Estimating treatment effects in studies of perinatal transmission of HIV

HEEJUNG BANG
Department of Biostatistics, University of North Carolina, 137E Franklin Street, Suite 400, Chapel Hill, NC 27599-8030, USA
heejung_bang@unc.edu

DONNA SPIEGELMAN
Departments of Biostatistics and Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

SUMMARY
Fetal loss often precludes the ascertainment of infection status in studies of perinatal transmission of HIV. The standard analysis based on liveborn babies can result in biased estimation and invalid inference in the presence of fetal death. This paper focuses on the problem of estimating treatment effects for mother-to-child transmission when infection status is unknown for some babies. Minimal data structures for identifiability of parameters are given. Methods using full likelihood and the inverse probability of selection-weighted estimators are suggested. Simulation studies are used to show that these estimators perform well in finite samples. Methods are applied to the data from a clinical trial in Dar es Salaam, Tanzania. To validly estimate the treatment effect using likelihood methods, investigators should make sure that the design includes a mini-study among uninfected mothers and that efforts are made to ascertain the infection status of as many babies lost as possible. The inverse probability weighting methods need precise estimation of the probability of observing infection status. We can further apply our methodology to the study of other vertically transmissible infections which are potentially fatal pre- and perinatally.

Keywords: AIDS; HIV; Logistic regression; Missing data; Perinatal transmission; Semiparametric efficiency; Selection bias; Vertical transmission.

1. INTRODUCTION
Perinatal transmission is a devastating cause of the spread of the HIV-1 infection in developing countries (Guay et al., 1999; De Cock et al., 2000; Fawzi et al., 2000). The developing world now bears more than 90% of the world’s HIV/AIDS burden. In sub-Saharan Africa, more than 13 million women of childbearing age are infected with HIV and nearly one million infants already have been seroconverted prior to or during birth, or through breastfeeding. More than 2.8 million children/infants have died of AIDS worldwide (http://www.unaids.org/). In 2002, President George W. Bush announced a new ‘International Mother and Child HIV Prevention Initiative’ which focuses on treatment and care for HIV infected pregnant women to reduce transmission of HIV/AIDS to infants.

The effects of relatively inexpensive treatments such as multivitamin and short-course AZT on HIV transmission from pregnant women to their babies have been studied by many investigators (Kigotho, 1997; Guay et al., 1999; Marseille et al., 1999; De Cock et al., 2000; Fawzi et al., 2000). The USA has
Table 1. Data from a trial of perinatal transmission of HIV

<table>
<thead>
<tr>
<th>Untreated</th>
<th>T = 0</th>
<th>Treated</th>
<th>T = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetal loss</td>
<td>Live birth</td>
<td>Total</td>
<td>Fetal loss</td>
</tr>
<tr>
<td>$D = 1$</td>
<td>$D = 0$</td>
<td>$n_0$</td>
<td>$D = 1$</td>
</tr>
<tr>
<td>Infected $I = 1$</td>
<td>$x_{01}$</td>
<td>$x_{10}$</td>
<td>$x_{11}$</td>
</tr>
<tr>
<td>Uninfected $I = 0$</td>
<td>$x_{00}$</td>
<td>$n_0$</td>
<td>$x_{10}$</td>
</tr>
</tbody>
</table>

$x_{TDI}$ is the number of subjects having $TDI$ where $T$, $D$, and $I$ denote treatment, fetal loss, and infection status, respectively.

Table 2. Data from a mini-study (among HIV-negative mothers)

<table>
<thead>
<tr>
<th>Untreated</th>
<th>T = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetal loss</td>
<td>Live birth</td>
</tr>
<tr>
<td>$D = 1$</td>
<td>$D = 0$</td>
</tr>
<tr>
<td>Uninfected $M = 0$</td>
<td>$x_{010}$</td>
</tr>
</tbody>
</table>

$y_{TDI}$ is the number of subjects having $TDI$ among HIV-negative mothers ($M = 0$), where $T$, $D$, and $I$ denote treatment, fetal loss, and infection status, respectively.

seen a 43% decline in perinatally acquired AIDS but, in contrast, 25% more babies less than one year of age are currently dying of this epidemic in Zimbabwe and Zambia (Morris, 1997). In addition, African countries show unacceptably high infant mortality rates of 5–15% and the mortality rate is expected to be even higher for babies born to HIV-positive mothers.

In Dar es Salaam, Tanzania, a double-blind randomized clinical trial was conducted to assess the possible preventive effects of micronutrient supplements on mother-to-child transmission of HIV among 1078 HIV-1-positive pregnant women between 12 and 27 weeks gestation (Fawzi et al., 2000). A $2 \times 2$ factorial design was used to simultaneously assess the effects of multivitamin (excluding vitamin A) and/or vitamin A. Thirty-one fetal deaths occurred among 521 women assigned multivitamin, compared to 51 losses among 520 women on no multivitamin (vitamin A or placebo) arm; the odds ratio of multivitamin compared to no multivitamin was 0.58 (95% CI 0.37, 0.93) for preventing fetal loss where CI denotes a confidence interval. Similarly, the odds ratio of vitamin A compared to no vitamin A was 0.79 (95% CI 0.50, 1.24) (Fawzi et al., 1998). Infection status was not known for 8% of babies in this trial due to miscarriages and stillbirth and 21% of babies due to other reasons (see Table 6 for the data). Missing infection data and/or high mortality are common problems in the studies of vertically transmissible life-threatening diseases such as HIV/AIDS. The typical data layouts observed in such studies are shown in Tables 1 and 3.

The primary goal of this and other similar studies is to ascertain the effect of treatment on pre- or perinatal transmission. The standard analysis of vertical transmission studies conditions on survival of the infant and therefore is subject to selection bias since it entails comparing groups that are defined as subsets of the original randomization groups by a post-randomization factor, namely survival of the infant beyond
Estimating treatment effects in studies of perinatal transmission of HIV

Table 3. Data from a trial of perinatal transmission of HIV with partial missingness

<table>
<thead>
<tr>
<th></th>
<th>Untreated $T = 0$</th>
<th>Treated $T = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fetal loss $D = 1$</td>
<td>Live birth $D = 0$</td>
</tr>
<tr>
<td>Infected $I = 1$</td>
<td>$x_{01}$</td>
<td>$x_{00}$</td>
</tr>
<tr>
<td>Uninfected $I = 0$</td>
<td>$x_{01}$</td>
<td>$x_{00}$</td>
</tr>
<tr>
<td>Unknown $I = ?$</td>
<td>$x_{01}$</td>
<td>$x_{00}$</td>
</tr>
<tr>
<td>Total</td>
<td>$x_{01}$</td>
<td>$x_{00}$</td>
</tr>
</tbody>
</table>

birth. Naturally, inference drawn without addressing the potential for selection bias may be erroneous. In the presence of a relatively high infant mortality rate which itself is a serious problem in developing countries and may vary by treatment arm, estimation and inference for the effects of treatments in relation to perinatal transmission of HIV becomes more difficult. Other approaches to analysis include the study of a combined outcome of HIV infection and death, and of HIV-free survival (Hughes and Richardson, 2000; The Ghent Group, 2001; Fawzi et al., 2002). Hughes and Richardson (2000) presented a method for studying HIV-free survival and the joint distribution of HIV-1 infection time and death time up to two years, using survival data analysis methods. Since we are interested in perinatal infection (e.g. up to 3–6 weeks) and because censoring generally does not occur during this short interval, logistic regression is used to study the relationship between the (binary) outcome, e.g. infection or death, and the explanatory variable(s). HIV-free survival is an important public health outcome and making it the primary endpoint obviates the dependent censoring/selection bias issues. However, it is also of great scientific interest to evaluate the effect of treatment on reducing perinatal transmission separately from the efficacy of the treatment on reducing child mortality and morbidity. Treatments which are efficacious for fetal loss may not be for perinatal transmission. For example, regions where fetal loss has been virtually eliminated, as in the United States, would be more interested in treatments with benefit to reduce vertical transmission exclusively. Our approach would enable the investigator to identify different biological mechanisms for early infection and mortality attributed to HIV.

In this paper, we propose methods which provide valid estimation of the treatment effect on vertical transmission when the outcome of interest, transmission status, is partially unobserved. In Section 2, we investigate the nature of the bias which results from the estimation of the treatment effect among liveborn babies only when the fetal loss is the unique cause of missing outcome data. Likelihood-based methods are proposed for unbiased estimation of the treatment parameter with typically occurring data structures in Section 3. Model identification is studied for data structures and model assumptions which occur in practice. In Section 4, we present the class of the inverse probability of selection-weighted estimators and the algorithm to find the most efficient one from the class. Section 5 illustrates applications to the randomized clinical trial of vitamins against perinatal transmission in Dar es Salaam, Tanzania. Simulation results are summarized in Section 6. Conclusions and some additional remarks are provided in Section 7. Methods investigated in this paper can be applied to studies of treatments for vertically transmissible infections which are potentially fatal pre- and perinatally.
2. Bias in the standard analysis

Let $T$ be the indicator for the treatment assigned to mother, with value 1 if treatment is applied and 0 otherwise, $I$ be a binary variable denoting HIV infection status of the baby, i.e. $I$ is 1 if the baby is infected pre- or perinatally or shortly after birth and 0 otherwise, and similarly $D$ denotes fetal loss such that $D$ is 1 if the baby dies pre- or perinatally or shortly after delivery, therefore making it difficult or impossible to ascertain HIV status, and 0 otherwise. We assume that information about treatment and death is available from all subjects but baby’s infection status is missing for lost babies (Table 1).

A simple marginal model of primary interest in such randomized clinical trials is

$$\logit\{P(I = 1|T)\} = \beta_0 + \beta_1 T$$

(2.1)

where $\logit(x)$ is defined as $\log(x/(1 - x))$. However, in the presence of missing data in $I$, the standard approach is to fit the model

$$\logit\{P(I = 1|T, D = 0)\} = \beta_{u_1} + \beta_T$$

(2.2)

among liveborn babies. In this section, we study the bias of $\hat{\beta}_u$ from the model (2.2) as an estimator of $\beta_1$, the true parameter of interest. The analytic relation between the limit of $\hat{\beta}_u$ in probability and $\beta_1$ will quantify the bias, if any, and, if possible, provide a means for correcting it. If $I$ and $D$ are independent conditional on $T$, it turns out that estimation and inference for $\beta_1$ are valid by simply fitting model (2.2). But this is not true generally.

We introduce a model for fetal loss

$$\logit\{P(D = 1|I, T)\} = \gamma_0 + \gamma_1 I + \gamma_2 T + \gamma_3 I T.$$  

(2.3)

Applying Bayes' theorem, we obtain the following relation:

$$P(I = 1|T, D = 0) = \frac{e(\beta_0 + \beta_1 T)}{1 + e(\beta_0 + \gamma_1 + \gamma_2 T)} + e(\beta_0 + \beta_1 T)$$

and

$$e(\beta_u) = \left\{ \frac{1 + e(\gamma_0 + \gamma_2)}{1 + e(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3)} \right\} \left\{ \frac{1 + e(\gamma_0 + \gamma_1)}{1 + e(\gamma_0)} \right\} e(\beta_1),$$

(2.4)

where $e(\beta_1)$ is the odds ratio for the effect of the treatment on perinatal transmission.

First, we set $\gamma_3 = 0$, i.e. we assume that treatment does not offer any additional protection against fetal loss among (un)infected babies. If $\gamma_1$ or $\gamma_2$ is close to 0 or if $\gamma_0$ dominates among $\gamma = (\gamma_0, \gamma_1, \gamma_2)$, i.e. if the effect of treatment or infection on fetal loss is weak or if the overall risk of fetal loss is very low, then the bias will be trivial. In these cases, $\logit\{P(I = 1|T, D = 0)\}$ will be approximately linear in $(1, T)$ and $\hat{\beta}_u$ will be approximately consistent for $\beta_1$. The bias is computed over the range of values for the parameters of practical importance and some results are presented in Table 4.

If a consistent estimator for $\gamma$ is available from a suitably designed and comparable study, an estimator of $\beta_1$ can be computed from a naive estimator of $\beta_u$ via equation (2.4). Rarely will this be possible. However, these relations can be used for sensitivity analysis, when only $\hat{\beta}_u$ is identifiable. In the next section, we will propose a method for obtaining consistent estimators for $(\beta, \gamma)$, where $\beta = (\beta_0, \beta_1)$, based upon an augmented study design and the full likelihood.

Now, consider a log linear model for risk of fetal death,

$$\log(1 - P(D = 1|I, T)) = \gamma_0 + \gamma_1 I + \gamma_2 T + \gamma_3 I T$$
as an alternative to the logistic regression model (2.3). Note that we are modelling $1 - P(D = 1)$ for a simpler derivation. Applying Bayes’ rule again, we can show that $\beta_u = \beta_1 + \gamma_3$ and $\beta_{bu} = \beta_0 + \gamma_1$, i.e. if $\gamma_1 = 0$, unbiasedness holds for $\beta_1$ without any bias correction. Therefore, the presence of bias depends on the form of the link function.

### 3. Identification, Estimation and Inference

#### 3.1 Case 1: unknown infection status due to fetal loss

Suppose that the available data from the clinical trial are laid out in Table 1 and that the joint model for $[I, D|T]$ is given by models (2.1) and (2.3), i.e.

$$P(I, D|T) = \frac{e[(\beta_0 + \beta_1 T) I]}{1 + e[(\beta_0 + \beta_1 T)]} \times \frac{e[(\gamma_0 + \gamma_1 I + \gamma_2 T + \gamma_3 IT) D]}{1 + e[(\gamma_0 + \gamma_1 I + \gamma_2 T + \gamma_3 IT)].}$$

In this study, since infection status is unknown when $D = 1$, $x_{011}$, $x_{010}$, $x_{111}$, and $x_{110}$ are not observed, but the marginal counts $x_{01}$ and $x_{11}$ are observed instead, where $x_{TDI}$ is the number of subjects on treatment $T$ with vital status $D$ and infection status $I$. With this data, the log likelihood function can be constructed as

$$L_1(\beta, \gamma) = x_{000} \log[P(D = 0, I = 0|T = 0)] + x_{001} \log[P(D = 0, I = 1|T = 0)]$$
$$+ x_{010} \log[P(D = 1|T = 0)] + x_{100} \log[P(D = 0, I = 0|T = 1)]$$
$$+ x_{101} \log[P(D = 0, I = 1|T = 1)] + x_{111} \log[P(D = 1|T = 1)],$$

where $P(D|T) = \sum_{i=0}^{1} P(D = i|T)$. Terms in the likelihood that do not involve unknown parameters are ignored. We can show that $L_1$ is not identifiable for $(\beta, \gamma)$ even when $\gamma_1 = 0$ (Rothenberg, 1971). Consequently, there exists no consistent estimator of $\beta_1$ in this design.

Suppose besides the data observable from Table 1, additional data from HIV-negative mothers are available as in Table 2. These data can be obtained from a companion substudy among mothers screened for HIV infection in the course of ascertaining eligibility for the trial, who are found to be uninfected. $L_1$ is then augmented by two additional terms,

$$L_2(\beta, \gamma) = L_1(\beta, \gamma) + y_{010} \log(p_{010}) + y_{000} \log(p_{000})$$
$$= x_{000} \log(p_{000}) + x_{001} \log(p_{001}) + x_{010} \log(p_{010})$$
$$+ x_{101} \log(p_{101}) + x_{111} \log(p_{111}) + y_{010} \log(p_{010}) + y_{000} \log(p_{000}).$$

For simple notation, we define $p_{d|i} = P(D = d, I = i|T = t)$, $p_{d} = P(D = d|T = t)$, $p_{01} = P(D = 1|I = 0, T = 0)$ and $p_{00} = 1 - p_{01}$. Each probability can be derived as a function of $(\beta, \gamma)$ using the equations (2.1) and (2.3).
When \( \gamma_3 = 0 \), it is easy to show that \( L_2 \) is identifiable for \((\beta, \gamma)\) under the following assumption: 
\[
P(D|I = 0, T = 0, M = 0) = P(D|I = 0, T = 0, M = 1),
\]
where \( M = 1 \) if the mother is HIV-positive and 0 otherwise. That is, the risk of fetal loss for uninfected babies from uninfected, untreated mothers and for uninfected babies from infected, untreated mothers is equal. The underlying idea is that as long as a baby is uninfected, maternal infection does not convey additional information about risk for fetal loss. Although this assumption is not empirically verifiable, in our case, there are a priori clinical reasons to believe this assumption to be reasonable. Because mothers in this study were in the earlier stages of HIV disease (i.e., about 85% of mothers were in WHO HIV disease stage one) where symptoms are minimal, their infection is unlikely to affect fetal and infant health beyond baby’s infection (Fawzi et al., 2000). In fact, this assumption is the fundamental rationale for studies of treatments to prevent vertical transmission of HIV, currently an enormous medical research endeavor.

The maximum likelihood estimators (MLEs) obtained from \( L_2 \) are consistent and asymptotically normal by standard theory. However, if \( \gamma_3 \neq 0 \), the model is not identifiable, even with this augmented design. Because fetal loss is an explicit function of infection status, missingness is not ignorable in the sense of Rubin’s definition (1976). To get consistent estimates, the missing data mechanism, which is equivalent to the model for fetal loss, is explicitly included in the likelihood.

### 3.2 Case 2: infection status known among some babies lost and/or unknown among some liveborn babies

We now consider the scenario where there is information on infection status among a subset of the babies who died perinatally. In addition, we allow missing infection status among a subset of the liveborn babies. With three categories for infection status: uninfected \((I = 0)\), infected \((I = 1)\) and unknown status \((I = ?)\), we obtain the general data layout shown in Table 3.

For this data structure, we introduce a new variable, \( \Delta \), that is 1 if the infection status is known and 0 otherwise. Each cell probability in Table 3 is modelled as the joint distribution of \( I, D, \Delta, S \), i.e., 
\[
\]
Randomization ensures that, on average, the parameter estimate which relates \( T \) to \( I \) and \( D \) jointly, marginal over \( S \), should be unbiased. Then, 
\[
P(D, \Delta|T, S) = \sum_{i=0}^{1} P(\Delta|i, D, T, S; \eta)P(D|i, T; \gamma)P(i|T; \beta)
\]
is the relevant model for the observed data, where \( \eta \) is the parameters indexing the missingness model.

Under the assumption that \( P(\Delta|i, D, T, S) = P(\Delta|D, T, S) \), that is, conditional on \( D, T \) and a set of covariates \( S \), \( I \) does not convey additional information for missingness of itself, 
\[
P(D, \Delta|T, S; \eta) \sum_{i=0}^{1} P(D|i, T; \gamma)P(i|T; \beta).
\]
Therefore, the relevant log likelihood function for \((\beta, \gamma)\) has the form
\[
L_3(\beta, \gamma) = x_{\Delta01} \log(p_{\Delta01}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10}) + x_{\Delta11} \log(p_{\Delta11}) + x_{\Delta10} \log(p_{\Delta10})
\]
where \( x_{T,D} \) is the number of subjects with treatment \( T \) and vital status \( D \) but unknown \( I \). Note that \( P(\Delta|D, T, S) \) does not need to be estimated explicitly and that model \( L_3 \) is identifiable for \((\beta, \gamma)\). Under the two assumptions above (i.e., separability of \((\beta, \gamma)\) and \( \eta \), and missingness at random), the multiple imputation algorithm can be used for estimation and inference about \((\beta, \gamma)\) by reducing the problem to a log-linear model for a \( 2 \times 2 \times 2 \) table (Schafer, 1997; Little and Rubin, 1987). We also applied this method to our example.

However, if it is theoretically impossible due to operational constraints to observe infection status among lost babies, i.e., if \( x_{\Delta11} \) and \( x_{\Delta10} \) are structural zeros for \( i = 0, 1 \) in Table 3, then \( L_3 \) is no longer identifiable. Model identifiability for the possible data structures which arise in this setting is summarized in Table 5.
Table 5. Summary of model identifiability

<table>
<thead>
<tr>
<th>Available data</th>
<th>$\gamma_3 = 0$</th>
<th>$\gamma_3 \neq 0$</th>
<th>restriction on data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>1 &amp; 2</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>3 &amp; 2</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>No</td>
<td>$x_{ti} \equiv 0$ for $t, i = 0, 1$</td>
</tr>
<tr>
<td>3 &amp; 2</td>
<td>Yes</td>
<td>No</td>
<td>$x_{ti} \equiv 0$ for $t, i = 0, 1$</td>
</tr>
</tbody>
</table>

The entry in the first column is table number and the entries in the second and the third columns denote identifiability.

Table 6. Data from Fawzi et al.’s study for multivitamin (vitamin A)

<table>
<thead>
<tr>
<th>$T = 0$</th>
<th>$T = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D = 1$</td>
<td>$D = 0$</td>
</tr>
<tr>
<td>$I = 1$</td>
<td>2(1)</td>
</tr>
<tr>
<td>$I = 0$</td>
<td>0(0)</td>
</tr>
<tr>
<td>$I = ?$</td>
<td>49(44)</td>
</tr>
<tr>
<td>Total</td>
<td>51(45)</td>
</tr>
</tbody>
</table>

Robustness of the MLEs with respect to small cell counts will be investigated by simulation. Maximization of likelihood was implemented using an IMSL subroutine, and analytic gradient and Hessian were provided by automatic differentiation (Bischof et al., 1992).

4. INVERSE PROBABILITY OF SELECTION WEIGHTED ESTIMATORS

If no data are missing for $n$ subjects, maximizing the standard logistic model (2.1) is equivalent to solving a score equation $\sum_{i=1}^{n} (1 - T_i) \{ I_i - e^{(\beta_0 + \beta_1 T_i)} / [1 + e^{(\beta_0 + \beta_1 T_i)}] \} = 0$. With data missing in $I$ as in the scenarios given by Tables 1 and 3, the Horvitz and Thompson (1952) inverse probability of selection-weighted (IPSW) estimating equation is suggested by

$$\sum_{i=1}^{n} \frac{\Delta_i}{\pi_i} \left( \frac{1}{T_i} \right) \left[ I_i - \frac{e^{(\beta_0 + \beta_1 T_i)}}{1 + e^{(\beta_0 + \beta_1 T_i)}} \right] = 0,$$

(4.1)

where $\Delta$ is a missingness indicator as defined in Section 3.2 and $\pi = P(\Delta = 1)$. If $\pi > \varepsilon > 0$ for a constant $\varepsilon$, the unbiasedness of (4.1) and therefore the consistency of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ is clear. Due to missingness in $I$, we will try to find a set of covariates $S$ which satisfy the relation

$$\pi \equiv P(\Delta = 1|I, D, T, S) \cong P(\Delta = 1|D, T, S)$$

(4.2)

as in the previous section. If $\Delta = 1 - D$ (i.e. Table 1), then risk factors that satisfy

$$\pi = 1 - P(D = 1|I, T, S) \cong 1 - P(D = 1|T, S)$$

(4.3)

will be good choices for $S$. This probability is easily estimated by logistic regression.
The resulting estimator follows by the unbiasedness of (4.1) along with the fact that \(|\pi - \hat{\pi}| \) converges to zero in probability. In practice, we would not expect (4.2) or (4.3) to be precisely true but, given a rich collection of prognostic factors, it may well hold approximately. In HIV trials, many clinical, immunological and virological risk factors for disease progression are recorded. Therefore, identifying such covariates \(S\) may be possible, or approximately so, although this assumption is not empirically verifiable. When it is approximately true, (4.4) will be an approximately unbiased estimating equation asymptotically giving consistent and asymptotically normal estimators.

However, despite its asymptotic properties, there is an important limitation in terms of the inefficiency of using only the data from complete cases. To address this limitation, Robins et al. (1994) introduced a class of inverse probability weighted estimators and performed an efficiency study. Their theory is easily modifiable to prove that any semiparametric estimator of \(\beta\) (i.e. for the parameter of the marginal model for the mean of \(I\)) from the observed data \(\{O_i = (\Delta_i, I_i, D_i, T_i, S_i); i = 1, \ldots, n\}\), with a known weight function is asymptotically equivalent to an estimator solving an augmented IPSW estimating function:

\[
\sum_{i=1}^{n} \frac{\Delta_i}{\pi_i} \left( \begin{array}{c} 1 \\ T_i \end{array} \right) \left\{ I_i - \frac{e(\beta_0 + \beta_1 T_i)}{1 + e(\beta_0 + \beta_1 T_i)} \right\} = 0.
\]  

(4.4)

Consistency of the resulting estimator follows by the unbiasedness of (4.1) along with the fact that \(|\pi - \hat{\pi}| \) converges to zero in probability. In practice, we would not expect (4.2) or (4.3) to be precisely true but, given a rich collection of prognostic factors, it may well hold approximately. In HIV trials, many clinical, immunological and virological risk factors for disease progression are recorded. Therefore, identifying such covariates \(S\) may be possible, or approximately so, although this assumption is not empirically verifiable. When it is approximately true, (4.4) will be an approximately unbiased estimating equation asymptotically giving consistent and asymptotically normal estimators.

However, despite its asymptotic properties, there is an important limitation in terms of the inefficiency of using only the data from complete cases. To address this limitation, Robins et al. (1994) introduced a class of inverse probability weighted estimators and performed an efficiency study. Their theory is easily modifiable to prove that any semiparametric estimator of \(\beta\) (i.e. for the parameter of the marginal model for the mean of \(I\)) from the observed data \(\{O_i = (\Delta_i, I_i, D_i, T_i, S_i); i = 1, \ldots, n\}\), with a known weight function is asymptotically equivalent to an estimator solving an augmented IPSW estimating function:

\[
\sum_{i=1}^{n} U_i(\beta, h, q) = \sum_{i=1}^{n} \frac{\Delta_i}{\pi_i} h(T_i) \left\{ I_i - \frac{e(\beta_0 + \beta_1 T_i)}{1 + e(\beta_0 + \beta_1 T_i)} \right\} - \frac{\Delta_i}{\pi_i} E[q(O_i)|F_i] + q(O_i)
\]

(4.5)

where \(h(T) \) and \(q(O) \) are vector functions with the same dimension as \(\beta\), and \(F = (I, T, S)\) is the full data. It is easy to see that (4.1) is a member of the class. Our efficient IPSW estimating equation is proposed as (4.5), but replacing \(\pi_i\) by \(\hat{\pi}_i\).

The optimal choices of \(h_{\text{eff}}(T)\) and \(q_{\text{eff}}(O)\) which give the locally semiparametric efficient estimator are calculated by successive approximation (Robins and Wang, 1998). Given any vector function \(B = b(F)\) of \(\text{dim}(\beta)\), define \(q(O, b) = E[B(O)|F] = \Delta B + (1 - \Delta)E[B(T, S)\text{ and } h(T, b) = E[m(B)|T, I = 1] - E[m(B)|T, I = 0]\), where \(m(B) = E[q(O, b)|F]\). \(B\) is updated by an iterative algorithm

\[
B^{k+1} = B^k + \left( \begin{array}{c} 1 \\ T \end{array} \right) \epsilon(\beta) - m(B^k) + E[m(B^k) - B^k] \epsilon(\beta)|T \left[ \frac{e(\beta_0 + \beta_1 T)}{1 + e(\beta_0 + \beta_1 T)^2} \right]^{-1} \epsilon(\beta),
\]

where \(\epsilon(\beta) = I - e(\beta_0 + \beta_1 T)/(1 + e(\beta_0 + \beta_1 T)) \) and \(B^0\) is set to be \(0\). Then, \(h_{\text{eff}}(T) = h(T, b^\infty)\) and \(q_{\text{eff}}(O) = q(O, b^\infty)\), where \(b^\infty\) is the value at convergence.

Note that each quantity above \((B, \epsilon, h \) and \(q)\) is also a function of the unknown parameter, \(\beta\). During updates, \(\beta\) can be fixed at the value of any consistent estimate. Implementation of the algorithm requires knowledge of the joint distribution of the full data. Hence, in practice, we specify and fit a fully parametric model with the observed data for \([T, I|T]\) and \([S|T, I]\) using, perhaps, the EM algorithm. An alternative specification, more conceptually appealing in epidemiology, would be \([S]\) and \([T|S]\) and \([I|T, S]\). The resulting estimator \(\hat{\beta} = \hat{\beta}(h_{\text{eff}}, q_{\text{eff}})\) is locally semiparametric efficient in the sense that it will attain the semiparametric variance bound if the fully parametric model for \(F\) is correctly specified. Even if the density of the full data is misspecified with respect to the nuisance densities, \([T]\) and \([S|T, I]\), the \(\hat{\beta}\) which solves (4.5) will still be a semiparametric estimator which is consistent and asymptotically normal with some efficiency. Note importantly that if \(\pi\) is misspecified, (4.5) is not unbiased. Thus, the strength of the association of \(S\) with \(\Delta\) plays an important role in the validity and efficiency of this approach.
Table 7. Vitamin effects on perinatal transmission of HIV

<table>
<thead>
<tr>
<th>Method</th>
<th>OR (95% CI)</th>
<th>p-value</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>1.54 (0.91, 2.61)</td>
<td>0.11</td>
<td>1.58 (0.93, 2.69)</td>
<td>0.09</td>
</tr>
<tr>
<td>$L_3$ with $\gamma_3 = 0$</td>
<td>1.39 (0.82, 2.36)</td>
<td>0.22</td>
<td>1.54 (0.91, 2.61)</td>
<td>0.11</td>
</tr>
<tr>
<td>$L_3$ with $\gamma_3 = 0^*$</td>
<td>1.22 (0.76, 1.94)</td>
<td>0.41</td>
<td>1.36 (0.85, 2.17)</td>
<td>0.20</td>
</tr>
<tr>
<td>HT</td>
<td>0.98 (0.61, 1.57)</td>
<td>0.92</td>
<td>1.57 (0.79, 3.12)</td>
<td>0.20</td>
</tr>
<tr>
<td>RRZ</td>
<td>0.93 (0.58, 1.48)</td>
<td>0.76</td>
<td>1.39 (0.78, 2.46)</td>
<td>0.26</td>
</tr>
<tr>
<td>Multiple Imputation*</td>
<td>1.04 (0.69, 1.56)</td>
<td>0.87</td>
<td>1.25 (0.81, 1.93)</td>
<td>0.31</td>
</tr>
</tbody>
</table>

*For $\gamma$, 2 was added to the cells in which the frequency is less than or equal to 2 in order to overcome potential numerical instability.

The asymptotic variance of $n^{1/2}(\hat{\beta}(h,q) - \beta)$ can be consistently estimated by $(\hat{\kappa}(h,q))^{-1} \Omega(h,q)(\hat{\kappa}(h,q))^{-1}$, where $\hat{\kappa}(h,q) = -n^{-1} \sum \delta U_i(\beta, h, q) / \delta \beta'$ and $\Omega(h,q) = n^{-1} \sum U_i(\beta, h, q) U_i(\beta, h, q)'$ evaluated at $\hat{\beta}(h,q)$.

5. AN ILLUSTRATIVE EXAMPLE FROM THE HIV CLINICAL TRIAL

In this section, we analyze the data collected from a randomized clinical trial conducted in Dar es Salaam, Tanzania to study the effects of vitamin supplements on transmission of HIV from mother to baby. Further details were given in Section 1 as well as in Fawzi et al. (1998). Since the publication of that paper, a small number of data updates occurred so our data reflect the most recent update. Perinatal transmission was defined as infection up to three weeks and the available data are shown in Table 6. Because infection status was known for only two babies among all perinatally lost babies, although the interaction parameter, $\gamma_3$, is technically identifiable, numerically the model is for all practical purposes not identifiable given the data.

For modelling the missingness probability, $\pi = P(\Delta = 1)$, we considered clinical information, including CD4+ and CD8+ lymphocyte counts, hemoglobin, erythrocyte sedimentation rate (ESR), white blood count, HIV stage at the times of randomization and delivery, and demographic and other background data, such as a mother’s age, weight, the number of sexual partners, loss of the index pregnancy, and maternal malaria and gonorrhea infection status. We fitted stepwise logistic regression models for missingness to these. The final model included baby’s vital status, defined previously as $D$ ($p$-value = 0.0001), mother’s age (years) ($p$-value = 0.001), ESR (mmh$^{-1}$) ($p$-value = 0.05) and hemoglobin (gdl$^{-1}$) ($p$-value = 0.05) measured at the time of delivery. The multiple correlation coefficient between the missingness indicator and the covariates in the final model was $R = \text{corr}(\Delta, S) = 0.52$ with $S = \{D, \text{age, ESR, hemoglobin}\}$. Note that we added 2 to extremely small cell counts ($\leq$2) for $L_3$ and multiple imputation. Although our method allows zero cells, unlike multiple imputation, we applied this rule to both methods to overcome potential numerical instability as well as to make a fair comparison. For multiple imputation, ten imputed data sets were used.

Results are summarized in terms of the odds ratio in Table 7. The naive approach suggested that both multivitamin and vitamin A appeared to increase the risk of perinatal transmission although neither of these effects were statistically significant. After adjusting for selection bias, the estimated effects are more compatible with no effect (all with increased $p$-values), relatively null or minimally protective effect for the multivitamin regimen, and the moderately adverse effect of vitamin A remained. Because $D$ was the most important determinant of missingness, we refitted the IPSW models with $S = \{D\}$. The result was very similar, further confirming our original finding.
The fit of $L_3$ assuming no interaction of treatment by infection status gave an estimated relative risk for fetal loss of 0.45 (95% CI 0.26, 0.78) for multivitamin and 0.62 (95% CI 0.38, 1.03) for vitamin A. The multivitamin regimen significantly reduced the fetal loss rate, and vitamin A also appeared beneficial in protecting against fetal loss but the effect was not statistically significant. In conclusion, further scientific efforts should be made in order to identify the roles of vitamin regimens on the biological mechanisms to determine the public health efficacy of HIV-related survival and infection.

6. SIMULATION STUDY

We conducted a numerical experiment to compare the performances of the proposed estimators in finite samples under various configurations of practical importance. We were especially concerned about small-sample bias when there are small cell counts, e.g. when $x_{111}$, $x_{110}$, $x_{011}$ and $x_{010}$ in Table 3 are small as would typically be the case.

We simulated the data according to Table 1 as follows; $T$ was generated from the Bernoulli distribution with probability 0.5 as in a balanced randomized design. First, we set $[\beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2, \gamma_3] = [-1.4, -0.7, -1.4, 2, -1.0]$. This specification generated a risk of perinatal transmission of 15% and 20% missing infection data due to fetal loss. To implement the two IPSW estimators, we created $S = [S_1, S_2]$ such as $S_i \sim N(c_1 I, c_2)$ and $S_2 \sim N(c_3 I + c_4 T, c_5)$ with constants $c = [c_1, c_2, c_3, c_4, c_5]$ chosen so that the multiple correlation coefficients between $D$ and $(T, S)$, $R = \text{corr}(D, (T, S))$, were 0.52 and 0.25, corresponding to accurate and poor surrogates, respectively. $c_3$ and $c_4$ were estimated by $\sum_i (1 - D_i) \hat{\pi}_i^{-1}(S_2 - c_3 I_i - c_4 T_i)(I_i, T_i)' = 0$, whilst $c_5$ was estimated as $n^{-1} \sum_i (1 - D_i) \hat{\pi}_i^{-1}(S_2 - \hat{c}_3 I_i - \hat{c}_4 T_i)^2$. $c_1$ and $c_2$ were estimated similarly. Note that $\pi_i = 1 - P(D_i = 1 | I_i, T_i)$ (according to Table 1). Next, the data were simulated according to the structure given by Table 3. We assumed the model $P(\Delta = 1 | D, T) = P(\Delta = 1 | D) = D \nu + (1 - D) \tau$ with $(\nu, \tau) = (0.8, 0.9), (0.3, 0.9), (0.03, 0.8)$ and $(0.03, 0.4)$. Our data had values of $(\nu, \tau)$ close to $(0.03, 0.8)$. Five hundred simulations were used for each scenario, with a sample size of 400 for the main study and 250 for the mini-study. For the second set of simulations, we set $\gamma_3 = 0.5$.

We summarize the results for the naive estimator $\hat{\beta}_u$ which is obtained from the live-born babies only, the MLEs, the Horvitz–Thompson estimator (HT) and the semiparametric efficient estimator proposed by Robins, Rotnitzky and Zhao (RRZ). All estimators will be compared to the ‘complete case estimator’, which uses the full data and can be computed only in simulation. In Table 8, the empirical relative bias ($\%$), mean squared error (MSE) and 95% coverage probability for $\hat{\beta}_1$, the treatment parameter, are given.

As expected, the naive estimator was severely biased but the MLEs were unbiased. We found through various experiments that the parameter estimates obtained from $L_3$ were robust to values of $\nu$ and $\tau$ which would lead to small cell counts. True variances were over-estimated leading to excessively large coverage probabilities as $\gamma_3$ moved away from zero and as fetal loss rate or missingness in $I$ increased. However, this problem nearly disappeared when the mini-study data was available ($n_x = 450$), as well as when we increased the sample size ($n = 800$). The mini-study data was used to avoid model non-identifiability for Table 1 and to overcome numerical instability for Table 3, as evidenced by the numbers in parentheses. We also repeated simulations near the null value of a treatment effect on infection and death, with similar results for bias and efficiency.

We suggest using the multiple correlation coefficient, $R$, to assess the quality of covariates for missingness in practice. With $R = 0.52$ as in the Fawzi et al. the HT and RRZ were virtually unbiased and the variance of the RRZ was very close to the variance of the complete case estimator. When the $\hat{\pi}$ is accurate ($R \geq 0.45$), the RRZ gave more efficient estimates than the HT, but when the weights were inaccurate ($R \leq 0.3$), we see that the HT and RRZ were biased.
Table 8. Summary of simulation results for treatment effect $\beta_1$

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>C.P.</th>
<th>Bias</th>
<th>MSE</th>
<th>C.P.</th>
<th>Bias</th>
<th>MSE</th>
<th>C.P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>61</td>
<td>0.37</td>
<td>0.80</td>
<td>27</td>
<td>0.24</td>
<td>0.93</td>
<td>24</td>
<td>0.12</td>
<td>0.93</td>
</tr>
<tr>
<td>ML-L2</td>
<td>(4)</td>
<td>(0.18)</td>
<td>(0.94)</td>
<td>not identifiable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(v, \tau) =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.8, 0.9)</td>
<td>1</td>
<td>0.09</td>
<td>0.96</td>
<td>1</td>
<td>0.09</td>
<td>0.96</td>
<td>-0.3</td>
<td>0.05</td>
<td>0.94</td>
</tr>
<tr>
<td>(0.3, 0.9)</td>
<td>3</td>
<td>0.11</td>
<td>0.97</td>
<td>1</td>
<td>0.10</td>
<td>1.00</td>
<td>-0.7</td>
<td>0.05</td>
<td>0.98</td>
</tr>
<tr>
<td>(0.03, 0.8)</td>
<td>3</td>
<td>0.13</td>
<td>0.97</td>
<td>1</td>
<td>0.10</td>
<td>1.00</td>
<td>-0.1</td>
<td>0.06</td>
<td>1.00</td>
</tr>
<tr>
<td>(0.03, 0.4)</td>
<td>-4</td>
<td>0.14</td>
<td>1.00</td>
<td>-51</td>
<td>0.40</td>
<td>1.00</td>
<td>6</td>
<td>0.14</td>
<td>0.99</td>
</tr>
<tr>
<td>(−41)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(−37)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td>$\pi_{\text{mis}}$</td>
<td>56</td>
<td>0.34</td>
<td>0.80</td>
<td>21</td>
<td>0.23</td>
<td>0.92</td>
<td>19</td>
<td>0.09</td>
</tr>
<tr>
<td>$\pi_{\text{cor}}$</td>
<td>6</td>
<td>0.18</td>
<td>0.94</td>
<td>-2</td>
<td>0.10</td>
<td>0.95</td>
<td>0.7</td>
<td>0.06</td>
<td>0.93</td>
</tr>
<tr>
<td>RRZ</td>
<td>$\pi_{\text{mis}}$</td>
<td>63</td>
<td>0.31</td>
<td>0.80</td>
<td>34</td>
<td>0.21</td>
<td>0.95</td>
<td>31</td>
<td>0.12</td>
</tr>
<tr>
<td>$\pi_{\text{cor}}$</td>
<td>1</td>
<td>0.09</td>
<td>0.95</td>
<td>1</td>
<td>0.09</td>
<td>0.95</td>
<td>0.4</td>
<td>0.05</td>
<td>0.93</td>
</tr>
<tr>
<td>Complete</td>
<td>1</td>
<td>0.089</td>
<td>0.94</td>
<td>1</td>
<td>0.089</td>
<td>0.94</td>
<td>-0.3</td>
<td>0.05</td>
<td>0.96</td>
</tr>
</tbody>
</table>

The results are based on $[\beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2] = [-1.4, -0.7, -1.4, 2, -1]$. ML, HT and RRZ denote maximum likelihood, the Horvitz–Thompson estimator and the semiparametric efficient estimator, respectively. 'Naive' and 'Complete' denote the estimators obtained from the observed data only and from the complete data, respectively. Bias, MSE and C.P. are the empirical relative bias (%), mean squared error and coverage probability with 5% type I error. $\pi_{\text{mis}}$ and $\pi_{\text{cor}}$ represent misspecified and precise weights.

7. DISCUSSION

We have studied the problem of estimating treatment effects on perinatal HIV transmission when infection status could be unknown for some babies, i.e. when fetal loss rates are relatively high as is common in many countries in transition. The behavior of bias in the treatment parameter was studied analytically and numerically in relation to values of the other parameters, when the missingness in the outcome is ignored.

We provided methods which give consistent and asymptotically normal estimators using maximum likelihood as well as semiparametric estimating equations. The former method enables us to find an asymptotically unbiased estimator without needing other auxiliary variables and then to use standard large-sample theory. However, possible non-identifiability of the models must be considered. Identifiability was studied for several important assumed models in relation to some typical data structures. A simulation study showed that the MLEs are highly stable and virtually unbiased even with small samples and high missingness in infection outcome. Numerically stable estimation of the variance and of the interaction parameter requires a larger sample size.

The applicability of the methodologies developed here is not restricted to marginal models which are of primary interest in this paper but can be extended to observational studies where conditional models may be considered for controlling confounding. The HT estimator requires the missing at random assumption and is easy to program. Therefore, this estimator has been widely adopted in various survival data problems such as medical cost and quality-adjusted life time analyses, although it is inefficient (Zhao and Tsiatis, 1997; Bang and Tsiatis, 2000). The RRZ estimator is motivated to bypass this limitation, but, if the dimension of $S$ is large, its implementation is computationally challenging. In addition, IPSW estimators do not suffer from identifiability problems. However, we found in our simulation study that
the ability to model the probability of missingness accurately plays an important role, unlike with the likelihood methods. No doubly robust estimator exists for the linear logistic model (Robins and Rotnitzky, 2001). In many cases, the model for the probability of missingness may be approximately correct when a rich collection of prognostic factors are available and when the resulting correlation coefficient between the missingness indicator \((1-D)\) or \(\Delta\) depending on the data structure) jointly with prognostic factors is as high as 0.45. Alternatively, and especially when \(\pi\) cannot be modelled well given the available data, the MLE using \(L_2\) or \(L_3\) may be the best choice for valid estimation of treatment effects under the two key assumptions: for \(L_2\), that risk of loss is the same for uninfected babies born to infected, untreated and uninfected, untreated mothers, and for \(L_3\), that missingness in \(I\) is conditionally independent of \(I\) given vital status, treatment and other measured covariates determining missingness, although this probability does not need to be known or modelled explicitly. However, if there exists an interaction between treatment and HIV status on fetal loss, i.e. \(\gamma_3 \neq 0\), if all fetal losses have unknown infection status and if there is no mini-study, the HT and RRZ estimators are the only estimators available.

Investigators conducting these sorts of trials should make sure that the design includes a mini-study among uninfected mothers and a collection of comprehensive covariate information, and that extraordinary efforts are made to ascertain the infection status of as many babies lost as possible. The methods discussed in this paper can also be applied to the study of other vertically transmissible infections.

Summary of proofs on model (non-)identifiability and a Fortran program to implement the analysis can be provided by the first author upon request.

ACKNOWLEDGEMENTS

This work is partially supported by the National Institute of Child Health and Human Development (NICHD R01 32257), and the Fogarty International Center (NIH D43 TW00004). The authors thank Dr Roger Logan for invaluable mathematical, computational and editorial assistance. They also thank Professor Wafaie Fawzi for providing the data and Professor James Robins for immense guidance in programming the efficient IPSW estimator.

REFERENCES


[Received November 11, 2002; revised June 26, 2003; accepted for publication July 2, 2003]