas, for example HIV prevention, or simply to preserve treatment-free time for the patient, which might have benefits in terms of toxicity and preserving future treatment options (delaying failure or resistance). Conversely, encouragement of treatment initiation slightly earlier than the optimal regime could have population advantages by reducing transmission risk; recently there has been a stronger interest and support in treatment as prevention (Cohen et al., 2011). As many HIV-infected persons worldwide who need treatment under the current treatment guidelines are not receiving it, earlier treatment initiation may be a luxury affordable only in high-income countries. Even then, timely treatment initiation is dependent on individuals presenting for care early in infection.

Of note, the issues discussed above may apply to other disease areas, if those areas have similar measurement error associated with the time-dependent covariate used to define the dynamic treatment regimes, low event rates, and/or ranges of the dynamic regimes across which the outcome is broadly constant.

The recent extension of these methods to incorporate grace periods is one attempt to address the more limited data typically available (Cain et al., 2010). This enables use of a greater number of observed treatment initiations, potentially resulting in less censoring and hence greater power. However, these extensions have been rarely applied in practice (Cain et al., 2010; Young et al., 2011), and their implications in realistic scenarios have not previously been explored. CD4 count observation frequencies and grace periods are clearly interrelated, since less-frequently

observed CD4 counts or longer grace periods may both result in CD4 counts dropping to lower levels while awaiting treatment initiation, and such lower levels are associated with higher risk of AIDS or death. We focussed initially on monthly observed CD4 counts as a first step, for pedagogic purposes to develop understanding of the methods and ensure they were working as anticipated before progressing to (perhaps more realistically) less-frequently observed CD4 counts. The choice of 3-monthly CD4 counts was driven by the typical visit schedule followed in resource-rich settings, and as observed in our subset of CASCADE patients. In resource-limited settings, 6-monthly measurement of CD4 counts may be more common. We found that in populations with less frequent measurement of CD4 counts, or faster treatment-naïve CD4 decline, the optimal CD4 count for treatment initiation was higher and the AIDS-free survival rates were lower, even on the optimal regime, as we might expect.

The treatment-naïve CD4 decline and CD4 count measurement frequency will in general be fixed in a given population; while the CD4 count frequency could be reduced by ignoring those recorded at intermittent time-points which do not fit into that schedule, this would surely only be for exploratory purposes since would typically reduce precision. In contrast, the grace period may be varied for analysis, and in the simulation of the observational studies, we considered the grace periods as a step in the analysis only, not the data generation. The minimum length of the grace period may only be limited by the observed data (for example, if the grace period was set as 1 day then there may be no patients compliant with any regime). There is no upper restriction to the length of the grace period, though of course the results must be interpreted accordingly; very lengthy grace periods may only serve to blur the distinction between regimes and are unlikely to be of much clinical relevance. We firstly used no grace period ($m = 1$), and extensions to 3- and 6-month grace periods corresponding to the CD4 count observation frequencies considered, and as used by other researchers (Cain et al., 2010; HIV-CAUSAL collaboration, 2011; Kitahata et al., 2009; Shepherd et al., 2010). As one may expect, we found, via the large realistic RCT simulations, that lengthening the grace period typically led to poorer 10-year AIDS-free survival. Although similar AIDS-free survival rates at 10 years could be achieved with grace periods up to 6 months, compared to no grace period (with CD4 counts observed 1- or 3-monthly), the optimal regime had to be raised accordingly.

There is a different aspect to grace periods which must be considered, aside from being a way of potentially reducing censoring of otherwise non-compliant treatment initiations: to permit a grace period is to ask a different question. Two main approaches have previously been defined (Cain et al., 2010); under the first approach, the regimes are defined by "do not initiate

treatment before the CD4 count is $<x$ cells/mm^3, and do initiate exactly m months after the CD4 count first drops below x cells/mm^3 if treatment has not already been initiated in the first $m-1$ months of the grace period". Under the second approach, regimes are defined as "initiate treatment within m months after the CD4 count first drops before x cells/mm^3, such that there is a uniform probability of starting in each of the months $1, 2, ..., m$". Different methods of estimation must be applied to these two approaches. While previous applications have focussed on the first approach (Cain et al., 2010; HIV-CAUSAL collaboration, 2011; Shepherd et al., 2010), this involves upweighting the potentially small and unrepresentative subset of patients observed to initiate in the last interval of the grace period. In addition, we have seen in the CASCADE data that the treatment initiation pattern across the grace period may differ by regime, with patients less likely to delay until later in the grace period if already at low CD4 counts. Therefore, we applied the second approach.

Of course, either approach should strictly then be interpreted in the appropriate context, both of which may be somewhat baffling to health care providers and patients. Both approaches may perhaps be loosely interpreted by clinicians who first observe a patient's CD4 count to drop below the given threshold x as "start treatment within the next m months". In fact, clinicians and patients alike may prefer this extra time in order to prepare for the initiation of treatment which will be life-long. Alternatively, the grace period may be ignored entirely and the dynamic regime simply interpreted as treatment initiation when CD4 count is first observed to be below the given threshold x. Public health policy makers may be keen to know the effect of such an interpretation. Via the simulation of small observational studies, we have investigated the bias-variance trade-off in permitting a grace period for the purposes of potentially increasing efficiency, at the risk of inducing bias for the inference of interest under no grace period. We found that under 3-monthly observed CD4 counts, permitting a 3-month grace period was beneficial over no grace period, in terms of increasing precision slightly with minimal penalty in terms of bias, but that the bias induced by extending to a 6-month grace period outweighed the gain in precision. We therefore recommend that a 3-month grace period be used in observational studies in similar resource-rich settings, which are likely to have a comparable CD4 observation frequency, but further research may be required for other settings.

4.6.2 Clinical findings

In light of the results from the simulation studies, we permitted a 3-month grace period to apply these methods to our CASCADE population, and consider that the resulting optimal regime

may be interpreted in the absence of a grace period, rather than precisely as per the somewhat complicated definition above.

A large proportion of patients were observed to initiate treatment immediately; this may in part be due to the nature of the study entry (all patients had a clinic visit). We were able to define a treatment regime to capture these treatment initiations, which otherwise would have been censored, since immediate treatment initiation has some meaning with respect to our study entry criteria (namely $\geq$ 500 cells/mm^3, at the first CD4 count within 1-5 years after seroconversion). We found under most scenarios that immediate treatment initiation was preferable, although perhaps delaying until CD4 count was observed to drop below 500 cells/mm^3 might offer some benefit. We cannot rule out the possibility that the subset of patients who initiated treatment immediately may be somewhat different to the remainder of the patients, although we did control for a number of confounders via the weights. While the subset of CASCADE patients included in these analyses were a selected subset, they are likely to constitute the population in whom the choice of when to start treatment uniquely applies; often patients who present later do so because of clinical symptoms and so in whom treatment is indicated. Of note, the pooled logistic regression approach predicted somewhat lower AIDS-free survival for the immediate treatment regime, especially at 6 years. The reasons for this are not clear. In addition, we observed a slight increase in the estimated 6-year AIDS-free survival for the regimes given by very low x (close to $x = 200$), compared to regimes around $x = 350$ to 400; the reason for this is not clear but there may be some residual confounding. Further, the results from the simulation studies illustrate the inherent uncertainties in these data. The shape of the curves derived from the pooled logistic regression models, when compared to those from the raw Kaplan-Meier approach, reassured us that the parameterisations we chose (for example, four knot spline for regime) were adequate.

Reassuringly, we found broadly consistent results across the different weighting strategies as determined in chapter 2, all suggesting clearer distinction in terms of AIDS-free survival between the regimes compared to without weighting. Of note, the uncertainty introduced in applying the estimated weights was visible in the somewhat "jagged" appearance of the weighted AIDS-free survival by regime curves, compared to the unweighted ones, particularly at 6 years. For brevity in this chapter we did not compare the effects of different weight truncations, but we know from chapter 2 that this will in general yield different results.

Our findings are broadly consistent with previous studies, in that early treatment initiation may be beneficial but that the differences in AIDS-free survival or overall survival are very small

at regimes given by high CD4 count. Due to this issue, combined with large measurement error in CD4 count and low event rates, we have demonstrated via the simulation studies that it is quite plausible for two studies with the order of thousands of participants to yield somewhat different estimated optimal regimes, even if the underlying distributions are the same. As an example, the HIV-CAUSAL collaboration (2011) allowed a 6-month grace period and estimated the optimal regime to be given by $x = 500$ in the set they considered ($x = 200$ to 500), but emphasised that the overall survival was very similar for regimes given by $x = 300$ to 500.

As outlined in section 4.2.4, it is possible to look at interactions of regime with baseline covariates, in order to tailor optimal treatment regimes to specific patients, but given the lack of power in our subset of CASCADE patients, it was not possible for us to address this.

4.6.3 Limitations

We considered regimes defined by 10 cells/mm^3 categories of CD4 count. This could lead to censoring of intermittent treatment initiations. For example, if a patient's nadir observed CD4 count to date was 489 cells/mm^3 and treatment was initiated in response to a subsequent observed CD4 count of 483 cells/mm^3, then this treatment initiation would be censored under all regimes given by $x = 200, 210..., 500$. In our CASCADE population, this occurred in 32 patients, but 12 of those 32 treatment initiations were permitted when allowing a 3-month grace period. The alternative would be to use a finer categorisation of CD4 count, the most extreme being defining regimes by 1 cell/mm^3 categories. This would not only be extremely computationally challenging, but the clinical relevance is questionable, given the known biological and measurement variation in CD4 count. Conversely, coarser categorisation of CD4 count could be applied, for example defining regimes by 150 cells/mm^3 categories, but this would result in the censoring of many observed treatment initiations. Therefore, the 10 cells/mm^3 categorisation was considered to be a good compromise (Cain et al., 2010).

One of the implications of the large measurement error incorporated into the simulations, and no doubt present in the CASCADE data, meant that large proportions of the observed treatment initiations were censored from all regimes due to initiation at a CD4 count above the nadir (lowest to date). If CD4 counts declined linearly while treatment-naïve (in the simulation studies, this means in the absence of Brownian motion and measurement error), this censoring would no longer occur. By permitting a grace period, we captured a greater number of the treatment initiations, although relatively large proportions were still censored. Alternative approaches, such as requiring two CD4 counts below the given threshold x for treatment initia-

tion, could perhaps help reduce the number of censored treatment initiations, and perhaps help mitigate to some extent the large measurement error, although would emulate to some extent what the grace periods are attempting to do. In addition, if confirmation of CD4 counts was not consistently performed in an observational study, then enforcing this in the analysis is unlikely to be beneficial. Given the large measurement errors evident in CD4 counts, such censoring of treatment initiations are inevitable. This was particularly visible in the RCT simulations, where patients often initiated treatment at random low observed CD4 counts while the true CD4 count (that is, incorporating Brownian motion but in the absence of measurement error) was much higher. However, this measurement error is likely to reflect what occurs in practice.

The simulation study models were based on previous modelling using CASCADE data. When treatment was initiated at high CD4 counts, the resulting mean slope from one year after initiation onwards was negative, due to the strong negative correlation indicated previously between CD4 count at treatment initiation and long-term slope thereafter. It may be that, in the data on which the previous modelling was performed, this correlation was driven by patients who were observed to initiate treatment early but subsequently stopped treatment (including, for example, in trials looking at short-course treatment in primary infection; SPARTAC Trial Investigators (2011)). Therefore, the overall decline in CD4 count from one year after treatment initiation observed in these data and hence incorporated into our simulation models may in fact be a consequence of those patients typically being off therapy subsequently. The implications of this are that we may have underestimated the benefit of early treatment initiation, with respect to CD4 count, assuming that treatment is continued once initiated. However, one could argue that this may mimic what would happen in practice, whereby patients feeling well may not be motivated to take their medications, or having to take treatment over such long periods of time may increase the cumulative risk of side effects, hence leading to poorer adherence. Of note, if a penalty for early treatment initiation had not been incorporated via this negative correlation, and CD4 counts increased continuously on treatment (or at least to some plateau), then it would always be optimal to initiate treatment immediately.

The determination of optimal treatment regimes is heavily dependent on time. When the regimes are defined by a biomarker which is on average monotonely decreasing, sufficient time must be allowed to pass for the biomarker to decrease and hence differences in the outcome emerge between the regimes. This is of particular importance when the patients enter the study with similar levels of the biomarker, as in both our simulation studies and analysis of CASCADE data. We considered follow-up to 10 years under the simulation studies, but the optimal regimes

may have been different if longer follow-up was considered. This is illustrated in Figure 4.8: if we had only considered up to 5 years, then the regime given by $x = 500$ would be preferable to that given by $x = 350$; this was reversed by 10 years. We also saw in our CASCADE population that different optimal regimes would be determined depending on whether 3- or 6- year follow-up was considered. We were only able to consider up to 6 years for defining the optimal regime, due to limited follow-up, but it may be that longer follow-up would indicate different optimal regimes. In addition, other metrics, for example a CD4- or quality of life-based metric (Robins et al., 2008; Shepherd et al., 2010), or restricted mean survival (Royston and Parmar, 2011), may well yield different optimal regimes.

4.6.4 Summary

We have investigated the impact of several aspects, perhaps most importantly grace periods, on the estimation of dynamic treatment regimes, and applied these methods to our CASCADE population. In our clinical setting, where CD4 counts were measured 3-monthly, permitting a 3-month grace period may offer efficiency benefits, with low bias, but lengthening to 6 months increased the bias substantially. In our population of CASCADE patients, immediate treatment initiation appeared to be most beneficial in terms of 6-year AIDS-free survival; with respect to the pre-specified regimes defined by CD4 count, treatment initiation when CD4 counts were first observed to drop < 500 cells/mm^3 was preferable delaying until CD4 counts were observed to drop further. In the next and final chapter, we discuss these results in relation to those from other chapters.

Chapter 5

Discussion

In this final chapter, we outline the main contributions of this research to the field of causal estimation, and in particular make comparisons across the three types of MSMs and draw some conclusions. We discuss some limitations and outline potential future work.

5.1 Construction of weights

Our first contribution to the application of MSMs for causal estimation is the development of a simple algorithm for the construction of the inverse probability of treatment weights. This process has been framed as a series of well-defined decisions, helping ensure transparency. This approach should enable future researchers to more clearly understand the steps involved and perhaps help identify reasons for any observed differences in estimated effects between studies. We have shown how a range of plausible strategies for constructing inverse probability weights may arise from these decisions. In our example, estimating the effect of treatment on time to AIDS or death in HIV-infected persons in CASCADE, these strategies consistently demonstrated a beneficial effect of treatment, although the point estimates and precision varied somewhat across the strategies. We recommend that researchers use a range of estimated weights to check the sensitivity of the results to their assumptions. Of course, other choices or strategies to those presented here are possible.

In addition, we have illustrated how a variable such as country or centre, across which broadly constant treatment effects may be expected, can be used in different ways. Firstly, separate treatment models, one for each country or centre, may be used to estimate the weights, although in our example we found this tended to be less efficient. Secondly, interactions between treatment and the country or centre covariate may be used to explore whether there is heterogeneity in the estimated treatment effect across different strata of that covariate. If so, this

could either be a true phenomenon, or may indicate that there remains residual confounding which has not been adequately captured by the weights.

5.2 Estimation of optimal dynamic treatment regimes

The second contribution of this work is related to the optimisation of dynamic treatment regimes using dynamic MSMs, and in particular jointly assessing the impact of grace periods (permitted delay for treatment initiation) and varying measurement frequencies, and evaluating the performance of these methods in realistically-sized observational studies. We recommend that both the (raw) Kaplan-Meier and pooled logistic regression model approaches are applied, and that the resulting estimated optimal dynamic treatment regimes are interpreted with respect to the shape of the outcome-by-regime curve and the precision.

Via the simulation of large realistic RCTs, we found that if CD4 counts are observed less frequently then the (true) optimal regime may be substantially higher, that is, given by earlier treatment initiation at higher CD4 counts. This has implications for the generalisability of results from both randomised trials and observational studies. For example, the findings from a randomised trial addressing the issue of when to start treatment with respect to CD4 count in a resource-rich setting, where CD4 counts are typically measured 3-monthly, may not be applicable to resource-limited settings, where CD4 counts are usually measured less frequently. Lengthening the grace period also indicated higher optimal regimes, but not to the same extent as CD4 count observation frequency. However, it is worth noting that, under the higher optimal regimes with CD4 count observation frequencies or grace periods of up to 6 months, the 10-year AIDS-free survival rates were similar to those under the optimal regimes with monthly observed CD4 counts and no grace period.

Via the simulation of corresponding realistically-sized observational studies ($n = 3000$), with CD4 counts observed 3-monthly, we found that permitting grace periods of up to 3 months in our clinical setting may offer benefits in terms of increased precision, at little expense of bias, for the estimation of the optimal dynamic treatment regime under no grace period, which may be easier to understand and implement in practice. However, for longer grace periods of 6 or 12 months, the bias induced outweighed the gain in precision. Across the different length grace periods considered, the efficiency gains were perhaps smaller than might have been anticipated.

5.3 Methodological comparison across the different MSMs

To our knowledge, this is the first time that standard, history-adjusted and dynamic MSMs have systematically been applied to the same data. Our third contribution is to examine differences between these approaches, and present a strategy and rationale for applying all three methods when interest lies in identifying optimal dynamic treatment regimes.

5.3.1 Weights

As highlighted in previous chapters, the principles of weight estimation are the same across the different types of MSM, although the weights finally applied are somewhat different. In particular, the weights used in the HAMSMs for the comparison of immediate versus no treatment may be considered inverse probability of censoring, rather than treatment, weights, since patients who initially deferred but subsequently initiated treatment are censored at treatment initiation and no longer contribute follow-up. A common cause of large weights under the standard MSMs is due to treatment initiations when the probability of treatment initiation is low, therefore, depending on the question asked, these large weights may no longer be used in the HAMSMs, hence potentially resulting in more stable weights, and perhaps more efficient estimation. Further, unlike standard MSMs, the stabilisation of the weights for the HAMSMs may be performed using time-updated (trial-baseline rather than true-baseline) covariates, potentially increasing efficiency.

The weight estimation for the dynamic MSMs with no grace period is also broadly similar to that applied for the standard MSMs. However, large weights arising from persons who persistently remain off treatment despite low CD4 counts will automatically no longer be incorporated, if there is no such pre-specified treatment regime under which that behaviour is permitted. Again, this may result in more stable weights and potentially more efficient estimation.

When incorporating a grace period in dynamic treatment regimes, adjustments must be made to the numerator of the weights, and different adjustments are required depending on the approach, related to different interpretations of the corresponding regimes. In order to avoid upweighting a potentially small and unrepresentative subset of patients who initiated in the last interval of the grace period, we applied an approach which assumes uniform treatment initiation across the grace period. Whilst this assumption may never exactly hold, moderate deviations from it are unlikely to have as large an impact in many applications as upweighting the small subset of people who initiated treatment in the last interval of the grace period. We

recommend investigating the observed distribution of treatment initiations over the grace period when applying these models.

5.3.2 Data expansion

One of the least transparent and potentially most influential steps in causal modelling is determining adequate weights; our algorithm was deliberately designed to delineate the choices required in this process. Once this step has been performed, the model fitting of the standard MSM follows fairly simply. The history-adjusted and dynamic MSMs have added complexity, requiring expansion of the data. This may be limited by computational capabilities, particularly for dynamic MSMs if a relatively large number of treatment regimes are to be compared.

5.3.3 Artificial censoring

After the data expansion required for the history-adjusted and dynamic MSMs, appropriate (artificial) censoring must be performed based on the observed history and compatibility with regimes. This is fairly straightforward under the HAMSMs, since the compatibility depends only on treatment. However, this step is more complex for the dynamic MSMs, since the censoring process depends on the relationship between regime, time-dependent CD4 count and treatment. Of note, in our example, the dynamic MSMs censor *all* treatment initiations which are not at the nadir (lowest to date) CD4 count (although with grace periods may permit delayed treatment initiation); this is not the case for HAMSMs, and therefore for this reason HAMSMs may potentially benefit from increased precision.

The censoring process is yet more complex for the dynamic treatment regimes if grace periods are incorporated, since patients must be allowed until the end of the grace period to initiate treatment, after it is indicated by the regime and time-dependent CD4 count, before applying any censoring due to non-initiation of treatment.

Of note, in our example, we found that incorporating weights for censoring due to LTFU or irregular CD4 count measurements had little impact on the estimated treatment effect estimates.

5.3.4 Strategy for causal estimation using MSMs

In summary, if one wishes to estimate optimal dynamic treatment regimes using dynamic MSMs, we recommend first implementing standard and history-adjusted MSMs. While standard and history-adjusted MSMs ask a different question compared to that addressed by dynamic MSMs, the reasons for our recommendation are: (*i*) to be satisfied that adequate weights have been

estimated, (*ii*) to demonstrate an effect of treatment in the population under study, and (*iii*) to gain understanding of the relationship between treatment and the time-dependent covariates of interest.

Whilst standard MSMs are limited to the estimation of static treatment regimes, the weight construction process is the same, and assessing the adequacy of the weights is substantially easier, since it is straightforward to obtain stabilised weights, whose sum should be close to 1. Further, if there is no evidence of a direct benefit of treatment, then the questions posed by optimal dynamic treatment regimes, for example relating to when to start treatment, may have little relevance.

The benefits of the additional complexity of HAMSMs are that treatment effect modifications by time-dependent covariates may be addressed. While the role of CD4 count in HIV disease epidemiology and treatment is well known in our example, the application of HAMSMs could aid identification of potential covariates for defining dynamic treatment regimes. Dynamic MSMs are considerably more complex to implement, and are computationally demanding.

5.4 Clinical comparison across the different MSMs

There are several comparisons which can be made across the application of standard, history-adjusted and dynamic MSMs to our CASCADE population, although it is important to recognise the differences between the three approaches and interpret the results in the light of these differences.

Standard MSMs estimate an "average" treatment effect, attempting to emulate a sequential randomised trial whereby patients at each given time-point who were previously treatment-naïve are randomised to initiate treatment or not. The treatment effect estimate is averaged across these sequential randomisations, that is, averaged across "sequential trials" with different follow-up times on and off treatment, and different CD4 counts at treatment initiation. For example, at later time-points, those patients initiating treatment will typically have lower CD4 counts, in whom we may expect to see a greater benefit of immediate treatment. Further, our primary models assumed an instantaneous and constant effect of treatment, regardless of the time spent on treatment. This assumes that current treatment is a good measure of treatment history.

HAMSMs similarly estimate an "average" treatment effect, but the (trial-baseline) CD4 count at treatment initiation is directly adjusted for in the model, and we condition on treatment history (patients must be previously treatment-naïve to contribute to a new "trial"). Therefore, the treatment effect estimate may differ from that obtained under the standard

MSMs. We can incorporate an interaction into the history-adjusted models to explore treatment effect modification by CD4 count. The resulting estimates are interpreted as the effect of immediate versus no treatment given CD4 count, conditional on having survived AIDS-free and off treatment to that time.

In contrast, dynamic MSMs estimate the cumulative effect of each regime defined by CD4 count. That is, the regimes are defined by treatment initiation when CD4 count is *first* observed to drop below a given threshold, and so also depend on CD4 count history beyond the current value (namely, the nadir). In addition, while we may anticipate that the effect of immediate versus no treatment given a CD4 count of z cells/mm^3 estimated from a HAMSM is most comparable to the dynamic regime given by $x = z$, this dynamic regime is somewhat different: it is defined by treatment initiation when the CD4 count is observed to drop *below* z cells/mm^3, and indeed could be substantially lower. In addition, we permitted 3-month grace periods under the dynamic treatment regimes, meaning that the CD4 count at treatment initiation may be even lower, although in practice we observed relatively minimal impact of such grace periods on the estimated optimal regimes in the simulation studies.

Having taken heed of these differences, it is informative to compare the estimates across the three approaches, since we might expect broad consistency.

5.4.1 History-adjusted and standard MSMs

The estimated ORs for the effects of immediate versus no treatment initiation under the HAMSMs were somewhat smaller (further from one) than the estimated effects of treatment under the standard MSMs (estimated ORs of around 0.2-0.3 under the HAMSMs compared to around 0.4-0.5 under the standard MSMs). This may be because the HAMSMs adjust for CD4 count at treatment initiation, and treatment history, unlike the standard MSMs.

Of note, excluding those with trial-baseline CD4 counts < 100 cells/mm^3 in the HAMSMs did not materially affect the estimated treatment effect, therefore this reassures us that the results from the standard MSMs were not unduly influenced by these "treatment refusers".

5.4.2 Dynamic and history-adjusted MSMs

Under the history-adjusted modelling, we saw evidence of a greater benefit of treatment at lower CD4 counts, with stronger estimated treatment effects at lower trial-baseline CD4 counts < 350 cells/mm^3. At higher CD4 counts, there was limited evidence of a benefit of treatment, and, comparing the ORs and associated confidence intervals, there was no significant difference

between the effects of treatment for patients with trial-baseline CD4 counts of ≥ 500 versus $350 - 499$ cells/mm^3. For example, under strategy Ia, the estimated ORs were 0.06, 0.38, 0.98 and 0.72 for CD4 counts < 200, 200−, 350− and ≥ 500 cells/mm^3, respectively. The evidence from the dynamic MSMs suggested that the optimal time to initiate treatment, in order to maximise 6-year AIDS-free survival, was immediately at study entry, or at least when CD4 counts were first observed to drop < 500 cells/mm^3, rather than further delay treatment.

Considering the AIDS-free survival rates, for illustration only under weighting strategies Ia/II/III and based on the pooled logistic regression model approach, the estimated 6-year AIDS-free survival probabilities were 0.90 (0.87, 0.93), 0.88 (0.84, 0.92), 0.93 (0.87, 0.96) and 0.93 (0.88, 0.97) under the regimes given by $x = 200$, 350, 500 and immediate treatment initiation, respectively. The confidence intervals overlap considerably, suggesting that the absolute benefits from early treatment initiation are likely to be small, and the results are probably not inconsistent with those from the HAMSMs. The lower 6-year AIDS-free survival rate under the regime given by $x = 350$ was seen consistently across the different weighting strategies and estimation approaches, but it does not concur with evidence from RCTs nor the results from our HAMSMs; we know that treatment initiation around CD4 counts of 350 cells/mm^3 is beneficial, compared to delay. The reasons for this apparent discrepancy are not clear, but this may illustrate a potential issue with few treatment initiations remaining uncensored after applying the artificial censoring process required for the dynamic MSMs. The evidence from our simulation studies based on real data, albeit a different population from the subset of CASCADE patients considered here, suggests that any potential benefits of early treatment initiation are likely to be small, and the AIDS-free survival probabilities may be very similar across high CD4 counts. In addition, the simulation studies indicated that the large measurement error in these data may be problematic.

5.4.3 Dynamic and standard MSMs

Lastly, we can compare the estimated 3- and 6-year AIDS-free survival rates from the standard and dynamic MSM chapters, for illustration in strategy Ia only. The 3-year AIDS-free survival under immediate treatment initiation was 0.97 (0.96, 0.98) under the standard MSM, compared to 1.00 (0.98, 1.00) and 0.99 (0.97, 0.99) under the dynamic MSM with the raw Kaplan-Meier and pooled logistic regression model approaches, respectively. At 6 years, the corresponding estimates were 0.95 (0.93, 0.97), 0.95 (0.90, 0.99) and 0.93 (0.88, 0.97). Of note, the precision of these estimates was slightly poorer under the dynamic compared to standard MSMs.

Although the regime given by $x = 200$ from the dynamic MSM setting is not the same as the regime of No treatment from the standard MSM approach, we may expect broadly similar results given that few patients should remain off treatment with CD4 counts < 200 cells/mm^3. At 3 years, and considering again only strategy Ia for illustration, the estimated AIDS-free survival rates were 0.95 (0.94, 0.96) under no treatment as estimated from the standard MSM, and 0.96 (0.94, 0.97) under regime $x = 200$ as estimated from the dynamic MSM (for both the raw Kaplan-Meier and pooled logistic regression model approaches). At 6 years, the corresponding estimates were 0.91 (0.89, 0.94), 0.91 (0.89, 0.94) and 0.90 (0.87, 0.93), respectively. Therefore the estimates and confidence intervals are very similar.

5.4.4 Summary

In conclusion, the results across all three approaches appear to be consistent, given the available precision.

5.4.5 In perspective

There have been a number of recent observational studies investigating when to start treatment in patients with HIV infection (HIV-CAUSAL collaboration, 2011; Kitahata et al., 2009; When to Start Consortium, 2009; Writing Committee for the CASCADE Collaboration, 2011). The overall suggestion from these studies is that early treatment initiation, with respect to CD4 count, may be optimal, but that the benefits of initiating at such high CD4 counts may be small in absolute terms (as discussed in section 1.5 and summarised in Table 5.1). Our results concur with these findings. These potentially small benefits should be balanced against the possible risks, which may not be captured in large observational studies which for pragmatic reasons collect a limited set of data, such as the development of drug resistance leading to more limited treatment options over the long-term. A large randomised trial is required to provide a more precise and unbiased estimate of the effect of earlier treatment across a range of prospectively evaluated outcomes, including those often not captured well in observational cohorts, such as serious non-AIDS events (see below). The START trial (INSIGHT (2009); EudraCT number 2008-006439-12) is currently underway to determine whether immediate initiation of treatment in patients with CD4 counts ≥ 500 cells/mm^3 is superior to deferral of treatment initiation until CD4 count drops to 350 cells/mm^3, however results are not expected until 2016. A major advantage of observational data is that it is possible to explore a broad range of dynamic regimes using causal methods; due to patient and resource limitations, it would not be feasible

to randomise patients to the wide spectrum of regimes which we have been able to consider here.

More recently, illnesses which were not originally considered to be directly associated with HIV infection, such as cardiovascular disease, have been recognised as a significant morbidity burden in HIV-infected persons, particularly following the SMART trial (SMART Study Group et al., 2006). However, information relating to serious non-AIDS events are not currently captured by CASCADE therefore we were unable to address this. It may be important to incorporate such information in future studies, and indeed such events are a component of the primary endpoint for the START trial.

At the population level, there may be additional benefits of earlier treatment initiation in terms of reduced transmission (Cohen et al., 2011). A further aspect which has not been considered, but would of course be of great interest to policy-makers, is the cost-effectiveness of earlier treatment initiation. Analysis of an RCT in a resource-limited setting (Haiti), comparing treatment initiation at CD4 counts between $200 - 350$ cells/mm^3 versus deferring until < 200 cells/mm^3, found that early treatment reduced mortality by 75% and was cost-effective (US$2050 per years of life saved, <3 times the gross product per capita; Koenig et al. (2011)).

5.5 Limitations and potential extensions

5.5.1 Our CASCADE population

Our population was constructed to capture patients early in HIV infection, where there is the greatest potential for early intervention and thus greatest potential benefit from early treatment. CASCADE participants have well-estimated dates of HIV seroconversion, and incorporating only those persons with CD4 counts ≥ 500 cells/mm^3 within 1-5 years after seroconversion at entry to the analysis meant that we did not include fast progressors who would be likely to start treatment anyway. While this led to the exclusion of approximately 11,000 patients, this ensured that our population may be considered the most appropriate in which to answer the question of when to initiate treatment, and in particular whether early initiation is beneficial.

Study	Outcome	Findings
Kitahata et al. (2009)	Death	Treatment initiation better than delay, even when CD4 > 500 cells/mm^3
When to Start Consortium (2009)	AIDS or death	At CD4 $351 - 400$ cells/mm^3, treatment initiation preferable to delaying until CD4 $251 - 350$ cells/mm^3
Shepherd et al. (2010)	Metric at 36 months[1]:	Optimal regime given by $x =$
	CD4-based	509 (95% CI $460 - 750$)
	Quality of life-based	475 (95% CI $287 - 750$)
Writing Committee for the CASCADE Collaboration (2011)	AIDS or death, or death alone	At CD4 < 500 cells/mm^3, treatment initiation preferable to delay (with no evidence of a benefit when CD4 $500 - 799$ cells/mm^3)
HIV-CAUSAL collaboration (2011)	AIDS or death	Beneficial to initiate when CD4 first < 500 cells/mm^3 compared to delay
	Death	No evidence of a benefit of treatment initiation while CD4 > 300 cells/mm^3
Our research	AIDS or death	Immediate treatment initiation (according to our study entry criteria), or at least when CD4 first < 500 cells/mm^3, preferable to any delay

Table 5.1: Comparison of our clinical findings with published research. [1] Also incorporating death, AIDS and serious non-AIDS events. CI=confidence interval.

Other recent work in this area has made use of seroprevalent cohorts, the benefit of which are the typically greater sample sizes (HIV-CAUSAL collaboration, 2011; Shepherd et al., 2010). These approaches include patients with no history of CD4 count below a given threshold, from the time when their CD4 count is first observed to be below this threshold. The treatment effect estimates from such studies might be considered to be closer to those which would be observed if such regimes were implemented in practice, where patients rarely present soon after infection, and therefore may be more pragmatically appropriate. In contrast, our estimates from the seroconverter cohorts may be considered to be closer to the true effects of different regimes defined by CD4 count thresholds, under a "best case" scenario where patients are identified soon after infection. Of note, the HIV-CAUSAL collaboration (2011) saw similarly large reductions in their patient numbers to us when restricting for the purposes of investigating causal effects (from $> 30,000$ to 8392 participants).

Shepherd et al. (2010) considered, as a sensitivity analysis, restricting to patients with a first CD4 count ≥ 500 cells/mm^3, which resembles our approach. The authors discuss the advantages and disadvantages of this, compared to their original approach as above. Clearly, a disadvantage is the ultimate restriction of patients to those with an observed CD4 count ≥ 500 cells/mm^3, which substantially limited the sample size in their seroprevalent cohort. The advantage of restricting to patients with an initial high CD4 count is to control for variation at the start of the trial. For example, if a patient entered the original analysis of Shepherd et al. (2010) with a CD4 count of 349 cells/mm^3 and initiated treatment immediately then this patient would be compliant with all regimes given by $x \geq 350$. Therefore, attempting to distinguish between the regimes given by higher x suffered from limited power in their analysis. In our approach, a patient would only be compliant with regimes $x = 350$ and 500 (and intermediate regimes) if their observed CD4 count dropped from > 500 cells/mm^3 to < 350 cells/mm^3, in response to which treatment was initiated; such patients are atypical. Further, their original approach compares patients compliant with the $x = 350$ regime who were never eligible for regime $x = 500$ (for example, a patient who remains treatment-naïve after a first CD4 count of 400 cells/mm^3), with those compliant with the $x = 350$ regime but who were (or still are) eligible and compliant with the $x = 500$ regime (for example, a patient who remains treatment-naïve after a first CD4 count of 550 cells/mm^3); in practice these patients may not be comparable.

5.5.2 Power

As discussed above, our stringent inclusion criteria led to only approximately 3000 CASCADE participants being included in our analyses. Our simulation studies have shown that the application of dynamic MSMs in such sample sizes to estimate optimal dynamic treatment regimes is likely to suffer from low power. In addition, our study based on CASCADE data suffered somewhat from limited follow-up. This is particularly pertinent to the application of the dynamic MSMs; as discussed in section 4.6.3, sufficient follow-up is necessary in order to be able to distinguish between the effects of different regimes on the outcome of interest.

5.5.3 Other dynamic treatment regimes

The focus in this thesis, and the majority of previous studies in this area (Hernán et al., 2006; HIV-CAUSAL collaboration, 2011; Robins et al., 2008; Writing Committee for the CASCADE Collaboration, 2011), has been on whether to initiate treatment early at CD4 counts of around 500 cells/mm^3 or later (lower). This is for pragmatic reasons, in that people rarely present for care earlier (with higher CD4 counts) and indeed it has recently been shown that nearly half of individuals have CD4 counts < 500 cells/mm^3 within just one year of seroconversion (Lodi et al., 2011). However, it may be that the optimal time to initiate with respect to CD4 count is above 500 cells/mm^3. If we had further restricted to patients with baseline CD4 counts above a higher threshold, then we would have had an even smaller subset of patients (upper quartile baseline CD4 count was 788 cells/mm^3). In their original analysis, Shepherd et al. (2010) did use a higher threshold, of 750 cells/mm^3, but in a sensitivity analysis found broadly consistent results when they applied an upper limit of 500 cells/mm^3.

The definition of dynamic treatment regimes need not be limited to just one time-dependent covariate. Time-independent covariates, such as sex or age, could be easily incorporated via interactions with treatment, as outlined in section 4.2.4, although we had limited power to address this. Further, other time-dependent covariates could be incorporated, such as HIV RNA levels or clinical events. For example, D Ford (personal communication, 25 March 2011) incorporated both previous clinical events and observed CD4 count to define dynamic treatment regimes related to switching from first- to second-line ART, and investigated their effects on mortality in HIV-infected persons in resource-limited settings. However, in our study, development of AIDS was part of the endpoint and therefore by definition could not be incorporated into the dynamic treatment regime. It has been shown in high-income settings that, in patients with CD4 counts > 350 cells/mm^3, higher levels of HIV RNA are known to be associated with

higher risk of AIDS and non-AIDS events (Reekie et al., 2011), but typically the only covariate consistently used in the treatment decision-making process would be CD4 count, in line with clinical guidelines (Gazzard and on behalf of the BHIVA Treatment Guidelines Writing Group, 2008; Panel on Antiretroviral Guidelines for Adults and Adolescents, 2009), therefore the value of extending the dynamic regime definition to include HIV RNA levels in our setting is not clear. Other disease areas may naturally have more complex regimes. For example, Taubman et al. (2009) incorporated a range of factors, such as BMI, exercise, alcohol and diet, to define a set of regimes and examine their collective impact on coronary heart disease.

5.5.4 Other causal methods

Other approaches such as the g-formula or g-estimation of SNMs could be used for the estimation of causal effects of treatment. Regardless of the method employed, such estimation in the presence of time-dependent confounding requires the assumption of no unmeasured confounders. This is similar to any observational analysis, except here this extends to time-dependent as well as time-independent confounders. All approaches also require correctly-specified models.

G-estimation of SNMs has the potential to be more efficient than MSMs and with fewer parametric assumptions than the g-formula (Daniel et al., 2011), but SNMs are less robust to model misspecification and are not intuitive to use. MSMs more closely resemble standard methods and so the implementation and interpretation of results using these models is more straightforward. For example, the hazard ratios obtained via the MSMs may be more familiar than the results from the AFT SNMs proposed in section 1.2.2. However, MSMs require the assumption of positivity, that is, at all levels of the covariate and treatment history, there is a non-zero probability of the possible future treatments (Cole and Hernán, 2008). This is not a requirement for g-estimation of SNMs nor the g-formula. In addition, the artificial censoring process required for the application of dynamic MSMs may result in the censoring of many treatment initiations, and hence potential loss of power. While the g-formula can easily incorporate highly complex dynamic regimes, it is computationally intensive and perhaps most useful when a small number of dynamic regimes are to be compared.

It should be noted that there are similarities between the application of the g-formula and the observational simulation studies we performed. Recall (section 1.2.1) that there are three steps to applying the g-formula: the first step is to estimate the parameters of the conditional distributions of each of the current covariates and the outcome, given covariate and treatment history; the second step requires simulation of a cohort based on the estimated distributions and

the treatment regime of interest; lastly, the simulated cohort is used to estimate the outcome under that treatment regime. For our observational simulation studies, we a priori defined the covariate, treatment and outcome distributions, which were conditional on covariate and treatment history. We then simulated a cohort using those distributions, similarly to the second step of the g-formula. However, in our simulation studies, consideration of the treatment regime of interest was not applied at this step, but rather after expansion of the simulated cohort, by censoring patients when no longer compliant with each regime. The final step of our simulation studies was to estimate the outcome, as in the third step of the g-formula, except that inverse probability weighting was applied to account for the potentially informative censoring of non-compliant patients. In addition, estimation of the outcome is performed separately for each regime under the g-formula, whereas the dynamic MSMs allow us to model the outcome across all regimes at once.

5.6 Final conclusions

Causal methods provide an opportunity to address many questions from observational studies, which it would otherwise not be possible to consider without potentially suffering major bias due to time-dependent confounding. It is infeasible to conduct sufficient randomised controlled trials to address all these questions. However, we have shown that answers from causal analyses may depend strongly on their implementation in ways which may not be obvious to a casual reader, particularly when attempting to compare results across different studies. Researchers conducting such analyses should be aware of these limitations and present multiple sensitivity analyses to delineate the effect of their assumptions on the results.

Appendices

Appendix A

Theory for simulation study

A.1 Conditional multivariate Normal distribution

A.1.1 Theorem

Let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ be a Normally-distributed n-dimensional random vector, where x_1 and x_2 have dimensions p and q respectively ($p + q = n$). Denote the mean vector and variance-covariance matrix for x by:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

Then the conditional distribution of x_2 given $x_1 = a$ is also Normally-distributed with mean vector and variance-covariance matrix given by:

$$\begin{aligned} \mu_{2|1} &= \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(a - \mu_1) \\ \text{and } \Sigma_{2|1} &= \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \end{aligned}$$

See for example Wang (2006) for proof.

A.1.2 Application of theorem for CD4 trajectory: simulating slope after treatment initiation, given CD4 count at treatment initiation

As in the main text, R is the square-root true CD4 count at treatment initiation, and S_1 and S_2 are the slopes during the first year and from one year after treatment initiation respectively. Our model states that these three are jointly Normally-distributed with mean vector and variance-

covariance matrix given by:

$$\mu = \begin{pmatrix} \mu_R \\ \mu_{S_1} \\ \mu_{S_2} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_R^2 & \sigma_{R,S_1} & \sigma_{R,S_2} \\ \sigma_{R,S_1} & \sigma_{S_1}^2 & \sigma_{S_1,S_2} \\ \sigma_{R,S_2} & \sigma_{S_1,S_2} & \sigma_{S_2}^2 \end{pmatrix}$$

Therefore, the conditional distribution of S_1 and S_2 given $R = \rho$ is also Normally-distributed with mean vector given by:

$$\begin{pmatrix} \mu_{S_1} \\ \mu_{S_2} \end{pmatrix} + \begin{pmatrix} \sigma_{R,S_1} \\ \sigma_{R,S_2} \end{pmatrix} \frac{1}{\sigma_R^2}(\rho - \mu_1)$$

and variance-covariance matrix:

$$\begin{aligned} & \begin{pmatrix} \sigma_{S_1}^2 & \sigma_{S_1S_2} \\ \sigma_{S_1S_2} & \sigma_{S_2}^2 \end{pmatrix} - \begin{pmatrix} \sigma_{R,S_1} \\ \sigma_{R,S_2} \end{pmatrix} \frac{1}{\sigma_R^2} \begin{pmatrix} \sigma_{R,S_1} & \sigma_{R,S_2} \end{pmatrix} \\ = & \begin{pmatrix} \sigma_{S_1}^2 & \sigma_{S_1S_2} \\ \sigma_{S_1S_2} & \sigma_{S_2}^2 \end{pmatrix} - \frac{1}{\sigma_R^2} \begin{pmatrix} \sigma_{R,S_1}^2 & \sigma_{R,S_1}\sigma_{R,S_2} \\ \sigma_{R,S_1}\sigma_{R,S_2} & \sigma_{R,S_2}^2 \end{pmatrix} \end{aligned}$$

A.1.3 Application of theorem for Brownian motion: simulating $W(t_2)$ given $W(t_1)$

Time was split into monthly intervals, therefore let $t_1 = t_2 - 1/12$. Then we have:

$$\begin{aligned} Var[W(t_1)] &= \delta t_1 = \delta t_2(1 - 1/12t_2) \\ Var[W(t_2)] &= \delta t_2 \\ corr[W(t_1), W(t_2)] &= \frac{t_2 - 1/12}{\sqrt{(t_2 - 1/12)t_2}} = \sqrt{1 - 1/12t_2} \\ cov[W(t_1), W(t_2)] &= \sqrt{(1 - 1/12t_2)\,\delta t_2\,(1 - 1/12t_2)\,\delta t_2} = \delta t_2\,(1 - 1/12t_2) \end{aligned}$$

and so $W(t_2)$ given $W(t_1) = w$ is Normally-distributed with mean vector simply w, since $Var[W(t_1)] = cov[W(t_1), W(t_2)]$, and variance-covariance matrix given by:

$$\delta t_2 - \frac{[\delta t_2(1 - 1/12t_2)]^2}{\delta t_2(1 - 1/12t_2)} = \frac{\delta}{12}$$

Appendix B

Example code

Here we provide some example code for estimating standard, history-adjusted and dynamic MSMs using Stata (StataCorp, 2009). The data are set up with one observation per patient per time period (month). Some key variables are defined in Table B.1. Throughout the code, text in brackets such as <xxx> indicates insertion of the variables xxx as appropriate.

B.1 Standard MSMs

```
*** WEIGHT ESTIMATION
* DENOMINATOR
/* fit treatment model for denominator of weights, in periods up to and
   including treatment initiation */
noi xi:logistic trt <time covariates> <baseline covariates> ///
   <time-dependent covariates> if period<=initperiod|initperiod>=.
gen insample=e(sample)
/* predicted probability of treatment based on the denominator model; after
   treatment initiation, pr(trt)=1 */
predict pred_ptrt if insample
replace pred_ptrt=1 if initperiod<. & period>initperiod
* predicted probability of OBSERVED treatment based on the denominator model
```

Variable name	Description
patient	Unique patient identifier
period	Time period
trt	Indicator for being on treatment in a given period
initperiod	Period in which the patient initiated treatment (missing if not observed to initiate treatment)
event_1	Lagged event indicator

Table B.1: Definition of key variables.

```
gen ptrt_denom=pred_ptrt*trt+(1-pred_ptrt)*(1-trt)
drop pred_ptrt
* NUMERATOR
noi xi: logistic trt <time covariates> <baseline covariates> if insample
* predicted probability of treatment based on the numerator model
predict pred_ptrt if insample
replace pred_ptrt=1 if initperiod<. & period>initperiod
* predicted probability of OBSERVED treatment based on numerator model
gen ptrt_num=pred_ptrt*trt+(1-pred_ptrt)*(1-trt)
drop pred_ptrt
/* at each point, probability of treatment/censoring history is product to
   that point. Last record will have predicted probability of treatment
   missing (since trt=missing then), but don't use last period anyway since
   always looking at Y(k+1). Keep original probabilities for use in the
   history-adjusted and dynamic MSM work */
gen ptrt_denomORIG=ptrt_denom
sort patient period
by patient: replace ptrt_denom=ptrt_denom*ptrt_denom[_n-1] if _n>1
by patient: replace ptrt_num=ptrt_num*ptrt_num[_n-1] if _n>1
* WEIGHTS
* non-stabilised
gen weightns=1/ptrt_denom
* stabilised
gen weights=ptrt_num/ptrt_denom
*** OUTCOME ESTIMATION
* using the stabilised weights (note, untruncated)
noi xi: logistic event_1 trt <time covariates> <baseline covariates> ///
   [pw=weights], robust cluster(patient)
```

B.2 HAMSMs

```
*** DATA EXPANSION
/* if a patient has X intervals of follow up, then need 1 copy of first
   interval, 2 copies of the second, ..., and X copies of the last. Then
   for each patient, have 1 trial starting at each month */
gen int exp=period+1
/* BUT for looking at effect of initiate vs defer, once patient has
   initiated treatment, don't need further "trials" */
```

```
replace exp=initperiod+1 if period>initperiod
/* because of the way the data is set up, trt will be missing for each of
   the last records, therefore don't actually want to consider that a new
   trial, and we're assuming the any censoring weighting has already been
   sorted out */
replace exp=exp-1 if trt>=.
expand exp
drop exp
sort patient period
* eg trial=12 means the trial starting at month 12 onwards
by patient period: gen int trial=_n-1
* trial time
gen trialtime=(period/12)-trial/12
replace trialtime=0 if trialtime<0.00001
/* for each trial, the treatment regime (or randomisation, rx) is
   determined by the trt in the first period*/
sort patient trial trialtime
by patient trial: gen byte rx=trt[1]
/* the "baseline" covariates are those at the start of that trial, ie in
   the first period */
foreach var of varlist <trial-baseline covariates> {
   by patient trial: gen b_'var'='var'[1]
}
/* flag for censoring due to initiation of treatment after deferring in
   first period of the trial (won't include any records from that
   initiation onwards [including the one where initiate]) */
gen byte censdef=trt==1 & rx==0
* generate indicator for the records to be used in the models
gen byte inmodel=(rx==0 & trialtime>0 & (period<=initperiod | initperiod>=.))
*** WEIGHT ESTIMATION
* NUMERATOR
noi xi: logistic trt <time covariates, for trial and trial-time> ///
   <true-baseline covariates> <trial-baseline covariates> if inmodel==1
/* trt weights only applied from the second month (ie period=1) onwards,
   since in the first month (period=0), that's when the "randomisation" is
   determined; so trt weights should be =1 in the first month. Also
   treatment weights should be missing after initiation in patients
   'randomised' to defer, and treatment weights should be =1 for patients
```

```
   'randomised' to initiate (just treatment weights applied to those
   'randomised' to defer) */
* predicted probability of treatment based on the numerator model
predict pred_ptrt if e(sample)
* predicted probability of OBSERVED treatment based on the numerator model
gen ptrt_numHA=pred_ptrt*trt+(1-pred_ptrt)*(1-trt)
drop pred_ptrt*
* DENOMINATOR
* same as for standard MSMs, with some adjustments below
gen ptrt_denomHA=ptrt_denomORIG
/* at each point probability of treatment history is product to that point,
   OVER PATIENT/TRIAL. In first month of each trial, set =1, and after
   treatment initiation in patients who initially Deferred, make treatment
   probabilities missing, and in patients 'randomised' to initiate, set
   treatment weights =1 */
replace ptrt_denomHA=1 if trialtime==0|rx==1
replace ptrt_numHA=1 if trialtime==0|rx==1
replace ptrt_denomHA=. if censdef==1
replace ptrt_numHA=. if censdef==1
sort patient trial trialtime
by patient trial: replace ptrt_denomHA=ptrt_denomHA*ptrt_denomHA[_n-1] if _n>1
by patient trial: replace ptrt_numHA=ptrt_numHA*ptrt_numHA[_n-1] if _n>1
* WEIGHTS
gen weightnsHA=1/ptrt_denomHA
gen weightsHA=ptrt_numHA/ptrt_denomHA
*** OUTCOME ESTIMATION
noi xi: logistic event_1 rx <time covariates, for trial and trial-time> ///
   <true-baseline covariates> <trial-baseline covariates> [pw=weightsHA] ///
   if censdef==0, robust cluster(patient)
```

B.3 Dynamic MSMs

```
*** DATA EXPANSION
* apply program dynexpr - see below
*** WEIGHT ESTIMATION
* if wish to stabilise the weights (only if no grace period):
noi xi: logistic censreg <time and regime rx modelled flexibly> ///
   <baseline covariates> if (time<=censregtime | censregtime>=.)
```

```
predict pred_pcens if e(sample)
gen puncens=1-pred_pcens
replace puncens=0 if censregtime<. & time>censregtime
sort patient rx time
by patient rx: replace puncens=puncens*puncens[_n-1] if _n>1
* apply program dynwt - see below
*** OUTCOME ESTIMATION
/* the rx and time covariates should be flexibly modelled, and include
   interactions; the results can then be used to predict and plot survival */
noi xi: logistic event_1 <rx and time covariates> if censreg==0 ///
   [pw=weightnsDYN], robust cluster(patient)
```

B.3.1 Program dynexpr

This program expands the data into one record per patient per regime (per time period).

```
prog def dynexpr
vers 10.1
syntax, cd4var(string) xu(integer) xl(integer) xj(integer) m(integer) ///
   [approach(integer 0) immed]
/* Expansion for dynamic MSM based on regimes defined by CD4 (given by
   variable cd4var) as: xl(xj)xu. Note: should have already fit treatment
   denominator models and got Pr(observed treatment|time-dependent
   covariates) in each interval
   - m = grace period (1=no grace period)
   - approach = approach 1 or 2 of Cain et al 2010, if using a grace period
     with m>1
   - immed should be specified if want to consider a regime of immediate
     treatment initiation */
qui {
   noi dib "dynexp: expansion based on `xl'(`xj')`xu' [`immed']"
   noi dib "GRACE PERIOD = `m' (1=no grace period); approach `approach'"
   * checks
   assert `xl'<`xu'
   assert `m'>=1
   assert `approach'==0 if `m'==1
   assert `approach'==1|`approach'==2 if `m'>1
   * expand the dataset, with variable rx representing faux randomisation
   compress
```

```
local nreg=(`xu'-`xl')/`xj' + 1
if "`immed'"!="" local nreg=`nreg'+1
confirm integer number `nreg'
expand `nreg'
sort patient time
local i=1
gen rx=0
assert `cd4var'<10000 if `cd4var'<.
local rxlist="`xl'(`xj')`xu'"
if "immed'"!="" local rxlist="`rxlist' 10000"
foreach x of numlist `rxlist' {
   by patient time: replace rx=`x' if _n==`i'
   local i=`i'+1
}
* indicator for when eligible for treatment initiation according to regime
gen elig_trt=(`cd4var'<rx)
sort patient rx time
by patient rx: replace elig_trt=sum(elig_trt)
replace elig_trt=1 if elig_trt>1 & elig_trt<.
assert elig_trt==0|elig_trt==1
/* grace variable (if applicable) =1 for first eligible interval, 2 for
   second, ..., m for mth; missing outside of the grace windows */
if `m'>1 {
   sort patient rx time
   by patient rx: gen grace=1 if elig_trt==1 & (_n==1 | _n>1 & ///
      elig_trt[_n-1]==0)
   local k=2
   while `k'<=`m' {
      by patient rx: replace grace=`k' if grace[_n-`k'+1]==1 & _n>`k'-1
      local k=`k'+1
   }
   assert grace>=1 & grace<=`m' if grace<.
   assert elig_trt==1 if grace<.
}
* censor if initiate before eligible
gen censreg=(trt==1 & elig_trt==0)
gen _censregind=censreg
/* censor if initiated too late
```

```
      - no grace period: if did not initiate in first eligible interval
      - grace period: if did not initiate in (by) mth eligible interval
      (since once on always on, can just look forward to mth) */
   if `m'==1 replace censreg=1 if trt==0 & elig_trt==1
   if `m'>1  replace censreg=1 if trt==0 & elig_trt==1 & grace==`m'
   replace _censregind=2 if censreg==1 & _censregind==0
   * remain censored after first censored from regime
   sort patient rx time
   by patient rx: replace censreg=sum(censreg)
   replace censreg=1 if censreg>1 & censreg<.
   assert censreg==0|censreg==1
   lab var censreg "cens, noncomp with dyn regime"
   sort patient rx time
   by patient rx: egen censregind=max(_censregind)
   replace censregind=0 if censreg==0
   drop _censregind
   assert censregind==0|censregind==1|censregind==2
   lab def censregindlab 1 "early" 2 "late"
   lab val censregind censregindlab
   * when censored
   sort patient rx time
   by patient rx: gen _censregtime=time if censreg==1 & ///
      (_n==1 | _n>1 & censreg[_n-1]==0)
   by patient rx: egen censregtime=max(_censregtime)
   drop _censregtime
}  /* end of qui */
end
```

B.3.2 Program dynwt

This program estimates the weights, assuming that the denominator probabilities within each time period have already been derived (and the numerator probabilities, if using stabilised weights).

```
prog def dynwt
vers 10.1
syntax, m(integer) approach(integer 0) [stab]
/* - m indicates the grace period length
   - approach indicates the approach as per Cain et al 2010
```

```
   - specify stab option if wish to stabilise the weights - only possible
     here with approach 1 */
qui {
   assert `approach'==1|`approach'==2
   if `approach'==2 assert "`stab'"==""
   gen ptrt_denomDYN=ptrt_denomORIG
   /* if have grace period, AND APPROACH 1, then no-one is censored in the
      first m-1 intervals of the grace period, irrespective of whether
      initiated treatment or not; therefore force DENOMINATOR probabilities
      =1 there */
   if `approach'==1   replace ptrt_denomDYN=1 if grace>=1 & grace<`m' & grace<.
   /* if have grace period, AND APPROACH 2, then need to amend the numerator
      of the weights during the grace period. Haven't taken inverse of
      treatment probabilities yet therefore multiply by (throughout grace
      period, including m): 1/[1/(m+1-j)]=(m+1-j) where initiate,
      1/[1-{1/(m+1-j)}]=(m+1-j)/(m-j) if don't initiate. NB this will create
      infinity (=missing) in mth grace period if treatment not initiated
      there, but doesn't matter since that (and all subsequent) interval(s)
      will be censored and so won't have weights anyway. No change if after
      treatment initiation - interval-specific-weights there will be =1
      (check for this just below) */
   if `approach'==2 {
      sort patient rx time
      by patient rx: replace ptrt_denomDYN= ///
         ptrt_denomDYN*(`m'+1-grace)/(`m'-grace) if grace<. & trt==0
      by patient rx: replace ptrt_denomDYN=ptrt_denomDYN*(`m'+1-grace) ///
         if grace<. & trt==1 & (_n==1 | _n>1 & trt[_n-1]==0)
      /* note that the 'probabilities' here may be >1 after adjustment of
         the numerator of the non-stabilised weights with grace period and
         under approach 2 - so the non-stabilised weights may be <1 */
   }
   /* probability of remaining uncensored at each time point (while
      uncensored) is the product of probabilities; this multiples over
      time when censored too but doesn't matter since we won't take inverse
      for weights there */
   sort patient rx time
   by patient rx: replace ptrt_denomDYN=ptrt_dnomDYN*ptrt_denomDYN[_n-1] ///
      if _n>1
```

```
   * weights
   gen weightnsDYN=1/ptrt_denomDYN if censreg==0
   if "`stab'"!="" gen weightsDYN=puncens/ptrt_denomDYN if censreg==0
}  /* end of qui */
end
```

Bibliography

Arribas, J., M. Mora, and J. Pascual-Pareja (2009). Early versus deferred antiretroviral therapy for HIV. *New England Journal of Medicine 361*(8), 823.

Bang, H. and J. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–972.

Bodnar, L., M. Davidian, A. Siega-Riz, and A. Tsiatis (2004). Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *American Journal of Epidemiology 159*(10), 926–934.

Breslow, N. (1970). A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika 57*, 579–594.

Brookhart, M. and M. van der Laan (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics & Data Analysis 50*, 475–498.

Buchbinder, S. and V. Jain (2009). Early versus deferred antiretroviral therapy for HIV. *New England Journal of Medicine 361*(8), 822.

Burton, A., D. Altman, P. Royston, and R. Holder (2006). The design of simulation studies in medical statistics. *Statistics in Medicine 25*(24), 4279–4292.

Cain, L., J. Robins, E. Lanoy, R. Logan, D. Costagliola, and M. Hernán (2010). When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *International Journal of Biostatistics 6*(2), Article 18.

CASCADE Collaboration (2009). Concerted Action of SeroConversion to AIDS and Death in Europe. http://www.ctu.mrc.ac.uk/cascade.

CDC (1992). 1993 revised classification system for HIV infection and ex-

panded surveillance case definition for AIDS among adolescents and adults. http://www.cdc.gov/mmwr/preview/mmwrhtml/00018871.htm.

Cohen, M., Y. Chen, M. McCauley, T. Gamble, M. Hosseinipour, N. Kumarasamy, J. Hakim, J. Kumwenda, B. Grinsztejn, J. Pilotto, S. Godbole, S. Mehendale, S. Chariyalertsak, B. Santos, K. Mayer, I. Hoffman, S. Eshleman, E. Piwowar-Manning, L. Wang, J. Makhema, L. Mills, G. de Bruyn, I. Sanne, J. Eron, J. Gallant, D. Havlir, S. Swindells, H. Ribaudo, V. Elharrar, D. Burns, T. Taha, K. Nielsen-Saines, D. Celentano, M. Essex, T. Fleming, and the HPTN 052 Study Team (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine 365*(6), 493–505.

Cole, S. and M. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology 168*(6), 656–664.

Cole, S., M. Hernán, K. Anastos, B. Jamieson, and J. Robins (2007). Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. *American Journal of Epidemiology 166*, 219–227.

Cole, S., M. Hernán, J. Margolick, M. Cohen, and J. Robins (2005). Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology 162*(5), 471–478.

Cole, S., M. Hernán, J. Robins, K. Anastos, J. Chmiel, R. Detels, C. Ervin, H. Feldman, R. Greenblatt, L. Kingsley, S. Lai, M. Young, M. Cohen, and A. Muñoz (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology 158*(7), 687–694.

Collaboration of Observational HIV Epidemiological Research Europe (COHERE) Study Group (2008). Response to combination antiretroviral therapy: variation by age. *AIDS 22*(12), 1463–1473.

Cook, N., S. Cole, and C. Hennekens (2002). Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the physicians' health study. *American Journal of Epidemiology 155*(11), 1045–1053.

D'Agostino, R., M. Lee, A. Belanger, L. Cupples, K. Anderson, and W. Kannel (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine 9*(12), 1501–1515.

Daniel, R., S. Cousens, B. De Stavola, M. Kenward, and J. Sterne (2011). Methods for dealing with time-varying confounding. *Statistics in Medicine [in press]*.

Dawid, A. and V. Didelez (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statistics Surveys 4*, 184–231.

Ewings, F., K. Bhaskaran, K. McLean, D. Hawkins, M. Fisher, S. Fidler, R. Gilson, D. Nock, R. Brettle, M. Johnson, A. Phillips, A. Johnson, and K. Porter (2008). Survival following HIV infection of a cohort followed up from seroconversion in the UK. *AIDS 22*(1), 89–95.

Fewell, Z., M. Hernán, F. Wolfe, K. Tilling, H. Choi, and J. Sterne (2004). Controlling for time-dependent confounding using marginal structural models. *Stata Journal 4*(4), 404–420.

Gazzard, B. and on behalf of the BHIVA Treatment Guidelines Writing Group (2008). British HIV association guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008. *HIV Medicine 9*, 563–608.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika 52*, 203–223.

Gran, J., K. Roysland, M. Wolbers, V. Didelez, J. Sterne, B. Ledergerber, H. Furrer, V. von Wyl, and O. Aalen (2010). A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine 29*(26), 2757–2768.

Greenland, S. (2003). Quantifying bias in causal models: classical confounding vs collider-stratification bias. *Epidemiology 14*, 300–306.

Greenland, S., S. Lanes, and M. Jara (2008). Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and g-estimation. *Clinical Trials 5*(1), 5–13.

Harrell, F. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hernán, M., A. Alonso, R. Logan, F. Grodstein, K. Michels, W. Willett, J. Manson, and J. Robins (2008). Observational studies analyzed like randomized experiments. An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology 19*(6), 766–779.

Hernán, M., B. Brumback, and J. Robins (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology 11*(5), 561–570.

Hernán, M., B. Brumback, and J. Robins (2001). Marginal structural models to estimate the joint causal effect of nonrandmized treatments. *Journal of the American Statistical Association 96*(454), 440–448.

Hernán, M., B. Brumback, and J. Robins (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine 21*(12), 1689–1709.

Hernán, M., S. Cole, J. Margolick, M. Cohen, and J. Robins (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety 14*(7), 477–491.

Hernán, M., E. Lanoy, D. Costagliola, and J. Robins (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology 98*(3), 237–242.

Hernán, M. and J. Robins (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health 60*, 578–586.

Hernán, M. and J. Robins (2009). Early versus deferred antiretroviral therapy for HIV. *New England Journal of Medicine 361*(8), 822–823.

HIV-CAUSAL Collaboration (2010). The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS 24*, 123–137.

HIV-CAUSAL collaboration (2011). When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries. *Annals of Internal Medicine 154*, 509–515.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

INSIGHT (2009). START 001 international randomized trial. http://insight.ccbr.umn.edu/start/.

Kitahata, M., S. Gange, A. Abraham, B. Merriman, M. Saag, A. Justice, R. Hogg, S. Deeks, J. Eron, J. Brooks, S. Rourke, M. Gill, R. Bosch, J. Martin, M. Klein, L. Jacobson, B. Ro-

driguez, T. Sterling, G. Kirk, S. Napravnik, A. Rachlis, L. Calzavara, M. Horberg, M. Silverberg, K. Gebo, J. Goedert, C. Benson, A. Collier, S. Van Rompaey, H. Crane, R. McKaig, B. Lau, A. Freeman, R. Moore, and the NA-ACCORD Investigators (2009). Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine 360*(18), 1815–1826.

Koenig, S., H. Bang, P. Severe, M. Jean Juste, A. Ambroise, A. Edwards, J. Hippolyte, D. Fitzgerald, J. McGreevy, C. Riviere, S. Marcelin, R. Secours, W. Johnson, J. Pape, and B. Schackman (2011). Cost-effectiveness of early versus standard antiretroviral therapy in HIV-infected adults in Haiti. *PLoS Medicine 8*(9), e1001095.

Lau, B., S. J. Gange, and R. D. Moore (2007). Interval and clinical cohort studies: epidemiological issues. *AIDS Research and Human Retroviruses 23*(6), 769–776.

Lebanon, G. (2006). Relative efficiency, efficiency, and the Fisher information. http://www.cc.gatech.edu/~lebanon/notes/efficiency.pdf.

Lefebvre, G., J. Delaney, and R. Platt (2008). Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine 27*(18), 3629–3642.

Lichtenstein, K., C. Armon, K. Buchacz, J. Chmiel, K. Buckner, E. Tedaldi, K. Wood, S. Holmberg, and J. Brooks (2010). Low CD4+ T cell count is a risk factor for cardiovascular disease events in the HIV Outpatient Study. *Clinical Infectious Diseases 51*(4), 435–447.

Little, R. and D. Rubin (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health 21*, 121–145.

Lodi, S., A. Phillips, G. Touloumi, R. Geskus, L. Meyer, R. Thiébaut, N. Pantazis, J. del Amo, A. Johnson, A. Babiker, and K. Porter (2011). 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. *Clinical Infectious Diseases 53*(8), 817 –825.

Loeys, T. and E. Goetghebeur (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics 59*(1), 100–105.

Loeys, T., E. Goetghebeur, and A. Vandebosch (2005). Causal proportional hazards models and time-constant exposure in randomized clinical trials. *Lifetime Data Analysis 11*(4), 435–449.

Lok, J., R. Gill, A. van der Vaart, and J. Robins (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica 58*(3), 271–295.

Madec, Y., F. Boufassa, K. Porter, L. Meyer, and the CASCADE collaboration (2005). Spontaneous control of viral load and CD4 cell count progression among HIV-1 seroconverters. *AIDS 19*, 2001–2007.

Mellors, J., C. Rinaldo Jr, P. Gupta, R. White, J. Todd, and L. Kingsley (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science 272*(5265), 1167–1170.

Moodie, E. (2009). Risk factor adjustment in marginal structural model estimation of optimal treatment regimes. *Biometrical Journal 51*(5), 774–788.

Moodie, E., T. Richardson, and D. Stephens (2007). Demystifying optimal dynamic treatment regimes. *Biometrics 63*, 447–455.

Mortimer, K. M., R. Neugebauer, M. van der Laan, and I. B. Tager (2005). An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology 162*(4), 382–388.

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical methodology) 65*(2), 331–355.

Neyman, J., D. Dabrowska, and T. Speed (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1990 translation of Polish original). *Statistical Science 5*(4), 465–472.

Panel on Antiretroviral Guidelines for Adults and Adolescents (2009). Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf.

Petersen, M., S. Deeks, J. Martin, and M. van der Laan (2007). History-adjusted marginal structural models for estimating time-varying effect modification. *American Journal of Epidemiology 166*(9), 985–993.

Petersen, M., S. Deeks, and M. van der Laan (2007). Individualized treatment rules: generating candidate clinical trials. *Statistics in Medicine 26*, 4578–4601.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2010). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research 21*(1), 31–54.

Reekie, J., J. Gatell, I. Yust, E. Bakowska, A. Rakhmanova, M. Losso, M. Krasnov, P. Francioli, J. Kowalska, A. Mocroft, and on behalf of EuroSIDA in EuroCoord (2011). Fatal and non-fatal AIDS and non-AIDS events in HIV-1 positive individuals with high CD4 counts according to viral load strata. *AIDS 25*(18), 2259–2568.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9-12), 1393–1512.

Robins, J. (1989a). The control of confounding by intermediate variables. *Statistics in Medicine 8*, 679–701.

Robins, J. (1989b). Estimation of the effect of AZT on time to AIDS in HIV-infected subjects from observational data. *American Journal of Epidemiology 130*(4), 798.

Robins, J. (1994). Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics 23*(8), 2379–2412.

Robins, J. (1998). Marginal structural models. In *1997 Proceedings of the American Statistical Association*, Section on Bayesian Statistical Science, pp. 1–10.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer.

Robins, J., D. Blevins, G. Ritter, and M. Wulfsohn (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology 3*(4), 319–336.

Robins, J. and D. Finkelstein (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics 56*, 779–788.

Robins, J., S. Greenland, and F. Hu (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome: rejoinder. *Journal of the American Statistical Association 94*(447), 708–712.

Robins, J., M. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Robins, J., M. Hernán, and A. Rotnitzky (2007). Invited commentary: effect modification by time-varying covariates. *American Journal of Epidemiology 166*(9), 994–1002.

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

Robins, J., A. Rotnitzky, and L. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Robins, J. and A. Tsiatis (1991). Correcting for non-compliance in randomized trials using rank-preserving structural failure time models. *Communications in Statistics - Theory and Methods 20*(8), 2609–2631.

Rosthoj, S., C. Fullwood, R. Henderson, and S. Stewart (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine 25*, 4197–4215.

Rotnitzky, A., A. Farall, A. Bergesio, and D. Scharfstein (2007). Analysis of failure time data under competing censoring mechanisms. *Journal of the Royal Statistical Society: Series B (Statistical methodology) 69*, 307–327.

Royston, P. and M. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine 30*(19), 2409–2421.

Royston, P. and W. Sauerbrei (2008). *Multivariable model-building*. Wiley series in probability and statistics. UK: Wiley.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Samet, J., K. Freedberg, J. Savetsky, L. Sullivan, and M. Stein (2001). Understanding delay to medical care for HIV infection: the long-term non-presenter. *AIDS 15*, 77–85.

Scharfstein, D. and J. Robins (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika 89*(3), 617–634.

Scharfstein, D., J. Robins, W. Eddings, and A. Rotnitzky (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics 57*, 404–413.

Shepherd, B., C. Jenkins, P. Rebeiro, S. Stinnette, S. Bebawy, C. McGowan, T. Hulgan, and T. Sterling (2010). Estimating the optimal CD4 count for HIV-infected persons to start antiretroviral therapy. *Epidemiology 21*(5), 698–705.

SMART Study Group, W. El-Sadr, J. Lundgren, J. Neaton, F. Gordin, D. Abrams, R. Arduino, A. Babiker, W. Burman, N. Clumeck, C. Cohen, D. Cohn, D. Cooper, J. Darbyshire, S. Emery, G. Fätkenheuer, B. Gazzard, B. Grund, J. Hoy, K. Klingman, M. Losso, N. Markowitz, J. Neuhaus, A. Phillips, and C. Rappoport (2006). CD4+ count-guided interruption of antiretroviral treatment. *New England Journal of Medicine 355*(22), 2283–2296.

SPARTAC Trial Investigators (2011, July). The effect of short-course antiretroviral therapy in primary HIV infection: final results from SPARTAC, an international randomised controlled trial. In *6th International AIDS Society conference on HIV pathogenesis, treatment and prevention*, Rome, Italy (Abstract number WELBX06).

StataCorp (2009). *Stata 11 Base Reference Manual*. College Station, TX: Stata Press.

Study Group on Death Rates at High CD4 Count in Antiretroviral Naive Patients (2010). Death rates in HIV-positive antiretroviral-naive patients with CD4 count greater than 350 cells per microlitre in Europe and North America: a pooled cohort observational study. *Lancet 376*, 340–345.

Taubman, S., J. Robins, M. Mittleman, and M. Hernán (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology 38*(6), 1599–1611.

Toh, S., S. Hernández-Díaz, R. Logan, J. Robins, and M. Hernán (2010). Estimating absolute risks in the presence of nonadherence. an application to a follow-up study with baseline randomization. *Epidemiology 21*, 528–539.

Tyrer, F., A. Walker, J. Gillett, K. Porter, and the UK Register of HIV Seroconverters (2003). The relationship between HIV seroconversion illness, HIV test interval and time to AIDS in a seroconverter cohort. *Epidemiology and Infection 131*(3), 1117–1123.

Van der Laan, M. and M. Petersen (2004). History-adjusted marginal structural models and statically-optimal dynamic treatment regimes. *U.C. Berkeley Division of Biostatistics Working Paper Series 158*.

Van der Laan, M., M. Petersen, and M. Joffe (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *International Journal of Biostatistics 1*(1), Article 4.

Velleman, P. and D. Hoaglin (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press.

Walker, A., I. White, and A. Babiker (2004). Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. *Statistics in Medicine 23*(4), 571–590.

Wang, R. (2006). Marginal and conditional distributions of multivariate normal distribution. http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html.

When to Start Consortium (2009). Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet 373*, 1352–1363.

White, I., A. Babiker, A. Walker, and J. Darbyshire (1999). Randomization-based methods for correcting for treatment changes: examples from the Concorde trial. *Statistics in Medicine 18*(19), 2617–2634.

WHO (2010). Antiretroviral therapy for HIV infection in adults and adolescents. Recommendations for a public health approach. http://whqlibdoc.who.int/publications/2010/9789241599764_eng.pdf.

Witteman, J., R. D'Agostino, T. Stijnen, W. Kannel, J. Cobb, M. de Ridder, A. Hofman, and J. Robins (1998). G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology 148*(4), 390–401.

Writing Committee for the CASCADE Collaboration (2011). Timing of HAART initiation and clinical outcomes in human immunodeficiency virus type 1 seroconverters. *Archives of Internal Medicine 171*(17), 1560–1569.

Young, J., L. Cain, J. Robins, E. O'Reilly, and M. Hernán (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences 3*(1), 119–143.

Young, J., M. Hernán, S. Picciotto, and J. Robins (2009). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis 16*(1), 71–84.

Zeger, S. and K. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 42*, 121–130.