2-25-2008

Multiple imputation of timing of mother-to-child transmission of HIV

Elizabeth Brown
elizab@u.washington.edu

Ying Qing Chen
Fred Hutchinson Cancer Research Center, yqchen@u.washington.edu

Multiple imputation of timing of mother-to-child transmission of HIV

Elizabeth R. Brown* and Ying Q. Chen

Department of Biostatistics, University of Washington, Seattle, WA

Abstract

In this paper, we present a model for imputing timing of mother-to-child transmission (MTCT) of HIV. The method reflects the three modes of MTCT of HIV: in utero, during delivery and via breastfeeding and can accomodate shapes for the baseline hazard that vary between infants. Additionally, it allows that the majority of infants do not experience MTCT of HIV. Final analyses from the imputed data sets are combined in a multiple imputation framework. The methods is illustrated on a large trial designed to assess the use of antibiotics in preventing MTCT of HIV and is validated using simulations. Additionally, we explore appropriate censoring techniques to account for weaning.

*Corresponding author: elizab@u.washington.edu, Tel: 206-667-6062, Fax: 206-667-4812

1 Introduction

There is great interest in understanding the dynamics of mother to child transmission (MTCT). This includes estimating MTCT rates during the three exposure periods (in utero, intrapartum and postnatal (while breastfeeding)), estimating the relationship between baseline covariates and the likelihood of transmission in each of the exposure periods, estimating the distribution of the timing of MTCT overall or within one of the exposure periods and relating distributions of the timing of MTCT to baseline covariates [e.g., 27, 12, 13, 15, 18, 21]. Unfortunately, infants are often lost to follow-up, frequently due to death of the infant or mother, moving away from the study site or unwillingness of the primary caregiver to allow the infant to continue in the study. These infants often have an unknown infection status at the end of the study. Also, because HIV-1 infection is only assessed at discrete points in time (usually during routine clinic visits), the time of infection is only known within an interval. Missed visits can also complicate analyses.

The HIV Prevention Trials Network (HPTN) 024 study was a multisite, placebo-controlled, double blinded randomized trial of antibiotics to prevent perinatal MTCT of HIV-1 [26]. In trials such as HPTN 024, the primary endpoint is often the cumulative transmission rate at a point in time shortly after birth. In HPTN 024 and most MTCT studies, the infants were scheduled to be tested within 48 hours after birth to assess in utero transmission. A second visit was scheduled to be between 4 and 8 weeks after birth to assess intrapartum transmission. Subsequent visits were also scheduled to evaluate late postnatal MTCT of HIV-1 via breast milk (transmission first detected after 6 weeks) and mortality. Specifically in HPTN 024, these visits were scheduled at 3, 6, 9 and 12 months. The usual

approach to estimate the distribution of the timing of MTCT of HIV-1 via breast milk is to subset the data to breastfed infants known to be negative at the 4-8 week visit because they tested negative at that or a later visit. Infants who missed the 4-8 week visit and subsequently tested positive would not be included in the analysis because their 4-8 week HIV-1 infection status is unknown. If interest is in estimation of the survival curve, then the origin is set to 8 weeks and the infant's time to event is taken to be the time of the first positive test or the midpoint between the last negative test and the first positive test. If the infant has no positive test, his time to event is censored at the last negative test, weaning or death. Hughes and Richardson [11] proposed nonparametric and semiparametric approaches for estimating the HIV-1 infection time in infants; however, these approaches do not allow for covariates. Other general interval-censoring techniques may also be used to estimate the survival distribution [29, 23] or hazard ratios [7, 22, 9]. However, none of these approaches account for uncertainty about inclusion of some infants in subset analyses nor do they reflect the unique features of the distribution of the timing of MTCT of HIV which we will detail later.

Multiple imputation [20] has previously been proposed to aid in the analysis of interval-censored data. Pan [17] proposed an imputation scheme based on the Poor Man's Data Augmentation algorithm [30]. Bebchuk and Betensky [1] used local likelihood methods for imputing interval censored observations. Neither of these approaches made use of auxilliary information nor did they impute right-censored observations. Glynn and Rosner [8] proposed a multiple imputation scheme for interval censored paired data based on a parametric frailty model. Most recently, Hsu et al. [10] proposed a non-parametric imputation approach that uses auxilliary variables and imputes both right and interval censored observations.

In this paper, we present a flexible model for imputing the timing of MTCT of HIV-1. This model allows that a proportion of infants are born with detectable HIV-1 infection from in utero transmission and that another significant proportion will never experience MTCT of HIV. It also allows for a flexible estimate of the hazard for postnatal transmission while still allowing for straight-forward computation using available software. Additionally, we demonstrate how to use this model to impute transmission times both for right and interval censored observations. Finally, we use multiple imputation methods to calculate the final estimates and their standard errors. An application of this approach to the HPTN 024 data set demonstrates the value of this approach to estimate the distribution of timing of late postnatal transmission. An extensive simulation explores the properties of the MI procedure.

2 Methods

2.1 Overview

Let s denote the time that an infant would first test positive for HIV-1 (referred to as timing of MTCT). Note that this is not the same as the time at which of MTCT of HIV-1 occurs due to the low sensitivity of HIV-1 PCR assays in the period immediately after transmission occurs [4, 5, 24, 31]. We do not observe s precisely. Instead, we observe the pair of times (L, R), where $R < s < L$. We define L as the time of the last negative test and R as the time of the first positive test. If the infant never has a negative test, without loss of generality,

we set $L = -\infty$. If instead the infant never has a positive test, we set $R = \infty$ and treat the observation as right-censored. Given the observed pair (L, R), we can estimate the distribution of s, $f(s)$, or its associated survival distribution, $S(s)$, using interval-censoring estimation techniques, but often we are interested in estimating this distribution over a specific time interval. For example, when examining the late postnatal transmission distribution, we might be interested in estimating the effect of a set of covariates, X, on the timing of transmission after a certain age, t_1, $g_1(s|X) = f(s|s > t_1, X)$, where t_1 is often taken to be 6 weeks. Many infants' observations of (L, R) will contain t_1 and therefore it is unclear if they belong to the analysis subset of interest thereby complicating estimation of $g_1(s|X)$. Alternatively, we might be interested in examining the effect of covariates on the timing of MTCT in that group of subjects who are infected late postnally, $g_2(s|X) = f(s|t_1 < s < t_2, X)$, where t_2 is usally taken to be the end of follow-up. Again, for many infants, (L, R) will contain either t_1 or t_2, and it will be unclear if they should be members of the analysis subset of interest.

To allow straight-forward estimation of both g_1 and g_2, we propose a multiple imputation (MI) technique for the actual random variable of interest, s. First, we specify a likelihood-based model for the complete data. Next, we set prior distributions for the parameters in the model. The imputations can then be generated using Markov Chain Monte Carlo (MCMC) methods, where, after a significant burn-in period, the missing data is imputed by taking draws from the posterior predictive distribution conditional on the current draws of the parameters from the posterior distribution of the paramters. Each data set created by this imputation technique is referred to as an augmented data set.

Our proposed imputation model reflects features seen in but not unique to MTCT studies. First, many infants are infected in utero and their infection can be detected immediately after birth. Second, because the at-risk time for MTCT of HIV is limited by exposure to breast milk, not all subjects will experience MTCT of HIV. This is in contrast to a usual time to event analysis where we assume that if we could follow a subject indefinitely and there were no competing risks, s/he would eventually experience the event. Third, s is not observed past some end of follow-up time, t_2. To accomodate this, we will assume that all infants are censored at t_2 if they have not yet experienced the event and account for this accordingly in the multiple imputation. This assumption will still allow for estimation of both g_1 and g_2.

2.2 Analyses models

To motivate the imputation model, we first describe two analyses of interest for estimating late postnatal transmission in HPTN 024. The first is estimation of the cumulative risk of MTCT of HIV-1 at the end of the study in those infants at-risk for late postnatal transmission. The second is estimation of proportional hazards models for late postnatal transmission. For the observed data, these analyses will be performed subsetting the data to those infants who have a negative test at 4-8 weeks and are still breastfeeding. For the MI analyses, the data will be subsetted to those infants with $s > 6$ weeks who are still breastfeeding at 6 weeks.

Censoring can be complex in these studies due to the different causes: death, weaning and loss to follow-up; therefore, we propose examining different censoring

rules in the analyses and in simulations to determine which censoring approach produces the best estimate of the survival distribution or association parameter of interest. If an infant dies, is lost to follow-up or reaches the end of the study without having a positive HIV-1 test, his/her time to event is censored at the time of the last negative test. If infant is weaned, there are three censoring options:

C1 An infant's event time is censored at his last negative test. This is a common approach that does not require information on weaning.

C2 An infant's event time is censored at the end of follow-up if there is a negative test after weaning in the observed data. This censoring approach reflects that these infants are no longer at risk after weaning and should produce an estimate of distribution of time to first positive test in the population under study.

C3 In the observed data analysis, if an infant has a negative test after weaning, his event time is censored at the time of weaning. Otherwise, it is censored at the time of his last negative test. In the imputed data, an infant is censored at the time of weaning if he has not already experienced the event. This approach estimates the late postnatal time to first positive distribution as if no weaning occured.

Scenarios C1 and C2 result in the same censoring scenario for the MI analysis, no censoring except at the end of the follow-up time. However, MI results under these scenarios will be presented as coming from Scenario C2. Under a frequent testing schedule, there should be little difference between Scenarios C1 and C3.

2.3 Imputation model

We now present the model for imputing individual times to detectable HIV-1 infection. Let $s_i,\ i = 1, \ldots, N$ denote the age at which the ith infant born to an HIV–infected mother first has detectable HIV infection. This step requires specifying a likelihood for the complete data. As stated in the Introduction, there are three features of the distribution of the timing of MTCT of HIV that the likelihood should reflect. First, most transmissions that occur in utero can be detected immediately after birth. One way to approach this is to treat the time to detectable infection for these infants as left censored at zero; however, we are not really interested in estimating the timing of detectable infection before birth. Instead, without loss of generality, we will assume the time of first positive test for these infants is 0. Second, all infants will be weaned at some point and will no longer be at risk; therefore, if they have not experienced MTCT before weaning, we do not expect that they will experience it all. The third feature is closely related to the second. Most studies, including HPTN 024 do not follow infants until the last infant is weaned, but instead follow them for 12-18 months. Because, in general, we do not observe events past the period of follow-up, we cannot expect to accurately impute event times past this time. Additionally, any analysis we might perform, even had we completely observed the outcome of interest according to the study design, would be limited to the period of follow-up, and it would be impossible to distinguish those infants who will never be infected from the few who are infected after the end of follow-up. Therefore, we propose an imputation model that is a mixture of three distributions: a point mass at zero that reflects the proportion of infants with detectable infection at birth, a continuous distribution for those infections that are first detectable after birth

and before the end of study time, t_2, and a point mass at a time greater than t_2 representing the proportion who experience the event after t_2 or never experience the event. The third distribution results in an overall distribution similar to a cure rate mixture model ([2, 6, 14] and others) without the medical concept of cure. Therefore, we can express the distribution of the ith infant's time to detectable infection as

$$f(s_i|Z_i, p_1, p_2, \Theta) = p_1\delta_0(s_i) + p_2 f_2(s_i|Z_i, p_1, p_2, \Theta) + (1 - p_1 - p_2)\delta_\infty(s_i), \quad (2.1)$$

where $\delta_x(s_i)$ denotes a point mass at $s_i = x$, p_1 and p_2 are mixing proportions, Θ is the set of parameters that define f_2 and Z_i is a vector of covariates (inlcuding an intercept term) of length q that includes any covariates of interest for the final analysis. To facilitate estimation, we introduce a latent (auxilliary) variable, d_i, where

$$d_i = \begin{cases} 1, & \text{if } s_i \sim \delta_0 \\ 2, & \text{if } s_i \sim f_2(s_i|Z_i, p_1, p_2, \Theta) \\ 3, & \text{if } s_i \sim \delta_\infty \end{cases} .$$

Therefore, we can rewrite (2.1) conditional on d_i as

$$f(s_i|Z_i, d_i, \Theta) = \delta_0(s_i)^{I(d_i=1)} f_2(s_i|Z_i, \Theta)^{I(d_i=2)} \delta_\infty(s_i)^{I(d_i=3)}, \quad (2.2)$$

where $I(x)$ is an indicator function that takes on the value 1 if x is true, 0 otherwise. Here, d_i is a partially observed latent variable. In order to completely specify the likelihood for imputation, we must specify a distribution for d_i. We take d_i to be a multinomial random variable and specify its mean vector as a

function of the set of covariates, Z_i, such that

$$Pr(d_i = 1|Z_i) = \text{expit}(\alpha' Z_i) \tag{2.3}$$

and

$$Pr(d_i = 2|Z_i, d_i > 1) = \text{expit}(\omega' Z_i) \tag{2.4}$$

where expit is the inverse-logit function and α and ω are sets of covariates linking Z_i to d_i. The probability mass function for d_i is then

$$\begin{aligned} p(d_i|\alpha, \omega) &= \text{expit}(\alpha' Z_i)^{I(d_i=1)} \{\text{expit}(\omega' Z_i)[1 - \text{expit}(\alpha' Z_i)]\}^{I(d_i=2)} \times \\ &\quad \{1 - \text{expit}(\alpha' Z_i) - \text{expit}(\omega' Z_i)[1 - \text{expit}(\alpha' Z_i)]\}^{I(d_i=3)}. \end{aligned}$$

Next, we specify f_2. For ease of computation, we restrict our options to parameteric distributions. The Weibull distribution allows for a wide range of shapes for the hazard function given by

$$h(t) = at^{a-1} \exp(\beta' Z_i), \tag{2.5}$$

where a is the shape parameter of the Weibull distribution and β is a vector of parameters linking the ith infant's covariate vector, Z_i, to the hazard and $\exp(-\beta' Z_i)$ is the ith infant's scale parameter. The hazard shown in (2.5) assumes a proportional hazards model.

A frailty model may define a common scale parameter within groups, thereby recognizing that some groups may inherently be at higher risk than other groups

throughout follow-up. Instead, we explore the possibility that different groups may follow hazards with different shapes without specifying membership in the groups a priori. This approach is motivated by the hypothesis that infants' underlying risks may follow different trajectories based on unobserved information. For example, mixed feeding (breastfeeding plus formula or other foods) may put infants at a higher risk of transmission [3], resulting in an underlying hazard that may remain constant or increase rapidly after irth; whereas, exclusively breastfed infants may be expected to have a hazard that decreases soon after birth then becomes roughly constant [16]. In HPTN 024, information about mixed feeding is not collected. Therefore, we allow the shape parameter to vary across infants to accommodate this potential variation in shape of the underlying hazards. We define γ to be a vector of length m and the ith infant's hazard function to be

$$h(t) = \gamma_{k_i} t^{\gamma_{k_i}-1} \exp(\beta' Z_i),$$

where k_i takes on values from 1 to m and indicates which of the m elements of γ determines the ith infant's hazard function. Additionally, we assume the latent variable k_i follows a multinomial distribution such that

$$k_i \sim \text{multinomial}(\pi_{\gamma 1}, \ldots, \pi_{\gamma m}),$$

where $\sum_{l=1}^{m} \pi_{\gamma l} = 1$. Therefore, conditional on k_i, f_2 is a Weibull distribution, where

$$f_2(s_i|Z_i, \beta, \gamma, k_i) = \gamma_{k_i} e^{\beta' Z_i} s_i^{\gamma_{k_i}-1} \exp(-e^{\beta' Z_i} s_i^{\gamma_{k_i}}). \tag{2.6}$$

Thus far, we have described the distribution of the true but unobserved time until

an infant would first test positive for HIV-1, s_i. We do not actually observe s_i but instead observe (L_i, R_i) and therefore need to express the likelihood in terms of (L_i, R_i). The ith subject's contribution to the likelihood is then

$$\begin{aligned} L_i(\gamma, \beta, \alpha, \pi, \omega | L_i, R_i, Z_i, d_i, k_i) &= Pr(L_i < s_i < R_i | Z_i, \Theta, d_i, k_i) p(d_i | Z_i, \alpha) p(k_i) \\ &= \int_{L_i}^{R_i} f(u | Z_i, p_1, p_2, \gamma, \beta, k_i) du \ \text{expit}(\alpha' Z_i)^{I(d_i=2)} \times \\ &\quad \{\text{expit}(\omega' Z_i)[1 - \text{expit}(\alpha' Z_i)]\}^{I(d_i=1)} \times \\ &\quad \{1 - \text{expit}(\alpha' Z_i) - \text{expit}(\omega' Z_i)[1 - \text{expit}(\alpha' Z_i)]\}^{I(d_i=3)} \\ &\quad \prod_{l=1}^{m} \pi_{\gamma_l}^{I(k_i=l)}. \end{aligned}$$

2.4 Estimation procedure

Before producing estimates of the parameters in Equation 2.7, we first specify prior distributions for these parameters as follows:

$$\begin{gathered} (\pi_{\gamma_1}, \ldots, \pi_{\gamma_m}) \sim \text{Dirichlet}(\mathbf{1}_m), \text{ where } \mathbf{1}_m \text{ is a vector of ones of length } m, \\ \beta_j \sim N(0, 1000), \ j = 1, \ldots, q, \\ \gamma_j \sim N(0, 10) I(\gamma_j > 0), \ j = 1, \ldots, m, \\ \omega_1 \sim N(-2, 1), \ \omega_j \sim N(0, 1000), \ j = 2, \ldots, q, \\ \alpha_1 \sim N(-2, 1), \text{and } \alpha_j \sim N(0, 1000), \ j = 2, \ldots, q. \end{gathered}$$

We implement the mulitple imputation scheme in BUGS [25], using OpenBUGS [28] and the BRugs package in R [19].

3 Data Analysis

In this section, we apply the multiple imputation model for timing of MTCT transmission of HIV-1 to data collected in HIV Prevention Trials Network (HPTN) 024 [26]. Although HIV testing was initially scheduled to be done at birth, 4-6 weeks and 3, 6, 9 and 12 months, the majority of 4-6 week visits occurred between 6 and 8 weeks, and the three month visit was dropped early in the study. Samples collected at 3, 6 and 9 months were only tested if the 12 month sample was positive or missing.

Infants born to HIV-1-infected mothers are only at risk for MTCT of HIV while breastfeeding. At one site, mothers were counseled to stop breastfeeding by the time their infants reached 6 months of age, and, by 6 months of age, over 90% of the the infants at this site had been weaned. In contrast, over 90% of the infants at the 3 remaining sites were still breastfeeding at six months. This difference in the underlying hazard between the sites will be accounted for by performing a stratified proportional hazards analysis.

We performed the multiple imputation as described previously. The values for L_i and R_i were discussed in general in the Methods section. Here, we discuss how they were set more specifically for HPTN 024. If the ith infant never had a negative test, we set $L_i = 0$ and R_i equal to the time of the first positive test. Because the earliest detection time is birth, $L_i = 0$ is as general as $L = -\infty$ in implementation. If the first positive test occured on the day of birth, we set $L_i = R_i = 0$. For these infants, we know that $d_i = 1$ and $s_i = 0$. If the infant had both a negative and positive test before weaning, we set L_i equal to the time of

the last negative test and R_i equal to the time of the first positive test. If weaning occured before the first positive test, we set R_i equal to the time of weaning plus 30 days (due to the sensitivity issue discussed perviously). For subjects who have both a positive and negative test, d_i is known to be 2. For subject's with only negative tests and no positive tests, we set L_i equal to the time of the last negative test unless weaning occured more than 30 days before the negative test. In that case, we set L_i equal to the time of weaning plus 30 days. Additionally, because follow-up was limited to approximately one year and therefore there was no information past this point in terms of observed events, R_i was set to 400 days. This would not impact the final analysis where the imputed data was censored at one year.

The following auxiliary variables were used in the imputation procedure: maternal CD4 count, hemoglogin, viral load, weight and age at 32 weeks gestation; enrollment site; whether the mother took nevirapine; an indicator of whether the infant was delivered at the study clinic; whether the infant took nevirapine; the duration of ruptured membranes; and the infant's birthweight and sex.

In each augmented data set, every infant has an imputed value for s_i. This s_i reflects the true time of detectable infection if other events, such as death or weaning, did not intervene. Also, because there was little information past one year in the original data set, we censor the infants' times to event at one year in the final analyses. Because s_i is now on a continuous scale in the augmented data set, we can perform time to breastfeeding transmission by subsetting to those subjects whose s_i is greater than 6 weeks. In contrast, the observed analysis must define the subset of interest as those infants with a negative test after 4 weeks and not

positive before 8 weeks, misclassifying those infants tests at 8 weeks who may have tested negative at 6 weeks and those infants who tested negative at 4 weeks may test positive by 6 weeks. Therefore, we expect some bias in the baseline number at risk. Additionally, when performing the observed data analysis, we assumed only right censoring and set the time to event for any infant with a positive test to be the midpoint between the last negative test and the first positive test. In the proportional hazards model, we studied the relationship between maternal CD4 and viral load, stratified by site.

Overall for the observed analysis, of the 1977 potential infants, 1317 tested negative after the 4-8 week visit and were still breastfeeding at 8 weeks. Infants were excluded because they were known to be positive by 8 weeks (N=298), were weaned before 8 weeks and therefore not at risk (N=70), had unknown infection status at 4-8 weeks due to missing 4-8 week test and later positive test (N=22), or had no test results after the 4-8 week visit (N=270). Analyses on the observed data were carried out under all three censoring scenarios (C1-C3). Analyses on the augemnted data sets were carried out under censoring scenarios C2 and C3.

[Table 1 about here.]

Table 1 shows results from the Kaplan-Meier (KM) and proportional hazards (PH) analyses on both the observed and imputed data. For the observed data, C1 and C2 produce similar results. If all infants were tested at the end of the study, we expect these results to be identical because all weaned infants who did not experience MTCT of HIV would be censored after the end of follow-up. For the observed data, censoring scenario C3 resulted in higher KM estimates of the

cumulative infection rates than C1 or C2. Because C3 treats weaned infants as if they were still at risk at the time of the weaning and therefore assumes that some would experience the event, we would expect the proportion to be higher. The same differences between C2 and C3 are seen in the multiple imputation analysis. Because censoring under C2 (not at risk after weaning) is usually of interest, we will focus the comparison between the observed and MI analyses under censoring scenario C2. Many infants who test negative at 4-8 weeks do not have another test result available until 12 months and that test result is positive; therefore, because the time of the first positive test for these infants is imputed to be approximately 7 months in the observed analysis, we expect the observed data analyses to underestimate the transmission rate at earlier times. The results indicate that the MI may be correcting this, producing higher estimates at 3 and 6 months than the observed analysis. MI produces lower estimates of transmission rates at 9 and 12 months, though. The simulations summarized in the next section show that we expect the observed analysis to overestimate the transmission rate at 12 months. Also, the MI analyses include the 294 above who have no test results after 4-8 weeks. Potentially, these infants were less likely to have experienced MTCT, thus increasing the number at risk disproportionately to the number of events. The MI results do not vary substantially over m.

[Figure 1 about here.]

To better understand the variability between imputations and how the estimates from the augmented data sets compare to the observed analysis, we plotted the KM estimates of the cumulative infection rate curves for each augmented data set ($m = 2$) and for the observed data (Figure 1). There is variability between the

estimates from the augmented data sets. At most points of interest the estimates are all contained within an interval of width approximately equal to 0.02. Before five months, the observed data analysis estimates of the survival curve are higher than all the estimates from the augmented data sets. From 5 to 8 months, the observed curve crosses all the augmented data set estimates. After 8 months, the observed curve is at the lower end of the augmented data estimates.

Also shown in Table 1 are proportional hazards regression models fit to the observed and MI data. The estimate of association was higher in the observed analyses than in the MI analyses. Additionally, the standard errors were lower for viral load and higher for CD4 count in the MI analyses. The MI results varied little over m or the censoring scenario. However, the observed analyses results varied more over censoring scenarios (C3 vs. C2 or C1), suggesting some interplay between timing of weaning and CD4 count and viral load.

4 Simulations

In this section, we describe simulations designed to assess the multiple imputation procedure and compare it to traditional analyses on the observed data. Additionally, we explored the three censoring approaches for weaning (C1-C3). We simulated d_i and s_i subject to the effects of 4 covariates, 2 binary ($X_1 \sim$ bernoulli(.5), $X_2 \sim$ bernoulli(.25)) and 2 continunous ($X_3 \sim$ uniform$(-1, 1)$, $X_4 \sim N(0, 1)$, with $m = 2$, $\gamma = (7, 0.9)$, $Pr(k_i = 1) = 0.26$, $\alpha = (-2.7, -1.0, 1.0, 0.5, -1.0)$, $\omega = (-0.5, -2.0, 2.0, 0.5, -1.0)$ and $\beta = (-1.6, -1.0, 1.0, -0.5, 0.5)$.

We then simulated visits according to visit schedule in HPTN 024 (birth, 4-8 weeks and 3, 6, 9 and 12 months) and according to the proportions seen in HPTN 024. Specifically, we sampled attendence at first and second visits with probabilities equal to 0.94 and 0.85 respectively. If the infant had a negative test at either birth or 4-8 weeks, then he had a test result at 12 months with probability equal to 0.80. If the infant missed both the birth and 4-8 week visit, he had a test result from the one year visit with probability equal to 0.20. If the infant missed the 12 month visit, he had a test result from the 9 month visit with probability equal to 0.50. If subject tested positive at the 12 month visit, he had a sample from the 9 month visit with probability equal to 0.80. If the infant missed the 9 and 12 month visits, he had a test result from the 6 month visit with probability equal to 0.60. If subject was positive at the 9 month visit, he had a test result from the 6 month visit with probability equal to 0.80. If the infant missed the 6, 9 or 12 month visit, he had a test result from the 3 month visit with probability equal to 0.20. If the infant had a positive test result from the 6 month visit, he had a test result from 3 months with probability equal to 0.40. At each visit, his visit time was simulated according to the observed distribution of visits in HPTN 024.

Next, we simulated times of death and weaning according to the distributions seen in HPTN 024 under the following three scenarios:

S1 Non-informative loss to follow-up and death and no weaning

S2 Non-informative loss to follow-up, death related to one of the covariates (X_3) in the same manner as CD4 count is related to infant death in HPTN 024 and no weaning.

S3 Non-informative loss to follow-up, death and weaning.

S4 Non-informative loss to follow-up and death. Time to weaning is related to X_2 similarly to the relationship between the HPTN 024 site with early weaning and time to weaning. Therefore, $X_2 = 1$ is associated with early weaning and increased risk of transmission.

Under each scenario, we simulated 100 data sets with 1000 observations each and fit the complete data analysis (no censoring except for death) which was used as the gold standard for comparing the observed data analysis and the MI analysis with $m = 1, 2, 3$ under S1 and S2. For S3 and S4, we compared the results to two gold standards designed to represent the best estimates possible if we had observed timing of detectable infection perfectly. The first gold standard (G1) censors infants at death. The second gold standard (G2) censors infants at death and weaning and estimates transmission rates and associations assuming there was no weaning. C2 is designed to estimate the first gold standard, and C3 is designed to estimate the second gold standard. C1 mimics what is usually done in practice.

We compared the results in terms of their bias compared to the gold standard, the variance ratio (variance of the estimate of the analysis divided by the variance of the gold standard analysis) and the coverage rates (frequency that the confidence interval contained the gold standard estimate).

The results for the simulations under S1 and S2 are shown in Table 2. The MI analyses performed the same or better in terms of bias under both scenarios for both estimators. The MI analyses had lower variance estimates than the observed

analyses for the estimate of cumulative transmission, but the opposite was true for the estimate of the hazard ratio. The coverage rates for the MI analyses were the same or better than the observed analyses under both scenarios.

The results for the simulations under S3 are shown in Table 3. First, we examine estimates for the late postnatal transmission rate at 12 months. The bias for the observed analyses was relatively high compared to the truth (0.1182) and twice that of most of the MI analyses. Additionally, the MI analyses were more efficient under all scenarios. The coverage rates for the MI analyses were better for G1; however, the coverage rates for the observed analyses were better for G2. Turning our attention to the PH analyses, the observed and MI analyses performed similarly for bias under G1 and G2. The lowest bias was the observed analysis under C3 for G1. In all cases, the observed analyses were more efficient than the MI analyses. Both the observed and MI analyses had similar coverage rates.

The results for the simulations under S4 are shown in Table 4. Under G1, MI performed better in terms of bias for both the KM and PH estimates. Under G2, the observed analysis performed better. Additionally, the observed analysis produced less biased estimates of G2 and G1, even under censoring scenarios designed to estimate G1. The MI KM analyses were more efficient than the observed analyses for G1 and G2. The opposite was true for the PH estimates. Recalling that C3 is designed to estimate G2, the low coverage rates for MI under G2 using C2 is not alarming; however, the coverage rates are higher than desirable under C3.

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

5 Discussion

We present an approach to imputing the timing of MTCT of HIV-1. Given the augmented data sets produced by the multiple imputation procedure, we can now perform many analyses that would not be possible otherwise without excluding large portions of the data set. Here, we showed an example estimating the cumulative late postnatal transmission rate at 12 months and the effect of covariates on the hazard of late postnatal transmission. Additional analyses are now also attainable. For example, investigators are also interested in estimating the distribution of timing of MTCT among those infants who experience MTCT for use in planning HIV-1 testing schedules. Potential analyses are not limited to late postnatal transmission. For example, we may want to assess how baseline covariates predict transmission during the three exposure periods. The MI approach allows us to include those infants whose timing could not previously be precisely categorized. This approach is flexible and can easily be implemented with OpenBugs and R.

The MI approach was validated in simulations and shown to be less biased in most situations than the traditional estimator. In the presence of weaning, the

traditional estimator proved to be a better estimator of the distribution of MTCT ignoring weaning which is seldom of interest.

Our goal here was to find a flexible MI model that could easily be implemented in available software. Although, the MI model we propose is flexible and mirrors the modes of MTCT of HIV, it could be improved in several ways. Here, we do not directly account for HIV-1-infected infants being at higher risk of death. In reality, it is likely that an infant who had a negative test long before death actually acquired HIV-1. Additionally, a mother's decision to wean may also be related to the health status of her infant. To reflect these issues, we could consider modeling the relationship between the risk of death, time to weaning and the risk of MTCT more directly using competing risk models. We chose mixtures of Weibull models for flexibility in the distribution of time to detectable infection after birth. Instead, we could explore other flexible baseline hazards; however, these models would likely require customized software and would not be easily fit.

Lastly, we chose to impute time to detectable infection and not the actual time to transmission. The sensitivity of the HIV-1 assays vary over time in way that is not clearly understood. To attempt to incorporate this in the model would have been to complex. Additionally, from a public health perspective, understanding the timing of detectable infection is more important.

References

[1] Bebchuk, J. and Betensky, R. [2000]. Multiple imputation for simple estimation of the hazard function based on interval censored data, *Statistics in*

Medicine **19**: 405–419.

[2] Berkson, J. and Gage, R. P. [1952]. Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**: 501–515.

[3] Coutsoudis, A., Pillay, K., Kuhn, L., Spooner, E., Tsai, W. Y. and Coovadia, H. M. [2001]. Method of feeding and transmission of hiv-1 from mothers to children by 15 months of age: prospective cohort study from durban, south africa, *AIDS* **15**(3): 379–387.

[4] Cunningham, C., Charbonneau, T., Song, K. and et al [1999]. Comparison of human immunodeficiency virus 1 DNA polymerase ch ain reaction and qualitative and quantitative RNA polymerase chain reaction in h uman immunodeficiency virus 1-exposed infants, *Pediatric Infectious Disease Journal* **18**: 30–35.

[5] Dunn, D., Simonds, R., Bulterys, M. and et al [2000]. Interventions to prevent vertical transmission of HIV-1: effec t on viral detection rate in early infant samples, *AIDS* **14**: 1421–1428.

[6] Farewell, V. T. [1982]. The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**: 1041–1046.

[7] Finkelstein, D. M. [1986]. A proportional hazards model for interval-censored failure time data, *Biometrics* **42**(4): 845–854.

[8] Glynn, R. J. and Rosner, B. [2004]. Multiple imputation to estimate the association between eyes in disease progression with interval-censored data, *Statistics in Medicine* **23**(21): 3307–3318.

[9] Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A. and Zaslavsk y, A. M. [1998]. A markov chain monte carlo em algorithm for analyzing interval-censored d ata under the cox proportional hazards model, *Biometrics* **54**(4): 1498–1507.

[10] Hsu, C. H., Taylor, J. M. G., Murray, S. and Commenges, D. [2007]. Multiple imputation for interval censored data with auxiliary variables, *Statistics in Medicine* **26**(4): 769–781.

[11] Hughes, J. P. and Richardson, B. A. [2000]. Analysis of a randomized trial to prevent vertical transmission of HIV-1, *Journal of the American Statistical Association* **95**(452): 1032–1043.

[12] Iliff, P. J., Piwoz, E. G., Tavengwa, N. V., Zunguza, C. D., Marinda, E. T., Nathoo, K. J., Moulton, L. H., Ward, B. J. and Humphrey, J. H. [2005]. Early exclusive breastfeeding reduces the risk of postnatal HIV-1 transmission and increases HIV-free survival, *AIDS* **19**(7): 699–708.

[13] Kourtis, A. P., Lee, F. K., Abrams, E. J., Jamieson, D. J. and Bulterys, M. [2006]. Mother-to-child transmission of hiv-1: timing and implications for prevention, *Lancet Infectious Diseases* **6**(11): 726–732.

[14] Kuk, A. Y. C. and Chen, C. H. [1992]. A mixture model combining logistic-regression with proportional hazards regression, *Biometrika* **79**(3): 531–541.

[15] Luzuriaga, K., Newell, M. L., Dabis, F., Excler, J. L. and Sullivan, J. L. [2006]. Vaccines to prevent transmission of hiv-1 via breastmilk: scientific and logistical priorities, *Lancet* **368**(9534): 511–521.

[16] Magoni, M., Bassani, L., Okong, P., Kituuka, P., Germinario, E. P., Giuliano, M., Vella, S. and Xo [2005]. Mode of infant feeding and hiv infection in children in a program for prevention of mother-to-child transmission in uganda, *AIDS* **19**(4): 433–437.

[17] Pan, W. [2000]. A multiple imputation approach to Cox regression with interval -censored data, *Biometrics* **56**: 199–203.

[18] Piwoz, E. G., Humphrey, J. H., Marinda, E. T., Mutasa, K., Moulton, L. H. and Iliff, P. J. [2006]. Effects of infant sex on mother-to-child transmission of HIV-1 according to timing of infection in zimbabwe, *Aids* **20**(15): 1981–1984.

[19] R Development Core Team [2006]. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

[20] Rubin, D. [1996]. Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**: 473–489.

[21] Saba, J., Haverkamp, G., Gray, G., McIntyre, J., Mmiro, F., Ndugwa, C., Coovadia, H. M., Moodley, J., Kilewo, C., Massawe, A., Kituuka, P., Okong, P., von Briesen, H., Goudsmit, J., Biberfeld, G., Grulich, A., Weverling, G. J. and Lange, J. M. A. [2002]. Efficacy of three short-course regimens of zidovudine and lamivudine in preventing early and late transmission of HIV-1 from mother to child in Tanzania, South Africa, and Uganda (Petra study): a randomised, double-blind, placebo-controlled trial, *Lancet* **359**(9313): 1178–1186.

[22] Satten, G. A. [1996]. Rank-based inference in the proportional hazards model for interval censored data, *Biometrika* **83**(2): 355–370.

[23] Sen, B. and Banerjee, M. [2007]. A pseudolikelihood method for analyzing interval censored data, *Biometrika* **94**(1): 71–86.

[24] Simonds, R., Brown, T., Thea, D. and et al [1998]. Sensitivity and specificity of a qualitative RNA detection assay to diagnose HIV infection in young infants, *AIDS* **12**: 1545–1549.

[25] Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. [2005]. *WinBUGS: User Manual, Version 2.10*, Medical Research Council Biostatistics Unit, Cambridge.

[26] Taha, T. E., Brown, E. R., Hoffman, I. F., Fawzi, W., Read, J. S., Sinkala, M., Martinson, F. E. A., Kafulafula, G., nga, G. M., Emel, L., Adeniyi-Jones, S., Goldenberg, R. and the HPTN024 T eam [2006]. A phase III clinical trial of antiobiotics to reduce chorioamn ionitis-related perinatal HIV-1 transmission, *AIDS* **20**: 1313–1321.

[27] The Breastfeeding and HIV International Transmission Study Group [2004]. Late postnatal transmission of HIV-1 in breast-fed children: An individual patient data meta-analysis, *Journal of Infectious Diseases* **189**(12): 2154–2166.

[28] Thomas, A., Hara, B. O., Ligges, U. and Sturtz, S. [2006]. Making BUGS open, *R News* **6**: 12–17.

[29] Turnbull, B. W. [1976]. Empirical distribution function with arbitrarily

grouped, censored and tr uncated data, *Journal of the Royal Statistical Society Series B-Methodological* **38**(3): 290–295.

[30] Wei, G. and Tanner, M. [1991]. Applications of multiple imputation to the analysis of censored regression data, *Biometrics* **47**: 1297–1309.

[31] Young, N., Shaffer, N., Chaowanachan, T., Chotpitayasunondh, T., Vanparapar, N., Mock, P., Waranawat, N., Chokephaibulkit, K., Chuachoowong, R., Wasinrapee, P., Mastro, T. and Simonds, R. [2000]. Early diagnosis of HIV-1-infected infants in Thailand using RNA and DNA PCR assays sensitive to non-B subtypes, *Journal of Acquired Immune Deficiency Syndromes* **24**: 401–407.

List of Figures

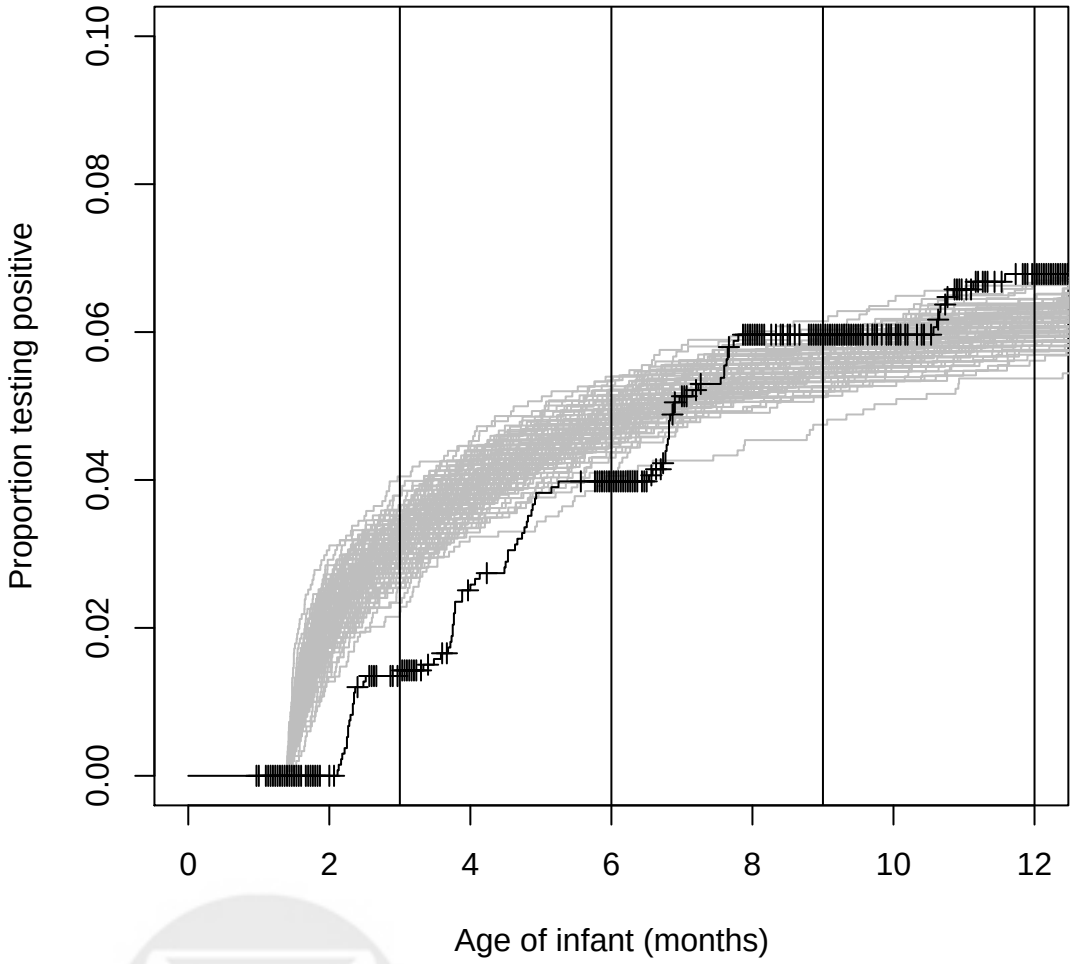

Figure 1: Curves of the cumulative proportion infected for the observed data analysis (black) and each of the augmented data sets (grey). The vertical lines represent the centers of the visit windows. The crosses on the solid line indicate where infants had a last negative test and then were subsequently lost to follow-up.

List of Tables

Table 1: Results from breastfeeding transmission analyses (C1=censored at last negative; C2=censored after end of follow-up if weaned before last negative; C3=censored at time of weaning). (* estimates correspond to change in one standard deviation)

	C1	C2	C3
Kaplan Meier analysis (% testing positive, std. error)			
Observed			
3 months	1.4 (0.32)	1.4 (0.32)	1.4 (0.33)
6 months	4.0 (0.54)	4.0 (0.54)	4.5 (0.61)
9 months	6.0 (0.66)	5.8 (0.65)	6.9 (0.77))
12 months	6.8 (0.72)	6.6 (0.70)	8.0 (0.85)
MI (m=1)			
3 months	–	2.8 (0.51)	2.9 (0.52)
6 months	–	4.6 (0.60)	4.8 (0.63)
9 months	–	5.3 (0.64)	5.8 (0.69)
12 months	–	5.9 (0.67)	6.4 (0.74)
MI (m=2)			
3 months	–	3.0 (0.55)	3.1 (0.56)
6 months	–	4.6 (0.61)	4.9 (0.64)
9 months	–	5.4 (0.65)	5.8 (0.70)
12 months	–	5.9 (0.68)	6.5 (0.75)
MI (m=3)			
3 months	–	3.2 (0.56)	3.2 (0.57)
6 months	–	4.8 (0.64)	5.0 (0.67)
9 months	–	5.6 (0.67)	6.0 (0.72)
12 months	–	6.1 (0.70)	6.6 (0.77)
Proportional hazards analysis (coefficient, std. error			
Observed			
CD4 count*	-0.695 (0.111)	-0.699 (0.111)	-0.757 (0.116)
viral load*	1.077 (0.163)	1.077 (0.162)	1.088 (0.163)
MI (m=1)			
CD4 count*	–	-0.659 (0.113)	-0.681 (0.114)
viral load*	–	1.039 (0.158)	1.043 (0.158)
MI (m=2)			
CD4 count*	–	-0.650 (0.115)	-0.661 (0.117)
viral load*	–	1.040 (0.160)	1.044 (0.160)
MI (m=3)			
CD4 count*	–	-0.647 (0.114)	-0.668 (0.116)
viral load*	–	1.041 (0.155)	1.045 (0.155)

Table 2: Simulation results when there is assumed to be no weaning during follow-up. S = simulation scenario, O = observed data analysis, MI = multiple imputation analysis.

Scenario		Bias		Variance ratio		Coverage rate	
S	m	O	MI	O	MI	O	MI
12 month estimate of MTCT							
1	1	0.0187	0.0018	1.39	1.19	0.87	0.97
1	2	–	-0.0046	–	1.29	–	0.93
1	3	–	-0.0040	–	1.23	–	0.93
2	1	0.0158	-0.0110	1.47	1.08	0.88	0.93
2	2	–	-0.0159	–	1.15	–	0.87
2	3	–	-0.0164	–	1.14	–	0.89
PH estimate							
1	1	0.0355	0.0037	1.05	1.19	0.96	0.96
1	2	–	0.0139	–	1.36	–	0.96
1	3	–	0.0018	–	1.33	–	0.97
2	1	0.0178	0.0040	1.22	1.41	0.98	0.97
2	2	–	0.0072	–	1.62	–	0.96
2	3	–	0.0003	–	1.62	–	0.97

Table 3: Simulation results when weaning is observed during follow-up.

m	12 month estimate					PH estimate				
	Observed			MI		Observed			MI	
	C1	C2	C3	C2	C3	C1	C2	C3	C2	C3
Bias										
gold standard censored at death (G1; KM=0.118, PH=-1.30)										
1	0.0213	0.0204	0.0380	-0.0080	0.0096	0.0044	0.0047	-0.0003	-0.0051	-0.0240
2	–	–	–	-0.0156	0.0014	–	–	–	0.0002	-0.0202
3	–	–	–	-0.0089	0.0026	–	–	–	0.0036	-0.0104
gold standard censored at death/weaning (G2; KM=0.137, PH =-1.32)										
1	0.0079	0.0070	0.0246	-0.0214	-0.0038	0.0249	0.0252	0.0202	0.0153	-0.0035
2	–	–	–	-0.0290	-0.0120	–	–	–	0.0298	0.0094
3	–	–	–	-0.0223	-0.0108	–	–	–	0.01685	0.0101
Variance ratio										
gold standard censored at death (G1)										
1	1.44	1.42	1.88	1.02	1.42	1.03	1.03	1.03	1.26	1.26
2	–	–	–	1.108	1.48	–	–	–	1.43	1.43
3	–	–	–	1.18	1.46	–	–	–	1.44	1.44
gold standard censored at death/weaning (G2)										
1	1.27	1.25	1.67	0.90	1.25	1.14	1.14	1.14	1.40	1.40
2	–	–	–	0.96	1.31	–	–	–	1.58	1.58
3	–	–	–	1.05	1.29	–	–	–	1.59	1.59
Coverage rates										
gold standard censored at death (G1)										
1	0.87	0.88	0.57	0.93	1.00	0.95	0.94	0.95	0.95	0.96
2	–	–	–	0.82	1.00	–	–	–	0.95	0.98
3	–	–	–	0.92	1.00	–	–	–	0.96	0.97
gold standard censored at death/weaning (G2)										
1	0.97	0.97	0.82	0.63	1.00	0.95	0.94	0.95	0.95	0.96
2	–	–	–	0.55	1.00	–	–	–	0.94	0.96
3	–	–	–	0.88	1.00	–	–	–	0.95	0.96

Table 4: Simulation results when weaning related to a covariate is observed during follow-up.

m	12 month estimate					PH estimate				
	Observed			MI		Observed			MI	
	C1	C2	C3	C2	C3	C1	C2	C3	C2	C3
Bias										
gold standard censored at death (G1; KM=0.113, PH=-1.51)										
1	0.0187	0.0179	0.0353	-0.0057	0.0057	0.0627	0.0613	0.0955	-0.0412	-0.0188
2	–	–	–	-0.0115	-0.0003	–	–	–	-0.0295	-0.0202
3	–	–	–	-0.0180	-0.0006	–	–	–	-0.0348	-0.0127
gold standard censored at death/weaning (G2; KM=0.136, PH=-1.33)										
1	-0.0046	-0.0054	0.0120	-0.0295	-0.0180	-0.1170	-0.1184	-0.0842	-0.2444	-0.2327
2	–	–	–	0.0352	0.0240	–	–	–	-0.2170	-0.1941
3	–	–	–	-0.0353	-0.0243	–	–	–	-0.214	-0.192
Variance ratio										
gold standard censored at death										
1	1.43	1.41	1.89	1.09	1.39	1.01	1.01	1.01	1.33	1.32
2	–	–	–	1.16	1.46	–	–	–	1.53	1.53
3	–	–	–	1.15	1.44	–	–	–	1.52	1.51
gold standard censored at death/weaning										
1	1.15	1.13	1.52	0.88	1.11	1.48	1.48	1.48	1.76	1.75
2	–	–	–	0.93	1.17	–	–	–	2.25	2.25
3	–	–	–	0.92	1.15	–	–	–	2.01	2.00
Coverage rates										
gold standard censored at death										
1	0.91	0.91	0.54	0.98	0.98	0.93	0.93	0.92	0.99	0.98
2	–	–	–	0.92	0.99	–	–	–	0.99	0.99
3	–	–	–	0.91	0.99	–	–	–	1.00	1.00
gold standard censored at death/weaning										
1	0.98	0.98	0.98	0.48	1.00	0.96	0.96	0.97	0.95	0.96
2	–	–	–	0.29	1.00	–	–	–	0.97	0.97
3	–	–	–	0.32	1.00	–	–	–	0.95	0.96