MATERNAL AND PERINATAL RISK FACTORS ASSOCIATED WITH VERTICAL TRANSMISSION OF HIV-1 INFECTION FROM MOTHER TO CHILD

MARK J. LAMIAS

ABSTRACT. In 2009, the US Centers for Disease Control and Prevention (CDC) estimated that almost 11,000 children younger than 13 years old in the United States were diagnosed with HIV infection. Of those, it is estimated that the vast majority (88%) acquired the HIV infection perinatally. Given that both perinatal transmission is the most common route of HIV infection in children under 13 in the US and that childbirth rates among HIV infected women have increased by approximately 30% in recent years, we can expect additional cases of mother to child HIV transmission (vertical) unless effective public health interventions are instituted. To do so, it is imperative to gain a better understanding of risk factors associated with vertical transmission of HIV-1 infection. We perform a secondary analysis of data from the Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted HIV-1 Infection Study to investigate these factors and to describe differences in maternal and perinatal characteristics between HIV-1 transmitters and non-transmitters. We employed a bootstrap aggregation procedure using stepwise logistic regression to identify the potentially important predictors of increased HIV-1 transmission. Of the 17 variables we investigated, only CD4-CD8 ratio, vaginal bleeding during pregnancy, AZT used during pregnancy, and ruptured membranes greater than 24 hours were identified as significant risk factors.

Date: December 5, 2014.
Key words and phrases. vertical HIV transmission, infant HIV risk factors, HIV-1.

Introduction

According to the 2010 HIV Surveillance Report published by the US Centers for Disease Control and Prevention,[1] an estimated 10,834 children younger than 13 years of age in the United States (US) were living with a diagnosed HIV infection in 2009. Of these an estimated 9,522 or 87.8% were infected via perinatal transmission. Furthermore, the same report indicated that of the 187 HIV diagnoses among children younger than 13 years of age in the US made in 2010, 77% of those cases were acquired perinatally. These statistics, coupled with the fact that the number of HIV infected women giving birth in the US has increased in recent years by as much as 30%,[2] suggests that we will likely continue to see additional cases of mother to child (or vertical) transmission of HIV unless effective interventions are applied. Identification of risk factors associated with vertical HIV transmission may help form the basis for public health interventions aimed at reducing pediatric HIV cases. Without preventative interventions based upon identified risk factors, it is estimated that 25% to 35% of the nearly 7,000 HIV positive women giving birth each year in the US would transmit infection to their children[3 4].

To elucidate potential risk factors, in this paper, we seek to identify maternal and perinatal risk factors associated with vertical HIV-1 transmission and to more generally describe the differences between mothers who vertically transmitted HIV-1 infection to their children and those who did not. We restrict our attention in this paper to HIV-1 infection as this is the most predominant type of HIV infection in the United States (and globally), while HIV-2 infection is mostly confined to West Africa[5].

Methods

Data Sources

To investigate the risk factors associated with the vertical HIV-1 transmission, we performed secondary analyses on a dataset obtained from the Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted HIV-1 Infection Study, which is a prospective study "initiated in 1990 to determine the prevalence, incidence, and types of cardiovascular and pulmonary complications in the fetus, newborn, and young child with vertically transmitted HIV infection and to describe the course and outcome of these disorders."[6 7] Additional particulars regarding the study and the data collection process are described elsewhere.[8] The dataset we analyzed consisted of 22 variables on a total of 508 singleton births with known HIV-1 status of the child among a unique set of 508 individual mothers. The complete dataset included variables: (1) describing the demographics of the mothers (mother's age and race), (2) indicating presence of pregnancy complications (vaginal bleeding during pregnancy, diabetes, preecalmpsia hypertension, ruptured membranes greater than 24 hours

during pregnancy, hospitalizations for any pregnancy complications), (3) describing characteristics of the baby (birth weight, sex), (4) describing characteristics of the delivery (vaginal vertex, vaginal-breech, or cesarean, gestational age), (5) indicating possible AZT usage by the mother (ACTG 076 protocol AZT use, AZT used during pregnancy), (6) indicating whether the mother participated in "risky" behaviors while pregnant (smoking, drugs, and consuming alcohol), and laboratory test results assessing the strength of the mothers' immune systems (CD4 and CD8 counts and percentages). A derived CD4-CD8 ratio variable was also computed and made available for analyses by taking the CD4 counts and dividing by the CD8 counts (or percentages if counts were missing).

Analysis Plan

Prior to performing any data analyses, a data analysis plan was completed that set forth data cleaning and recoding procedures; addressed the handling of missing values and sparse observations/cases; detailed table shells and specific analyses to be performed, including methods to assess analysis assumptions; and software to be used (SAS 9.3, Cary, NC and R 3.1.1). In addition, the plan specified *a priori* the level of significance to be used for each of the analyses ($\alpha = 0.05$). We briefly summarize here some of the particular details of the analysis plan.

Our analysis plan first required descriptive statistics and graphics depicting the distributions of all variables be prepared to identify problematic cases or variables (e.g. impossible values for range-restricted variables, invalid binary codings, etc.). The plan also called for creating a derivative variable CD4-CD8 ratio formed by dividing the CD4 count (or percentage) by the CD8 count (or percentage). The plan dictated an *a priori* assessment of data completeness of observations and variables. It was determined that sparse cases – those containing more than 40% missing values would be dropped from analyses if affecting only a small number of cases. Imputation of missing values would be considered if tests for missing completely at random (MCAR) failed and only on those variables that had less than 30% missing values. Specifics of these assessments are detailed in the *Statistical Analysis* section along with the specifics of the statistical tests/algorithms used and methods employed for checking necessary test and model assumptions.

Statistical Analysis

At the outset of our analyses, we examined the sparsity of the dataset as detailed in the statistical analysis plan. We counted the number of missing values for each case in the dataset and dropped those cases from subsequent analysis where more than 40% of the variables were missing ($n = 9$)* as dictated in the analysis plan. This left 499 cases available for subsequent analyses (Figure 1). Variables were examined for completeness as well, and missingness

*Of the lab test values, only CD4-CD8 ratio was used in the sparsity computations.

ranged from 0 to 10.82%. To determine if imputation was necessary to minimize missing value bias, we performed Little's chi-squared test (Appendix 1) to assess the assumption of Missing Completely at Random (MCAR) for multivariate data.[9] We did not find statistically significant evidence to suggest that the missing data patterns were anything but MCAR ($\chi^2 = 380.347$; $df = 342$; $p = 0.08$). As a result, no imputation was performed for missing values, and all subsequent analyses were completed using pairwise deletion to maximize the number of cases available for analysis and to increase power of statistical tests.

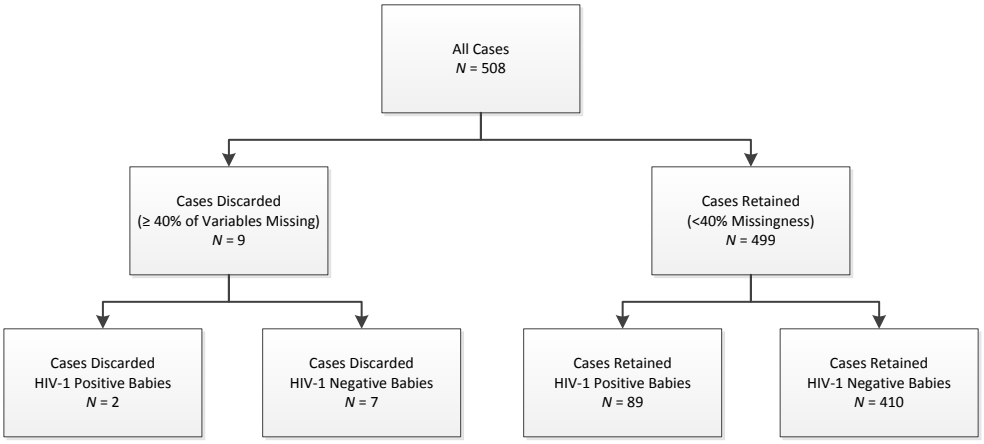

Figure 1: Final disposition of cases included in study

Basic descriptive statistics of the possible risk factors for all cases are shown in Table 1 and are used to summarize the data. In addition, the table displays measures of central tendency and variability stratified by the final HIV status of the child. Comparisons between the two groups were made using two-sample t-tests, chi-squared tests, Fisher's exact tests, Wilcoxon rank-sum tests, and a Monte Carlo exact chi-square test with 1 million simulations where appropriate.[†]

We employed a logistic regression procedure to identify risk factors associated with vertical HIV-1 transmission. However, before risk factor identification commenced, we examined our data for problems of multicollinearity by regressing (logistic) only first order terms of the potential risk factors onto the HIV transmitting status of the mother and then performing the following steps. We used the predicted values from the logistic regression model as the weights in a weighted multiple linear regression of the same risk factors on HIV transmitting status of the mother. We then calculated the tolerance and variance inflation factors (VIFs) associated with each variable as described in Allison.[10] Variables with tolerance values less

[†]This test was used only in contingency table analysis with cells containing zeros.

than 40% (VIFs > 2.5) were deemed potentially highly correlated and potentially problematic during the model building process. Only the CD4 and CD8 variables displayed large tolerance and VIF scores so they were deemed to have significant multicollinearity. As a result, these variables were dropped from analysis and the composite index, CD4-CD8 ratio, was used instead as it did not display significant collinearity with any remaining potential risk factors.

18 potential risk factors remained after dropping the CD4 and CD8 variables and creating the CD4-CD8 ratio. These variables were entered into a stepwise logistic regression procedure within a bootstrap aggregation (bagging) algorithm.[11] This algorithm (Appendix 2) created 1,000 bootstrapped datasets, and then performed Firth's penalized-likelihood logistic regression[12] on each dataset to avoid problems associated with quasi-complete separation of data points.[13] We took a liberal approach and only variables with coefficients significant at the 80% confidence level were entered into the model as recommended by Hosmer and Lemmeshow.[14] Factors appearing in 50% of the models (reliability) were retained as likely risk factors as recommended by Blackstone[15].

To begin assessing model assumptions, interaction terms from of all likely risk factors were successively entered into our logistic regression model and none were found to significantly contribute to the model. Therefore, our final model was computed using only the first order risk factors. We performed an assessment of the overall measure of the fit of the model using the Hosmer-Lemeshow Goodness of Fit Test.[13] The Hosmer-Lemeshow Goodness of Fit Test indicated that the logistic response function of the final model was appropriate ($\chi^2 = 2.3258$, $df = 8$, $p = 0.97$). Finally, we prepared index plots of leverage values, Cook's distance plots, and proportional-influence plots of delta deviance statistics to identify possible influential observations. Two mothers (ID Numbers 35484 and 55147) seemed to have relatively large influence on the logistic regression fit. These particular observations were dropped from the model and the logistic regression was performed again to examine any changes to model parameters and corresponding confidence intervals. The model estimates remained stable and nearly unchanged when the potentially influential observations were excluded. As a result, the excluded cases were added back into the final model. All odds ratios (ORs) and confidence intervals reported in the *Results* section are based on $n = 438$ observations. 61 observations were dropped from the final model due to missing values on at least one of the covariates appearing in the final model.

Results

The potential maternal and perinatal risk factors for all mothers and separately for HIV-1 transmitting (17.8%) and non-transmitting mothers are shown in Table 1. Of the 21 characteristics listed in Table 1, only gestational age, maternal CD4 lymphocyte percentage, maternal CD4 cell count (cells/mm3), maternal CD8 lymphocyte percentage, CD8%, maternal CD4-CD8 ratio, and ruptured membranes during pregnancy lasting longer than 24

hours, were found to be significantly different between transmitting and non-transmitting mothers. HIV-transmitting mothers had a median gestation age of 38 weeks compared to 39 weeks for their non-transmitting counterparts ($p = 0.016$). Mothers had a median maternal CD4 percentage of 27. Mothers that vertically transmitted HIV to their children had a median CD4 percentage of 23 while mothers fortunate enough to have prevented transmission had a CD4 percentage of 28 ($p = 0.002$). Maternal CD4 count was higher by nearly 100 cells per cubic millimeter among those mothers who prevented HIV transmission to their children ($median = 436$, $p = 0.022$). The average maternal CD8 lymphocyte percentage was 53.7 among transmitting mothers and somewhat lower at 50.3 among non-transmitting mothers ($p = 0.031$). Mothers that failed to prevent transmission of the virus to their babies had a median maternal CD4-CD8 ratio 11 percentage points lower than those who prevented transmission (0.44 vs. 0.55, $p = 0.003$). Not surprisingly, ruptured membranes lasting longer than 24 hours during pregnancy was more common among transmitting mothers than those who did not transmit (10.1% versus 3.2%, $p = 0.009$). Vaginal bleeding during pregnancy was nearly significant at the 95% confidence level, but did not quite achieve statistical significance ($p = 0.078$). Non-transmitting mothers were about half as likely (3.5%) as transmitting mothers (7.9%) to have experienced a vaginal bleeding event during their pregnancy. From a univariate perspective, mothers in both groups were nearly identical with respect to average age, percent white/black, child birth rates, and alcohol use during pregnancy.

Table 1. Maternal and Perinatal Descriptive Statistics of Mothers by HIV-1 Infection Status of Child

	ALL (n=499)		HIV-1 Infected (n=89)		HIV-1 Uninfected (n=410)		Infected vs. Uninfected
	N Missing	Summary Statistics	N Missing	Summary Statistics	N Missing	Summary Statistics	p-value
Age of Mother (years)[1]	14	27.2±5.4	2	27.8±6.0	12	27.0±5.3	0.271
Gestational age (weeks)[2]	4	39 (37, 40)	1	38 (37, 40)	3	39 (37, 40)	0.016*
Birth Weight (kg)[2]	3	3.07 (2.70, 3.47)	0	3.03 (2.69, 3.32)	3	3.09 (2.70, 3.49)	0.257
Maternal CD4 Percent[2]	50	27.0 (30.0, 34.0)	13	23.0 (16.5, 31.0)	37	28.0 (21.0, 35.0)	0.002*
Maternal CD4 count (cells/mm^3)[2]	52	423 (259, 628)	13	343 (212, 580)	39	436 (277, 634)	0.022*
Maternal CD8 Percent[1]	52	50.8± 12.8	13	53.7± 11.2	41	50.3± 13.0	0.031*
Maternal CD8 count (cells/mm^3)[2]	54	794 (530, 1087)	13	854 (566, 1145)	41	783 (524, 1073)	0.329
Maternal CD4/CD8 ratio[2]	52	0.54 (0.35, 0.76)	13	0.44 (0.26, 0.64)	39	0.55 (0.36, 0.80)	0.003*
Mother's Race[3]	16		1		14		0.290
White		80 (16.3)		14 (15.9)		66 (16.5)	
Black		239 (48.6)		39 (44.3)		197 (49.4)	
Hispanic		164 (33.3)		35 (39.8)		127 (31.8)	
Other		9 (1.8)				9 (2.3)	
Gender of Baby[3]	0		0			0	0.217
Male		265 (53.1)		43 (47.2)		223 (54.4)	
Female		234 (46.9)		47 (52.8)		187 (45.6)	
Vaginal bleeding during pregnancy[3,4]	4	21 (4.2)	0	7 (7.9)	4	14 (3.5)	0.078
Diabetes[3]	6	14 (2.8)	0	1 (1.1)	6	13 (3.2)	0.482
Preecalmpsia hypertension[3,4]	4	18 (3.6)	0	5 (5.6)	4	13 (3.2)	0.342
Ruptured Membranes > 24 hrs.[3,4]	3	22 (4.4)	0	9 (10.1)	3	13 (3.2)	0.009*
Hospitalized for preg. Complications[3]	6	83 (16.8)	0	19 (21.4)	6	64 (15.8)	0.209
Delivery Type[3,5]	0		0		0		0.510
Vaginal vertex		399 (80.0)		75 (84.3)		324 (79.0)	
Vaginal breech		5 (1.0)		0.00 (0)		5 (1.2)	
Cesarean		95 (19.0)		14 (15.7)		81 (19.8)	
AZT used during pregnancy[3]	8	166 (33.8)	2	27 (31.0)	6	139 (34.4)	0.547
Smoked during pregnancy[3]	18	181 (37.6)	2	34 (39.1)	16	147 (37.3)	0.758
Alcohol use during pregnancy[3]	24	93 (19.6)	4	17 (20.0)	20	76 (19.5)	0.914
Illicit drug use during pregnancy[3]	18	148 (30.8)	4	30 (35.3)	14	118 (29.8)	0.319
ACTG 076 protocol AZT use[3]	0	62 (12.4)	0	13 (14.6)	0	49 (12.0)	0.491

*p ≤ 0.05; 1 – mean ± standard deviation; 2 – median (q1, q3); 3 – n (%); 4 Fisher's Exact Test; 5 Monte Carlo Exact Chi-Square test with 1M simulations

Of the 17 variables included in the bootstrap aggregation stepwise regression procedure, only four risk factors were found to be associated with mother to child HIV transmission. Those factors are displayed in Table 2 along with their corresponding odds ratios, 95% confidence intervals, and the proportion of models finding the risk factor as a significant predictor of HIV transmission (reliability). Maternal CD4-CD8 ratio was identified as a risk factor in nearly every one of our bootstrap samples (96.3%). A ten percent increase in the CD4-CD8 ratio was found to correspond to an approximate 13.2% increase in the odds of HIV-transmission from mother to child (OR per 10% increase, 0.884; 95% CI, 0.810–0.964; $p = .0006$).

Table 2. Risk Factors Associated with Vertical HIV-1 Transmission

Risk Factor	Odds Ratio (95% CI)	P-value	Reliability (%) (N = 1,000)
Maternal CD4/CD8 ratio (per 10% increase)	0.884 (0.810, 0.964)	0.006*	96.3
Vaginal bleeding during pregnancy	3.375 (1.247, 9.134)	0.017*	75.8
AZT used during pregnancy	0.650 (0.370, 1.141)	0.134	73.5
Ruptured Membranes > 24 hrs.	3.121 (1.161, 8.385)	0.024*	72.2

n = 438

Mothers who experienced a vaginal bleeding event during their pregnancy could expect about a 238% increase in the odds of transmitting their HIV infection to their child (OR, 3.375; 95% CI, 1.247–9.134; $p = 0.017$). 75.8% of our bootstrap samples identified this risk factor as a statistically significant predictor of vertical transmission of HIV. Reliably identified in nearly 74% of our boostrap samples, use of AZT during pregnancy was found to be associated with a 35% decrease in the odds of HIV transmission (OR, 0.650; 95% CI, 0.370–1.141;$p = 0.134$), even though it was not found to be statistically significant in our final model at the 95% confidence level. Last, mothers experiencing at least one ruptured membrane during their pregnancy were associated with an approximate 212% increase in the odds of HIV transmission (OR, 3.121; 95% CI, 1.161–8.385;$p = 0.024$). This risk factor was identified in 72.2% of the bootstrap samples.

Discussion

Our analyses revealed important differences in the maternal and perinatal characteristics among women who transmitted HIV-1 infection to their infants during birth. The largest increases in the odds of mother-infant HIV transmission were associated with ruptured membranes greater than 24 hours and vaginal bleeding, which suggest that interventions targeting

the reduction in these types of pregnancy complications might directly decrease opportunities for HIV transmission to occur. The introduction of pharmacological treatments such as AZT and other medications aimed at boosting maternal CD4-CD8 ratio could also prove fruitful in reducing the odds of infant infection based on our analyses.

There are several limitations to our study. First, our methods do not allow us to assess the generalizability of our findings to larger populations since our subjects were enrolled in only one of five clinical centers and other studies were not examined for comparability of our findings. Furthermore, differences between clinical center sites may have been related to HIV transmission, but unfortunately, our dataset did not identify the clinical center where each mother was enrolled in the study. Mothers within sites may have correlated outcomes due to geographical or quality of health care similarities (or differences), but without clinical care identifiers associated with each mother, assessing these correlations and adjusting for them was not possible. In addition, although our missing data analysis revealed the data was MCAR, our analyses suffered from decreased power due to pairwise exclusion of cases with missing values appearing in variables used during any given analysis. Important variables to consider for future research might include length of time of bleeding and ruptured membranes as well as duration of AZT medication use. Our analyses were limited in that those receiving a single dose of AZT could not be differentiated from those receiving years of AZT treatment.

Despite our limitations we believe we have identified important risk factors associated with mother to child HIV transmission that can be further investigated in hopes of creating public health interventions targeting reductions in transmission. Our findings suggest that reduction in pregnancy complications like ruptured membranes and vaginal bleeding may reduce the odds of HIV transmission from mother to child. In addition, treatments with AZT medications and drugs used to boost maternal CD4-CD8 ratio may have moderate effects in decreasing the risk of vertical HIV transmission.

References

[1] US Centers for Disease Control and Prevention. HIV Surveillance Report, 2010; vol. 22. Published February 2012.

[2] US Centers for Disease Control and Prevention. HIV among Pregnant Women, Infants, and Children in the United States. Published December 2012. Accessed November 11, 2014. URL: http://www.cdc.gov/hiv/pdf/risk_WIC.pdf.

[3] Cooper ER, Charurat M, Mofenson L, et al. Combination antiretroviral strategies for the treatment of pregnant HIV-1-infected women and prevention of perinatal HIV-1 transmission. *J Acquir Immune Defic Syndr*. 2002;29(5):484-94.

[4] Human Immunodeficiency Virus (HIV) Infection in Children: Viral Infections in Infants and Children. *Merck Manual Home Edition*. Accessed December 4, 2014. URL: http://www.merckmanuals.com/.

[5] De Cock K, Adjorlo G, Ekpini E et al. Epidemiology and transmission of HIV-2 – why there is not an HIV-2 pandemic. *JAMA* 1993; 270: 2083–6.

[6] Demmler GJ, Istas A, Easley KA, Kovacs A. Results of a quality assurance program for detection of cytomegalovirus infection in the pediatric pulmonary and cardiovascular complications of vertically transmitted human immunodeficiency virus infection study. *Journal of Clinical Microbiology*. 2000; 38:3942.

[7] P2C2 HIV Study Group. Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted Human Immunodeficiency Virus (P2C2 HIV) infection study: design and methods. *Journal of Clinical Epidemiology*. 1996; 49:1285–1294.

[8] Biologic Specimen and Data Repository Information Coordinating Center. Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted HIV Infection (P2C2). Published Prepared on June 7, 2011. Accessed November 11, 2014. URL: https://biolincc.nhlbi.nih.gov/studies/p2c2/.

[9] Little, Roderick J. A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*; Vol. 83, No. 404 (Dec., 1988), pp. 1198-1202.

[10] Allison, Paul David, and Paul David Allison. *Logistic Regression Using SAS Theory and Application, 2nd Edition*. Cary, N.C.: SAS Institute, 2012, p. 62-3.

[11] Breiman L. Bagging predictors. *Machine Learning* 26:123-140, 1996.

[12] Firth, D. (1993), "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80, 27–38.

[13] Heinze, G. (2006), "A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data," *Statistics in Medicine*, 25, 4216–4226.

[14] Hosmer, D. W., Jr. and Lemeshow, S. (2), *Applied Logistic Regression, 3rd Edition*, New York: John Wiley & Sons.

[15] Blackstone EH. Breaking down barriers: Helpful breakthrough statistical methods you need to understand better. The Journal of Thoracic and Cardiovascular Surgery; 122: 430-439, 2001.

Appendix 1

```
# Adapted from code provided by Yale Pepper Center (2008)  author: Yuming Ning
#
/*****************************************************************************/
*
*
*  This SAS program implements the test for missing completely at random (MCAR) missing
*  data outlined in Little's
*  (1988) JASA article.  Note that this macro requires SAS version 8.2 (or higher),
*  because PROC MI is used to
*  obtain ML estimates of the covariance matrix and mean vector.
*
*
*
*  Little's test for MCAR on CHS data
*
*
*
/*****************************************************************************/
;


options center nofmterr ps=50;

*ASSIGN LIBRARY NAMES FOR DATA INPUT;
libname hiv 'D:\HIVStudy';

%macro mcartest;

/* PROVIDE VALUES FOR THE FOLLOWING MACRO VARIABLES */

%let numvars = 17;
* NUMBER OF VARIABLES IN DATA FILE;

*SPECIFY MISSING VALUE CODE;
%let misscode = .;

*SPECIFY VARIABLE NAMES;
%let varnames = mothage gestage birthwt mcd4cd8ratio
moth_rac sex bleed diabetes PREECLAM ruptmem hospcomp
deltype aztmeds momsmk momdrk momdrug actg076;

/* INPUT DATA */

data Ltest;
set hiv.Hiv_tran2;
keep &varnames;

data one;
      *infile &datafile ;
      *input &varnames ;
    set Ltest;

/* DO NOT ALTER THE CODE BELOW */

array m[&numvars] &varnames ;
array r[&numvars] r1 - r&numvars ;

do i = 1 to &numvars;
      if m[i] = &misscode then m[i] = .;
end;
drop i;
```

```
do i = 1 to &numvars;
      r[i] = 1;
      if m[i] = . then r[i] = 0;
end;
drop i;

proc sort;
      by r1-r&numvars;

proc mi data = one nimpute = 0 noprint;
      var &varnames;
      em maxiter=5000 converge=1e-3 outem = emcov;
      mcmc initial=em(maxiter=5000 converge=1e-3);

proc iml;

use one;
read all var {&varnames} into y;
read all var {%do i = 1 %to &numvars; r&i %end;} into r;
use emcov;
read all var {&varnames} into em;

mu = em[1,];
sigma = em[2:nrow(em),];

/* ASSIGN AN INDEX VARIABLE DENOTING EACH CASE'S PATTERN */

jcol = j(nrow(y), 1 , 1);

do i = 2 to nrow(y);
      rdiff = r[i,] - r[i - 1,];
      if max(rdiff) = 0 & min(rdiff) = 0 then jcol[i,] = jcol[i - 1,];
      else jcol[i,] = jcol[i - 1,] + 1;
end;

/* NUMBER OF DISTINCT MISSING DATA PATTERNS */

j = max(jcol);

/* PUT THE NUMBER OF CASES IN EACH PATTERN IN A COL VECTOR M */
/* PUT THE MISSING DATA INDICATORS FOR EACH PATTERN IN A MATRIX RJ */

m = j(j, 1, 0);
rj = j(j, ncol(r), 0);

do i = 1 to j;
      count = 0;
            do k = 1 to nrow(y);
                  if jcol[k,] = i then do;
                        count = count + 1;
                  end;
                  if jcol[k,] = i & count = 1 then rj[i,] = r[k,];
                  m[i,] = count;
            end;
end;

/* COMPUTE D^2 STATISTIC FOR EACH J PATTERN */

d2j = j(j, 1, 0);

do i = 1 to j;
```

```
/* OBSERVED VALUES FOR PATTERN J */

yj = y[loc(jcol = i),loc(rj[i,] = 1)];

/* VARIABLE MEANS FOR PATTERN J */

ybarobsj = yj[+,]/nrow(yj);

/* D = P X Pj MATRIX OF INDICATORS (SEE P. 1199) */

Dj = j(ncol(y), rj[i,+], 0);

count = 1;
do k = 1 to ncol(rj);
      if rj[i,k] = 1 then do;
            Dj[k, count] = 1;
            count = count + 1;
      end;
end;

/* REDUCE EM ESTIMATES TO CONTAIN OBSERVED ELEMENTS */

muobsj = mu * Dj;
sigmaobsj = t(Dj) * sigma * Dj;

/* THE CONTRIBUTION TO THE D^2 STATISTIC FOR EACH OF THE J PATTERNS */

d2j[i,] = m[i,] * (ybarobsj - muobsj) * inv(sigmaobsj) * t(ybarobsj - muobsj);

end;

/* THE D^2 STATISTIC */

d2 = d2j[+,];

/* DF FOR D^2 */

df = rj[+,+] - ncol(rj);
p = 1 - probchi(d2,df);

/* PRINT ANALYSIS RESULTS */

file print;
put "Number of Observed Variables = " (ncol(rj)) 3.0;
put "Number of Missing Data Patterns = " (j) 3.0; put;
put "Summary of Missing Data Patterns (0 = Missing, 1 = Observed)"; put;
put "Frequency | Pattern | d2j"; put;
do i = 1 to nrow(rj);
  put (m[i,]) 6.0 "    | " @;
    do j = 1 to ncol(rj);
      put (rj[i,j]) 2.0 @;
  end;
    put " | " (d2j[i,]) 8.6;
end;
put;
put "Sum of the Number of Observed Variables Across Patterns (Sigma psubj) = " (rj[+,+])
5.0; put;
put "Little's (1988) Chi-Square Test of MCAR"; put;
put "Chi-Square (d2)      = " (d2) 10.3;
put "df (Sigma psubj - p) =    " (df) 7.0;
put "p-value              = " (p) 10.3;

%mend mcartest;
```

```
%mcartest;

run;
```

Appendix 2

```
#*************************************************************************************#
#
#
#  This R 3.1.1 program implements the the bootstrap aggregation algorithm used to
#  Identifiy potential risk factors for vertical transmission of HIV-1.
#
#*************************************************************************************#


library(logistf)

bootsample<-lapply(1:1000, function(x) NULL)
bootstrap<-lapply(1:1000, function(x) NULL)
for (i in 1:1000){
  sample_i<-sample(499, replace=TRUE)
  bootsample[[i]]<-hiv.comp[sort(sample_i),]
}



for (i in 1:1000){
  my.df<-bootsample[[i]]
  bt.basemodel<-logistf(data=my.df, formula=SUBGRP~1, pl=FALSE, firth=TRUE)
  bt.model<-forward(bt.basemodel, trace=FALSE, printwork=FALSE, slentry = 0.20)
  bootstrap[[i]]<-names(coefficients(bt.model))
}

save.image("C:\\SavedResults.RData")

#Print reliability estimates
cbind(sort(sort(table(unlist(bootstrap))/1000), decreasing=T))
```