2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Analysis of time to HIV infection and determinants of mother-to-child transmission of HIV

Promotor :
Mevrouw Liesbeth BRUCKERS

Promotor :
Prof. FRANK CHIBWANDIRA

Emmanuel Saka

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Maastricht University

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Analysis of time to HIV infection and determinants of mother-to-child transmission of HIV

Promotor :
Mevrouw Liesbeth BRUCKERS

Promotor :
Prof. FRANK CHIBWANDIRA

Emmanuel Saka

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

ABSTRACT

Human Immunodeficiency Virus (HIV) causes Acquired Immunodeficiency Syndrome (AIDS) and is transmitted through unprotected sexual intercourse, unscreened blood transfusion, mother - to - child transmission and other forms of exposure to bodily fluids that carry the virus.

The primary objective of the thesis is to identify factors that determine mother - to - child transmission of HIV. The data constitutes of 576 mother baby pair. Mothers were followed during pregnancy and they accessed PMTCT management services. At delivery infants also accessed PMTCT management services. Follow up continued after delivery until HIV status of the infant was determined. Infants were tested for HIV using virologic test, HIV HIV DNA PCR. Logistic and parametric survival (accelerated failure time) models were fitted to answer the objects. In fitting the parametric survival model, failure time was assumed to follow Weibull and Exponential distributions. Likelihood ratio test was used to choose between Exponential and Weibull model. In addition to these two distributions, log-normal and log-logistic distributions were also considered.

Results of the logistic model show that Place of delivery, type of delivery, feeding option, and clinical danger signs are significantly, at 5% level of significance, related to MTCT. Furthermore results of the AFT model, assuming Exponential and Weibull distribution, show that feeding option and clinical danger signs are important factors for MTCT. Eventhough place of delivery and type of delivery were not significant, they were still kept in the models because they are important MTCT determining factors (WHO, 2010). AFT results further show that the Exponential model fits the data well as compared to the Weibull model.

From the results of the exploratory data analysis and model fits, it can be concluded that Place of delivery, Type of delivery, feeding option and clinical danger signs are factors that are related to MTCT and accelerate the survival time of the infants that are HIV exposed to become HIV infected. Health personnel therefore should not only take heed on how to take care of HIV positive pregnant women and those that are lactating in order to minimise MTCT but also consider other PMTCT management scenarios like putting pregnant women on short term ARVs, HAART, providing infants with sdNVP or other combination regimen and counselling on feeding options.

Key words: **AIDS, accelerated failure time, mother - to - child transmission, hazard, HIV.**

DEDICATION

To
Anna and Bongani

ACKNOWLEDGEMENTS

Many people at the Malawi Ministry of Health-HIV and AIDS Department, Thyolo District Hospital and Hasselt University in Belgium contributed in one way or another in this research. I would like to thank them all. It is impossible to list them all but a lot of injustice would be made if I failed to mention my special gratitude to Liesbeth Bruckers and Dr. Frank Chimbwandira for their excellent supervision and constructive criticism right from the kickoff day. Also I would like to thank Mr. Michael Eliya, Mrs. Chindiwo and Dalitso Midian of Ministry of Health, MSF-Belgium team at Thyolo District Hospital, Dr. Peter Kazembe of Baylor College of Medicine Children's Foundation, Dr. Carol Porter and Dr. Mark Kabue of Howard University Technical Assistance Project and Dr. Kondwani Ng'oma of United Nations Children's Fund for the initial set up of the Early Infant Diagnosis of HIV from which a section of the data used in this report comes from.

I am also thankful to the National Commission for Science and Technology (NCST) - Health Research Capacity Strengthening Initiative for the financial support all through my stay in Belgium.

In a special way I would like to thank my family for their moral and spiritual support. To my wife, Anna and son, Bongani - you were the wind beneath my wings. I sincerely apologise to Bongani for my absence on your 4^{th} and 5^{th} birthday.

A word of thanks should also go to Elasma Milanzi and Dalitso Mfunga for their moral support and encouragement. To all housemates in the "HOUSE WITHOUT A NAME" I say thank you for the company and sharing different kinds of food from your respective countries. I enjoyed every bite of it. To my groupmates, Leandro, Nolen and Veronica, I cherished every minute that I spent with you discussing numerous homeworks day in day out.

Finally credit should go to the Censtat Secretariat especially Martine Machiels, you were always there when I needed your help.

May God be with you all.

Emmanuel Saka

September, 2011

Diepenbeek

Table of Contents

List of tables

List of figures

List of abbreviations

AFT : Accelerated Failure Time

AIDS : Acquired Immunodeficiency Syndrome

ANC : Antenatal Care

CPT : Cotrimoxazole Prophylaxis

DBS : Dry Blood Spot

DNA : Deoxyribonucleic Acid

HAART : Highly Active Antiretroviral Therapy

HIV : Human Immunodeficiency Virus

TBA : Traditional Birth Attendant

MTCT: Mother - to - child Transmission

PMTCT : Prevention of mother - to - child transmission of HIV

PCR : Polymerase Chain Reaction

sdNVP : single dose Nevirapine

UNAIDS: Joint United Nations Programme on HIV/AIDS

Chapter 1. Epidemiology of HIV

1. Introduction

The human immunodeficiency virus (HIV) which causes acquired immunodeficiency syndrome (AIDS) has reached devastating proportions especially in Sub-Saharan Africa. UNAIDS estimate that there were 33.3 million people in 2009 living with HIV/AIDS worldwide out of which about 25.5 million were from the Sub-Saharan Africa. Even though the estimated number of infected adults has been decreasing, the estimated number of children below the age of 15 living with HIV increased from 1.8 million in 2001 to 2.5 million in 2009, UNAIDS, (2010).

The Republic of Malawi has a population of 13.1 million, NSO, (2010). The first serological evidence for HIV in Malawi was collected in the early 1980s. HIV prevalence increased sharply in the late 1980s and 1990s, and has stabilized around 12%, UNGASS (2010). A total of 840,156 adults and 111,510 children were estimated to be living with HIV in Malawi in 2007 (MoH, HIV and Syphilis Sero –Survey and National HIV Prevalence and AIDS Estimates Report, 2008). With an estimated 540,000 live births per year and an antenatal prevalence of 13%, approximately 81,000 pregnant women are infected annually.

AIDS pandemic is undermining Malawi's prospects for economic growth and poverty reduction. Valuable energies are being directed from productive uses to the care of people living with HIV, irreplaceable human capital is being lost, and hundreds of thousands of children and adults are being left destitute. However, in spite of the scale of the crisis, there are a few encouraging signs of reduction of the national HIV prevalence.

1.1 Mode of HIV transmission

HIV can be transmitted during unprotected sexual intercourse, during pregnancy (i.e., from mother to fetus), childbirth, breastfeeding, and other forms of exposure to bodily fluids that carry the virus. The main mode of HIV transmission in Malawi is through heterosexual sex. It is estimated that heterosexual sexual contact account for 90%, vertical transmission (mother-to-child) is estimated at 8% and blood transfusion accounts for only 2% (MOH, 2009). The major paths of mother - to - child transmission of HIV occur during delivery or during breastfeeding. MTCT can also occur in the uterine (*in utero*), Barber *et al.* (2005). Different factors influence HIV transmission during each period, and hence intervention to reduce transmission during each of these periods is of paramount importance. Since there are no national surveys to assess the HIV prevalence rate among the Malawian population, the most commonly used estimate of HIV prevalence in the reproductive age group is based on women visiting antenatal clinics (ANC).

1.2 Paediatric HIV diagnosis, care and treatment

More than 90% of HIV infection in children is acquired from the mother during pregnancy, labour and delivery, and after childbirth through breastfeeding. Early diagnosis of HIV infection in infants and children is important since the disease progresses rapidly in children with 50% mortality by the age two. According to Abrams *et al*. (2007), the process emphasises identification of HIV positive pregnant women at ANC and these mothers are followed up closely up until the status of the infant is determined. Passively transferred maternal HIV antibodies make interpretation of a positive HIV antibody test difficult in children under the age of 18 months. In order to diagnose HIV infection definitively in children under the age of 18 months, assays that detect the virus or its components (i.e. virologic tests) are required. Polymerase chain reaction (PCR) techniques for the amplification of HIV DNA or RNA are currently employed in the diagnosis of HIV infection particularly in infants under the age of 18 months. Antibody test may also be used to diagnosis HIV in infants provided that maternal antibodies have completely wane off which is usually at 9 or 12 months and exposure due to breastfeeding is ruled out. In Malawi the use of Dried Blood Spots (DBS) for HIV DNA PCR was adopted since it is easier to obtain, store and transport for centralized testing. Abrams *et al* (2007).

1.3 HIV prevention in adults and children

1.3.1 HIV prevention in adults

HIV prevention initiatives constitute primary response to HIV. This predominantly involves attempts to prevent sexual transmission to men and women, boys and girls through education programmes designed to promote abstinence, mutual faithfulness and the use of condoms. With the development of anti retroviral drugs/therapy (ART), which reduces the viral load and delays the progression from the HIV infection to AIDS, the number of deaths has reduced after the Government of Malawi, with funding from the Global Fund, introduced and scaled up provision of free anti retroviral drugs (ARVs) to those eligible. However, not everyone eligible has access to the life prolonging drugs due to several limitations, the prominent one being inadequate funds, (MOH, 2009). The evolution of HIV and AIDS in communities has raised a number of human rights issues. Stigma and discrimination based on HIV status is widespread. At the most basic level, this is manifested even in the prayer houses, i.e. churches and mosques, in the stigmatisation of people living with HIV as individuals who are deemed immoral, sinful or cursed.

1.3.2 General description of Prevention of mother – to – child transmission of HIV (PMTCT)

According to Barber *et al.* (2005), the most effective way to tackle paediatric HIV is to reduce mother-to-child transmission (MTCT) through early identification of HIV positive pregnant women at Antenatal Care, Labour and delivery ward and identification of HIV

exposed infants in the paediatric wards, underfive clinics Anteretroviral clinics, nutritional rehabilitation units and outpatient departments and putting them on care and treatment.

According to MoH (2007), pregnant women are counselled and tested for HIV when they report for Antenatal at any stage of pregnancy. Those that are HIV positive are closely followed up and access PMTCT services which include measure of CD4 count and or staged into the one of the 4 World Health Organisation HIV/AIDS stages, also they are put on CPT. Women who have low CD4 count (250 cells/mm^3) or are in either WHO HIV/AIDS stage III or IV are put on highly active antiretroviral therapy (HAART). Putting eligible pregnant women on HAART and short course ARVs, (single dose nevirapine (sdNVP) or other combination regimen), can significantly decrease vertical transmission of the virus. In the labour and delivery women are also offered HIV counselling and testing and if HIV infected they access PMTCT services. When the infant is born, it also receives sdNVP or other combination regimen.

Infants born from an HIV positive mother is described to be HIV exposed up until the status (HIV free or HIV infected) is determined. Follow-up of HIV-exposed infants provides an opportunity to provide life-saving cotrimoxazole prophylactic therapy and early HIV diagnosis, Barber *et al.* (2005). When the results of the virologic test come out to be positive, it implies that the infant is HIV infected and are put on ART irrespective of the stage of the disease or CD4 percentage whilst when the result is negative, it may imply that the infant is not HIV infected provided the infant has not been breastfeeding 6 weeks prior to the test otherwise the infant is still HIV exposed through breast milk and hence a confirmatory test to determine final status should be done 6 weeks after weaning as argued by Barber *et al.* (2005). Diagram 1 describes a general 'flow' of PMTCT processes.

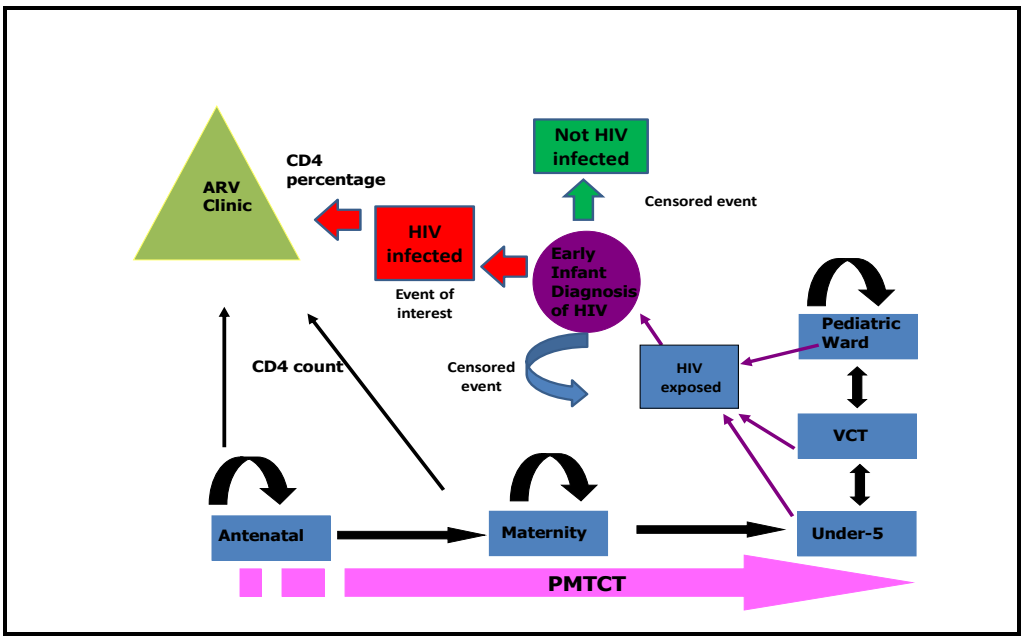

Diagram 1: Schematic presentation of PMTCT

1.4 Objective of the study

The primary objective of the study is to identify factors determining mother - to - child transmission (MTCT) of HIV in infants that are below 18 months old and to analyse the time to HIV infection in these infants.

Chapter 2. Methodology

2 Methodology

2.1 Exploratory data analysis

In order to get an insight of the data, exploratory data analysis was performed. Cross tabulation of PMTCT outcome (no HIV infection, HIV infection and HIV exposed) and covariates was performed. In order to test whether the observed counts differ from the expected counts, a Chi-square test was used. The Chi-square is given as:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where: O_{ij} is the observed cell counts and E_{ij} is the expected counts under the null hypothesis in the $ij - th$ cell of a (IxJ) table. The degrees of freedom are given as $df = (I-1)(J-1)$.

The expected cell counts under the null hypothesis are given as:

$$E_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$$

or

$$E_{ij} = \frac{[i^{th}row\ total][j^{th}column\ total]}{[total\ sample\ size]}$$

According to Agresti (2002), the Chi-squared test of independence merely indicates the degree of evidence of association and it requires large samples (minimum of 5 counts per cell). In cases where the counts in one cell are less than 5 then methodologies for small sample size have to be applied. One of the tests that can be applied when the sample size is small is the Fisher's Exact test. The Fisher's Exact test uses the hypergeometric distribution to calculate the probability of getting the observed data and all data sets with more extreme deviations, under the null hypothesis that the proportions are the same.

$$p(t) = p(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$$

Where t are observed values. The formula expresses the distribution of $\{n_{ij}\}$ in terms of only n_{11} and given the marginal totals, n_{11} determines the other cell counts. According to Agresti (2002) for the $2\ X\ 2$ tables, independence is equivalent to the odds ratio $\theta = 1$. In order to test $H_0: \theta = 1$, the p-value is the sum of certain hypergeometric probabilities. For the given marginal totals, tables having larger than n_{11}have larger sample odds ratios and hence stronger evidence in favour of the alternative. Thus the P-value equals $p(n_{11} \geq$

t_0) where t_0 is the observed value of n_{11}. More about the Fisher's Exact test can be found in Agresti (2002).

2.2 Basic survival analysis

In follow-up studies the exact survival time is only known for those study participants or units who show the event of interest during the follow-up period. For the others, what one can say is that they did not experience the event of interest during the follow-up period. These study participants or units are called censored observations. Individuals can be right censored, left censored or interval censored.

2.2.1 Censoring

According to Bakoyannis and Touloumi, (2011) subjects are right censored if it is known that the event of interest occurs some time after the recorded follow-up time whilst left censoring is when it is known that the event of interest happened sometime before the recorded follow up time. Interval censoring is when the exact time when the event occured is not known precisely, but an interval bounding this time is known. If interval is very short, it is common to ignore this form of censoring and pick one end point of the interval consistently. Interval-censored survival data frequently arise in clinical trials and follow up studies such as AIDS and cancer studies, DeGruttola and Lagakos (1989).

When occurrence times of the event of interest are observed exactly or are right-censored, meaning that it is only known that the event occurred after the last observation, the product limit estimator (Kaplan-Meier estimator) is frequently used in describing time to event experience of the subjects under study. When there is no competing risks, the classical survival data are usually presented as a bivariate random variable (T, C). The censoring variable, C, is 1 if the event of interest was observed, and is 0 if the observation was censored. When $C = 1$ the first member of the pair, T, is the time at which the event occurred and when $C = 0$, T is the time at which the observation was censored.

In some situations, however, the times of the events of interest may only be known to have occurred within an interval of time. The exact time of the change of state (such as HIV seroconversion) is not known exactly, only that it occurred sometime within a specific time interval or occurred sometime in the past, before the test was done. According to Lindsey *et al*. (1998) a common approach is to assume that the event occurred at the end (or at the beginning or midpoint) of the interval and then one can apply the standard survival methodologies of time to event data. However Lindsey *et al*. (1998) argue that such an approach can lead to invalid inferences, and in particular will tend to underestimate the standard errors of the parameter estimates.

Adopting Corrente *et al*., (2003) notation, interval censoring is presented as follows: let the lifetimes $T_l, l = 1, \ldots\ldots, n$ where n is the sample size, be grouped into k intervals, $I_i = (a_{i-1}, a_i), i = 1, \ldots, k, where\ a_0 < a_1 < \ldots\ldots < a_k = \infty$, (left and right). It is assumed that every censoring takes place at the end of the interval. Let D_i be the set of the subjects

that failed in the interval I_i; R_i the set of subjects at risk at the beginning of the interval I_i and Δ_{li} an indicative variable for the subject's lifetime l at the interval I_i.

$$\Delta_{li}=\begin{Bmatrix} 0 & if\ \ subject\ l\ is\ censored\ at\ I_i \\ 1 & Otherwise\ (the\ subject\ experienced\ the\ event) \end{Bmatrix}$$

This implies that Δ_{li} is a binary variable with probability $p_i(Z_l)$, which is defined as the probability for the l^{th} infant failing at a_i given it survived until $a_i - 1$ in presence of explanatory variable Z, that is either the subject fails or not in the interval I_i) and such being the case, the probability can be modelled using Cox proportional hazard model or logistic model by introducing the effect of the interval I_i as a latent variable. In most cases the interval is predetermined i.e. scheduled time.

Lindsey and Ryan (1998) argue that left censoring, exact and right censoring are a special cases of interval censored data. When left censoring occurs, the only information available to a statistician is that the survival time is less than or equal to the observed left censoring time. This implies that the event happened before the first visit, in this case before the first test of the HIV HIV DNA PCR. With right censored data, the event of interest did not occur until the last visit but it can happen at anytime from that moment. Data with both right and left censored observations are also known as '*doubly censored*' data.

Adopting Klein and Moeschberger (2002) notation, a lifetime X associated with a specific individual in a study is considered to be *left censored* if it is less than a censoring time C_l (C_l for "left" censoring time), that is, the event of interest has already occurred for the individual before that person is observed in the study at time C_l. For such individuals, we know that they have experienced the event sometime before time C_l but their exact event time is unknown.

The exact lifetime X will be known if, and only if, X is greater than or equal to C_l. The data from a left-censored sampling scheme can be represented by pairs of random variables (T,ε), where T is equal to X if the lifetime is observed and ε indicates whether the exact lifetime X is observed ($\varepsilon = 1$) or not ($\varepsilon = 0$). According to Lindsay and Tyan (1998) if T_i are not observed directly, but instead lies in the interval $[L_i, R_i]$, the general likelihood is given as:

$$\mathcal{L} = \prod_{d\epsilon D} f(x_d) \prod_{r\epsilon R} S\,(x_r) \prod_{l\epsilon L} \big(1 - S(x_l)\big) \prod_{i\epsilon I} ((S(U_i) - S(V_i))$$

Where D denote a set of HIV infection times, R as the set of right censored times. L as the set of left censored times and I as the set of interval censored times.

A schematic diagram of the data at hand is presented in figure 1. The longest time (X) it took for an infant to have the first PCR test is 15 months. The true survival is not known since we only have one PCR test for all the infants in the data set.

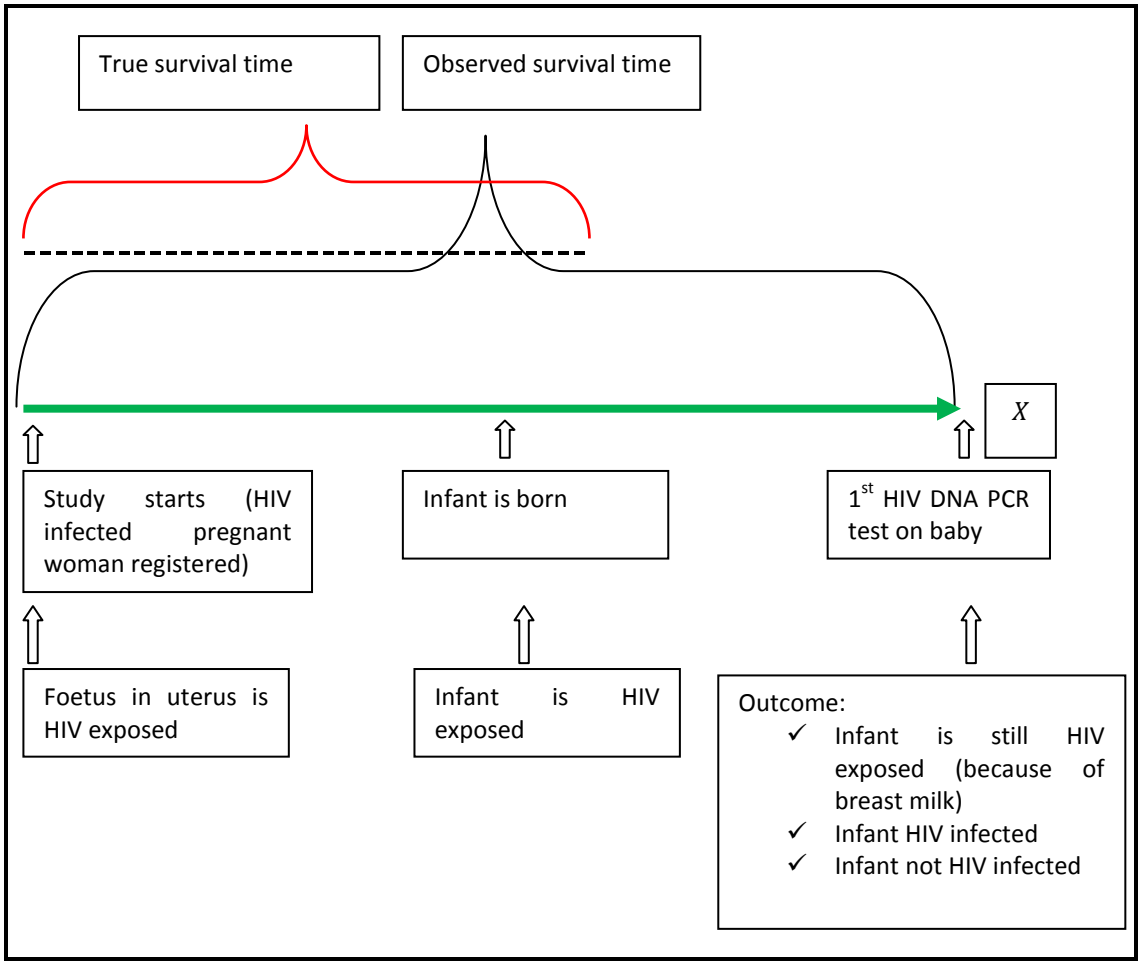

Figure 1. Follow up time of HIV exposed infants up to first HIV DNA PCR

2.3 Statistical Models

In this section statistical models that ignore censoring and the survival models are discussed. In order to model the probability of an infant to be HIV infected, a logistic regression model was fitted. In addition to the logistic regression model, parametric survival techniques were also applied.

2.3.1 Multiple logistic model

In a logistic model the dependent variable is usually dichotomous, that is it can take the value one (1) with a probability of success π or the value zero (0) with the probability of failure, $1-\pi$. This type of variable is called a Bernoulli (or binary). Logistic regression models have been extended to cases where a dependent variable is of more than two levels, known as multinomial or polytomous (Agresti, 2002). The logistic regression makes no assumption about the distribution of the independent variables. The relationship between the predictor and response variables is not a linear function but instead the logistic regression function is used, which is the logit transformation of π. The logistic regression model is a particular case of the generalized linear models family and presents logit link function and binomial random component (Agresti, 2002). Other links can as well

be used. Also known as the maximum likelihood models because the estimation of the β parameters uses maximum likelihood method (ML) (Molenberghs & Verbeke, 2005).

The saturated model, equation 1, was fitted to the data as a starting point.

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 HAART_i + \beta_2 MsdNVP_i + \beta_3 Pdelivery_i(home) + \beta_4 Pdelivery_i(healthcenter) + \beta 5Pdeliveryi(hospital) + \beta 6Tdeliveryi(vaginally) + \beta 7Tdeliveryi(caesarean) + \beta 8Tdeliveryi(vaccum\ extraction) + \beta 9Mcpti + \beta 10CD4i<250 + \beta 11CD4i250-500 + \beta 12CD4i(>500) + \beta 13IsdNVPi + \beta 14Agei + \beta 15Sexi + \beta 16Icpti + \beta 17Feediexclusive\ Breastfeeding + \beta 18Feedimixed\ feeding + \beta 19Feediexclusive\ formula\ feeding + \beta 20Cfeedi + \beta 21Nstatinormal + \beta 22Nstatimoderate + \beta 23Nstati(severe) + \beta 24Bweighti + \beta 25Cdangi \quad 1.$$

Where π_i is the probability of infant i to be HIV infected.

It is worth mentioning that the data under consideration is serological data, however the interpretation of the negative result is based on the breastfeeding status at the time of diagnosis. The outcome of the test is categorised as HIV infected, not HIV infected and still HIV exposed. The still HIV exposed group are infants that are still breastfeeding at the time of diagnosis, implying that they are still at risk to get HIV from the mother through breast milk. In order to properly incorporate the still HIV exposed group of infants in the logistic model analysis three scenarios were assumed, namely: a) Sixteen percent (16%) of the infants are HIV infected. This percentage was reported by MoH (2009), as the prevalence of paediatric HIV for infants that are born form mothers that are HIV positive. b) best case scenario: all the still HIV exposed infants are not infected, and finally c) Excluding in the analysis infants that are still HIV exposed. This approach was suggested by Soleiman *et al*. (2004), in the study of diagnosis of *Myasthenic* disease. However one has to bear in mind that doing such an analysis may lead to biased results since information is lost.

2.3.2 Parametric Survival models

Since the data is survival in nature, survival data analysis methodologies were applied. It is worth mentioning that only parametric survival techniques especially AFT model were applied for the reason that the failure time for the infants is not exactly known and most of the non or semi parametric methodologies assume exact failure time. In order to accommodate left, right and or interval censored data, it is natural to apply AFT survival analysis techniques (Kleinbaum and Klein, 2005) since failure time is being modelled. Even though this is the case non parametric e.g. Kaplan Mier or semi parametric e.g. Cox model techniques can be applied but one has to assume that the event happened to the left, midpoint or right time points. However Giola (2004) warns that this may lead to biased results.

According to Corrente *et al*., (2003) when it is not possible to observe the exact time when an event occured, but only its interval, tied observations are common. In the case of low number of ties, the analysis of these data can be done by the Cox proportional hazards

model by means of partial likelihood. When the number of tied observations is large, Corrente *et al.,* (2003) propose that time can be considered as discrete and a model can be fitted to the probability of occurrence of an event since it did not occur in the previous interval. Such fits can be made using the Cox proportional hazard model for grouped data or logistic model.

For the data at hand, what is known is that infants experienced the event of interest (HIV infection) before their first PCR test leading to unobserved true failure time and infants that were still HIV exposed or were HIV negative are right censored. Such being the case parametric survival methodologies especially the AFT model was sort in order to model the failure time.

2.3.2.1 AFT models

Parametric survival models can be proportional hazard (PH) or accelerated failure time (AFT). What distinguishes the two is not only in terms of the assumptions that they make about the shape of the hazard rate but also in terms of the specification and interpretation. In the PH, the effect of covariates is assumed to be fixed across time while AFT model assumes a linear relationship between log of (latent) survival time T and characteristics of observations. The general form of the AFT model is given as:

$$logT_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_p Z_{ip} + \sigma\varepsilon_i \qquad 2.$$

Where ε_i is the random error or the noise and is assumed to be independent and identically distributed (*i.i.d*) as per some distribution, $\beta_0,\ \beta_1, \beta_2, \ldots\ldots, \beta_p$ and σ are parameters to be estimates and Z are covariates and σ is the shape parameter which determines the shape of the hazard function. Failure time T_i falls between Right and Left or Upper and Lower since the failure time in this case is not exact. AFT model is similar to the classical linear regression; the only difference is the inclusion of censored observations in the data.

One can assume a variety of distributions for T of which the common ones are Weibull, Exponential, gamma, log-logistic and log-normal. These models can be fitted in the SAS procedure LIFEREG or in R. However, one has to do some data manipulation in order to conform to SAS procedure and R functions. In this report the Weibull, Exponential, log-normal and log-logistic were fitted. The Weibull and Exponential distributions are preferred because they can be expressed as a proportional hazard whilst for the case of log-logistic, the model is a Proportional Odds (PO) and the odds ratios are assumed to be constant over time.

2.3.2.1.1 AFT model - Weibull distribution

From equation 2, the Weibull model is given as:

$$log\lambda_i(\mathrm{t}) = \alpha log\mathrm{t} + \beta_0^* + \beta_1^* Z_{i1} + \beta_2^* Z_{i2} + \cdots + \beta_p^* Z_{ip} \qquad 3.$$

$$\text{where } \beta_j^* = \frac{-\beta_j}{\sigma} \text{ for all } j \text{ and } \alpha = \frac{1}{\sigma} - 1$$

Where σ is the shape parameter and $\lambda_i(\mathrm{t})$ is the hazard rate and is assumed to be monotonic. The parameter β_j is as a result of fitting the parametric PH model. When $\sigma > 1$, the hazard decreases with time. When $0.5 < \sigma < 1$ the hazard is increasing at a decreasing rate whilst when $0 < \sigma < 0.5$ the hazard is increasing at an increasing rate and finally when $\sigma = 0.5$ the hazard is an increasing straight line through the origin. The parameter σ has been defined as the acceleration factor. In a situation where $\sigma = 1$, the Weibull model reduces to an Exponential model. The error terms are assumed to have an extreme value, two parameters. The hazard function is given as $h(t) = \lambda\sigma t^{\sigma-1}$ and the survival time is given as $S(t) = \exp(-\lambda t^{\sigma})$. Having defined the hazard and survival functions, the Weibull model maximises the likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \left\{\lambda\sigma(\lambda t)^{\sigma-1} e^{(-\lambda t)^{\sigma}}\right\}^{d_i} \left\{e^{(-\lambda t)\sigma}\right\}^{1-d_i}$$

Where d_i indicates if an infant is censored (HIV negative or still HIV exposed) or has experience the event of interest (HIV positive).

For the Weibull model, if the AFT assumption is satisfied, then the proportional hazard assumption also holds and vice versa. The Weibull model also has another key property: the $log(-log)$ of $S(t)$ is linear with the log of time. This allows a graphical evaluation of the appropriateness of a Weibull model by plotting the $log(-log)$ of the Kaplan–Meier survival estimates against the log of time.

2.3.2.1.2 AFT model - Exponential distribution

From equation 2, the Exponential regression accelerated failure time model is given as

$$log\lambda_i(\mathrm{t}) = \beta_0^* + \beta_1^* Z_{i1} + \beta_2^* Z_{i2} + \cdots + \beta_p^* Z_{ip} \qquad 4.$$

The hazard for the Exponential model is given as ($h(t) = \lambda$) and is assumed to be constant overtime which is contrary to the Cox model where the hazard is proportional and not necessarily constant. In model 4, $\lambda_i(\mathrm{t})$ is the hazard rate. The survival time for the Exponential model is given as: $S(t) = exp(-\lambda t)$. The error terms are assumed to have an extreme value, one parameter. According to Kleinbaum and Klein (2005) constant hazard assumption pattern for each covariate is a strong assumption than the proportional hazard assumption. Kleinbaum and klein (2005) further argue that if the hazards are constant, then the ratio of the hazards is constant. However, the hazard ratio being constant does not

necessarily mean that each hazard is constant. Having defined the hazard function and the survival function, the Exponential model maximizes the likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \{\lambda e^{-\lambda t}\}^{d_i} \{e^{-\lambda t}\}^{1-d_i}$$

Where d_i indicates if an infant is censored (HIV negative or still HIV exposed) or has experience the event of interest (HIV positive).

The likelihoods, survival functions, density functions and the hazard functions for the log-logistic and log-normal can be found in Jenkins, S.P. (2008). The log-normal and log-logistic assume non-monotonic hazard and the error terms are assumed to follow normal and logistic distributions respectively.

2.4 Model selection

There are several ways in which covariates can be selected to be in the model and the methods not only include purposeful selection and stepwise selection (backward or forward selection procedures) but also best subset selection procedure. According to Hosmer *et al*. (1999), purposeful selection of covariates starts with a multivariable model that contains all variables significant in the univariate model at 20 to 25% as well as covariates with clinical or biological relevance. Following the fit of the initial multivariable model, covariates that are insignificant should be deleted from the model, starting with the most insignificant covariates. However Pintilie, (2006) warned that deleting too many non-significant variables at once should be done with care. The final step in covariate selection process is to determine whether interactions are needed in the model by forming a set of biologically plausible interaction terms from the main effects model.

After coming up with the model it worth to check the goodness- of - fit of the model. According to Agresti (2000) one can use the deviance or Pearson's Chi-square if the strata sample sizes are sufficient ($\geq$ 75% of the $n_j \geq 10$) however in a situation where the sample strata sizes are small ($\geq$ 25% of the $n_j \leq 10$) Hosmer and Lameshow's statistic may be used. A non significant p value indicates that the model fits well.

2.5 Model selection

Since AFT models constitute of submodels, it clearly means that one needs some way to decide on which model fits better to the data. Corrente *et al*. (1998) suggests using likelihood ratio test for comparing nested models. However this is not the case when one wants to make a choice between log-normal and Weibull, log-normal and Exponential since log-normal is not nested in the Weibull nor Exponential and hence they should be compared using Akaike's Information Criteria (AIC). The Akaike's method penalises each model's log likelihood to reflect the number of parameters that are being estimated. The lower the AIC the better the model.

The AIC is calculated as:

$$AIC = -2ln\mathcal{L} + 2(k + c)$$

Where k is the number of model covariates and c is the number of model specific distributional parameters.

In order to choose between the Weibull and Exponential model, a likelihood ratio test was used. The null hypothesis is that the Exponential model fits better whilst the alternative is the Weibull model.

$$H_0: Exponential\ model$$
$$H_1: Weibull\ model$$

The difference in the -2 log-likelihood between the two models follows a Chi-square distribution with 1 degree of freedom.

2.6 The data

The data set constitutes of 576 HIV positive mothers and their HIV exposed babies from Thyolo District Hospital in the Southern Region of Malawi. The mother baby pair were followed up until the infant had the first HIV HIV DNA PCR test. The mother baby pair had an access to PMTCT services. The event of interest is HIV infection, implying that the infant has acquired HIV from the mother. The event is assumed to have occurred sometime before the first HIV DNA PCR test such being the case their LEFT (lower time limit) is indicated as missing. This implies that these particular infants are left censored, 95 in total. Those particular infants that were still HIV exposed and or HIV negative (481) are right censored by setting their RIGHT (upper time limit) as missing as suggested by Allison (1997). Censoring was also due to death of the infant or lost to follow up. This data therefore can be considered to be both left and right censored and not interval censored data since only a single HIV DNA PCR test was performed on the infant. Summary of the variables are indicated in table 1.

Table 1. List of variables

Variable	Description	Code
pid	Patient identification number	
MAGE	Age of the mother	
CD4	CD4 count for the mother	1 = Less than 250 2 = Between 250 and 500 3 = More than 500
MCPT	Cotrimoxazole prothylaxis	0 = No, 1 = Yes
HAART	Highly antiretroviral therapy for the mother	0 = No, 1 = Yes
MsdNVP	Single dose nevirapine for Mother	0 = No, 1 = Yes
PDELIV	Place of delivery	1 = Home/TBA, 2 = health center 3 = Hospital
TDELIV	Mode of delivery	1 = Vaginally 2 = Caesarean section 3 = Forceps/vaccum extraction
IAGE	Age of the infant (in months)	
BWGHT	Birth weight (grams)	
SEX	Gender of the infant	1 = Male, 2 = Female
IsdNVP	Single dose nevirapine for infant	0 = No, 1 = Yes
ICPT	Cotrimoxazole prothylaxis for the infant	0 = No, 1 = Yes
CDANG	Clinical danger signs for the infant	0 = No, 1 = Yes
FEED	Feeding option	1 = Exclusive breastfeeding, 2 = Exclusive formula feeding 3 = Mixed feeding, 4 = Not breastfeeding
CFEED	Counselling on feeding option	0 = No, 1 = Yes
NSTAT	Nutritional status for the infant	1 = Normal, 2 = Moderate malnutrition 3 = Severe malnutrition
Censoring variable		
FTIME	Time it took for a first DNA PCR test (months)	
LEFT	Lower time bound	
RIGHT	Upper time bound	
Censoring	Indicating whether an infant is Censored or not	0 = Censored (death, HIV negative, Still HIV exposed) 1 = Not Censored (HIV positive)
Event	Indicator (whether the infant experienced the event or not)	0=Not HIV infected 1= HIV infected

Chapter 3. Results

3 Results

3.1 Exploratory Data Analysis

The data constitutes of 576 HIV positive women who were followed up during pregnancy. Follow up of these HIV positive mothers continued after delivery following with their HIV exposed babies. The maximum period it took for an infant to get the first HIV DNA PCR is 15 months with the minimum as early as 6 weeks. Out of these infants, a total of 48.95%, 51.05% were boys and girls respectively. It is worth mentioning that 9%, 3%, 3% and 2% of the infants had their first PCR between 0 to 6 months, 6 to 9 months, 9 to 12months and 12 to 15 months respectively experienced the event of interest, HIV infection. A larger number of infants had their definite PMTCT outcome not known because they were still breastfeeding. Table 2 summarises events by time it took for the first HIV HIV DNA PCR.

Table 2. Summary of first PCR diagnostic

Time when PCR test was done (months)	Infants who experienced the event	Censored infants (HIV negatives, Exposed, Lost to follow up)
0 to 6	54	521
6 to 9	17	474
9 to 12	15	411
12 to 15	9	371

The result of the exploratory data analysis also show that most (57%) of the infants had a birth weight of less 2500 grams.

3.1.1 Exploratory data analysis of outcome by maternal PMTCT management

Figure 2 and 3 indicate exploratory data analysis of outcome of HIV diagnosis by maternal PMTCT management. Figure 2 *panel a*, shows that 66% of the infants that were delivered at the hospital are still HIV exposed, 22% are not HIV infected whilst 11% are HIV infected. Figure 1 *panel a,* further shows that 64% of the infants that were delivered at home or TBA are still HIV exposed, 17% are not HIV infected while 19% are HIV infected. The results of the Chi-square test shows that there is a significant difference among place of delivery in relation to PMTCT outcome, $Chi - square = 19.9513, df = 4, p - value < 0.0001.$

The results in figure 2, *panel b*, show that 16% of the infants that were delivered vaginally were HIV infected while 21% were not HIV infected, the results also portray that 69% of the infants that were delivered through caesarean section are still HIV exposed, 6% are not HIV infected and 25% are HIV infected. The Fisher's Exact test was applied since 56% of the cell counts had expected frequency below 5. The results show that there is no significant difference among place of delivery with respect to PMTCT outcome. $(P\ value = 0.4700)$

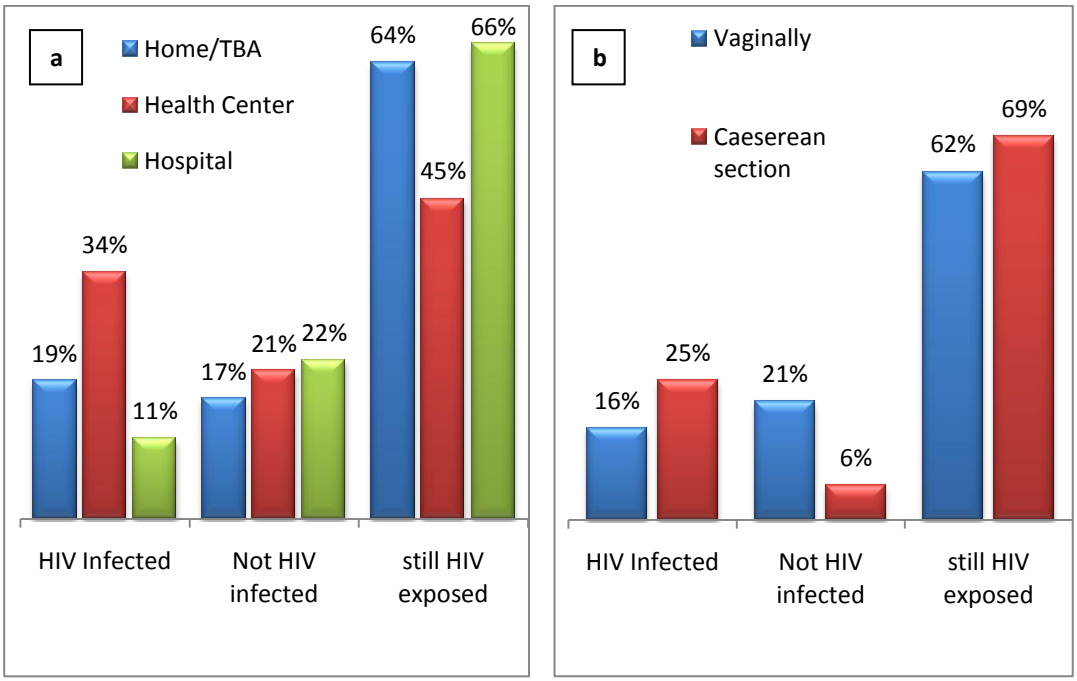

Figure 2. Outcome by PMTCT management for the mother (Place of delivery and Type of delivery in panels a and b respectively)

Results in figure 3, *panel a*, portray that infants whose mothers are on HAART, 66% were still HIV exposed at their first HIV HIV DNA PCR, 25% were not HIV infected while 8% were HIV infected. The results further show that infants who were born from mothers that were not on HAART, 26% were HIV infected and 53% were still HIV exposed while 21% were not HIV infected. The results of the Chi-square test shows that there is a significant difference, at 5% level of significance, between various PMTCT outcome for infants that their mothers were on HAART or not, $Chi - square = 14.9410, df = 2, p - value = 0.0006$.

PMTCT outcome results when the mother received sdNVP or not are in *Panel b* of figure 3 and the results show that 55% of the infants whose mother received sdNVP were not HIV infected while 30% were still exposed and 15% acquired HIV from their mother. Figure 3 *panel b* further shows that 75% of infants whose mother did not receive sdNVP were still HIV exposed, 8% were not HIV infected while 17% were HIV infected. The Chi-square test results show that there is a significant difference, at 5% level of significance, between these groups, $Chi - square = 8.3943, df = 2, p - value = 0.015$.

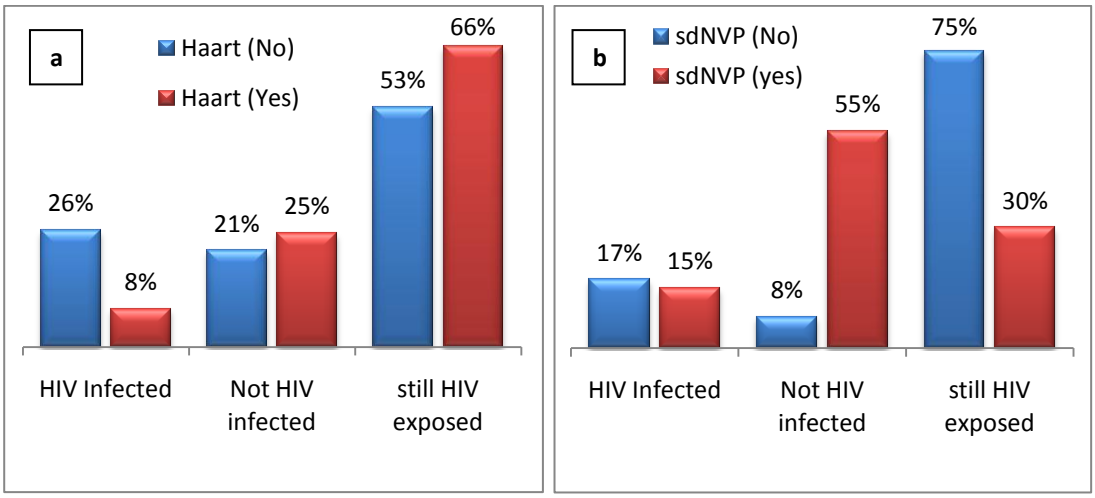

Figure 3. Outcome by PMTCT management for the mother (HAART and sdNVP in panels a and b respectively)

3.1.2 Exploratory data analysis of outcome by infant PMTCT management

Results in figure 4 and Figure 5 indicate exploratory data analysis of the outcome by infant PMTCT management (sdNVP, feeding option, counselling on feeding option and nutritional status). Results in figure 4 *panel a* indicate whether the infant received sdNVP or not versus the HIV status and the results show that 48% of infants that received sdNVP within 72 hours after birth are still HIV exposed, 22% were not infected while 13% were infected. For infants that did not received sdNVP, 48% were still HIV exposed, 37% were HIV infected and only 16% were not HIV infected. The results of the Chi-square test shows that there is a significant different, at 5% level of significance, between infants that received sdNVP versus those that did not received sdNVP in terms of their PMTCT outcome, $Chi - square = 28.0251, df = 2, p - value < 0.0001$.

Now turning to feeding option, figure 4 *panel b,* the results show that when infants are mixed fed, 45% are HIV infected while 55% were still HIV exposed. When infants are on exclusively breastfeeding, 11% were HIV infected while 89% were still HIV exposed. It is worth mentioning that biologically a situation can not arise for an infant who is exclusively breastfeed and be HIV negative. This is because of continued exposure due to breast milk. The Chi-square results show that there is a highly significant difference, at 5% level of significant, among the groups, $Chi - square = 498.5274, df = 6, p - value < 0.0001$.

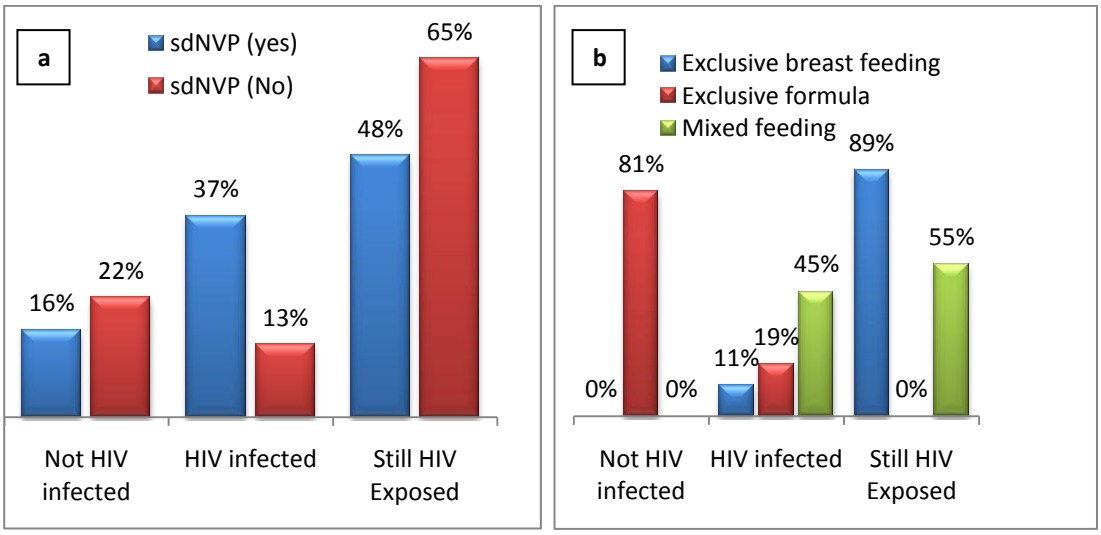

Figure 4. Outcome by PMTCT management for the Infant (sdNVP and feeding option *panels a* and *b* respectively)

Exploratory data analysis in figure 5, *panel a*, shows that a higher proportion (90%) of the infants that were severe malnourished were HIV infected while 6% were not infected. *Panel b* also shows that 72% of infants that had a normal nutrition status were still HIV exposed, 23% were not HIV infected while only 5% were HIV infected. For infants that were moderately malnourished, 9% were HIV infected while 6% and 3% were not HIV infected and still HIV exposed respectively. The results further show that 22% of the cells counts have expected counts less than 5 and such being the case Fisher's Exact test was applied and the results show that there is a significant difference among the groups, $p - value < 0.0001$.

Figure 5 *panel b* shows that mothers who were counselled on what feeding option to follow, 4% of their babies were HIV infected while 81% were still HIV exposed and also 15% were not HIV infected. For mothers that were not counselled on feeding option, equal percentage (36%) of their babies acquired HIV or not from their mothers while 27% were still HIV exposed. This calls for proper counselling of HIV positive mothers on issues to do with breastfeeding and other PMTCT management. However this is just an exploratory analysis and it has to be backed up by statistical analysis. The Fisher's Exact test was applied since 33% of the cell counts have expected counts less than 5. The results indicate that there is a significant difference between the groups, $p - value = 0.0004$.

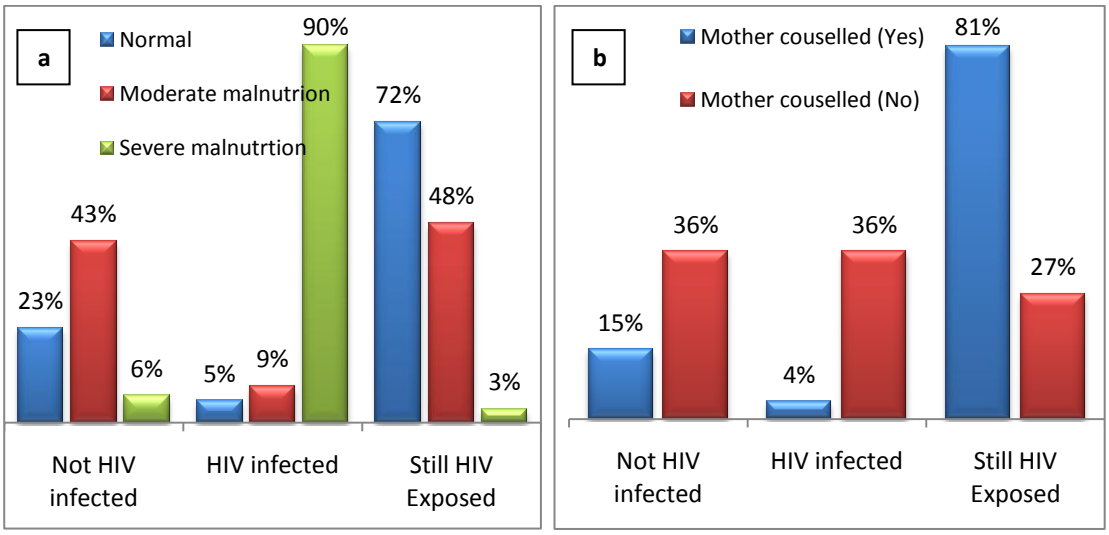

Figure 5. Outcome by PMTCT management for the Infant (nutritional status and counselling on feeding option in *panels a* and *b*)

The results in figure 6 *panel a*, show that most (84%) infants that are below 3 months old were still HIV exposed, the panel also shows that most of the infants that are HIV infected are above the age of 12 months and about 10% are aged below 3 months. Those infants infected and are aged below 3 months could have possibly acquired the virus *in-utero* (IU) or through delivery though postnatal transmission may not be ruled out. A bigger (59%) of the infants that are aged between 6 and 9 months were not HIV infected while 22% were infected. The chi-square test results indicate that there is a significance difference between the age groups in relationship to PMTCT outcomes, $Chi - square = 215.8657, df = 8, p - value < 0.0001$.

The results, *panel b* of figure 6 further show that infants that had a birth weight of below 2500 grams were not HIV infected, 20.7% were HIV infected and while 12.5% were still HIV exposed. Also 57% of infants that had a birth weight of above or equal to 2500 grams were not HIV infected. The Chi-square for test of difference shows that there is a significant difference, at 5% level of significance, in PMTCT outcome for infants that had a birth weight of below 2500 grams versus those that had a birth weight above 2500 grams, $Chi - square = 9.4064, df = 4, p - value = 0.0091$.

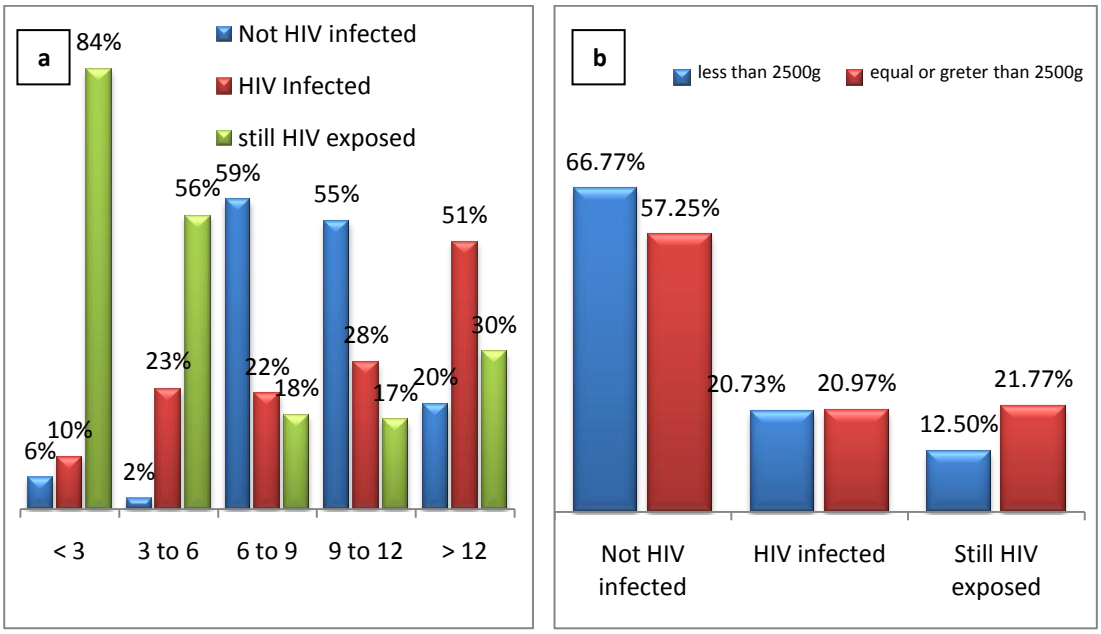

Figure 6. Outcome by Age and birth weight in *panels a* and *b* respectively

In summary, the results of the exploratory data analysis show that there is a significant difference in PMTCT outcome for HAART, mother single dose nevirapine, infant single dose nevirapine, place of delivery, feeding option, counselling on feeding, nutritional status, and birthweight. The results further show that there is no significant difference in terms of PMTCT outcome for type of delivery and sex of the infant. To ascertain this, statistical analysis using various models were applied.

3.2 Incomplete data

Figure 6 shows missingness of the data by covariate. The figure shows that more than half (52%) of the women are missing their CD4 count. Since the missingness is too high it was not included in the modeling processes. Also about 18% of the mothers had their information on whether they received sdNVP or not was missing. Missingness for the infant PMTCT management are in *appendix 1*. The results show that most of the infants were missing counselling data on feeding which is also an important covariate of PMTCT.

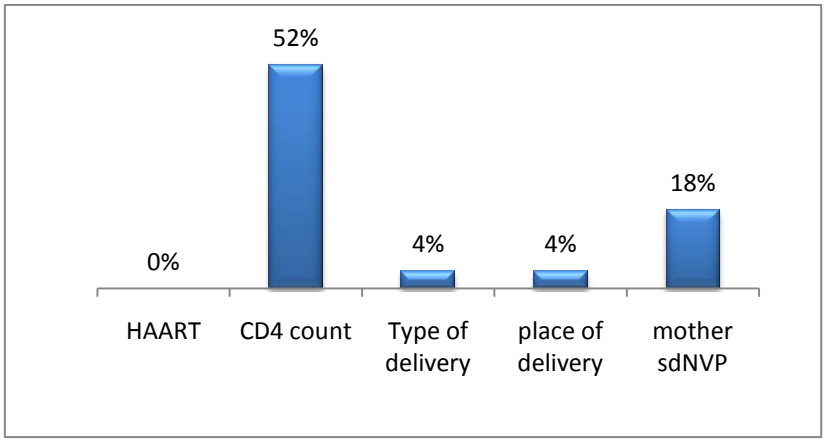

Figure 7. Missingness by covariate (maternal factors)

3.3 Statistical Analysis

Statistical analysis composed of fitting multiple logistic model (equation 1) and acceleration time failure model (equation 2). The results of the two model fits are discussed in this section. For the AFT model, Weibull, Exponential, log-logistic and log-normal distributions were considered.

3.3.1 Multiple logistic model

3.3.1.1 Logistic model (assuming 16 percent of the infants that their status is undetermined as infected)

Before fitting the model, the data was manipulated. First those infants that were still HIV exposed after their first HIV DNA PCR, were removed from the data set and 16% of these infants were randomly selected using the SAS procedure proc survey select. These infants were assigned the HIV positive status and the ones that were not selected were assigned the HIV negative status. The data set were then merged with the one that contains confirmed HIV status.

The results of the logistic model fit in table 3 show that place of delivery, clinical danger signs, feeding option are related to mother - to - child transmission of HIV. Type of delivery and age were not significant at 5% level of significant, however they were still kept in the model since they are important factors related to mother - to - child transmission and age may be a possible confounder. The results of the model show that the odds of an infant who was delivered vaginally to acquire HIV from the mother are 1.29 with 95% confidence interval of [0.2802 ; 5.9655]. This implies that the odds of an infant who was delivered vaginally to be HIV infected are 29% higher as compared to the odds of an infant who was delivered through caesarean section. The odds of an infant to acquire HIV from the mother who was delivered at the Health center are 2.44 times the odds of an infant that was delivered at a hospital, with a 95% confidence interval of [1.3820 ; 4.3004] whilst for home/TBA deliveries the odds are 1.33 and the 95% confidence interval is [0.8540 ; 2.5000]. This implies that the odds for an infant who was delivered at home/TBA to become HIV infected are 33% higher than the odds when delivery took place at the hospital.

The results of logistic model further show that the odds of an infant showing clinical danger signs to acquire HIV from the mother are almost 4 times the odds of an infant who is not showing clinical danger signs (odds ratio=3.79 with 95% confidence interval of [2.0867 ; 6.8719]. Now turning to breastfeeding, the odds of an infant to acquire HIV from the mother while on exclusive breastfeeding are 2.45 times, 95% Confidence interval of [1.1568 ; 5.1691], the odds of the infant who is neither breastfeeding. The results further show that the odds of an infant to acquire HIV from the mother when on mixed feeding are 3.01 times, 95% confidence interval [1.4327 ; 6.3593], the odds of an infant who is not breastfeeding. The covariates mother HAART, mother sdNVP, Infant single dose nevirapine,

child on CPT, counselling on feeding option and birth weight were not significantly related to MTCT at 5% level of significant and hence were removed from the model.

Table 3. Parameter estimates for the logistic model (16% of the still HIV exposed group is HIV infected)

Effects		Parameter	Estimate	Std. Error	Pr > ChiSq
	Intercept	β_0	-2.7113	0.8631	0.0017
Marternal factors	Place of delivery (Home/TBA)	β_1	0.3792	0.2740	0.1664
	Place of delivery (Health Center)	β_2	0.8911	0.2896	0.0021
	Type of delivery (Vaginaly)	β_3	0.2568	0.7802	0.7420
	Type of delivery (vacuum extraction)	β_4	1.0981	0.9559	0.2506
Infant factors	Age (in months)	β_5	0.0556	0.0457	0.2239
	Child showing clinical danger signs (Yes)	β_6	1.3316	0.3040	<.0001
	Feeding option (exclusive breastfeeding)	β_7	0.8942	0.3819	0.0192
	Feedng option (formula feeding)	β_8	0.3931	0.7411	0.5958
	Feeding option (mixed feeding)	β_9	1.1048	0.3802	0.0037

3.3.1.2 Logistic model (Excluding all unconfirmed tests)

Table 4 shows the results of the logistic model after excluding unconfirmed HIV test results (infants that are still HIV exposed after their first PCR test). The results show that place of delivery, type of delivery clinical danger signs, age and feeding option are highly significant at 5% level of significant. However one can see that the parameter estimates are larger as compared to when 16% percent proportion is applied to categorised infants that are still HIV exposed as either infected or not. It is also worth mentioning is that the standard errors are in most cases higher than when the 16% proportion is applied. One can also see that the parameters for some covariates have changed sign, this might be a sign of sensitivity of the data.

Table 4. Parameter estimates for the logistic model (excluding HIV exposed infants)

	Effects	Parameter	Estimate	Std. Error	Pr > ChiSq
	Intercept	β_0	-22.8783	1.3541	<.0001
Marternal	Place of delivery (Home/TBA)	β_1	1.3024	0.4730	0.0059
	Place of delivery (Health Center)	β_2	1.7280	0.4573	0.0002
	Type of delivery (Vaginally)	β_3	23.2834	1.3340	<.0001
	Type of delivery (vacuum extraction)	β_4	24.7879	0.0000	<0.0734
Infant	Age (in months)	β_5	- 0.2563	0.0582	<.0001
	Child showing clinical danger signs (Yes)	β_6	2.3062	0.4793	<.0001
	Feed option (exclusive breastfeeding)	β_7	- 2.2592	0.7838	0.0039

3.3.1.2 Logistic model assuming that all the HIV exposed infants are not HIV infected

The results of the logistic regression model considering the best case scenario are in *appendix 2*. The results show that place of delivery, type of delivery, child showing clinical danger signs, age of infant and feeding option (mixed feeding) are determining factors for MTCT. The model parameter estimates are generally larger as compared to when a 16% percent proportion of the HIV exposed infants are assumed to be HIV infected and the sign of the parameter estimate is consistent which is contrary to when infants that are still HIV exposed are excluded from the analysis.

3.3.2 Comparison of the logistic models and testing the goodness of fit for the models

From fitting the logistic model assuming these scenarios, one can see that the data is sensitive to the assumptions as it can be seen by the changes in standard errors, though the differences are not that large. However in most cases all the scenarios lead to the same factors that are related to MTCT. When the interest is in estimating odds ratio, one has to be mindful of the sensitivity of the models when different scenarios are assumed about infants with an unconfirmed PMTCT outcome.

In order to test the goodness of fit for the models, Hosmer and Lameshow test for goodness of fit was used. The results in table 5 show that all the model fits well since the results of the test provides a non significant p-value. Table 5 also shows that the model where infants that are still HIV exposed are excluded from the analysis fits better, since it has the smallest AIC as compared to other scenarios. However one has to be mindfull that excluding infants in the analysis may lead to loss of valuable information.

Table 5. Hosmer and Lameshow test for goodness of fit and AICs

Model	Chi-square	df	p-value	AIC
Assuming 16% of the HIV exposed infants are HIV infected	3.0228	7	0.8829	249.8396
Assuming all the HIV exposed infants are HIV negative	7.6012	7	0.3691	216.7082
Excluding HIV exposed infants in the analysis	5.0487	7	0.6540	118.0890

3.3.2 Parametric survival analysis model

Before fitting parametric survival models, data was manipulated in order to conform to the SAS procedure LIFEREG. The manipulation involves creating two time variables, LEFT and RIGHT or LOWER and UPPER. If the data are right censored, the upper or RIGHT is set to a missing (“.” in SAS and “NA” in R) and if the data are left censored, the lower or LEFT is set to missing. In a situation where you have interval censoring, which is not the case with the data at hand, the lower is less than the upper. In a situation where the lower is equal to the upper the patient is uncensored. Table 6 shows a section of the actual data set and a hypothetical situation for uncensored and interval censored patients.

Table 6. A sample of the reorganised data in SAS

Pid	Time to first PCR test	LEFT	RIGHT	Type of censoring	HIV status	Censoring indicator
3121-0-029-1	2	2.0	.	Right censored	Negative or exposed	0
3121-9-276-1	8	.	8.0	Left censored	Infected	1
3121-9-978-1	15	15	.	Right censored	Negative or exposed	0
3121-9-013-1	14	.	14.0	Left censored	Infected	1
	7	7	7	Uncesored*		
	3	1	2	Interval censored*		

*Not observed in the data at hand

After this data manipulation, the procedure LIFEREG will recognise that the data is both left, right and interval censored. The model statement is given as: $model\ (LEFT, RIGHT) =$

$covariates$, SAS codes are in *appendix 5*. The model statement indicates that the response time, T_i, for infant i is known to be between the time variables left and right. The procedure uses the Turnbull's algorithm as suggested by Turnbull (1976), reported by Giolo, (2004).

Results of the AFT model fits are in table 7 and 8. The results show that both the Weibull and Exponential models lead to the same factors that are related to MTCT i.e. child showing clinical danger signs and feeding option, at 5% level of significance. These results are different from the logistic model where place of delivery, type of delivery, child showing clinical danger signs, age of infant and feeding option (mixed feeding) were found to be determining factors for MTCT.

Comparing the Exponential and Weibull models, one can see that the parameter estimates are generally higher in most cases for the Exponential model. The results of the Weibull model further show that the hazard is increasing at decreasing rate as the survival time increases since the scale parameter is between 0 and 1.

The acceleration factors for the respective covariates for the Weibull model are discussed in this section. The acceleration factor is found by taking the exponent of the parameter estimate. The results of the acceleration factor for type of delivery (vaginal) suggests that the median survival (or any other quantile) time stretches by a factor of 0.76 times ($e^{-0.2775}$) as compared to when delivery is through caesarean section. When place of delivery is considered, the acceleration factor for Home/TBA delivery is 0.79 ($e^{-0.2301}$) times as compared to hospital deliveries whilst health center deliveries the acceleration factor for the survival time is 0.69 ($e^{-0.3657}$) times as compared to hospital deliveries. For the variable child showing clinical danger signs, the survival time is stretched 0.19 times ($e^{-1.6589}$) times for the infants that are showing clinical danger signs as compared to infants that are not showing clinical danger signs.

The estimated acceleration factor results also show that for infants who are exclusively breastfeed, the survival time is accelerated 11.20 times ($e^{2.4163}$) as compared to infants that are not breastfeeding and or are weaned, similarly infants that were mixed fed their survival time is accelerated 8.18 times ($e^{2.1020}$) times as compared to infants that are not breastfeeding and or are weaned. This is in agreement with Coovadia *et al*., (2007) who indicated that mixed feeding carries a higher risk of HIV transmission than exclusive breastfeeding. However one has to note that the risk when mixing breast milk with formula milk or solids is substantially higher than the risk from adding water or other non-food fluids. Although place of delivery and age are not significant at 5% level of significance, they were still kept in the model because they are important MTCT determining factors.

Table 7. Parameter estimates (s.e.) for the AFT models (Weibull distribution)

	Effect	Parameter	Estimate (s.e.)	Pr > ChiSq
	Intercept	β_0^*	1.1350 (0.9825)	0.2480
Maternal	Place of delivery (Home/TBA)	β_1^*	-0.2301 (0.3649)	0.5283
	Place of delivery (Health Center)	β_2^*	-0.3657 (0.3720)	0.3255
	Type of delivery (Vaginally)	β_3^*	-0.2775 (0.9507)	0.7704
	Type of delivery (vaccum extraction)	β_4^*	0.3740 (1.2177)	0.7587
Infant	Age (months)	β_5^*	-0.0079 (0.0909)	0.9310
	Child showing clinical danger signs (Yes)	β_6^*	-1.6589 (0.8394)	0.0481
	Feeding option (exclusive breastfeeding)	β_7^*	3.3549 (1.5845)	0.0342
	Feeding option (formula feeding)	β_8^*	-1.4568 (0.8142)	0.0736
	Feeding option (mixed feeding)	β_9^*	2.8702 (1.3260)	0.0304
	Scale		1.3759 (0.6145)	
	Shape		0.7268 (0.3246)	

For the Exponential model in order to get the hazard, one has to take the reciprocal of the acceleration factor, Kleinbaum and klein (2005). The hazards for place of delivery are 1.17 [$1/(e^{-0.1553}$)] and 1.33 [$1/(e^{-0.2826}$)] times as compare to deliveries that were done at Hospital for home/TBA and health center deliveries respectively. It can be noticed that the hazard for an infant to be HIV infected when delivery took place at health center are 33% higher as compared to when delivery took place at the hospital whilst home/TBA deliveries the hazard is 17% higher as compared to hospital deliveries. The result of the Exponential model also show that the hazard for feeding options are 0.09 [$(1/e^{2.4163}$)], 3.36 [$(1/e^{-1.2130}$)] and 0.12 [$1/e^{2.1020}$)] times as compared to infants that are not breastfeed/weaned for exclusive breastfeeding, formula and mixed feeding options respectively. For a month increase in age the hazard is 56% higher 1.56 [$(1/e^{0.0397}$)].

Table 8. Parameter estimates (s.e.) for the AFT models (Exponential distribution)

	Effect	Parameter	Estimate (s.e.)	Pr > ChiSq
	Intercept	β_0^*	0.9807 (0.7021)	0.1625
Maternal	Place of delivery (Home/TBA)	β_1^*	-0.1553 (0.2494)	0.5336
	Place of delivery (Health Center)	β_2^*	-0.2826 (0.2509)	0.2600
	Type of delivery (Vaginally)	β_3^*	-0.1616 (0.6884)	0.8145
	Type of delivery (vaccum extraction)	β_4^*	0.2722 (0.8879)	0.7591
Infant	Age (months)	β_5^*	0.0397 (0.0337)	0.2396
	Child showing clinical danger signs (Yes)	β_6^*	-1.2079 (0.2839)	<.0001
	Feeding option (exclusive breastfeeding)	β_7^*	2.4163 (0.2707)	<.0001
	Feeding option (formula feeding)	β_8^*	-1.2130 (0.5301)	0.0221
	Feeding option (mixed feeding)	β_9^*	2.1020 (0.3002)	<.0001

3.4 Model selection

To select between the Exponential and Weibull model, the likelihood ratio test was used. The -2 log-likelihood for the Exponential model is 325.430 and the log-likelihood for the Weibull model is 324.722. Hence the likelihood ratio Chi-square statistic is $325.430 - 324.722 = 0.708$ which is less than $\chi_{1,0.95} = 3.84$. Clearly we fail to reject the null hypothesis at 5% level of significance, imply that the Exponential model fits the data well.

3.5 Testing the Exponential model.

In order to test if the Exponential model is feasible to the data at hand, the hypothesis to be tested is given as: $H_0: \sigma = 1$ versus $H_1: \sigma \neq 1$. The scale parameter for the Exponential model is forced to 1 and the results of the model fit indicate that there is no enough evidence to reject the null hypothesis (Chi-square equal to 0.4586 with a p value of 0.4983) and it can be conclude that the Exponential model is feasible for the data.

3.6 Other distributions for the AFT models

Results of the Log-logistic and log-normal model are in *appendix 3* and *4* respectively. The results show that feeding option is the only significant covariate in both models. Table 9 shows the AIC values for the various models and results show that the Log-logistic model has the lowest AIC. For the Exponential and Weibull their AIC values are close.

Table 9. AIC values for the models

Model	AIC
Weibull	346.72
Exponential	345.43
Log-logistic	339.92
Log-normal	341.48

.

Chapter 4. Discussion and Conclusion

4 Discussion and Conclusion

4.1 Discussion

The results of the AFT model (assuming Weibull and Exponential distributions) and multiple logistic model assuming several scenarios are in most cases in agreement leading to more or less the same factors that are related to mother - to - child transmission of HIV. The results of the multiple logistic model indicate that place of delivery especially at Home/TBA, type of delivery (vaginally), child showing clinical danger signs and feeding option are factors significantly related to MTCT. Home deliveries are associated with MTCT for the reason that most, if not all, of the deliveries are handled by unqualified traditional birth attendants who have not undergone through any formal nursing and midwifery training. Such being the case they have limited knowledge on how to handle HIV positive pregnant women during delivery leading to higher chances of MTCT. Also these TBAs lack good quality equipment and may be reusing the same equipment from one woman to the other without properly sterilising them and such being the case increasing the transmission of HIV from one woman to the other in addition to MTCT. It is worth mentioning that deliveries at home/TBA may also deny infants and the mother short term ARVs (sdNVP and other combination regimen) and cotrimoxazole to prevent mother - to - child transmission of HIV and malaria respectively.

Mode of delivery is also an important variable in preventing mother - to - child transmission of HIV. The results show that infants that were born through vaginal mode of delivery are at high risk of MTCT. The reason might be that vaginal deliveries are associated with loss of too much blood and if good obstetric practices are not followed or if midwifes perform unnecessary episiotomy chances of MTCT increase.

Feeding option is one of the most important factors in determining prevention of mother - to - child transmission. In the first month of life between 0 to 6 months WHO recommends that infants have to be exclusively breastfed or mothers can adopt formula feeding so long as it is Acceptable, Feasible, Affordable, Sustainable and Safe (AFASS). The results of models show that infants that are on mixed feeding and who are exclusively breast feed have a higher chance of acquiring HIV from their mother. The reason might be that if infants are mixed fed there is a higher risk of MTCT since their intestines are not fully developed and hence when fed solid foods it may rapture the walls of the Gastrointestinal tract (GIT) and hence making the virus to easily transmit from mother - to - child through breast milk.

The results also show that fitting fully parametric survival model, between, the Exponential and Weibull distributions lead to the same factors that are related to MTCT and the Exponential model fits better than the Weibull whilst among the non nested models the Log-log-logistic model fits the data better as compared to the Log-normal, Exponential, and

Weibull. The results further show that there is a higher hazard for HIV infection in infants that are delivered at home and health center as compared to those that are delivered at hospital. Similarly there is a higher hazard of HIV infection for infants that were delivered vaginally as compare to those infants that were delivered through caesarean section. It is worth mentioning that feeding option also play a bigger role in MTCT as infants that are on mixed feeding their hazard is higher as compared to infants that are not breastfeeding or are weaned.

4.2 Conclusion

From the various methodologies used to analyse the data, it can be concluded that place of delivery, type of delivery, child showing clinical danger signs and feeding option are determinant of mother - to - child transmission. The other PMTCT management methods are not related to mother - to - child transmission of HIV, these include HAART, mother sdNVP and infant sdNVP. Also birth weight and sex of the infant are not related to MTCT. Other covariates were not included in the model because of missingness. It is worth mentioning that other PMTCT management like putting women on HAART, putting women at short term ARVs (sdNVP, or other combination regimen) in preventing MTCT. These PMTCT management should be done as early as possible since most of the infants get infected in the early days of their life and also diagnosis should be done as early as possible so that those particular infants that are HIV positive are put on life prolonging drugs.

References

Abrams, E.J, Fayorsey, R., Gonzalez, L.F., (2007). Diagnosis of HIV infection in Infants. A Comprehensive Implementation Clinical Manual. Columbia University Milman School of Public Health.

Agresti, A., (2002). Categorical Data Analysis, 2nd Edition. New York: John Wiley & Sons, Inc.

Allison D.P., (1997). Survival Analysis using SAS. A Practical Guide. SAS Institute Inc. Cary, NC, USA.

Bakoyannis, G. and Touloumi, G., (2011). A Practical Guide in Modelling Competing Risk Data. CASCADE Collaboration. Available at: www.cascade-collaboration.org. Date accessed: June 20, 2011.

Barber, B., Cooper, E., Donohue, M., Jobarteh, K., Robison, L., Serwanga, R., and Vibbert, M. (2005). Pediatric care and Treatment: HIV and AIDS in resource limited settings. Columbia University Milman School of Public Health.

Cleves, M., Gould, W.M., & Gutierrez, R.G. (2002). *An Introduction to Survival Analysis using Stata*. College Station, Tex: Stata Press, pp.7- 34.

Coovadia, H. M., Rollins, N.G, Bland, R., Little, K., Coustsoudis, A., Bennish, M., and Newell, M.L. (2007). Mother-to-child transmission of HIV-1 infection during exclusive breastfeeding: the first six months of life. Lancet, **369**:1107-1116.

Colosimo, E.A., Chalita, L.V.A. and Demetrio, C.G.B. (2000). Tests of proportional hazards and proportional odds models for grouped survival data, *Biometrics*, 56, pp. 1233–1240

Corrente, J.E, Chalita, L. V.A.S, Moreira, J.A. (2003). Choosing between Cox proportional hazards and logistic models for interval censored data via bootstrap. Journal of Applied Statistics. **30**: 37-47.

DeGruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. Biometrics 45, 1-12.

Fitzmaurice, G., Laird, N., and Ware, J. (2004) .*Applied Longitudinal Analysis*. New Jersey: John Wiley & Sons, Inc.

Giolo, S. R., (2004). Turnbull's Nonparametric Estimator for interval Censored Data. Department of Statistics, Federal University of Parana. Brazil.

Goggins, W.B, Finkelstein, D.M., Zaslasky, A.M., (1999). Applying the Cox Proportional hazards Model for Analysis of latency data with interval censoring. Statistics in Medicine **18**: 2737-47

Gomez, G., Calle, M.L, Oller, R., and Langohr,K., (2009). Tutorial on methods for interval-censored data and their implementation in R . Statistical Models. 9(4): 259-297

Hosmer, D. and Lemeshow, S. (1999). *Applied Survival Analysis*. New York: Wiley, 58-65, 90.

Jenkins, S.P. (2008). Survival Analysis. Institute for Social and Economic Research, University of Essex, Colchester.

Kleinbaum, D.G. and Klein, M. (2005). Surviavl Analysis: A Self Learning Text. Newyork: Springer

Klein, J.P. and Moeschberger, M.L. (2003). Survival Analysis: techniques for censored and truncated data. New York: Springer.

Lindsey, J.C. and Ryan, L.M, (1998). Tutorial in biostatistics. Methods for interval censored data. Statistics in Medicine, **17**, pp. 219-238

Ministry of Health, (2009). HIV and syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report for 2007.

Ministry of Health, (2009). Infant Diagnosis of HIV using HIV DNA PCR. Implementation protocol. Lilongwe.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

National Statistical Office, (2010). Malawi Demographic Health Survey. Available at: http://www.nso.malawi.net/images/stories/data_on_line/demography/MDHS2010/MDHS2010MalawiPreliminaryReport2-28-11LTR.pdf. Accesed on July 2, 2011.

Pintilie, M. (2006). *Competing risks: A practical perspective*. Statistics in Practice. John Wiley & Sons, Ltd. West Sussex. England

UGASS (2010). Country Progress Report: Malawi AIDS Monitoring and Evaluation Report: 2008-2009.

UNAIDS, (2010). AIDSinfo Country fact sheet. Available at: http://www.unaids.org/en/dataanalysis/tools/aidsinfo/countryfactsheets/. Accessed on July 15, 2011.

World Health Organisation (WHO), (2010). Available at: http://www.who.int/hiv/pub/mtct/PMTCTfactsheet/en/index.html. Accessed on August 2, 2011

Soleimani, A., Moayyeri, A., Akhondzadeh, S., Sadatsafavi, M., Shalmani, H.T. and Soltanzadeh, A., (2004). Frequency of myasthenic crisis in relation to thymectomy in generalized myasthenia gravis: A 17-year experience. BMC Neurology, **4**:12.

Appendix 1. Missingness by covariate (Infant factors)

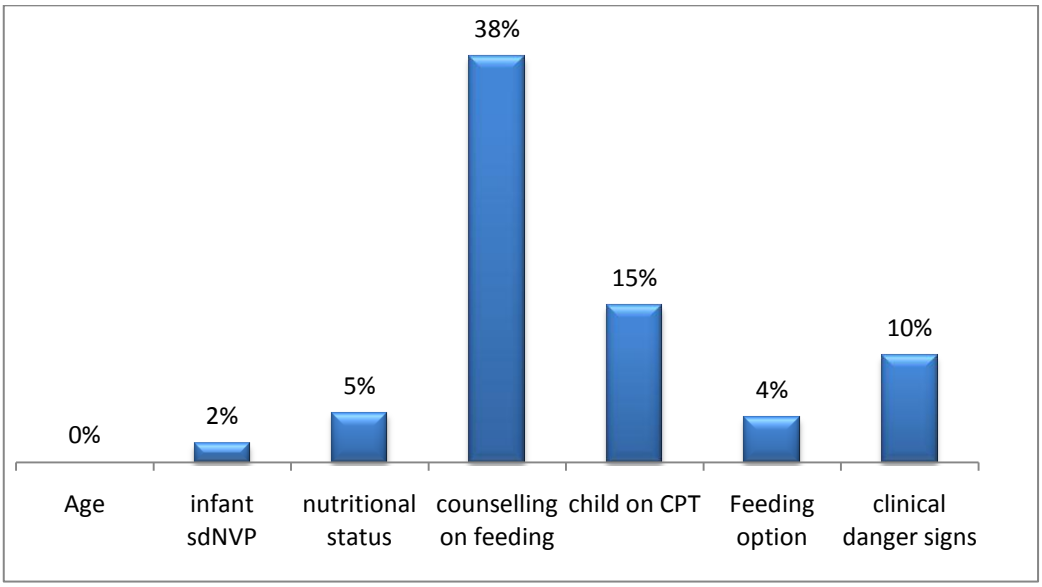

Figure 8. Missingness by covariate (Infants factors)

Appendix 2. Parameter estimates for the logistic model assuming all infants with unconfirmed HIV status as HIV positive

Effect	Parameter	Estimate (s.e.)	Pr > ChiSq
Intercept	β_0	-25.1401 (0.7749)	<.0001
Place of delivery (Home/TBA)	β_1	0.6654 (0.3518)	0.0586
Place of delivery (Health Center)	β_2	1.5243 (0.3416)	<.0001
Type of delivery (Vaginaly)	β_3	21.5321 (0.7011)	<.0001
Type of delivery (vacuum extraction)	β_4	22.5663 (0.0000)	<0.0001
Age (in months)	β_5	0.1230 (0.0542)	0.0233
Child showing clinical danger signs (Yes)	β_6	1.8718 (0.3271)	<.0001
Feed (exclusive breastfeeding)	β_7	0.5698 (0.4577)	0.2131
Feed (formula feeding)	β_8	1.0120 (0.7919)	0.2013
Feed (mixed feeding)	β_9	0.9020 (0.4191)	0.0314

Appendix 3. Parameter estimates for the Log-logistic model

Effects	Parameter	Estimate	Pr > ChiSq
Intercept	β_0^*	0.6064 (1.0877)	0.5771
Place of delivery (Home/TBA)	β_1^*	-0.3468 (0.4403)	0.4309
Place of delivery (Health Center)	β_2^*	-1.3034 (0.7587)	0.0858
Type of delivery (Vaginally)	β_3^*	-0.7794 (1.1729)	0.5064
Type of delivery (vaccum extraction)	β_4^*	-0.8334 (1.5266)	0.5851
Age (months)	β_5^*	-0.0495 (0.1268)	0.6964
Child showing clinical danger signs (Yes)	β_6^*	-1.6545 (0.9429)	0.0793
Feeding option (exclusive breastfeeding)	β_7^*	3.9715 (2.0272)	0.0501
Feeding option (formula feeding)	β_8^*	-1.3657 (1.3144)	0.2988
Feeding option (mixed feeding)	β_9^*	4.0353 (2.0342)	0.0473
Scale		1.0643 (0.5277)	

Appendix 4. Parameter estimates and standard errors for the Log-normal model

Effect	Parameter	Estimate	Pr > ChiSq
Intercept	β_0^*	0.6188 (1.1956)	0.6047
Place of delivery (Home/TBA)	β_1^*	-0.2810 (0.4374)	0.5206
Place of delivery (Health Center)	β_2^*	-1.1539 (0.7061)	0.1022
Type of delivery (Vaginally)	β_3^*	-0.8947 (1.2733)	0.4822
Type of delivery (vaccum extraction)	β_4^*	-0.8457 (1.5366)	0.5821
Age (months)	β_5^*	-0.0546 (0.1320)	0.6790
Child showing clinical danger signs (Yes)	β_6^*	-1.7853 (1.0283)	0.0825
Feeding option (exclusive breastfeeding)	β_7^*	4.2532 (2.2023)	0.0535
Feeding option (formula feeding)	β_8^*	-1.6197 (1.3590)	0.2333
Feeding option (mixed feeding)	β_9^*	4.1838 (2.1376)	0.0503
Scale		2.0206 (1.0089)	

Appendix 5. Section of the SAS codes used

```
*Creating dummy variables;
data thesis1; set thesis;
if pdeliv=1 then pdeliv1=1; else pdeliv1=0;
if pdeliv=2 then pdeliv2=1; else pdeliv2=0;
if pdeliv=3 then pdeliv3=1; else pdeliv3=0;
if tdeliv=1 then tdeliv1=1; else tdeliv1=0;
if tdeliv=2 then tdeliv2=1; else tdeliv2=0;
if tdeliv=3 then tdeliv3=1;else tdeliv3=0;
if nstat=1 then nstat1=1; else nstat1=0;
if nstat=2 then nstat2=1; else nstat2=0;
if nstat=3 then nstat2=2; else nstat3=0;
if feed=1 then feed1=1; else feed1=0;
if feed=2 then feed2=1; else feed2=0;
if feed=3 then feed3=1; else feed3=0;
if feed=4 then feed4=1; else feed4=0;
if bwght<=2.5 then bwgt=1; if bwght > 2.5 then bwgt=2;
if age < 3 then age1=1; if age =>3 <=6 then age1=2;
if age > 6 <=9 then age1=3; if age > 9 <=12 then age1=4;
if age >12 then age1=5; run;

*Chi-square test and Fisher's Exact test for the sex variable;
proc freq data=thesis1;
table sex*outco/ chisq fisher;run;

*Fitting AFT model, by Turnbull Algorithm: Final Model for Weibull distrubution;
proc lifereg data=thesis1;
model (left, right)= pdeliv1 pdeliv2 tdeliv1 tdeliv2 age cdang
          feed1 feed2 feed3/dist=weibull;
probplot ppout maxitem=4000; run;

*Fitting logistic model: Final Model;
proc genmod data=thesis1 desc;
model status = pdeliv1 pdeliv2 tdeliv1 tdeliv2 age cdang
          feed1 feed2 feed3/ link=logit dist=bin aggregate; run;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Analysis of time to HIV infection and determinants of mother-to-child transmission of HIV

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Saka, Emmanuel

Datum: **12/09/2011**