Published in final edited form as:

Stat Methods Med Res. 2017 August ; 26(4): 1854–1866. doi:10.1177/0962280215590432.

Accounting for dropout reason in longitudinal studies with nonignorable dropout

Camille M Moore¹, Samantha MaWhinney¹, Jeri E Forster^{1,2}, Nichole E Carlson¹, Amanda Allshouse¹, Xinshuo Wang^{1,3}, Jean-Pierre Routy⁴, Brian Conway⁵, and Elizabeth Connick⁶

¹Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Aurora, CO, USA

²Veterans Integrated Service Network 19, Mental Illness Research Education and Clinical Center, Denver VA Medical Center, Denver, CO, USA

³Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA

⁴Division of Hematology and Chronic Viral Illness Service, McGill University, Montreal, Quebec, Canada

⁵Vancouver Infectious Diseases Centre, Vancouver, British Columbia, Canada

⁶Division of Infectious Diseases, University of Colorado Denver, Aurora, CO, USA

Abstract

Dropout is a common problem in longitudinal cohort studies and clinical trials, often raising concerns of nonignorable dropout. Selection, frailty, and mixture models have been proposed to account for potentially nonignorable missingness by relating the longitudinal outcome to time of dropout. In addition, many longitudinal studies encounter multiple types of missing data or reasons for dropout, such as loss to follow-up, disease progression, treatment modifications and death. When clinically distinct dropout reasons are present, it may be preferable to control for both dropout reason and time to gain additional clinical insights. This may be especially interesting when the dropout reason and dropout times differ by the primary exposure variable. We extend a semi-parametric varying-coefficient method for nonignorable dropout to accommodate dropout reason. We apply our method to untreated HIV-infected subjects recruited to the Acute Infection and Early Disease Research Program HIV cohort and compare longitudinal CD4⁺ T cell count in injection drug users to nonusers with two dropout reasons: anti-retroviral treatment initiation and loss to follow-up.

Keywords

B-spline; dropout; HIV/AIDS; longitudinal data; nonignorable missing data; varying-coefficient model

Corresponding author: Samantha MaWhinney, Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Aurora, CO 80045, USA. sam.mawhinney@ucdenver.edu.

Conflict of interest Dr Brian Conway is conducting research for and currently receives honoraria from: AbbVie, Bristol Meyers Squibb, Janssen, Gilead, and Viiv.

1 Introduction

Longitudinal studies are ubiquitous in the medical and epidemiological literature. However, dropout is common, raising concerns of bias due to nonignorable missing data. In the presence of nonignorable dropout, estimates from traditional analyses are biased towards subjects with longer follow-up, which can obscure important relationships.^{1,2} There has been much research on selection,^{3,4} frailty,^{5,6} and mixture models^{7,8} to account for potentially nonignorable missingness by relating the longitudinal outcome to time of dropout.

In addition, many longitudinal studies encounter multiple types of missing data or reasons for dropout, such as loss to follow-up, disease progression, treatment modifications, and death. When clinically distinct dropout reasons are present, it is not clear that adjustment for reason is needed to improve bias or efficiency in the marginal estimate of the outcome or change in the outcome over time. However, investigators are often uncomfortable averaging results over dropout reason, particularly when dropout reasons and times differ by the primary exposure variable. Pauler et al.⁹ considered classifying missing data patterns by dropout reason in a pattern mixture model (PMM) and reported dropout reason specific estimates. In this example, dropout reasons included loss to follow-up, such that the patient was still alive but did not return for study visits, and death, where the additional visits did not exist in theory. Accounting for dropout reason may also allow for additional clinically important insights that increase understanding of the dropout mechanism and what drives differences between exposure groups. For example, within dropout reasons, there may be no differences between the exposure groups, but given the differing dropout reason distributions, an exposure effect may be realized in the marginal estimate. Therefore, it is clinically useful to adapt existing dropout approaches to account for dropout reason. In addition, considering dropout reason in the analysis may improve model fit, particularly if the forms of the dropout varying slopes are quite different across dropout reasons or are very complex. In this paper, we extend a semi-parametric varying-coefficient model^{10,11} that allows regression coefficients to vary smoothly according to unknown functions of dropout time to also account for dropout reason. Our approach allows for estimation of both dropout reason specific and marginal effects of time and differences in the effect of time by exposures or treatments.

2 Example

Longitudinal HIV cohorts provide an opportunity for studying the impact of illicit drug use on disease progression. After initiating antiretroviral therapy, it is well established that hard drug users have poorer outcomes, in part due to less faithful adherence to treatment regimens.¹² Illicit drug use has also been hypothesized to accelerate disease progression by directly enhancing virus replication and impairing immune responses; however, among untreated subjects, there are inconsistent findings for the effect of drug use. While *in vitro* and animal studies suggest that drugs and alcohol impair immune function and increase HIV replication,^{13–15} epidemiological studies of drug use and longitudinal CD4⁺ count and HIV-1 RNA in untreated subjects have been mixed, with some studies showing no

association,^{16–20} others finding accelerated disease progression among drug users,^{21–23} and some even finding slower progression.^{24,25} These conflicting results may be linked to differential dropout mechanisms between drug users and nonusers. In illicit drug users, multiple factors could lead to adverse outcomes, including consequences of substance abuse and liver disease, as well as sub-optimal engagement in HIV care, which could contribute to differential dropout.²⁶ Thus, study completers may have improved outcomes and be less likely to engage in drug use compared to those who drop out.²⁷

In this paper, we consider the effect of injection drug use on disease progression in untreated, HIV-infected subjects enrolled in the multicenter Acute Infection and Early Disease Research Program (AIEDRP) cohort. Given subjects are recruited with early HIV infection, dropout is primarily due to anti-retroviral treatment initiation or loss to follow-up. In the AIEDRP cohort we have the potential for both the dropout reason and the dropout time distribution to vary by the exposure variable, injection drug use, providing motivation to accommodate both dropout reason and dropout time in the analysis.

3 Statistical methods

The use of mixtures of random effects models to adjust for potentially nonignorable dropout in longitudinal studies has been described by several authors.^{8,28–30} Mixture models account for dropout by factoring the joint distribution of the outcome, y , and dropout time, u , as the product of the conditional distribution of the outcome for a given dropout time and the distribution of dropout times, $f(y, u) = f(y|u)f(u)$, and the full-data response distribution $f(y)$ is given by $\int f(y|u)dF(u)$.

Varying coefficient models (VCM)³¹ provide a general framework for fitting the conditional distribution $f(y|u)$ by modeling regression coefficients as functions of dropout time. Wu and Bailey's conditional linear model (CLM), which assumes regression coefficients are polynomial functions of dropout time, as well as PMM's, which assume regression coefficients depend on a discrete set of dropout patterns, can be viewed as special cases of VCMs.¹ Semi-parametric VCM's utilizing penalized splines and natural cubic B-splines have also been proposed.^{11,32} Unlike PMMs, these models allow for dropout at any continuous point in time and are also more flexible than the CLM, only requiring that regression coefficients are smooth, continuous functions of dropout time. Since outcomes for a subject are not observed after his or her dropout time, u , mixture models require assumptions about the behavior of the unobserved outcomes that occur after u . It is usual to assume that the relationship between the outcome and time is the same prior to and after u , and sensitivity analyses are required to determine if results are robust to reasonable violations of this assumption.

3.1 Extending varying-coefficient models to account for dropout reason

We extend semi-parametric VCMs that account for dropout time to adjust for dropout reason as well. Let $h = \{1, \dots, H\}$ denote dropout reason, $g = \{1, \dots, G\}$ represent group, such exposure or treatment groups that are to be compared, m_g be the number of subjects in group g , and $m_{h|g}$ be the number of subjects with dropout reason h in group g , where the i th subject has n_i observations and dropout time u_i . The joint distribution of y , u , and h , is $f(y, u, h|g) = f$

$(y|g, h, u) f(u|g, h) p(h|g)$, where the first term is the outcome model given dropout time and reason, the second is the model of dropout times given dropout reason, and the third the probability of dropout reason h . We follow a traditional longitudinal modeling framework, where y , given dropout time, reason, and group is assumed to be normally distributed, and within group, dropout reason is multinomial. The distribution of dropout times given reason and group is unspecified.

The subject-specific model of a continuous outcome is

$$(Y_i|g_i, h_i, u_i) = \mathbf{1}_i \beta_{g_i h_i 0}(\mathbf{u}_i) + \mathbf{t}_i \beta_{g_i h_i 1}(\mathbf{u}_i) + \mathbf{C}_i \beta_c + \mathbf{1}_i \alpha_{i0} + \mathbf{t}_i \alpha_{i1} + \epsilon_i \quad (1)$$

where $\mathbf{Y}_i$, $\mathbf{1}_i$, $\mathbf{t}_i$ and ϵ_i are $n_i \times 1$ vectors of outcomes, 1's, observation times and normally distributed errors, respectively. $\beta_{g_i h_i 0}(u_i)$ and $\beta_{g_i h_i 1}(u_i)$ are the dropout-varying coefficients for subjects in group g with dropout reason h . $\mathbf{C}_i$ is an $n_i \times p$ matrix of covariates, which may also include covariate interactions with time, and β_c are the associated coefficients. The

random intercept, α_{i0} , and slope, α_{i1} , are distributed $N(0, D)$, where $D = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$. This model reduces to a standard linear mixed model, which does not depend on dropout time or reason, if $\beta_{g_i h_i 0}(u) = \beta_{g_i 0}$ and $\beta_{g_i h_i 1}(u) = \beta_{g_i 1}$.

3.2 Calculation of dropout reason-specific and marginal effects

The expected value of the outcome at time t given dropout reason, group and the covariates is

$$E[Y(t)|h, g, \mathbf{C}] = \int [\mathbf{C} \beta_c + \beta_{g h 0}(u) + t \beta_{g h 1}(u)] dF(u|g, h, \mathbf{C}) \quad (2)$$

$$= \mathbf{C} \beta_c + \int \beta_{g h 0}(u) dF(u|g, h, \mathbf{C}) + t \int \beta_{g h 1}(u) dF(u|g, h, \mathbf{C}) \quad (3)$$

$E[Y(t)|g, h, \mathbf{C}]$ is also a linear function of time, so that the marginal coefficients for each dropout reason and group combination, $\beta_{gk}(h)$, are given by

$$\beta_{gk}(h) = \int \beta_{g h k}(u) dF(u|g, h, \mathbf{C}) \quad (4)$$

where $k = 0$ for the intercept and $k = 1$ for the slope.

If $R(u|g, h, \mathbf{C}) = R(u|g, h)$ meaning that the distribution of dropout times depends only on group and dropout reason and not on additional covariates, then $\beta_{gk}(h) = E[\beta_{g h k}(u)|g, h]$ and the dropout reason specific slope can be estimated from the distribution of dropout times

in group g with dropout reason h . Define $\mathbf{u}_{h|g}^0 = (\mathbf{u}_{1|h|g}^0, \dots, \mathbf{u}_{R_{h|g}}^0)^T$ as the vector of $R_{h|g}$ ordered dropout times for subjects with dropout h in group g , $\Pi_{\mathbf{u}|h,g}$ the vector of $R_{h|g}$

proportions of subjects with dropout reason h in group g with each dropout time $u_{r_{h|g}}^0$ (vector of proportions with denominator $m_{h|g}$), and $\Pi_{h|g}$ as the vector of H proportions of subjects with each dropout reason h in group g (vector of $\frac{m_{h|g}}{m_g}$). The marginal coefficients for each dropout reason are then estimated

$$\hat{\beta}_{gk}(h) = \int \hat{\beta}_{ghk}(u) d\hat{F}(u|g, h) \quad (5)$$

$$= \hat{\Pi}_{u|h,g}^T \hat{\beta}_{ghk}(u_{h|g}^0) \quad (6)$$

This is a weighted average of the dropout varying coefficients over the unique dropout times for dropout reason h in group g . It is equivalent to taking the average of the dropout varying coefficients for each subject with dropout reason h in group g

$$\hat{\beta}_{gk}(h) = \frac{1}{m_{h|g}} \sum_{i=1}^{m_{h|g}} \hat{\beta}_{ghk}(u_i) \quad (7)$$

Marginal coefficients averaged over dropout reason for each group, β_{gk}^* , can be obtained as well

$$\beta_{gk}^* = \sum_{h=1}^H p(h|g) \int \beta_{ghk}(u) dF(u|g, h) \quad (8)$$

$$= \sum_{h=1}^H p(h|g) \beta_{gk}(h) \quad (9)$$

$$\hat{\beta}_{gk}^* = \hat{\Pi}_{h|g}^T \hat{\beta}_{gk}(h) \quad (10)$$

$$= \frac{1}{m_g} \sum_{i=1}^{m_g} \hat{\beta}_{gk}(h_i) \quad (11)$$

where $\hat{\beta}_{gh}(h)$ is the $H \times 1$ vector of estimated dropout reason specific coefficients for group g . These coefficients can be interpreted as the population average of the outcome at baseline

$(\hat{\beta}_{g0}^*)$ and the population average change in the outcome per unit time $(\hat{\beta}_{g1}^*)$ for group g given other covariates in the model.

If the assumption that the distribution of dropout times does not depend on the covariates is inappropriate, it may not always be possible to easily estimate marginal coefficients,^{1,33} particularly in more complex cases where the distribution of dropout times may depend on continuous covariates or several different covariates. However in simple cases, where the distribution of dropout times depends on a few categorical covariates, using the empirical distribution of dropout times is a straightforward method that does not require distributional assumptions.

3.3 Differing forms of the dropout time and reason varying-coefficient model

Depending on assumptions, varying coefficient models can account for group, dropout time and reason in several different frameworks (Figure 1). $H = 1$ provides the traditional varying-coefficient model accounting for dropout time, but not reason (Figure 1(a)). In equation (1), each dropout reason and group combination is allowed to have a distinct functional form for the dropout-varying coefficients (Figure 1(b)). Special cases include permitting the dropout-varying coefficients to include a group effect that depends on dropout reason but not dropout time or assuming the effect of group does not depend on dropout reason (Figure 1(c) and (d), respectively). These models may be advantageous when sample sizes are low in certain dropout reason and group combinations, which may make it unreasonable to estimate a distinct functional form of the dropout varying coefficients for all dropout reason and group combinations.

3.4 Fitting the varying-coefficient model with natural cubic B-splines

As in Forster et al.,¹¹ this method utilizes natural cubic B-spline basis functions. Here, the lower boundary knot can be shifted inwards to increase the stability of coefficients for early dropout times where data may be sparse. Let $u_{h|g}$ be the set of dropout times observed in group g with dropout reason h . The conditional varying-coefficient model accounting for dropout time and reason is

$$(Y_i | g_i, h_i, u_i) = 1_i \sum_{j=0}^{J_{gh0}} \theta_{ghj0} \tilde{u}_{ghij0} + t_i \sum_{j=0}^{J_{gh1}} \theta_{ghj1} \tilde{u}_{ghij1} + C_i \beta_c + 1_i \alpha_{i0} + t_i \alpha_{i1} + \epsilon_i \quad (12)$$

where $\tilde{u}_{ghijk} = \tilde{B}(u_{h|g}, J_{ghk})_{[i,j+1]}$ for $k = 0, 1$. For $j > 0$, $\tilde{B}(u_{h|g}, J_{ghk})$ is the matrix of natural cubic B-spline basis functions with J_{ghk} degrees of freedom and $\tilde{B}(u_{h|g}, J_{ghk})_{[1,1]} = 1$. This matrix can be calculated in R using the `ns()` function in the `splines` package.

$\tilde{B}(u_{h|g}, J_{ghk})_{[i,j+1]}$ indicates the element in the i th row and $j + 1$ th column of the $\tilde{B}(u_{h|g}, J_{ghk})$ matrix. The i th subjects dropout-varying-coefficients are given by

$$\beta_{ghik} = \sum_{j=0}^{J_{ghk}} \theta_{ghjk} \tilde{u}_{ghijk} = \sum_{j=0}^{J_{ghk}} \theta_{ghjk} \tilde{B}(u_{h|g}, J_{ghk})_{[i,j+1]} \quad (13)$$

3.5 Estimation

These models can be fit in SAS Proc Mixed or R packages nlme and lme4 using restricted maximum likelihood estimation or maximum likelihood estimation. As the marginal estimates depend on an empirical dropout distribution, a bootstrap is used to obtain estimates of the standard errors and 95% confidence intervals (CI). R and SAS code to fit this model on a simulated dataset are available in the online Web Appendix (available at: <http://smm.sagepub.com/>).

4 Adjusting for dropout in the association of injection drug use and CD4⁺ T cell decline

To demonstrate the impact of accounting for dropout time and reason on the analysis of CD4⁺ count in untreated HIV, we analyze data from untreated subjects enrolled in the Acute Infection and Early Disease Research Program (AIEDRP). Investigating disease progression in untreated populations eliminates the well-described problem of less faithful adherence to treatment regimens among injection drug users compared to others.³⁴ We hypothesized that after accounting for dropout time and reason, untreated HIV seropositive injection drug users experience steeper declines in CD4⁺ count compared to untreated nonusers.

AIEDRP was a multicenter, observational cohort study of subjects identified during early HIV infection.³⁵ Inclusion criteria have been described elsewhere.³⁵ Subjects were recruited between 1997 and 2007 at sites throughout the United States, Australia, Canada, and Brazil and, with physician guidance, self-selected when and whether to initiate anti-retroviral therapy. The study closed in 2008. Subjects were classified at enrollment with either acute or recent HIV infection if they presented within 2 weeks or 3 weeks to 12 months of seroconversion, respectively.³⁵ Since all subjects were enrolled during early infection, the AIEDRP dataset presents a unique opportunity to compare longitudinal outcomes between injection drug users and nonusers while avoiding differences in disease progression due to varying infection duration. CD4⁺ counts generally decline during the high titer viremia of acute infection, and subsequently recover, at least partially, in concert with development of the immune response and declines in viremia.³⁶ Subsequently, CD4⁺ counts decline over time in untreated individuals. Subjects in both the acute and recent HIV infection groups demonstrated these patterns of early CD4⁺ count decline and recovery followed by long term declines. Fluctuations ceased for the majority of subjects by 5 weeks after enrollment (data not shown). Therefore, only observations obtained after week 5 were included in the analysis. Nadir CD4⁺ count from this initial 5-week period was adjusted for in all analyses.

Data from 1059 AIEDRP subjects who did not immediately initiate antiretroviral therapy and who had at least one qualifying HIV-1 RNA measurement were evaluated. Baseline data were collected on race/ethnicity, age, sex, HIV risk factors, HIV-1 RNA, and CD4⁺ count.

Descriptive statistics are presented in Table 1. Similar to other HIV seroconverter studies,³⁷ the AIEDRP cohort was biased towards white, homosexual men with a smaller proportion of women, non-whites, and injection drug users than the overall population of HIV-infected individuals in the AIEDRP countries. Many AIEDRP subjects were also recruited into clinical trials, such that subjects were also more likely to initiate therapy. Seventy-six (7.2%) subjects were classified as injection drug users according to self-reported HIV-risk behaviors at study entry. As subjects were recent seroconverters, this indicates recent exposure; however, since self-reported injection drug use was collected only at baseline, individuals who initiated or stopped injecting drugs after study entry, or misrepresented their drug history may be incorrectly categorized. In addition, we are unable to consider other illicit drug use categories as these data were not collected.

HIV-1 RNA and CD4⁺ count were measured at 2, 4, and 12 weeks, and then every 12 weeks until week 168, and every 24 weeks thereafter. Since few subjects had greater than 3 years of untreated follow-up, the analysis was limited to visits occurring within 3 years of enrollment, and subjects that had a full 3 years of follow-up were considered “completers” and assigned a dropout time of 1096 days. The most recent log₁₀(HIV-1 RNA) between enrollment and the first CD4⁺ count included in the model (viral load) and the nadir CD4⁺ count from the initial 5-week period were considered as predictors in a model of log_e(CD4⁺).³⁸

4.1 Methods

Subjects had incomplete untreated CD4⁺ count data if they were lost to follow-up or initiated antiretroviral therapy. Overall, injection drug users tended to dropout earlier than nonusers and were less likely to be removed from the analysis due to treatment initiation (Table 1). Among those who started treatment, injection drug users and other subjects had a similar range of dropout times; however, among those lost to follow-up, injection drug users dropped out earlier. Graphical analyses of subject-specific trajectories of log_e(CD4⁺) suggested that earlier dropouts had steeper declines in CD4⁺ count than completers (Figure 2). Since the outcome is related to dropout time, dropout may be nonignorable. As there are two distinct reasons for dropout, dropout reason is also considered.

Log_e(CD4⁺) was modeled as a function of injection drug use and time with dropout reason and dropout time-varying B-spline bases for the slopes in the fixed effects and a random intercept and slope. A separate dropout-varying slope was assumed for subjects lost to follow-up and those who started treatment. An interaction between injection drug use, dropout reason, and time was included to allow injection drug users to have different changes in CD4⁺ count over time compared to nonusers who dropped out for the same reason. Additional covariates included age, race, sex, disease status (acute vs. recent HIV infection groups), log_e(nadir CD4⁺), log₁₀(viral load) and time interactions with disease status, log_e(nadir CD4⁺), and log₁₀(viral load). For B-splines, a maximum of 5 degrees of freedom was considered for each dropout-varying effect, with knots placed at the quantiles of the dropout distribution. Models with varying degrees of freedom were fit using maximum likelihood and a model was chosen based on Akaike information criterion (AIC). A final model was fit using restricted maximum likelihood estimation. For stable slope

estimation, the lower boundary knot was set to week 24, corresponding to approximately four observations per subject. The final model had the following form

$$\log_e(\text{CD4}^+)_i = \mathbf{1}_i \beta_{g_i h_i 0} + \beta_{g_i h_i 1} \mathbf{t}_i + \beta_{h_i}(\mathbf{u}_i) \mathbf{t}_i + \mathbf{C}_i \beta_c + \mathbf{1}_i \alpha_{i0} + \mathbf{t}_i \alpha_{i1} + \epsilon_i$$

This is similar to the model shown in Figure 1(c). To test the hypotheses, 1000 bootstrap samples were drawn. Standard errors, estimated as the standard deviation of the bootstrap estimates, were used to perform Z-tests, and calculate 95% CIs. A traditional mixed model that did not account for dropout ($\log_e(\text{CD4}^+)_i = \mathbf{1}_i \beta_{g_i 0} + \beta_{g_i 1} \mathbf{t}_i + \mathbf{C}_i \beta_c + \mathbf{1}_i \alpha_{i0} + \mathbf{t}_i \alpha_{i1} + \epsilon_i$) and a model that accounted only for dropout time but not reason ($\log_e(\text{CD4}^+)_i = \mathbf{1}_i \beta_{g_i 0} + \beta_{g_i 1} \mathbf{t}_i + \beta_2(u_i) \mathbf{t}_i + \mathbf{C}_i \beta_c + \mathbf{1}_i \alpha_{i0} + \mathbf{t}_i \alpha_{i1} + \epsilon_i$) were fit to compare results.

5 Results

5.1 Accounting for dropout time and reason

For $\log_e(\text{CD4}^+)$, dropout-varying slopes for representative subjects are shown in Figure 3 with steeper declines (more negative slope) associated with earlier dropout. Controlling for age, sex, race, $\log_e(\text{nadir CD4}^+)$, viral load, and acute infection, injection drug use was associated with more rapid declines in CD4^+ count (Table 2 and Figure 4). For an injection drug user with recent infection, median nadir CD4^+ count of 500 and median viral load of 4:45 $\log_{10}(\text{copies/mL})$, CD4^+ counts declined by 31.2% per year (95% CI: 24.3%, 37.4%), compared to 24.4% per year (95% CI: 21.0%, 27.6%) for a nonuser.

Consistent with concurrent treatment guidelines, steeper CD4^+ count decline was also associated with treatment initiation. Subjects that were removed from the analysis at treatment initiation demonstrated a more rapid CD4^+ count decline of 38.3% per year (95% CI: 31.8%, 44.2%) than those who were lost to follow-up with a decline of 15.4% per year (95% CI: 12.1%, 18.5%), again assuming median nadir CD4^+ count, viral load and recent infection. Among subjects lost to follow-up, injection drug users had significantly greater declines in CD4^+ count than nonusers. Among subjects who initiated therapy, the same trend was observed, although failed to reach statistical significance despite having a larger estimated difference (Table 2). This difference could potentially reflect a treatment bias towards initiating treatment earlier in disease progression in nonusers.^{39,40}

Limitations of this analysis include the relatively small number of injection drug users ($n = 76$) and that the AIEDRP study only collected injection drug use as a risk factor for HIV and did not collect information on noninjection drug use, such as cocaine and morphine. We are therefore unable to quantify injection drug use exposure or identify whether a subset of subjects also had utilized stimulants and/or opiates. Thus, it is possible that a subgroup of injection drug using subjects had increased exposure or used additional drugs that would also exhibit an effect on disease progression. In addition, only subjects that had a viral load measurement were included in this analysis as viral load is an important predictor of CD4^+ decline;³⁸ however, a sensitivity analysis without adjusting for baseline viral load including all subjects found similar declines and differences between injection drug users and others.

5.2 Comparison to other models

Using a model that does not account for dropout time or reason, changes in $CD4^+$ count are reduced and the effect of injection drug use is attenuated and no longer statistically significant; drug use is associated with only a $0.065 \log(\text{cells}/\text{mm}^3)$ steeper decline in $\log_e(CD4^+)$ per year (95% CI: $-0.135, 0.005$; $p = 0.07$). Using the model that accounts for dropout time and reason, the magnitude of the drug use association is 45% larger (Figures 5 and 6) than the unadjusted model. Additionally, in the extended model estimated changes in $\log_e(CD4^+)$ for nonusers and drug users are 59.8% and 56.0% greater, respectively. Using the standard mixed model, injection drug users are estimated to have declines of only 21.3% per year (95% CI: 15.4–26.8%) and nonusers were estimated to have 16.0% per year (95% CI: 14.2%, 17.8%). These results suggest that models that do not account for dropout may be overly optimistic and underestimate changes in $CD4^+$ count over time for both injection drug users and nonusers. Using the standard mixed model that does not account for dropout, the average subject would be estimated to have a $CD4^+$ count of $< 200 \text{ cells}/\text{mm}^3$, the clinical definition of AIDS, at 4.2 years and 5.8 years for drug users and nonusers respectively; while using the model that accounts for dropout time and reason, this amount of decline would be expected in just 2.8 years and 3.7 years for injection drug users and nonusers, respectively.

Estimates from the model that accounts for only dropout time are slightly attenuated compared to the extended model that accounts for dropout reason as well. Differences in the change in $\log_e(CD4^+)$ between drug users and nonusers are 13.7% larger using the extended model, and the estimated changes in $\log_e(CD4^+)$ are 5.8% and 7.7% larger for nonusers and drug users, respectively. The extended model that accounts for dropout time and reason has increased flexibility compared to the model that only accounts for dropout time, since each dropout reason is allowed a separate natural cubic B-spline. In this example, the single natural cubic B-spline used in the model that accounted for dropout time only was flexible enough to fit the mean change in $\log_e(CD4^+)$ at each dropout time, and produced fairly similar results to the extended model. However, the extended model may improve model fit, particularly if the forms of the dropout varying slopes are quite different across dropout reasons or are very complex.

5.3 Sensitivity analysis

These models assume that subjects continue on the same linear trajectory of $\log_e(CD4^+)$ after dropout. If rather than continuing on the same $CD4^+$ trajectory after dropout, subjects had more rapid declines in $CD4^+$, then $CD4^+$ declines and differences between injection drug users and others would be expected to be more extreme than predicted by this analysis. However, if declines in $CD4^+$ count attenuate after dropout, differences between injection drug users and nonusers may not remain statistically significant. The sensitivity of the results to this assumption was evaluated by determining the impact of a proportional attenuation, denoted $\delta(u)$, of $\log_e(CD4^+)$ decline after dropout.^{11,32} For example, for $\delta(u) = 0.5$, a subject with an estimated decline of $-0.4 \log_e(CD4^+)$ per year would be assigned a slope of $-0.2 \log_e(CD4^+)$ per year after dropout. $\delta(u) = 0$ is equivalent to the estimated $CD4^+$ count at a subject's last observation carried forward, while $\delta(u) = 1$ results in the assumption that a subject continues on the same trajectory after dropout. $\delta(u) = 0$ serves as

an unrealistic lower bound, as $CD4^+$ counts are known to decline over time in untreated subjects. For $\delta(u) = 0.5, 0.25, 0$, injection drug users have steeper declines in $CD4^+$ count than nonusers. For $\delta(u) = 0.5$, differences between injection drug users and nonusers at 2 and 3 years remain statistically significant and differences at 3 years remain significant with $\delta(u) = 0.25$ (Table 3).

6 Discussion

Longitudinal studies are common in the literature with the potential for nonignorable dropout. In practice, we often encounter clinically distinct dropout reasons with the dropout time distribution depending on both the primary exposure variable and dropout reason. We propose a method to accommodate both dropout reason and dropout time that allows for estimation of both dropout reason specific and marginal effects for each exposure group. While accounting for dropout reason in addition to dropout time may not necessarily result in large changes to the marginal estimates compared to a model that accounts for only dropout time, adjusting for reason does allow for additional clinically important insights that may increase understanding of the dropout mechanism and what drives differences between exposure groups. For example, in our analysis of the effect of injection drug use on $CD4^+$ decline in untreated subjects in the AIEDRP, accounting for dropout time and reason revealed steeper declines and larger differences between injection drug users and others compared to a standard linear mixed model. By accounting for dropout reason in addition to time, we found that subjects that were removed from the analysis at treatment initiation demonstrated more rapid declines than those who were lost to follow-up, consistent with the concurrent treatment guidelines. Among subjects lost to follow-up, injection drug use was associated with significantly greater $CD4^+$ declines. Among subjects who initiated treatment, the same trend was observed, although failed to reach statistical significance despite having a larger estimated difference; this difference could potentially reflect a treatment bias towards initiating treatment in nonusers.

This approach can be implemented using standard software without imposing distributional assumptions on the dropout mechanism, and is broadly applicable to the analysis of longitudinal cohort data, where nonignorable dropout and different dropout reasons may obscure important relationships. In addition, this approach could also be utilized to treat administratively censored subjects and study completers separately from subjects that dropped out.^{33,41} Increasingly, longitudinal cohort data are being mined to investigate health-related questions and ultimately to develop public policy. Therefore, optimizing analysis of longitudinal cohort data is of the utmost importance.

Acknowledgements

The authors would like to thank Dr Martin Markowitz, Dr Susan Little, Dr Richard Hecht, Dr Eric S Daar, Dr Ann C Collier, Dr Joseph Margolick, Dr Michael Kilby, Dr John Kaldor, Dr Jay Levy, Dr Robert Schooley, Dr David Cooper, Dr Bruce Walker, and Dr Douglas Richman for their contributions to the AIEDRP study. The authors would also like to thank the reviewers for providing valuable suggestions and insights that helped improve this article.

We acknowledge the participating sites: United States: University of Minnesota, Minneapolis, MN; University of Cincinnati, Cincinnati, OH; Northwestern University, Chicago, IL; Rush University, Chicago, IL; SUNY Downstate, Brooklyn, NY; Columbia University, NY, NY; Fenway Community Health, Boston, MA; Community

Research Initiative of New England, Boston, MA; Brigham and Women's Hospital, Boston, MA; Beth Israel Medical Center, Boston, MA; University of Pennsylvania, Philadelphia, PA; Vanderbilt University, Nashville, TN; Duke University Medical Center, Durham, NC; University of Rochester, Rochester, NY; University of Texas Southwestern, Dallas, TX; University of Colorado Denver, Aurora, CO; Aaron Diamond AIDS Research Center, Rockefeller University, New York, NY; University of California, San Diego, San Diego, CA; Cedars-Sinai, Los Angeles, CA; University of California, San Francisco, San Francisco, CA; Los Angeles Biomedical Research Institute at Harbor-University of California, Los Angeles Medical Center, Torrance, CA; University of Washington Primary Infection Clinic, Seattle, WA; Johns Hopkins University, Baltimore, MD; University of Alabama, Birmingham, AL; Partners AIDS Research Center, Boston, MA; Canada: McGill University Health Centre, Montreal; University of British Columbia, Vancouver; Australia: The Centre Clinic, St Kilda, VIC; Prahran Market Clinic, St Kilda, VIC; Carlton Clinic, Carlton, VIC; Taylor Square Private Clinic, Darlinghurst, NSW; 407 Doctors, Darlinghurst, NSW; Holdsworth House General Practice, Darlinghurst, NSW; St Vincent's Hospital, Darlinghurst, NSW; The Alfred Clinic, Prahran, VIC; Melbourne Sexual Health Clinic, Carlton, VIC; AIDS Research Initiative, Darlinghurst, NSW; Brazil: Centro de Referencia Estadual de AIDS, Salvador, Bahia.

Funding This work was supported by the National Institutes of Health, National Institute on Drug Abuse (grant numbers DA026743, DA030495, DA037778) and National Institute of Allergy and Infectious Diseases (grant numbers AI41532, AI41531, AI41535, AI43638, AI41535 AI57005, AI41536, AI43271, AI41530, AI41534, AI52403, and AI57005).

References

1. Daniels, MJ., Hogan, JW. Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis. Chapman and Hall/CRC; Boca Raton: 2008.
2. Little, RJA., Rubin, DB. Statistical analysis with missing data. John Wiley and Sons, Inc; New York: 2002.
3. Diggle, PJ., Liang, KY., Zeger, SL. Analysis of longitudinal data. Clarendon Press; Oxford: 1994.
4. Kenward MG. Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Stat Med*. 1998; 17:2723–2732. [PubMed: 9881418]
5. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995; 51:151–168. [PubMed: 7766771]
6. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Stat Med*. 1992; 11:1861–1870. [PubMed: 1480878]
7. Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *JASA*. 1977; 72:538–543.
8. Wu MC, Bailey K. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*. 1989; 45:939–955. [PubMed: 2486189]
9. Pauler DK, McCoy S, Moynour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Stat Med*. 2003; 22:795–809. [PubMed: 12587106]
10. Liang H, Wu H, Carroll RJ. The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*. 2003; 4:297–312. [PubMed: 12925523]
11. Forster JE, MaWhinney S, Ball EL, et al. A varying-coefficient method for analyzing longitudinal clinical trials data with nonignorable dropout. *Contemp Clin Trials*. 2012; 33:378–385. [PubMed: 22101223]
12. Lucas GM, Cheever LW, Chaisson RE, et al. Detrimental effects of continued illicit drug use on the treatment of HIV-1 infection. *J Acquir Immune Defic Syndr Hum Retrovirol*. 2001; 27:251–259.
13. Amedee AM, Nichols WA, Robichaux S, et al. Chronic alcohol abuse and HIV disease progression: Studies with the non-human primate model. *Curr HIV Res*. 2014; 12:243–253. [PubMed: 25053367]
14. Banerjee A, Strazza M, Wigdahl B, et al. Role of muopiods as cofactors in human immunodeficiency virus type 1 disease progression and neuropathogenesis. *J Neurovirol*. 2011; 17:291–302. [PubMed: 21735315]
15. Wang X, Ho WZ. Drugs of abuse and HIV infection/replication: Implications for mother-fetus transmission. *Life Sci*. 2011; 88:972–979. [PubMed: 21056582]
16. Margolick JB, Munoz A, Vlahov D, et al. Direct comparison of the relationship between clinical outcome and change in CD4+ lymphocytes in human immunodeficiency virus-positive

- homosexual men and injecting drug users. *Arch Intern Med.* 1994; 154:869–875. [PubMed: 7908795]
17. Italian Seroconversion Study. Disease progression and early predictors of AIDS in HIV-seroconverted injecting drug users. The Italian Seroconversion Study. *AIDS.* 1992; 6:421–426. [PubMed: 1616637]
 18. Pezzotti P, Galai N, Vlahov D, et al. Direct comparison of time to AIDS and infectious disease death between HIV seroconverter injection drug users in Italy and the United States: Results from the ALIVE and ISS studies. *J Acquir Immune Defic Syndr Hum Retrovirol.* 1999; 20:275–282. [PubMed: 10077177]
 19. von Overbeck J, Egger M, Smith GD, et al. Survival in HIV infection: Do sex and category of transmission matter? Swiss HIV Cohort Study. *AIDS.* 1994; 8:1307–1313. [PubMed: 7802985]
 20. Chaisson RE, Keruly JC, Moore RD. Race, sex, drug use, and progression of human immunodeficiency virus disease. *N Engl J Med.* 1995; 333:751–756. [PubMed: 7643881]
 21. DesJarlais DC, Friedman SR, Marmor M, et al. Development of AIDS and HIV seroconversion and potential co-factors for T4 cell loss in a cohort of intravenous drug users. *AIDS.* 1987; 1:105–111. [PubMed: 2896511]
 22. Weber R, Ledergerber B, Opravil M, et al. Progression of HIV infection in misusers of injected drugs who stop injecting or follow a programme of maintenance treatment with methadone. *BMJ.* 1990; 301:1362–1365. [PubMed: 1980219]
 23. Meijerink H, Wisaksana R, Iskandar S, et al. Injecting drug use is associated with a more rapid CD4 cell decline among treatment naïve HIV-positive patients in Indonesia. *J Int AIDS Soc.* 2014
 24. Farzadegan, H., Levy, D., Astemborski, et al. Effect of gender, race, injecting drug use and disease stage on infectious viral load among IDUs and gay men. XI international conference on AIDS; Vancouver, Canada. 1996.
 25. Donahoe RM, Vlahov D. Opiates as potential cofactors in progression of HIV-1 infections to AIDS. *J Neuroimmunol.* 1998; 83:77–87. [PubMed: 9610676]
 26. Murray M, Hogg RS, Lima VD, et al. The effect of injecting drug use history on disease progression and death among HIV-positive individuals initiating combination antiretroviral therapy: Collaborative cohort analysis. *HIV Med.* 2012; 13:89–97. [PubMed: 21819529]
 27. Lanoya E, Mary-Krausea M, Tattavinb P, et al. Predictors identified for losses to follow-up among HIV-seropositive patients. *JCE.* 2006; 59:829–835. [PubMed: 16828676]
 28. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative censoring by modeling the censoring process. *Biometrics.* 1988; 44:175–188.
 29. Mori M, Woodworth GG, Woolson RF. Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Stat Med.* 1992; 11:621–631. [PubMed: 1594805]
 30. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Stat Med.* 1997; 16:239–257.
 31. Hastie T, Tibshirani R. Varying-coefficient models. *J Roy Stat Soc Ser B.* 1993; 55:757–796.
 32. Hogan JW, Lin X, Herman B. Mixtures of varying-coefficient models for longitudinal data with discrete or continuous nonignorable dropout. *Biometrics.* 2004; 60:854–864. [PubMed: 15606405]
 33. Su L, Hogan JW. Varying-coefficient models for longitudinal processes with continuous-time informative dropout. *Biostatistics.* 2010; 11:93–110. [PubMed: 19837655]
 34. Westergaard R, Hess T, Astemborski J, et al. Longitudinal changes in engagement in care and viral suppression for HIV-infected injection drug users. *AIDS.* 2013 Epub ahead of print.
 35. Meditz AL, MaWhinney S, Amanda Allshouse A, et al. Sex, race, and geographic region influence clinical outcomes following primary HIV-1 infection. *J Infect Dis.* 2011; 203:442–451. [PubMed: 21245157]
 36. Schacker T, Collier AC, Hughes J. Clinical and epidemiologic features of primary HIV infection. *Ann Intern Med.* 1996; 125:257–264. [PubMed: 8678387]
 37. Connick E, MaWhinney S, Wilson CC, et al. Challenges in the study of patients with HIV type 1 seroconversion. *Clin Infect Dis.* 2005; 40:1355–1357. [PubMed: 15825039]

38. Mellors JW, Munoz A, Giorgi JV, et al. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann Intern Med.* 1997; 126:946–954. [PubMed: 9182471]
39. Peterson PK, Gekker G, Hu S, et al. Morphine amplifies HIV-1 expression in chronically infected promonocytes cocultured with human brain cells. *J Neuroimmunol.* 1994; 50:167–175. [PubMed: 8120138]
40. Celentano DD, Galai N, Sethi AK, et al. Time to initiating highly active antiretroviral therapy among HIV-infected injection drug users. *AIDS.* 2001; 15:1707–1715. [PubMed: 11546947]
41. Li J, Schluchter MD. Conditional mixed models adjusting for non-ignorable drop-out with administrative censoring in longitudinal studies. *Stat Med.* 2004; 23:3489–3503. [PubMed: 15505888]

Figure 1.

Examples of relationships between dropout time, reason, and group. Panel A depicts a varying-coefficient model with a group effect that does not account for dropout reason. In Panel B, dropout reason is accounted for and a different functional form of the slope is allowed for each dropout reason and group combination. In Panel C, the functional form of the slope depends only on dropout reason and not group. This is the model used to account for dropout time and reason in the application. In Panel D, the functional form of the slope depends only on dropout reason, and in addition, the effect of group is assumed to be the same across dropout reasons.

Figure 2.

Subject-specific least squares estimates of change in $\log_e(\text{CD4}^+$ T cell count) per year by dropout time with loss fit, untreated subjects, Acute Infection and Early Disease Research Program, 1997–2007.

Figure 3. Predicted change in $\ln(\text{CD4}^+ \text{ T cell count})$ by dropout time assuming recent infection, Nadir $\text{CD4}^+ \text{ T Cell Count}$ of 500 and median $\log_{10}(\text{HIV-1 RNA})$ of 4.45, untreated subjects, Acute Infection and Early Disease Research Program, 1997–2007.

Figure 4.

Predicted CD4⁺ T cell count over time for an untreated 35-year-old White male with Nadir CD4⁺ T cell count of 500 and baseline log₁₀(HIV-1 RNA) of 4.45 by Dropout Reason and Drug Use, Acute Infection and Early Disease Research Program, 1997–2007. Panel A depicts estimates for recent infection and panel B for acute infection.

Figure 5. Predicted CD4⁺ T cell count for a 35-year-old White male with recent infection with Nadir CD4⁺ T cell count of 500 and baseline log₁₀(HIV-1 RNA) of 4.45 by Drug Use, Accounting for and Ignoring Dropout Time and Reason, Acute Infection and Early Disease Research Program, 1997–2007.

Figure 6. Estimated percent reduction in $CD4^+$ T cell count for injection drug users compared to nonusers and 95% point-wise confidence limits, Accounting for and Ignoring Dropout, Acute Infection and Early Disease Research Program, 1997–2007.

Table 1

Characteristics of untreated HIV Seroconverters Acute Infection and Early Disease Research Program, 1997–2007: Mean(Standard Deviation) or %(No.).

Characteristic	Noninjection drug users			Injection drug users		
	Overall <i>n</i> = 983	Started treatment <i>n</i> = 375	Lost <i>n</i> = 608	Overall <i>n</i> = 76	Started treatment <i>n</i> = 25	Lost <i>n</i> = 51
Dropout time (days) ^{a,b}	451 (182–841)	290 (123–533)	587 (254–1078)	322 (145–693)	203 (127–429)	527 (161–802)
Age (years)	35.4 (8.8)	36.3 (8.9)	34.9 (8.7)	37.5 (7.6)	38.3 (7.4)	37 (7.7)
Number of observations ^a	6 (3–10)	4 (2–8)	7 (4–12)	5 (2.75–8)	5 (2–7)	5 (3–9)
Nadir CD4 ⁺ :						
Baseline to week 5 ^{a,c}	504 (401–669)	450 (334–583)	545 (435–716)	460 (334–576)	460 (264–520)	460 (384–677)
Baseline log ₁₀ (HIV-1 RNA) ^d	4.27 (0.96)	4.73 (0.74)	3.98 (0.98)	4.53 (0.81)	4.90 (0.67)	4.34 (0.82)
Sex						
Male	94.7 (931)	94.7 (355)	94.8 (576)	85.5 (65)	88.0 (22)	84.3 (43)
Female	5.3 (52)	5.3 (20)	5.2 (32)	14.5 (11)	12.0 (3)	15.7 (8)
Race						
White	79.1 (778)	81.1 (304)	78.0 (474)	77.6 (59)	68.0 (17)	82.4 (42)
Minority	20.9 (205)	18.9 (71)	22.0 (134)	22.4 (17)	32.0 (8)	17.6 (9)
Infection time						
Acute	11.3 (111)	13.6 (51)	9.9 (60)	14.5 (11)	4.0 (1)	19.6 (10)
Recent	88.7 (872)	86.4 (324)	90.1 (548)	85.5 (65)	96.0 (24)	80.4 (41)

^a Median and interquartile range.

^b 49 noninjection drug users and 7 injection drug users remained on study for 3 years were considered completers and assigned a dropout time of 1096 days.

^c Cells per mm³.

^d log₁₀(copies/mL).

Table 2

Change in $\log_e(\text{CD4}^+ \text{ T cells/mm}^3)$ per year in HIV Seroconverters Acute Infection and Early Disease Research Project, 1997–2007.

Group	Difference	95% Confidence interval		<i>p</i>
Injection Drug Users vs. Nonusers	−0.094	−0.177	−0.012	0.02
Started Treatment vs Lost to Follow-up	−0.316	−0.423	−0.208	< 0.0001
Started treatment				
Injection Drug Users vs. Nonusers	−0.167	−0.372	0.039	0.11
Lost to Follow-up				
Injection Drug Users vs. Nonusers	−0.083	−0.156	−0.010	0.03

Table 3

Sensitivity analysis.

$\delta(u)$	Year	CD4 ⁺ T Cells per mm ³		Difference	<i>p</i>
		Injection drug users	Nonuser		
1	1	388.7	428.5	39.8	0.10
	2	324.1	267.5	56.6	0.05
	3	245.1	184.1	61.0	0.04
0.5	1	451.4	418.0	33.3	0.08
	2	380.5	332.8	47.7	0.03
	3	326.9	272.8	54.1	0.03
0.25	1	463.2	433.5	29.7	0.11
	2	412.3	371.2	41.1	0.06
	3	377.5	332.1	45.4	0.05
0	1	475.4	449.6	25.8	0.18
	2	446.7	414.1	32.7	0.14
	3	435.9	404.2	31.7	0.17