The Effect on Treatment Comparisons of Different Measurement Frequencies in Human Immunodeficiency Virus Observational Databases

James T. Griffin¹, Christophe Fraser¹, Luuk Gras², Frank de Wolf¹,², and Azra C. Ghani³

¹ Department of Infectious Disease Epidemiology, Faculty of Medicine, Imperial College, London, United Kingdom.
² HIV Monitoring Foundation, Amsterdam, the Netherlands.
³ Infectious Disease Epidemiology Unit, Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom.

Received for publication July 15, 2005; accepted for publication October 27, 2005.

Data collected in a routine clinical setting are frequently used to compare antiretroviral treatments for human immunodeficiency virus (HIV). Differences in the frequency of measurement of HIV RNA levels and CD4-positive T-lymphocyte cell counts introduce a possible source of bias into estimates of the difference in effectiveness between treatments. The authors investigated the size of this bias when survival analysis methods are used to compare the initial efficacy of antiretroviral regimens. Data sets of clinical markers were simulated by use of differential equations that model the interaction between HIV and human T-cells. Cox proportional hazards and parametric models were fitted to the simulated data sets to evaluate the bias and coverage of 95% confidence intervals for the difference between regimens. The authors’ results demonstrate that differences in the frequency of follow-up can substantially bias estimated treatment differences if methods do not correctly account for the intervals between measurements and if the statistical model chosen does not fit the data well. Analyses using methods applicable to interval-censored data reduce the bias. In the Athena cohort of HIV-infected individuals in the Netherlands from 1999 to 2003, there are differences in measurement frequency between current regimens that are of sufficient magnitude to conclude incorrectly that some regimens are more effective than others.

antiretroviral therapy, highly active; bias (epidemiology); cohort studies; HIV; survival analysis

Abbreviations: ART, antiretroviral therapy; CD4⁺, CD4 positive; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus.

The current recommended therapy for people infected with human immunodeficiency virus (HIV) is highly active antiretroviral therapy (HAART) that typically comprises a combination of three antiretrovirals from two different drug classes. Such regimens have been found to substantially extend life expectancy and to reduce morbidity (1–3). There are currently 20 antiretrovirals licensed for use with over 100 different combinations of the drugs in common clinical practice. While the ideal method for comparing the efficacy of the different regimens remains the randomized clinical trial, such trials are able to compare only a limited number of these many combinations of drugs. Furthermore, most randomized clinical trials aim to assess the short-term efficacy and side effects of a new combination therapy and hence are typically run for 48–72 weeks, which may be insufficient to determine the risk of long-term side effects. Hence, observational data collected in a routine clinical setting are increasingly being used to compare the short- and long-term effectiveness and risks of antiretroviral therapy.

The low rate of progression to acquired immunodeficiency syndrome and death means that it is often not practical to
use clinical outcomes to compare regimens. Therefore, most studies rely on two surrogate markers, the HIV RNA level and the CD4-positive (CD4$^+$) T-lymphocyte cell count in the blood. In observational studies, these measurements can occur at various times following the initiation of treatment, depending on the frequency of clinic visits. To compare regimens in such data, one must therefore necessarily compare the times to events using survival analysis methods rather than compare absolute changes at a given time. Common measures of the success of antiretroviral therapy include the time taken from initiation of treatment until the HIV RNA level falls below the lower limit of quantification of the assay (4–7), the time taken until viral rebound (8, 9), and the time taken for the CD4$^+$ T-cell count to rise by a given amount from its baseline value (4, 5).

During analysis of the time until one of these endpoints occurs, the true date of the endpoint is known to be between the times of only two measurements (e.g., for HIV RNA levels, the last detectable measurement and the first undetectable measurement) (figure 1). Statistically, these data are termed “interval censored.” In many of the analyses undertaken to date, the time of the first undetectable measurement is assumed to be the true time of the event (4–6). An alternative approach is to take the midpoint of the measurement times before and after the event occurred as the true time. Finally, statistical methods are available to account for the interval-censored nature of the data. In other contexts, it has been shown that the first two methods can produce biased parameter estimates (10).

In an observational clinical cohort, there is no set follow-up schedule for measurements as there would be in a traditional cohort study. Thus, the frequency of follow-up may differ between regimens. This could lead to biased estimates of the difference in effectiveness between treatments. Survival analysis methods looking at events occurring on a timescale that is of comparable size or not many times longer than the typical gap between measurements are likely to be particularly susceptible to this source of bias. In this paper, we investigate the extent to which estimates of differences between treatments could be biased because of differences in the frequency with which HIV RNA levels and CD4$^+$ T-cell counts are measured.

**MATERIALS AND METHODS**

**Data**

To parameterize our simulation model, we use data from the Athena cohort that is coordinated by the Dutch HIV Monitoring Foundation. As of March 31, 2005, the cohort contained data on 10,208 HIV-positive patients, including the majority of treated patients in the Netherlands. Patients’ data are collected at each clinic visit, which typically occurs 1 month after starting treatment and then every 3 months, although there is a great deal of variation in this schedule. The database includes information on the antiretroviral treatments taken over time, comedications used, demographic information, clinical marker data, information about clinical symptoms, side effects and toxicity, and causes of death. Parameters and characteristics of the simulated data sets were based on the 2,141 treatment-naïve patients in the Athena cohort starting a HAART regimen between 1999 and 2003.

**Simulated HIV patient data**

Data sets of HIV RNA levels and of CD4$^+$ T-cell counts for patients starting their first HAART regimen were simulated by use of a dynamic model of the interaction between HIV and human T-cells in the body (11, 12). The data were generated by numerically solving the set of differential equations that assume no difference in effectiveness between treatments ($\theta = 0.98$ for both groups, with $\theta$ as defined in the Web Appendix). (This information is described in supplementary material referred to in the text as the “Web Appendix,” which is posted on the Journal’s website (http://aje.oxfordjournals.org/).) Random variation in parameters between individuals, stochastic variation over time, and measurement error were included. Parameter values were based on the Athena cohort and on published data. The equations and parameters are listed in the Web Appendix. All our results are based on 2,000 simulated data sets.

For the first part of the study, we investigated the effect of differences between regimens in measurement frequency by simulating measurement times. To guide our choice of measurement frequency, the distribution of times between successive HIV RNA measurements in the Athena cohort was estimated by fitting separate lognormal models to the time to the first measurement, between the first and second, and so on. We then assumed that there were two subgroups, one of which visited their clinic twice as often as the other, and scaled the mean time between visits for each of these distributions accordingly. The data sets were then simulated.

**FIGURE 1.** Simulated human immunodeficiency virus (HIV) RNA data illustrating interval censoring.
with measurements sampled from these lognormal distributions. For simplicity, we assumed that CD4$^+$ T-cell counts were measured at the same visits as were the HIV RNA levels.

To investigate the effect of measurement frequency on the estimated treatment differences, we gradually increased the imbalance of the treatment assignment from these subgroups. If each subgroup has an equal probability of being assigned to either treatment, then the measurement frequencies will be the same for both treatments. However, if the probability differs between subgroups, then the measurement frequency will differ between treatments. Each data set comprised two treatment groups of 200 patients each, who were followed up for 2 years.

For the second part of the study, we used the Athena cohort data to investigate differences in measurement frequency for four comparisons defined by baseline characteristics or treatment. The mean gap between measurements may be right censored during the last gap and, hence, the overall mean was obtained from the area under the Kaplan-Meier survival curve for gap length, extending the survival curve down toward zero beyond the longest observed gap with an exponential tail (13). For each comparison, data sets were simulated of the same size as the real data sets, and the simulated data were sampled at the actual measurement times observed in the Athena cohort, which are not necessarily the same for HIV RNA and CD4$^+$ T-cell measurements. Again, there was no difference in efficacy between treatments in the simulated data. We compared four groups: nonnucleoside reverse transcriptase inhibitor-containing HAART versus protease inhibitor-containing HAART; efavirenz-containing HAART versus nevirapine-containing HAART; antiretroviral therapy (ART)-naive versus ART-experienced patients starting HAART; and patients with a baseline CD4$^+$ T-cell count of greater than or equal to 200 cells/$\mu$L versus less than 200 cells/$\mu$L. All the comparisons apart from the third were restricted to ART-naive patients.

### Analyses of outcomes

A number of survival analysis models were fitted to each simulated data set, and the accuracy of the estimates of the parameter representing the difference between treatments was assessed by calculating the bias of the parameter estimates (the extent to which they deviated from the true value of zero) and the coverage of 95 percent confidence intervals for the parameter. The coverage is defined as the proportion of times the confidence interval contains the true parameter value, which should be 95 percent. The most widely used time-to-event model is the Cox proportional hazards model. In this model, right censoring, where the event is known to have occurred only after a certain time, can be accommodated but not interval censoring. Parametric models do allow interval censoring to be modeled. Two distributions widely used in survival analysis were fitted, the Weibull and log-logistic. These are two examples of accelerated failure models, in which the log of the time to the event is written as a linear function of the covariates. The exponentiated coefficients are known as acceleration factors. The acceleration factor between two levels of a covariate is equal to the ratio of the predicted median times to the event. The Weibull model may also be written in proportional hazards form.

An alternative to Cox proportional hazards is to fit natural cubic splines to the baseline cumulative hazard in a proportional hazards model, as proposed by Royston and Parmar (14). We refer to this as the “flexible hazard” method, with 2 df used for the cumulative hazard (one more parameter than for the Weibull model). Allowing more than 2 df made little difference to the results for either endpoint.

The Cox proportional hazards models were fitted taking either the next visit after the endpoint occurred or the midpoint as the true time; for the other models, either the next visit was used (when it was first observed) or an interval-censored analysis was carried out. The analyses of the time taken for the HIV RNA level to fall below 50 copies/ml were adjusted for the baseline HIV RNA level, while the analyses of the time taken for the CD4$^+$ T-cell count to rise by 100 cells/$\mu$L from the baseline level were adjusted for both the baseline HIV RNA and the CD4$^+$ T-cell count.

### Software

Interval-censored models may be fitted using standard statistical software. The routines PROC LIFEREG in SAS software (SAS Institute, Inc., Cary, North Carolina) or SURVREG in S-PLUS software (Insightful Corporation, Seattle, Washington) fit parametric models including the Weibull and log-logistic models. Our STATA (StataCorp LP, College Station, Texas) program for interval-censored survival analysis can be downloaded from the Statistical Software Components archive (15) by keying the command “ssc install intcens” from within STATA when connected to the Internet. The STATA program for flexible proportional hazards and proportional odds models (14) can be downloaded by keying “ssc install stpm.”

### RESULTS

#### Simulated measurement times

For analyses of both the time taken for HIV RNA levels to fall below 50 copies/ml and the time taken for the CD4$^+$ T-cell count to rise by 100 cells/$\mu$L from the baseline level, taking the time of the next measurement after the event has occurred as the true time results in biased estimates of the treatment difference between the groups (figure 2, parts a–d). Using the midpoint also leads to bias, although with this method the bias is smaller. The interval-censored methods give estimates closest to the true treatment difference (in this case, no difference between the treatment arms).

When interval censoring is ignored, the bias in estimated treatment differences leads to poor coverage of the confidence intervals when there is a difference in measurement frequency (figure 2, parts e and f). For example, when one treatment group visits twice as often as the other and when a Cox proportional hazards model is fitted with the time of the next measurement after the endpoint occurs taken as the true time, the coverage of the confidence interval is only 44 percent for analyses of the time taken for HIV RNA levels to fall below 50 copies/ml and the time taken for the CD4$^+$ T-cell count to rise by 100 cells/$\mu$L from the baseline level.
FIGURE 2. Simulation results as the imbalance in visit frequency changes. An imbalance of 0.5 means that the visit frequency is the same in both groups, and a value of 1.0 means that one group visits twice as often as the other. Unless “midpoint” or “interval censored” is stated, the next measurement after the event was assumed to be the true time. The referent category is the treatment with less frequent measurements. Analyses of the human immunodeficiency virus (HIV) RNA endpoint were adjusted for baseline HIV RNA, while those of the CD4 endpoint were adjusted for both baseline HIV RNA and the CD4-positive T-cell count.
fall below 50 copies/ml and 31 percent for analyses of the time taken for CD4+ T-cell counts to rise by 100 cells/µl. For the HIV RNA endpoint, an interval-censored log-logistic model has a coverage of 90 percent even when one treatment group visits twice as often as the other.

These analyses were repeated using simulated data sets with two different values for the treatment effectiveness, \( \theta = 0.97 \) and 0.99 (refer to the Web Appendix). These values give posttreatment basic reproductive ratios of 0.75 and 0.25, respectively. The biases of estimated treatment differences were little different from the results in figure 2.

**Measurement times from the Athena cohort**

The comparisons chosen are listed in table 1, along with the frequency of clinic visits at which there was an HIV RNA measurement. Groups are designated “1” and “2,” with group 1 being the group with a longer mean gap (less frequent measurements) in each case. For all four comparisons, there are important differences in measurement frequency, particularly in the mean time to the first measurement after starting therapy.

There is no true underlying difference in the treatment efficacy in the simulated data between the two groups, so any estimated treatment difference represents a bias introduced because of differences in measurement frequency. Interval-censored analyses using a proportional hazards model with flexible baseline hazard (figure 3, parts a and b) or a log-logistic model (figure 3, parts c and d) are the most reliable, with little bias in the estimated treatment difference for the time taken for HIV RNA levels to fall below 50 copies/ml (figure 3, parts a and c). As before, these methods do not completely remove the bias in estimated treatment differences obtained by fitting the models to the time taken for CD4+ T-cell counts to increase by 100 cells/µl from baseline (figure 3, parts b and d). For all the models, interval censoring was less biased than using the next measurement after the endpoint occurred. If a Cox proportional hazards model is fitted to the time taken for HIV RNA levels to fall below 50 copies/ml (using the time of the first undetectable HIV RNA measurement as the time of event), then a boosted protease inhibitor would appear to be more effective than a nonnucleoside reverse transcriptase inhibitor, with a hazard ratio of 1.13, when in fact both treatments were equally effective. The bias resulting from ignoring the

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean time to first visit (days)</th>
<th>Mean time from first to second visit (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>NNRTI*-containing HAART*</td>
<td>Boosted protease inhibitor-containing HAART</td>
<td>52.5</td>
</tr>
<tr>
<td>Nevirapine-containing HAART</td>
<td>Efavirenz-containing HAART</td>
<td>63.1</td>
</tr>
<tr>
<td>Not ART* naïve</td>
<td>ART naïve</td>
<td>81.1</td>
</tr>
<tr>
<td>Baseline CD4 count of ≥200 cells/µl</td>
<td>Baseline CD4 count of &lt;200 cells/µl</td>
<td>42.8</td>
</tr>
</tbody>
</table>

* NNRTI, nonnucleoside reverse transcriptase inhibitor; HAART, highly active antiretroviral therapy; ART, antiretroviral therapy.

**DISCUSSION**

When there is a difference in measurement frequency between treatments, our results demonstrate that analyses taking the first measurement after the event has occurred as the true time of that event will result in biased estimates of the relative effectiveness of the two treatments. In such settings, the regimen used by the group with less frequent measurements will appear less effective than that used by the group with more frequent measurements, simply because the endpoint appears to happen later for the former group. Using the midpoint between visits as an estimate of the time of the event also results in biased estimates. Statistical methods that are able to correct for the interval-censored nature of the data give less biased results. However, they may not remove the bias completely. Two possible causes of residual bias are poor model fit and detection bias. The latter would occur because the endpoint may occur between two clinic visits but not be detected at the next visit if, for example, the HIV RNA level had gone back above the limit of detection. The event is more likely to be missed if there are less frequent measurements, resulting in bias. If the simulations are repeated with survival models fitted to the visit times on either side of the true time of the event, then interval-censored models give almost unbiased estimates (results not shown). Therefore, the residual bias in our results is mainly detection bias.

In the Athena cohort, there are substantial differences in the frequency of clinic visits between regimens. For example, the mean time to the first HIV RNA measurement was 1.39 times longer for patients initiating a HAART regimen containing a nonnucleoside reverse transcriptase inhibitor than for those initiating a HAART regimen containing a boosted protease inhibitor. Similarly, patients initiating a HAART regimen including nevirapine visited less frequently than did those initiating a HAART regimen including...
Bias of log hazard ratios (HR) (Cox, Weibull, and flexible hazard)

a) Time to HIV RNA below 50 copies/ml

b) Time to CD4 count increase of 100 cells/µl

Bias of log acceleration factors (AF) (Weibull and log-logistic)

c) Time to HIV RNA below 50 copies/ml

d) Time to CD4 count increase of 100 cells/µl

Coverage of 95% confidence intervals (CI)

e) Time to HIV RNA below 50 copies/ml

f) Time to CD4 count increase of 100 cells/µl

FIGURE 3. Simulation results using the measurement times observed from 1999 to 2003 in the Dutch Athena cohort. Unless “midpoint” or “interval censored” is stated, the next measurement after the event was assumed to be the true time. The referent category is the treatment with less frequent measurements. Analyses of the human immunodeficiency virus (HIV) RNA endpoint were adjusted for baseline HIV RNA, while those of the CD4 endpoint were adjusted for both baseline HIV RNA and the CD4-positive T-cell count. PI, protease inhibitor; ART, antiretroviral therapy; NNRTI, nonnucleoside reverse transcriptase inhibitor.

Am J Epidemiol 2006;163:676–683
efavirenz, with 1.45 times as long to the first measurement. Differences in measurement frequency of this size result in biases of 0.1 or 0.2 in estimated hazard ratios and therefore may erroneously indicate a difference in effectiveness between treatments, favoring the treatment group with more frequent measurements. The bias also leads to incorrect confidence intervals. For example, comparing those initiating a boosted protease inhibitor-containing HAART regimen with those initiating a nonnucleoside reverse transcriptase inhibitor-containing HAART regimen for the time taken for the CD4⁺ T-cell count to rise by 100 cells/μl from baseline gives a coverage of 5 percent for the hazard ratio in a Cox proportional hazards model when the next measurement is taken as the true time of the event (the usual method). This means that 95 percent of the time, such an analysis would imply that there was a statistically significant difference between the treatments, all in favor of the boosted protease inhibitor, when they were in fact equally effective. Thus, although the bias in the estimated hazard ratio of 0.1–0.2 may appear modest, it is sufficient to indicate a significant difference between the treatments in fairly moderately sized databases (1,355 patients in total for the comparison between a nonnucleoside reverse transcriptase inhibitor and a boosted protease inhibitor).

There are also differences in the measurement frequency in the Athena cohort between patients with high or low baseline CD4⁺ T-cell counts, with the mean time to the first HIV RNA measurement 1.20 times as long for the former, and between the ART experienced and the ART naïve, with the mean time to the first measurement 1.54 times as long for the former. Therefore, analyses that aim to discover the relation between either of these factors and the endpoints will be susceptible to bias if interval censoring is ignored, either if the effect of the factor is our primary interest or if we want to adjust for confounding by the factor. The estimated hazard ratio would be around 1.15 when comparing the ART naïve with the ART experienced in a Cox proportional hazards model of HIV RNA levels falling below 50 copies/ml using the next visit as the true time, if the factor had no real association with the outcome.

Although we have looked at only one possible biologic model for HIV RNA levels and CD4⁺ T-cell counts, the form of the model is unlikely to materially alter the conclusions presented here, since the bias of the different statistical methods depends mainly on three factors: the measurement schedules (derived from the Athena cohort), the timescales over which each endpoint occurs (which match the observed timescales in the Athena data), and the shape of the hazard function (which closely resembles those observed in the Athena data). However, when investigating other questions such as how successful control for confounding is or how to model treatment failure, the form of the model is likely to be more important. It is for those questions that dynamic models will be particularly valuable, as they can capture relations between variables that are nonlinear or that evolve over time, as well as inform our choice of statistical model.

Survival analysis methods have also been used to compare the longer-term outcomes of therapy, such as the time until HIV RNA levels become detectable after previously being undetectable (known as viral rebound). These events happen on a longer timescale than the endpoints investigated in this study, and the differences in measurement frequency between regimens are greater for the time to the first measurement than for the gaps between later measurements (table 1). Thus, analyses of long-term outcomes should be more robust to differences in measurement frequency. The statistical models applied in this study assume that measurement times are independent of a patient’s treatment outcome, but it is likely that the frequency of clinic visits is related to the patient’s current HIV RNA level and CD4⁺ T-cell count, which will depend on the effectiveness of treatment. Statistical methods that have been developed to deal with such informative visit times in the context of survival analysis (16) could be applied. Further research is required to evaluate the extent to which this affects the estimated treatment differences.

There are many other difficulties inherent in drawing conclusions on the relative effectiveness of HAART regimens from HIV observational databases. These include informative dropout (when the patient is more likely to discontinue therapy if it is not succeeding) and the fact that HIV RNA measurements and CD4⁺ T-cell counts are outcome measures in statistical analyses and also influence the decisions to initiate or change treatment. Statistical methods are increasingly being developed to deal with many of the issues commonly encountered in observational data (17–20). The simulation approach presented here will provide a useful tool to evaluate the impact of these biases on the conclusions drawn from naïve statistical methods, as well as the ability of the newly developed methods to overcome these biases.

ACKNOWLEDGMENTS

Dr. Ghani has received payments from Abbott Pharmaceuticals in the past for work as a consultant.

REFERENCES


Am J Epidemiol 2006;163:676–683