



University of Fort Hare
Together in Excellence

**TOPIC : STATISTICAL MODELLING OF MOTHER-TO-CHILD
TRANSMISSION RATE OF HIV INFECTION AND
INCUBATION PERIOD OF PAEDIATRIC AIDS USING
DATA FROM AN EASTERN CAPE HOSPITAL,
SOUTH AFRICA**

By

James Onyango Ndege

(200113178)

*A Dissertation Submitted to the University of Fort Hare,
Faculty of Science and Agriculture, Department of Statistics
For the fulfillment of the requirements of the degree of Doctor of
Philosophy in Biostatistics*

SUPERVISOR : Prof Y. Qin

CO - SUPERVISOR : Prof J. C. Tyler

JANUARY 2015

DECLARATION

I, **James Onyango Ndege** hereby declare that the content of this research work is my own and information extracted from other sources have been duly acknowledged as such. I further confirm that the research work has not been submitted in part or in full for an award of any other degree or in any other institution of higher learning apart from University of Fort Hare, South Africa.

Signature :

The image shows a handwritten signature in black ink, which appears to be 'J. Onyango Ndege', written over a faint, light blue crest of the University of Fort Hare. The crest features a sun with rays and an open book with the Latin motto 'VIVIT OMNIA'.

JAMES ONYANGO NDEGE

University of Fort Hare
Together in Excellence
January 2015

CERTIFICATION

This is to certify that this research work was carried out by James Onyango Ndege and submitted to the University of Fort Hare, Faculty of Science and Agriculture, Department of Statistics for the fulfillment of the requirements of the degree of

Doctor of

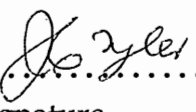
Philosophy in Biostatistics

Signature:  1/09/2015
Date

James Onyango Ndege

University of Fort Hare
Together in Excellence

SUPERVISOR: Prof Y. Qin  1/9/2015
Signature Date

CO - SUPERVISOR: Prof J. C. Tyler  1/9/2015
Signature Date

JANUARY 2015

ACKNOWLEDGEMENTS

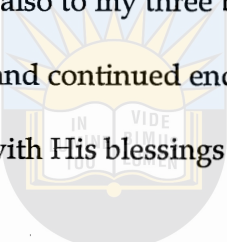
My first gratitude and deepest appreciation is to God Almighty for His love, mercy, protection and kindness for what he has done for me in life. When I look back to where I started, the hardships I have gone through and where I am today, I know that it is only through God's blessings that all these achievements have been made possible. All I have, I have received from you God, not because of my hard work or exceptional personal qualities but because it was Your gift and mercy to me. Thank you very much and I will forever praise your name Almighty God for everything.

My second gratitude goes to Prof Joanne Tyler, the former Head of Statistics Department at University of Fort Hare who made it possible for me to come and pursue my Doctoral studies in the Department and who has also been my supervisor and mentor throughout these years. Thank you Prof for your patience, kindness and mentorship. God will bless you abundantly. I am also very grateful to Prof Y Qin, also my supervisor and his invaluable advice, help and suggestions throughout my research work. I would also like to thank Govern Mbeki Research and Development Centre (GMRDC) who gave me some financial assistance during the initial stages of my research.

I also appreciate very much the help given to me by administrators, doctors and nurses who rendered to me invaluable help and assistance during my data collection at

Cecilia Makiwane (Mdantsane) Hospital in East London. Their explanations regarding some of the medical technical terms used in this study were very helpful.

I am very grateful to my family; my Wife Grace (Betty), the Love of my life and the mother of my children. She has been a wonderful mother and a great pillar of strength, support and encouragement. Thanks very much for your love, patience and understanding. Special thanks goes also to my three beloved Children; Peggy, Quinter and Brian for their encouragement and continued endurance during long hours of stay in the office. God will shower you with His blessings to help you all fulfill your endeavours.



University of Fort Hare
Together in Excellence

I cannot fail to thank my fellow colleagues in the Department of Statistics; Mr. R Mutambayi, Mr. R Chiruka, Mr. I Jubane and Mr. C Marange. Very Special thanks goes to my friend and colleague Mr. A S Odeyemi, whom we have travelled this very difficult journey together. Our constant interactions and discussions were very helpful and may our dreams be fulfilled in God's name.

ABSTRACT

HIV/ AIDS pandemic has caused many deaths and it has also impacted negatively on households, communities, and entire societies. This disease has posed many health challenges in Sub-Saharan Africa, particularly in South Africa, which is still believed to have one of the highest number of HIV/ AIDS infections in the world (South African National HIV survey, 2012).

The impact of this disease on children is equally devastating. About 30% of pregnant women in South Africa are living with HIV. Maternal and infant mortality in the country are still high with about 700 babies being infected with HIV daily. Also, HIV accounts for 35% of under-five child mortality, and nearly 44% of maternal deaths are due to AIDS-related causes especially in the rural communities (UNIAIDS, 2012).

Paediatric HIV/ AIDS infection therefore still poses a big challenge in South Africa.

This Study discusses the estimation of the rate of vertical human immunodeficiency virus (HIV) transmission and the pediatric acquired immune deficiency syndrome (AIDS) incubation period using birth cohort data collected from Cecilia Makiwane (Mdantsane) Provincial Hospital in East London, Eastern Cape. A total of 1000 children were sampled for the study using their file records from the Hospital. These were (children) cases presented to the hospital between 1st January 2003 to 30th June 2010. Standard methods of analysis, which ignore children whose infection status is not established, sometimes lead to biased estimates. Methods based on modeling the disappearance of HIV antibody or appearance of virus are inefficient, as they mostly

rely on a single variable. The study examines an alternative model that takes into account clinical, immunological, and virological data. Maximum likelihood estimates of rate of vertical transmission, sensitivity of virus tests, pediatric AIDS incubation period, and age distribution of antibody loss are obtained by EM algorithm. This new approach, with possible modifications, can also be used in analysis of data from randomized controlled trials of interventions to reduce vertical transmission. It may also be applied to other vertically transmissible infectious diseases.

Using this method yielded maximum likelihood (ML) estimates of 0.173 with a standard error (SE) of 0.021 for the vertical transmission rate; 0.798 with SE of 0.036 for sensitivity of viral culture; and 0.947 with SE of 0.049 for sensitivity of PCR. An estimated 0.154 (90% confidence interval of 0.099- 0.204) of infected children were diagnosed with AIDS by age 6 months, with a fairly constant hazard thereafter (approximately 0.091 per year). The median duration of anti-body persistence among uninfected children was 11 months, with all children losing antibody by 22 months of age.

The study also indicates that more women should be encouraged and incentivized to attend antenatal clinics during their early pregnancy period and be tested so that infected mothers can start early ART treatment to reduce the rate of mother-to-child transmission.

KEY WORDS: Statistical modelling ; Acquired immune deficiency syndrome ;
Paediatric AIDS ; Incubation period ; Mother-to-child transmission rate.

TABLE OF CONTENTS

CHAPTER ONE : INTRODUCTION	page
1.1 The Country's Profile	1
1.2 The HIV/ AIDS problem	2
1.3 Statement of the problem	14
1.4 Objectives of the study	17
1.5 Justification for the study	17
CHAPTER TWO : LITERATURE REVIEW	19
2.1 Introduction	19
2.2 Review of UNAIDS / WHO methods	20
2.2.1 HIV Prevalence estimation	20
2.2.2 Delphi survey method	21
2.2.3 Mathematical and computer simulation methods	21
2.2.4 Scenario modeling method	22
2.2.5 Epimodel	23
2.2.6 Estimation and projection package (EPP)	24
2.2.7 Spectrum method	28
2.2.8 HIV Infection among children	29

2.3	Review of other methods	31
2.3.1	Extrapolation models	31
2.3.2	Epidemic theory models	32
2.4	Back-Calculation models	33
2.4.1	The infection curve	34
2.4.2	The incubation density function	35
2.4.3	AIDS data series	36
2.4.4	Incubation distribution model	38
2.4.5	Weibull distribution model	39
2.4.6	Gamma distribution model	41
2.4.7	Log-Logistic distribution model	41
2.4.8	Gompertz distribution model	42
2.4.9	The staging models	42
2.4.10	The mixture models	43
2.5	Comments about the Back-calculation methods	44
2.6	Improvements in the Back-calculation methods	46
2.7	Parametric back projection method	49
2.7.1	Estimation when G is a basis of indicator function	51

2.7.2	The Multinomial maximum Likelihood estimate	52
2.7.3	The Quasi-Likelihood estimate	54
2.7.4	Poisson regression model	54
2.7.5	The incubation distribution model	54
2.7.6	Estimation when G is a spline function	55
2.8	Non-Parametric Back-projection	56
2.8.1	The EM algorithm method	57
2.8.2	Convergence	60
2.9	Modification of the Non-parametric Back-projection	60
2.10	Conclusion	65
CHAPTER THREE : METHODOLOGY		66
3.1	Introduction	66
3.2	Research design and Data collection	66
3.3	Ethical considerations	68
3.4	Challenges in some statistical estimation methods	69
3.5	The proposed model for the study	73
3.6	Identifiability	77
3.7	Estimation of the Parameters	77

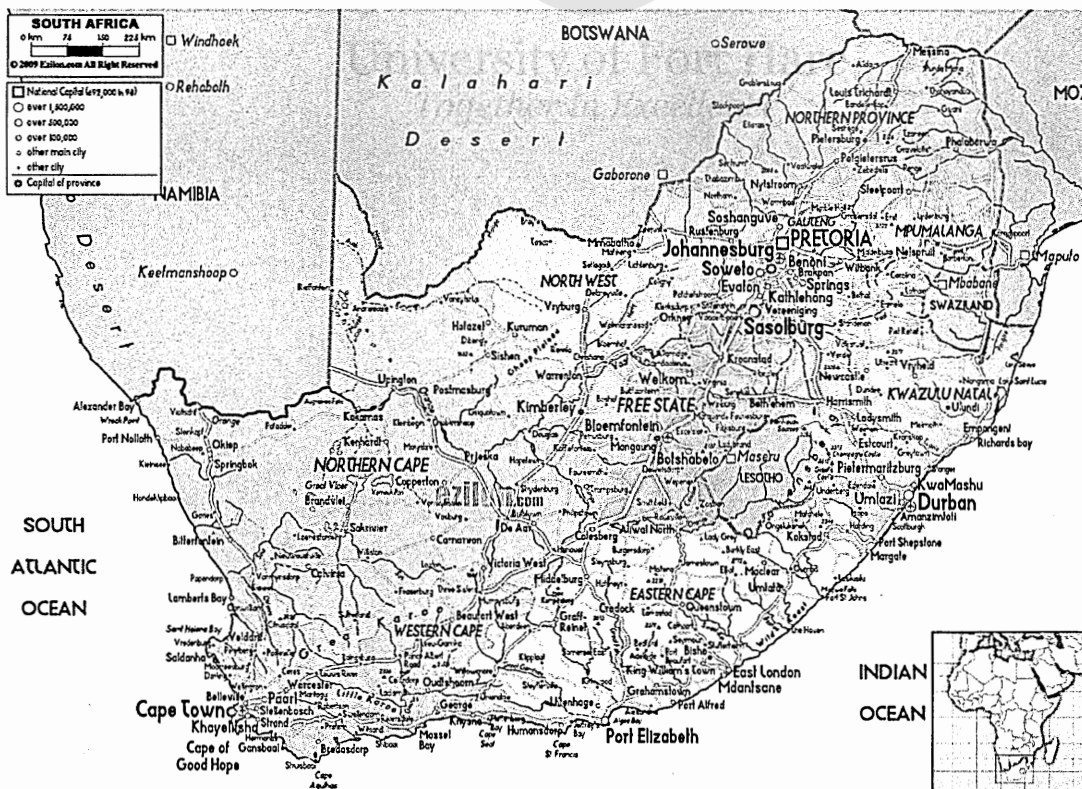
3.7.1	The E step	78
3.7.2	The M step	78
3.8	Precision of the estimates	81
3.9	Covariates	82
CHAPTER FOUR : DATA ANALYSIS		84
4.1	Introduction	84
4.2	Data Analysis and Discussion	87
	Figure 4.1	88
	Figure 4.2	89
	Figure 4.3	90
4.3	Comparison with crude estimators	91
	Table 4.1	92
4.4	Computing	93
CHAPTER FIVE : SUMMARY AND CONCLUSIONS		94
5.1	Introduction	94
5.2	Summary and Conclusions	94
5.3	Recommendations	98
	References	100
	Appendix 1	122
	Appendix 2	123

CHAPTER ONE

INTRODUCTION

1.1 A brief outline of the Country's Profile

The Country South Africa is derived from the country's geographic location at the southern tip of Africa. South Africa, officially known as the Republic of South Africa (RSA) has a total area of about 1,221 037 km^2 and an estimated population of 54 million. It has 2,798 kilometers of coastline that stretches along the South Atlantic and Indian oceans. To the north lie the neighbouring countries of Namibia, Botswana and Zimbabwe ; to the east are Mozambique and Swaziland; and within it lies Lesotho, an enclave surrounded by South African territory.



The country is divided into nine provinces namely; Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape.

The Eastern Cape where the study was carried out, occupies the south-eastern part of the South African coast. It has a total area of 168 966 km^2 with a population of 6.5 million people. The hospital where the data was collected is based in Mdantsane township in East London, the second largest City in the province.

1.2 The HIV / AIDS Problem

The problem of rapid population growth rate in developing countries, particularly in Africa, was well known before the advent of the HIV / AIDS pandemic in the late 1970's and early 1980's. However, the HIV / AIDS pandemic is threatening to reverse this scenario in the Continent, especially in Sub-Saharan Africa. The disease is a major public health concern and one of the leading causes of death in the region. Although Africa is inhabited by just over 14.7% of the world's population, it was estimated to have more than 88% of people living with HIV and 92% of all AIDS deaths in the world in 2007 (UNAIDS, 2008). Even though AIDS deaths has slowed down slightly in the recent past, it is still a major cause of death in many countries in the continent.

Sub-Saharan Africa in particular, has the highest prevalence rate of HIV and AIDS infection rate than any other region in the world. An estimated 22.4 million people

are living with HIV in the region which is around two thirds of the global total. In 2008 around 1.4 million people died from AIDS in Sub-Saharan Africa and 1.9 million people became infected with the virus. Since the beginning of the epidemic, more than 14 million children have also lost one or both parents to HIV/AIDS (UNAIDS, 2009).

Although there are massive expansions of prevention and treatment campaigns by many Governments and Non-Governmental organizations to help curb the disease, stigmatization and deeply rooted cultural practices are still major drawbacks in Sub-Saharan Africa and HIV/AIDS cases may still continue to rise. It is even worse for children whose parents may fear revealing their HIV status. This means that the impact of the HIV/AIDS pandemic in the region will continue to be felt for a considerable time to come. Its social and economic consequences are already widely felt, not only in the health sector but also in other sectors such as education, industry, agriculture, transport, human resources and other sectors of economy in general. This pandemic threatens to roll back decades of development and progress made in many spheres of life in the region.

Mother-to-child transmission (MTCT) in Sub-Saharan Africa is particularly alarming. Around 400,000 children in the region became infected with HIV in 2008 (UNAIDS, 2009). This figure is now considerably higher. The vast majority of these children have been infected with HIV during pregnancy, childbirth or breastfeeding, as a result of their mothers having been infected with the virus.

Without interventions, there is a 20 - 45% chance that an HIV-positive mother will pass infection on to her child in a developing country. If a woman is supplied with antiretroviral drugs, however, this risk can be significantly reduced (UNAIDS, 2009). Before these measures can be taken the mother must be aware of her HIV status and so testing plays a vital role in the prevention of MTCT. In many developed countries, these steps have helped to virtually eliminate MTCT. Yet Sub-Saharan Africa continues to be severely affected by this disease, partly due to lack of drugs, services and information as well as the shortage of testing facilities in some countries. In 2006, preventive drugs had only reached 31% of HIV-infected pregnant women in Eastern and Southern Africa (UNICEF/UNAIDS, 2009). However, the greatest impediment to HIV testing in developing countries and more so in Sub-Saharan Africa is the social and cultural stigma attached to and associated with the virus which make many people shy away from testing to know their HIV status even at the risk of their own lives and to that of their unborn babies which increases the risk of MTCT.

Over three decades has now passed since the first HIV/ AIDS cases were noted in the early 1980's. Researchers and the Medical community have brought to fore the concerns regarding the rapid development, speed and scale of the epidemic, and proposed various ways to combat the disease, yet it continues to ravage many parts of Africa, especially Sub-Saharan Africa (Caldwell, 1998 ; Bonnel, 2000). What distinguishes the HIV/ AIDS epidemic for example, from TB and malaria are the

different risk groups and the time lag present for costs and losses due to the long incubation period (Gaffeo, 2003).

It is estimated that between five and ten percent of the total population in Africa is infected. In 2007, Sub-Saharan Africa accounted for 67 percent of all people living with HIV worldwide, equivalent to 22 million people and culminating in about 75 percent of all AIDS related deaths worldwide. This is alarming, considering that Sub Saharan Africa only accounts for about ten percent of the world's population (UNAIDS, 2008). By 2013, over 35 million people worldwide had succumbed to the disease and some 2.1 million became newly infected while 1.5 million died of AIDS-related deaths. Of these 35 million people, more than two-thirds (about 25 million people) were living in Sub-Saharan Africa including 91% of world's HIV-positive children. An estimated 1.5 million people in the region became newly infected (UNAIDS, 2014)

Some of the consequences of HIV/ AIDS in Sub-Saharan Africa are reduced human capital, lower economic growth, increased medical costs for the individual and the state and reduced spending in other public areas. Various socioeconomic, health, demographic and political factors related to inequality, poverty, labour migration and political instability are believed to have an impact on the spread of the epidemic.

In Southern Africa and South Africa in particular, HIV and AIDS is a serious health and social concern because South Africa is believed to have the highest

number of people living with HIV/AIDS than any other country in the world.

Furthermore, the other top five countries with the highest HIV/AIDS prevalence rates are all neighbours of South Africa (UNAIDS, 2014)

AIDS was diagnosed for the first time in two patients in South Africa in 1983. The first recorded death owing to AIDS occurred in the same year. By 1986, there were 46 recorded AIDS diagnoses. Estimates from 2000 indicated that 5% of actual infections and only 1% of actual deaths due to AIDS were reported prior to 1990. During that period, AIDS was more common among homosexual people. By 1990, less than 1% of South Africans had AIDS. By 1996, the figure stood at around 3% and by 1999 the figure had reached 10%. AIDS infection started reaching pandemic proportions around 1995 (Simmon et al. 1983 ; Govender et al. 2000).

The HIV/AIDS scourge has affected all categories of age groups, particularly children. Recent declines in infant and child mortality rates across many countries in Sub-Saharan Africa have also been reversed by this pandemic. About 25% of children in this region are HIV positive.

Let us first consider the statistics of the HIV/AIDS pandemic in South Africa as this would enable us to understand the enormity of the pandemic and its impact on children in the country.

Health experts now agree that South Africa is facing very severe HIV/AIDS pandemic. According to South Africa National Health Survey (2006), it was

estimated that 10.8% of all South Africans were living with HIV. Among those between 15 and 49 years old, the estimated HIV prevalence was 16.2% in the same year. It was estimated that about 5.7 million South Africans were living with HIV/AIDS or just under 12% of the country's 48 million people in 2007 (UNIADS, 2008). This number of infection was larger than in any other single country in the world at the time. It was estimated that this number was likely to continue to rise in the years that follow unless a major behavioral change took place that could change the course of the epidemic.

In the same report, only 28% of people in South Africa with advanced HIV/AIDS were receiving anti-retroviral treatment. In 2004, 2005 and 2006 the figures were 4%, 15% and 21% respectively. In 2001, 180 000 South Africans died of HIV/AIDS. By 2007, the figure had reached 350,000 (UNIADS, 2008). This is an average increase of about 16% annually between 2001 and 2007. This suggests that anti-retroviral treatment had not had any significant positive impact on reducing HIV/AIDS.

Although new infections among mature age groups in South Africa remain high, new infections among teenagers seem to be slightly on the decline. HIV/AIDS prevalence figures in the 15-19 year age group for 2005, 2006 and 2007 were 16%, 14% and 13% respectively. In 2007, it was estimated that between 4.9 and 6.6 million of South Africa's 48 million people of all ages were infected with HIV virus, which is between 10.2% and 13.8% of the total population (UNIADS, 2008).

The prevalence rates by provinces in 2008 showed that HIV and AIDS was still a major challenge in the country. More than thirty percent of all South Africans were infected with HIV virus in KwaZulu-Natal and Mpumalanga provinces. Between 2005 and 2008, the total number of people infected with HIV/AIDS had increased in all of South Africa's provinces except KwaZulu-Natal and Gauteng. Nevertheless, KwaZulu-Natal still had the highest infection rate at 15.8 % . The other province were ; Mpumalanga 15.4 % , Free state 12,6 % , North West 11.3%, Gauteng 10.3 % , Eastern Cape 9.0 % , Limpopo 8.8 % , Northern Cape 5.9 % and Western Cape 3.8 % (South Africa HIV and AIDS survey, 2008).

HIV and AIDS epidemic is still therefore a major challenge throughout the country. Nicolay in her paper "Aids and Demographic model, 2008" argues that Eastern Cape is one of the provinces with the highest HIV and AIDS prevalence rates in the country. She estimates that about 730,000 people were living with the virus and about one in every five adults were HIV positive in 2008 culminating in 120 deaths per day with approximately 223 new infections daily (Nicolay, 2008).

The situation in 2012 was not any different. It still showed a continuation of the high infection rate. According to the South African National HIV survey , the national estimate for HIV prevalence among South Africans in 2012 was 17.9%. This estimate is statistically significantly different ($p < 0.001$), from the 2008 national estimate of 10.6%. In absolute numbers, it means that there was a significant increase of almost 1.2 million

more people living with HIV in South Africa, an increase from 5,253,493 in the 2008 to 6,422,179 in 2012 (South African National HIV survey, 2012).

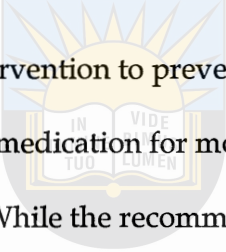
With this kind of scenario in South Africa, it is not surprising that approximately 30% of pregnant women are living with HIV. Although the country has made tremendous progress in reducing mother-to-child transmission (MTCT) of HIV, the infection rate is still very high. In 2011, approximately 2.7% of infants tested positive for HIV six weeks after birth, down from 8% in 2008 (UNAIDS, 2012). Even though a considerable number of women are aware of ANC at health care centres and hospitals, maternal and infant mortality in rural South Africa still remain high with about 700 babies being infected with HIV daily in the country. Also, HIV accounts for 35% of under-five child mortality, and nearly 44% of maternal deaths are due to AIDS-related causes (UNAIDS, 2012). These statistics highlight the need for a greater engagement and actual participation of women in the country's healthcare system. Paediatric HIV/ AIDS infection and deaths still poses an enormous challenge in the country.

When the AIDS epidemic started in the early 1980s and 1990s the high risk groups were identified as homosexuals, hemophiliacs, and heroin users. Today, the high risk group of the AIDS epidemic has changed especially in developing countries and women have the highest rates of infection. As more women become infected, the potential for vertical transmission from mother to child increases. It is now estimated that over 90% of the HIV-infected children in Sub-Saharan Africa acquired the infection through vertical transmission (UNAIDS, 2012).

While a combination of intervention methods can reduce the risk of vertical transmission to less than 2%, it still persists in many developing countries in Sub-Saharan Africa. It is the major cause of HIV infection in children below age 15 . Without intervention or treatment, the possibility of HIV transmission from mother-to child is 15-30% in developed countries and 30-45% in developing countries. While 50-80% of infants are vertically infected during delivery, HIV transmission can also occur during pregnancy and after birth. During pregnancy, the fetus can become infected by contacting maternal blood through a placental hemorrhage or by swallowing infected amniotic fluid. There are other several maternal factors which increase the chance of mother-to-child transmission such as ; maternal seroconversion during pregnancy, high viral load, malnutrition, other sexually transmitted diseases, and lack of or poor compliance with antiretroviral drug therapy (UNAIDS, 2009).

During birth, factors that increase the risk of mother-to-child HIV transmission include: vaginal delivery, rupture of vaginal tissue, contact with maternal blood and vaginal secretions, and chorioamnitis. Pregnant women with chorioamnitis have a potentially increased white blood cell count that acts as a target for the HIV virus. Higher maternal viral load is positively correlated with vertical HIV transmission. Postnatally, the most significant risk factor is breastfeeding. The HIV virus has been isolated from breast milk, demonstrating the risk of long-term breastfeeding in infants. Transmission of HIV through breastfeeding occurs in 16-29% of cases. Specific risk factors during breastfeeding include: cracked nipples, mastitis (breast inflammation), breastfeeding for

extended time periods, postnatal maternal seroconversion, high viral load, and low CD4 cell count. Mixed feeding of breast milk and other food sources has been shown to increase the risk of HIV transmission. Scientists hypothesize that an infant's immune response is triggered by the introduction of new foods, attracting white blood cells to the gastrointestinal tract and increasing targets for the HIV viruses to spread infection (UNAIDS, 2009; 2012).



The most successful methods of intervention to prevent mother-to-child HIV transmission include: antiretroviral medication for mother and child, caesarian section, and refraining from breastfeeding. While the recommended type and regimen of antiretroviral drugs for mothers varies, mothers on antiretroviral drugs have less than a 2% chance of transmitting HIV to their babies. The Pediatric AIDS Group Protocol 076 (PACTG 076), (Conner et al.1994) was an important study that demonstrated the effectiveness of using zidovudine in reducing vertical HIV transmission. This study showed that zidovudine given during pregnancy and labor and given to the baby during the first six weeks of life decreased the risk of vertical transmission by 66%. Another drug called nevirapine, a nonnucleoside reverse transcriptase inhibitor, is also effective in reducing vertical HIV transmission when given during pregnancy and after birth to the child. Aside from drug therapy, pregnant HIV positive women are also advised to have a cesarean section at 38 weeks to further reduce the risk of vertical transmission. Cesarean delivery reduces exposure of the infant to maternal fluids and tissues which is high during vaginal delivery. Finally, HIV positive women are

encouraged to refrain from breastfeeding to reduce the risk of spreading HIV to their babies postnatally.

While the methods of preventing vertical HIV transmission are fairly effective, implementing these interventions in Sub-Saharan Africa and other developing areas of the world is difficult due to cultural and economic barriers. While encouraging HIV-positive mothers to refrain from breastfeeding is a logical preventive measure against vertical HIV transmission, most women in poverty-stricken areas and in many areas of rural South Africa may not afford to pay for formula food or do not have access to clean drinking water to prepare the formula. Also, women who stop breastfeeding to protect their children from HIV risk the stigma of being labeled as HIV positive and being discriminated against. These obstacles, combined with the lower level of access to antiretroviral drugs in some rural parts of the country create obvious barriers to decreasing vertical HIV transmission. In Africa, studies suggest that one in three newborn children infected with HIV die before the age of one, over half die before reaching their second birthday, and most are dead before they are five years old (Newell et al. 2004).

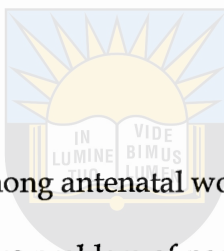
Statistics available on mother to child transmission of HIV virus in the country is shocking. A study by love life (Love life, 2001) showed that about 6.5 % of all children born at the time in South Africa were HIV infected. Another study by the Department of Health Survey (2008) showed that the infection rates among children

aged 2 to 14 years was 5.6% , 3.3 % and 2.5% in 2002, 2005 and 2008 respectively (South Africa Health survey, 2008).

The National Antenatal Sentinel HIV and Syphilis Prevalence Survey (2008) revealed that although there was some evidence that the estimated national HIV prevalence amongst women was beginning to stable at about 29% (UNAIDS 2008) this was not indicative of major behavioral change but was rather due to natural progress of long wave epidemic. Moreover, this was not even throughout the country. The results showed that the highest HIV prevalences were recorded on the North-Eastern side of the country, and the lowest prevalences in the Western parts of the country. While the national average of HIV-positive women attending antenatal clinics in 2008 was 29.3%, the province of Kwa-Zulu Natal had the highest prevalence rate of 38.7% followed by Mupumalanga at 35.5%, Eastern Cape at 27.6% and the Western Cape had a comparatively low prevalence rate of 16.1% which was still high by world standards (National Antenatal Sentinel HIV and Syphilis Prevalence Survey, 2008). It was estimated that there were 66,000 new paediatric infections in 2008 alone, of which about 38,000 were either born HIV positive or infected at birth and 28,000 were infected through breastfeeding (UNAIDS, 2009).

The National Antenatal Sentinel HIV Prevalence Survey (2012) showed that the overall HIV prevalence rate was still quite high. According to the survey, the HIV estimate among antenatal women was 29.5% (95% CI: 28.8 - 30.2%), whereas the

national HIV prevalence estimate was the same at 29.5% (95%CI: 28.7 – 30.2%) as it was in 2011. The UNAIDS EPP and Spectrum model estimated that in 2012 the South African national HIV prevalence in the general population was 17.9%. UNAIDS also estimated that 6.2 million people were living with HIV with an estimated 370, 000 new infections. The same model also estimated that about 410 000 children under 15 years were HIV positive nationally (National Department of Health SA, 2013 ; UNAIDS , 2013).



These high HIV prevalence rates among antenatal women and the large number of infected children indicates the serious problem of paediatric Aids in the country. Therefore, estimating vertical HIV transmission rate and the distribution of incubation period of paediatric AIDS in children is useful for testing and designing preventive and therapeutic interventions as well as forecasting health care needs for the country especially for infants and children.

1.3 Statement of the problem

HIV/ AIDS pandemic is a serious health issue that has caused many deaths and has also impacted negatively on households, communities and entire societies across different age groups. Without a strong universal dedication and strong intervention methods to fight this disease, especially the mind sets of most people living in Sub-Saharan Africa, the disease will continue to cause many deaths, erode human development, deepen poverty, and further hinder access to education, health and viable livelihoods.

It is also known that Sub-Saharan Africa has the largest burden of pediatric HIV/ AIDS in the world. Global target was set for eradication of pediatric HIV by 2015 but there are still many complex issues and challenges facing HIV infected and affected children in the sub-continent. According to the 2014 progress report on Global HIV/ AIDS, an estimated 35 million people were living with HIV at the end of 2013; 2.1 million new infections and 1.5 million deaths occurred in the same year alone. Sub-Saharan Africa is occupied by 12% of the global population, but disproportionately has more than 91% of children younger than 15 years and 70% of adults that are living with HIV. There are approximately 16.6 million AIDS orphans (children who have lost one or both parents to HIV), most of whom live in Sub-Saharan Africa and South Africa in particular has the highest number of people living with HIV in the world (6.4 million) (UNAIDS, 2014).

Women account for 56–63% of the HIV infection in the region. This has far reaching implications because most children acquire HIV from their HIV-infected mothers during pregnancy, birth or breastfeeding. An estimated 90% of pediatric HIV occurs through mother-to-child transmission (MTCT). The failure to implement prevention programs against MTCT on an appropriate scale has resulted in hundreds of thousands of preventable HIV infections among newborns. Each day, more than 1000 children are newly infected with HIV, and an alarming 700 die from AIDS-related complications (US Global Health Policy, 2013). This is particularly true in South Africa where the HIV infection rate among women is still high.

Although significant progress has been made in the fight against the prevention of HIV/AIDS in adults as well as in children in Sub-Saharan Africa and even in South Africa, there are still specific challenges militating against its total elimination. These include insufficient and delayed uptake of HIV services, drug-related and psycho-social issues, and dysfunctional health systems especially in the rural areas among other challenges.

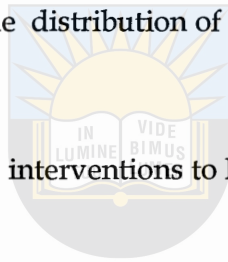
It is also a fact that HIV/AIDS still continues to contribute a great deal to Infant and child mortality in South Africa, but most available studies and statistical models which have been used to analyse HIV/AIDS data, its vertical and transmission rate as well as its incubation distribution period in particular, have been applied mostly to adults. Very few studies have been carried out to examine the impact of the disease on children

in the country in order to recommend the necessary interventions to reduce this scourge among infants and children. This study proposes to address these gaps.

1.4 OBJECTIVES OF THE STUDY

The objectives of the study are :

- i) To analyse the rate of vertical HIV transmission in children.
- ii) To model and to estimate the distribution of incubation period of paediatric AIDS in children.
- iii) To make recommendations for interventions to help in reducing the pandemic in children.



University of Fort Hare
Together in Excellence

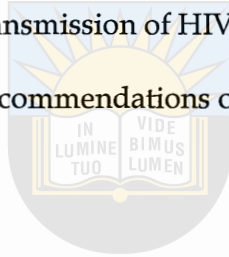
1.5 JUSTIFICATION FOR THE STUDY

Several studies have been carried out to estimate the vertical transmission rate and incubation period of adults infected with HIV/ AIDS in many countries. Such studies have been quite useful in predicting the life span of HIV infected people and to suggest some remedies in helping to reduce the scourge . However, very few studies have been carried out to estimate the vertical transmission rate and incubation period of paediatric AIDS in Africa, especially in South Africa where the HIV infection rate among children is very high.

Secondly, the study of vertical transmission rate and incubation period can help health-care providers with timely interventions to offer optimal care and treatment to HIV

infected mothers and their children and also to assist them in decision-making as well as advise to mothers and families of the children involved. An HIV-positive mother with an HIV-uninfected baby can be counselled and supported to stop breastfeeding if replacement feeding is acceptable, feasible, affordable, sustainable and safe (AFASS) and therefore assist families in planning better and healthy lives for their children.

This study therefore proposes a statistical model which is discussed in Chapter Three to help estimate the rate of vertical transmission of HIV and the incubation period of paediatric AIDS and also to give recommendations on necessary interventions to help reduce this pandemic in children.



University of Fort Hare
Together in Excellence

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

The identification of acquired immunodeficiency syndrome (AIDS) as a very unique and lethal infectious disease and the spread of the human immunodeficiency virus (HIV) continues to be an enormous challenge to many parts and regions in the world, especially in Sub-Saharan Africa. Statisticians have contributed a great deal to the HIV/ AIDS knowledge base. Statistical methods have been modelled and used to estimate both the projected and actual challenges of the epidemic and its devastating effects as well as analysis to give insight into how some of these challenges can be resolved.

However, review of available literature reveals that more work has been done on modeling paediatric HIV/ AIDS epidemic in the developed world compared to Sub-Saharan Africa. Nevertheless, the UNAIDS/WHO over the years have developed some general models which have been applied to various countries depending on the nature of the epidemic especially on data dealing with the population aged 15 - 49 years. But some of these methods can also be applied to paediatric HIV/ AIDS data. We shall start by reviewing some of the important methods used by UNAIDS/WHO obtaining and projecting estimates and prevalence rates of HIV/ AIDS data and later on discuss some of the current methods as well.

2.2 - REVIEW OF THE UNAIDS/WHO METHODS

2.2.1 HIV prevalence estimation

The HIV prevalence rate is estimated by applying a general formula on data from the HIV positive women aged 15-49 years who attend antenatal clinics (ANC). The sample estimate of the prevalence rate is obtained by dividing the number of positive cases found in the sample by the sample size. To estimate the HIV prevalence in the adult population (15-49 years), the prevalence rate for sampled women is multiplied by the total adult (15-49 years) population.

This method was also used by UNAIDS/WHO in estimating the seroprevalence rate in high-risk groups or any specific group of interest, even for children. The method is based on the assumption that the HIV prevalence obtained from the antenatal clinic attendees, with adjustment for the male to female ratio, is a surrogate for HIV prevalence in the total population aged 15-49 years. Recent studies (Saphonn et al. 2002; Fowler and Mishra 2008 ; Montana et al. 2008), have confirmed that the ANC estimates generally overestimate the population-based survey prevalence, especially for younger women and men. Gouws et al (2008) found that the ANC overestimates the population-based survey prevalence by about 20% and recommended that HIV prevalence derived from the ANC surveillance data be multiplied by about 0.8 to adjust for overestimation in countries where population-based HIV surveys have not been conducted. This method can also be applied to estimate HIV/ AIDS prevalence in children.

2.2.2 Delphi survey method

This projection method was widely used in many countries during the early stages of the epidemic in the late 1980s to obtain projected estimates of the HIV/ AIDS prevalence rate including South Africa. The method involves soliciting opinions in a systematic way from knowledgeable experts in field of HIV/ AIDS both for adults and children (through their mothers who attend ANC). The average and range of their responses are then used as projections. Whereas this method has the advantage of quick gathering of information at low cost, it is highly subjective and may produce estimates that range widely. It may also be difficult to find experts in quantitative epidemiology who are very familiar with the demographics of the country. This method was mostly fashionable at the time when data on HIV/ AIDS were rare in some countries and not many data gathering techniques and projections had been developed. Results which were obtained using this technique were generally inaccurate and unreliable. Better, accurate and more reliable projection methods have since been developed.

2.2.3 Mathematical and computer simulation methods

This method has been used to estimate short and long term projections of HIV prevalence and to study HIV/ AIDS data on both adults and children. The input parameters needed are the size of the risk groups, the number of both their current partners and sex partner exchange rates as well as the rate of mixture between the risk population and the not-at-risk population for adults. In the case of children the data

needed are children born from HIV positive mothers who attend ANC. Initially there were general uncertainties surrounding the accuracy of the input parameters which made estimation and projection of HIV/ AIDS incidence and prevalence by this method very unreliable. Therefore the model outputs were treated with caution especially where specific programme and policy developments were concerned. However, with more advances in statistical computer packages, simulation methods are now widely used by most statisticians with confidence and high degree of accuracy.

2.2.4 Scenario modeling

This method or approach was used for the short term estimation and projection of AIDS cases and deaths in the late 1980s and early 1990s. It was developed by the Surveillance, Forecasting and Impact Assessment unit of the former WHO Global Programme on AIDS (GPA). A scenario is defined as an outline of any series of events, real or imagined. The HIV/ AIDS scenario can be constructed with or without models to fit the observed HIV/ AIDS data and trends. The procedure below gives a general guideline as outlined in the WHO document (WHO 2003) of this scenario/ modelling approach to develop working estimates and projections of HIV infections and AIDS cases and deaths both for adults and children.

- i) Assembled available HIV seroprevalence data are analyzed and used to estimate the most recent pattern(s), prevalence and trends of HIV infection for a specific population.
- ii) Different HIV patterns and prevalence levels can be constructed with some

confidence based on these data and other epidemiological observations.

- iii) An AIDS model can then be used to derive the annual and cumulative estimates and projection of AIDS cases/ deaths and other HIV-related conditions, using the general HIV scenario(s) constructed as outlined.

2.2.5 Epimodel

This model is a microcomputer programme used to estimate past and current prevalence and make short-term predictions of HIV/ AIDS cases and AIDS deaths for both adults and children. It was developed particularly for countries where AIDS case reporting was incomplete and unreliable. It uses estimates of HIV prevalence at a selected point in time and distributes this prevalence by annual HIV-infected cohorts back to the estimated start of the epidemic along a selected epidemic curve. The model then applies the annual progression rates from HIV infection to the development of AIDS to each of the annual HIV cohorts to calculate the annual number of AIDS cases and deaths. The use of a single point prevalence and time may yield biased AIDS estimates. If high HIV prevalence estimates are used then this will produce high estimates of AIDS cases. The stage of the HIV epidemic will also affect the HIV prevalence used. Estimates of HIV prevalence at the increasing phase of the epidemic will be higher than that at the declining stage of the epidemic. Thus the stage and duration of the modelled HIV epidemic will have a major impact on the estimated cumulative incidence of HIV infections as well as AIDS cases and deaths.

This model can, with the additional input of a population denominator, calculate annual incidence and prevalence rates for HIV infection. Other modules of EPIMODEL include a *Child module* that estimates and projects annual numbers of HIV-infected and uninfected infants born to HIV-infected women. This paediatric module also calculates paediatric AIDS cases and deaths, and the age of maternal orphans during the year of their mothers' death.

2.2.6 Estimation and Projection Package (EPP)

This method is used to estimate and project HIV prevalence and incidence from surveillance data. The input to EPP in countries with generalized epidemics is surveillance data from various sites and years showing HIV prevalence among pregnant women, as well as data from national population-based surveys. The input in countries with concentrated epidemics is data on prevalence in populations with high risk behaviours and populations at low risk, as well as estimates of the size of these populations. EPP is then used to fit a simple epidemic model to the data (UNAIDS/WHO 2001, 2003, 2004). Curves are created for each of the sub-epidemics which have been identified and these can either be urban or rural as is often done in generalized epidemics or by the different sub-populations as is often done in concentrated epidemics. The 2009 version of EPP uses ART data to improve the estimation of incidence from the prevalence over time. The national prevalence and incidence projections produced by EPP can then be imported into Spectrum to calculate the

number of adults or children living with HIV, new HIV infections, AIDS cases, AIDS deaths, treatment needs, AIDS orphans, and other variables.

In countries with a generalized HIV epidemic that have not conducted a national population-based survey, prevalence data from pregnant women attending antenatal clinics (ANC) is adjusted downward for both rural and urban ANC, since HIV prevalence from ANC has been shown to overestimate HIV prevalence in the general population. Four major parameter inputs required are

t_0 : the start year(month) of the HIV/AIDS epidemic

r : the force of infection. It is the summary of transmission probability (in the case of children)

f_0 : the initial proportion of the adult/child population that is exposed to the risk of infection

ρ : the behaviour or high risk adjustment parameter which determines the extent to which susceptible adults or children who die of AIDS are replaced by other adults or children who were not at risk.

The EPP model partitions the population(children) at time t into *three* groups which are:

$X(t)$ = not-at-risk population

$Z(t)$ = at-risk population

$Y(t)$ = infected population

$N(t) = X(t) + Y(t) + Z(t)$ = Total population

The change in these population groups is given by the following three differential equations

$$\frac{dX}{dt} = 1 - f\left(\frac{X}{N}\right) E_t - \mu X$$

$$\frac{dZ}{dt} = f\left(\frac{X}{N}\right) E_t - \left(\mu + \frac{rY}{N} + l\right) Z$$

$$\frac{dY}{dt} = \left(\frac{rY}{N} + l\right) Z - \int_0^t \left(\frac{rY_x}{N_x} + l_x\right) Z_x g(t-x) dx$$

where $f\left(\frac{X}{N}\right)$ is the fraction of those entering the adult population (E_t) and the at-risk group Z which is defined as

$$f\left(\frac{X}{N}\right) = \frac{\exp\left[\phi\left(\frac{X}{N} - (1 - f_0)\right)\right]}{\exp\left[\phi\left(\frac{X}{N} - (1 - f_0)\right)\right] + \frac{1}{f_0} - 1}$$

$l = 1$ for the first year of the epidemic and 0 for other years, f_0 = proportion of those entering the population at the At-Risk group. If $\phi > 0$, the proportion of people entering the at-risk group is increased. g = density function describing the progression to AIDS death since HIV infection and is given as

$$g(x) = \left(\frac{\mu + \alpha x^{\alpha-1}}{\beta}\right) \exp\left[-\mu x - \left(\frac{x}{\beta}\right)^\alpha\right]$$

where α is the shape parameter of the Weibull distribution fitted to the HIV survival times and β is the position parameter defined in terms of the median survival time, m , as

$$\beta = \frac{m}{[\ln(2)^{1/\alpha}]}$$

It is important that α be predefined based on available empirical data and that three values of m be used each corresponding to slow, medium and rapid progression respectively. The fixed parameters of the model are

- i) crude child/infant (≤ 15) death rate μ
- ii) number entering the infant population at time t , E_t
- iii) force of mortality due to AIDS, x years/months after infection

For a heterosexual epidemic like South Africa, E_t can be defined in terms of HIV negative children

$$E_t = B_{t-15}^- l$$

$$B_{t-15}^- = b [X_{t-15} + Z_{t-15} + (1-\nu)\varepsilon Y_{t-15}]$$

l is the cohort survival proportion to age 15, b is the birth rate and ν is the probability of vertical transmission. ε is fertility reduction due to HIV infection. It is assumed that most HIV positive births do not survive to adulthood. However, the projected number of HIV positive births is given as

$$B_t^+ = \nu \varepsilon Y_t$$

In the use of the EPP model, the parameters are assumed to be fixed. EPP searches for the best values of t_0 , f_0 , and r that best fit the observed surveillance data. The best fit is obtained by minimizing the sum of squared errors between the model curve and the surveillance estimates.

2.2.7 Spectrum

This programme uses the prevalence projection produced by EPP to calculate other estimates like numbers of people (adults and children) infected, new infections, AIDS cases and AIDS deaths. These calculations are based on population estimates provided by the UN population Division and model patterns prepared by UNAIDS/WHO Reference groups that describe progression from infection to death as well as distribution of infection by age and gender, transmission from mother to child and the effect of HIV infection on fertility.

The vital registration system in South Africa is not very efficient in some regions such as the Eastern Cape where this study is based. Most vital events are not well recorded in this region, especially in the rural areas. Hence the estimates of birth and death rates if used in this model may be inaccurate. The estimates of other parameters such as ϕ and the size of the at-risk population may also not be reliable. However, the EPP and Spectrum models have undergone several improvements since 2001

(Brown et al. 2008) aimed at overcoming some observed limitations. The improved version, EPP2007, incorporated uncertainty estimation (Alkema et al. 2008) for generalized epidemic using the technique of Bayesian Melding. Prior distribution for the model parameters are specified using expert knowledge. Other major improvement on the program include changes in the urban-rural population ratio, calibration of HIV prevalence measured at ANC in countries with generalized epidemics, based on a comparison of HIV prevalence from ANC to national population-based surveys, longer survival of HIV patients due to antiretroviral therapy (ART) and the estimation of

number of people including children eligible for ART (Ghys et al. 2008 ; UNAIDS 2006 ; UNAIDS 2008). These improvements has enabled the Spectrum software to calculate HIV indicators of interest on the basis of previously prepared prevalence estimates. New calculations and assumptions have been added in order to incorporate the latest data and respond to the international commitment for expanded treatment for adults and children (Stover et al. 2012).

2.2.8 HIV infection among children

In the case of children, previous versions of Spectrum estimated new HIV infections among children as occurring around delivery, through the combined probability of infection in utero, peripartum and intrapartum, and through breastfeeding. The survival pattern is usually modelled as the proportion of children infected at birth who survive to each future age and using a double Weibull curve to fit the data (NIH 2008). The double Weibull curve is used to account for the fact that some HIV-infected children are fast progressors (mostly those infected in utero, peripartum and intrapartum), and some are slow progressors (mostly those infected after birth through breastfeeding). In the latest version of Spectrum, infections around delivery are tracked separately from those occurring through breastfeeding in order to follow fast and slow progressors separately.

The probability of HIV infection at birth for a child born to an HIV-positive mother is assumed to be 20% in the absence of prophylaxis, 11% with single-dose Nevirapine, 4%

with dual ARV prophylaxis and 2% with triple ARV prophylaxis or treatment (Gelman and Rubin 1992). The probability of infection through breastfeeding is assumed to be 1.5% per month for mixed feeding during the first 6 months, 0.75% per month for exclusive feeding during the first 6 months, 0.75% per month for months 7 and later, and 0.3% per month when the mother is on triple therapy (Snowflake 2007 ; AIDS 2013).

New survival curves for children infected at birth or through breastfeeding were constructed using data from 12 sub-Saharan African clinical trials and cohort studies (Velayati et al. 2007). HIV-infected children were divided into categories of early or late infection, and non-AIDS mortality was removed to focus on net AIDS mortality. Double Weibull curves were fitted to each set of data to determine the pattern of survival. As there was little follow-up beyond 2.5 years, there is limited information to inform the survival pattern beyond age 2.5. Therefore, the survival pattern was extended by assuming that children that have survived at least 2.5 years would have survival similar to young adults aged 15–24, with a median survival time of about 20 years (Whiteside 2002). The resulting survival curves indicated a median survival of 1.1 years for early infection (Gelman and Rubin 1992).

2.3 REVIEW OF OTHER METHODS

The discussion below gives a review of other methods which have been used in the HIV/ AIDS prevalence estimates.

2.3.1 Extrapolation models

One of the key assumptions in this model is that sentinel data from pregnant women may be extrapolated to the entire population. There have been a handful of studies that have addressed the question of whether prevalence rates in antenatal clinic(ANC) sites were representative of the population prevalence rates. This question can be considered on three levels: firstly, do prevalence rates in antenatal clinic sites represent the general population rates in these areas among women of the same age as clinic attendees? Secondly, do prevalence rates among women in these age groups represent the overall prevalence rates in the population (adult and children) in these areas? Thirdly, do prevalence rates in the sentinel areas represent national prevalence rates? The answers to these questions are likely to vary widely across different settings because of the stage of the epidemic, along with a host of other factors. However, the studies that have addressed these questions, have provided valuable results using this model.

Using this method, AIDS incidence is modelled as a function of time. Time may assume any mathematical form; linear, polynomial, log-linear function, etc and extrapolated into the future (McEvoy and Tillet 1985 ; Morgan and Curran 1986 ; Healy and Tillet 1988). The problem with this model is that the extrapolated estimates are greatly

influenced by the mathematical function chosen for time. It also assumes that the trends will remain unchanged which is unlikely.

2.3.2 Epidemic theory models

Epidemic theory model uses both empirical and experimental observations and advanced stochastic mathematical methods. It enables epidemiologists to construct mathematical models to study a variety of diseases. It generally makes use of a partitioned population (the susceptible, infective and not-at-risk groups). The deterministic approach could be used to describe an epidemic in large populations with large number of people infected including children. The problem with this method is that it is not easy to determine precisely the proportions of individuals in each category, the extent of mixing between the susceptible and the infective, the partner exchange rate, concurrent partners, behaviour change parameters in the case of adults etc., but with modifications in the case of children. Also, some of these deterministic models assume that once the initial conditions and parameters are specified, the prevalence function and infection curves are uniquely determined and can be found by recursive methods. The stochastic version of these models may be useful in describing epidemic in limited geographic areas or sub-groups. The models are usually complicated with so many unknown parameters and involve assumptions that are often not verifiable. A detailed account of the procedures for deterministic and stochastic modelling of AIDS/HIV epidemiology can be found in Wai-yuan (2000).

2.4 BACK-CALCULATION MODELS

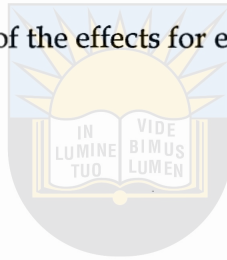
The method of back-calculation estimates the number of HIV infections from AIDS incidence data and projects future AIDS incidence both for adults and children. It was first proposed by Brookmeyer and Gail (1986). The back-calculation method is used to reconstruct the historical infection rates that may have occurred and generated the observed pattern of AIDS diagnosis. That is, the historical pattern of HIV infection curves or rates is reconstructed using information on observed AIDS incidence and knowledge of the incubation period distribution, defined as the duration between HIV infection and AIDS diagnosis. The reconstruction of the infection curve and the estimation of the time since infection, in turn, makes it possible to predict the future number of AIDS cases (Brookmeyer and Gail 1986 ; 1988 ; Rosenberg and Gail 1991).

The back-projection equation formulated is based on the fact that for an individual to be diagnosed with AIDS at time t , he/she must have been infected with HIV at some time s in the past and hence the incubation period is the time between s and t , that is $(t - s)$. This technique is widely accepted because of its efficient application and requires fewer assumptions and parameter inputs when compared with other modelling approaches like the epidemic theory models.

Generally, the back-projection model is given as

$$\mu_t = \sum_{s=1}^t \lambda_s f_{t-s,s}$$

μ_t is the mean AIDS incidence at time t , λ_s is the mean HIV incidence at time s and $f_{t-s,s}$ is the probability density function for someone infected at time s and diagnosed at time t . μ_t is known from the AIDS diagnosis and $f_{t-s,s}$ is known from other epidemiological studies (Brookmeyer and Goedert 1989 ; Lui et al. 1986). λ_s is then estimated using a deconvolution process. However, the nature of each of these components of the back-projection model has a far reaching effect on the outcome of the back-calculation procedure. Some of the effects for each component are discussed below.



2.4.1 The infection curve (λ)

The choice of the infection curve for λ_s may have some consequences on projected estimates (Day et al. 1989 ; Tylor 1989 ; DeGruttola and Lagakos 1989 ; Rosenberg and Gail 1991). For instance, DeGruttola and Lagakos (1989) fitted four different shapes of parametric infection curve to the US AIDS incidence data for the period 1981-1987 and found that the curves differ dramatically in the most recent years but give similar AIDS incidence in the distance past. The reason for this is that the use of a strong parametric model, such as the exponential model, as the infection curve will produce estimates that fit the distant past of the AIDS data very well but will portray an exponential growth in the most recent portion even when the AIDS incidence data suggest otherwise. To overcome this, Brookmeyer and Gail (1986 ; 1988), Rosenberg and Gail (1991) suggest the use of flexible models such as step functions with about four or five steps. The

justification of this is that these models are sufficiently flexible so that later portions of the infection curve can vary independently of early values of the curve. Rosenberg et al (1991) investigated the performance of step function models and found that they yielded the smallest percentage root mean square error and bias in the short-term projection of AIDS incidence and estimating cumulative HIV infections. The estimates of the cumulative HIV infections and the projection of the AIDS incidence are based on the integrals of the estimated infection curve, and given the discontinuities in step function models, the integral over the infection curve appear not to be feasible. Consequently, Rosenberg and Gail (1991) suggested the use of spline functions for the infection curve.



2.4.2 The incubation density $f(t)$

Most of the work done earlier on HIV/ AIDS using back-calculation (Brookmeyer and Gail 1986 ; 1988 ; Rosenberg and Gail 1991) assumed that the incubation distribution is stationary over the time period. Also, the incubation distribution adopted for any model has an influence on the HIV incidence estimation (Baachetti et al. 1993 ; Gigli and Verdecchia 2000 ; Tylor 1989 ; Rosenberg and Gail 1990 ; Rosenberg et al. 1991). To investigate this Baachetti et al (1993) used four different incubation period distributions to estimate HIV incidence through 1990 using AIDS diagnosis data and found that the estimates vary substantially in the most recent past (1987-90). According to Rosenberg and Gail (1992) and Brookmeyer and Gail (1994), assuming a shorter incubation distribution leads to lower estimates of cumulative number of infections and a slow

incubation distribution requires large values of infection curve to fit the AIDS incidence series while fast incubation distribution require small values of the infection curve. However, most authors (Brookmeyer and Gail 1988 ; Brookmeyer and Damiano 1989 ; Tylor 1989) agree that use of back-calculation for short-term projection of AIDS incidence and estimates of AIDS incidence at the earlier part of the epidemic is very reliable and is relatively insensitive to the choice of incubation distribution. This is because in the first few years or months following infection, the estimated incubation distribution has increasing hazard and flexible models of infection curve can adapt to a particular incubation distribution to fit the AIDS incidence.

2.4.3 AIDS data series

AIDS incidence is an important component of the back-calculation model, hence it should be reliable, up to date and accurate. Unfortunately, the accuracy of the AIDS incidence series is affected by reporting delay and underreporting and missing data. Reporting delay is measured as the time between AIDS diagnosis and time when report of such diagnosis is received by the AIDS statistics coordinating agency. Often, this delay may range from few a months to a year or more. In South Africa, it takes quite a bit of time although the HIV/ AIDS coordinating agencies are improving the communication channels but reporting delays is still a problem. Therefore, the use of such data in back-calculation calls for some form of adjustment on the data if back-calculated estimates and projections are to be reliable. Some authors (Haris 1987; 1990 ; Zeger et al. 1989 ; Brookmeyer and Damiano 1989 ; Brookmeyer and Liao 1990 ;

Rosenberg 1990 ; Lawless and Sun 1992), have suggested various adjustment procedures. These adjustments are performed on the data before back-calculation is applied. Brookmeyer and Damiano (1989) suggested that data delayed for up to 7 to 9 months be increased by 22% and those delayed for 16-24 months by 6%. The data proposed for back-calculation in this research is free from the limitations of reporting delay because they were extracted retrospectively from the records of health and laboratory facilities in the hospitals and grouped based on the dates of diagnosis.

What is of great concern in the data is underreporting. In some countries such as the United States and, it is estimated that about 15% of AIDS cases are never reported and consequently, the delay-adjusted AIDS data is further inflated by $1/0.85$ (Centre for Disease Control 1990).

The proportion of unreported AIDS cases in South Africa is much higher than that of the US. Some studies carried out regarding the extent of under-reporting of AIDS cases in South Africa (Hollister and Deanery 2007 ; Nojilana, et al. 2009 ; Yudkin, et al. 2009) shows that it is a major problem in the country. It is therefore important that underreporting be taken into consideration when generalizing AIDS data for the whole country.

2.4.4 Incubation distribution Model

The incubation distribution is an important component of the back projection model. The uncertainty surrounding the choice of this distribution has attracted the interest of many statisticians. Various attempts have been made to derive an appropriate distribution model that could reflect the true incubation distribution. The first of such attempts was made by Lui et al (1986) who studied 100 transfusion-associated AIDS cases reported to the Centers for Disease Control (CDC) as at April 1, 1985. The dates of infection were assumed to be the dates of transfusion with infected blood and they defined incubation period as the period from transfusion to the diagnosis of the first opportunistic disease associated with AIDS. Using a simple average, 2.6 years was estimated as the average incubation period for transfusion related AIDS cases. In order to correct for length-biased sampling resulting from the fact that the sample does not include those exposed persons who have long incubation periods and have not yet been diagnosed, the authors assumed a family of probability densities to describe the incubation distribution and obtained a maximum likelihood estimate of mean incubation period as 54 months with 90% confidence bound of (2.6 years, 14.2 years). The study had the limitation that the dates of infection were determined retrospectively and it is difficult to estimate the probability that a member of an infected cohort would develop AIDS in a specified time period. Also, those with long incubation period were selectively excluded from the study. These limitations in the work of Lui et al (1986) led to the need to follow up a cohort of individuals. Goedert et al (1986) studied a cohort of individuals who were already infected with HIV but whose dates of

seroconversion were not known and estimated that the cumulative probability of developing AIDS within 3 years of follow up is 0.36 . This type of study is widely known as a prevalent cohort study. Brookmeyer and Gail (1987) considered the bias inherent in such a study and showed that failure to adequately adjust for duration of infection will under some conditions, bias relative risk estimates toward 1.0. Other improved studies that led to interval censoring in the estimation of infection times are the hemophiliacs cohort who were regularly seen at a treatment centre in Hershey, Pennsylvania and their serum samples stored since the mid 1970's (Eyster et al. 1987 ; Goedert et al. 1989), and the San Francisco City Clinic Cohort study involving homosexual men enrolled in hepatitis B vaccine trial (Hessol et al. 1987 ; 1989). Most of the derived estimates of the incubation distribution were obtained using the data from the three studies above. In this study we shall use interval left censoring.

The incubation period distribution $F(t)$ is defined as the probability that an HIV infected individual (child) develops AIDS within t years after infection. That is, if I is the random variable representing the incubation period, then,

$$F(t) = P(I \leq t)$$

and its probability density function is $\frac{dF}{dt} = F'(t) = f(t)$ and survival function

$S(t) = 1 - F(t)$. Its hazard function $\frac{(f)}{(s)}$ is the risk of developing AIDS at time t after

infection conditional on not having AIDS just before t . The hazard of progression to AIDS increases with time.

A discussion on some parametric distributions that have been used to model the incubation distribution is considered below.

2.4.5 The Weibull distribution model

The Weibull density function is given by

$$f(t) = \alpha \lambda (\lambda t)^{\alpha-1} \exp - (\lambda t)^\alpha$$

where $\lambda = \frac{1}{\sigma} > 0$, $\alpha > 0$, $t > 0$, and

the incubation period distribution is given as

$$F(t) = 1 - e^{-\lambda t^\alpha}$$

while the hazard function is given by

$$\lambda(t) = \lambda \alpha t^{\alpha-1}$$

which is monotonically increasing if $\alpha > 1$ and decreasing otherwise. This model was used in deriving estimates of incubation periods among haemophiliacs (Brookmeyer and Goedert 1989 ; Darby et al. 1990), among homosexuals and blood transfusion recipients (Lui et al 1988). It can also be used to study paediatric infections (Lui et al. 1986 ; Medley et al. 1987 ; Kalbfleish and Lawless 1989).

The limitation of the model however, is that it assumes that the hazard function increases indefinitely and is proportional to the power of time since infection which may not be true in the long run.

2.4.6 The Gamma distribution model

The Gamma distribution is another distribution that many authors (Medley et al. 1987 ; Boldson et al. 1988 ; Blythe and Anderson 1988 ; Freund and Book 1990) have applied in estimating the incubation period of HIV/ AIDS for both adults and children.

Its *pdf* is given by

$$f(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)}$$

where k is the stages which infected individual/child must pass through before developing AIDS and $\lambda > 0$ is a constant representing the hazard function of transiting from one stage to the next. The cumulative distribution function is

$$F(t, k, \lambda) = \int_0^t f(u; k, \lambda) du = \frac{\gamma(k, \lambda t)}{\Gamma(k)}$$

2.4.7 Log-Logistic distribution model

Some authors such as Lui et al (1988), Lawless and Sun (1992) have used the log-logistic distribution to estimate the incubation distribution. The simple form of the distribution is given as

$$F(t) = 1 - (1 + (\lambda t)^\beta)^{-1}$$

$\lambda > 0$, $\beta > 0$. If $\beta > 1$, the hazard function increases initially until it reaches maximum and then decreases as $t \rightarrow \infty$. This distribution has the same behaviour as the Lognormal distribution in that it assumes that the logarithm of the incubation period

follows a normal distribution. Rees (1987) and Boldson et al (1988) used a Lognormal distribution to study incubation distribution for both adults and children.

2.4.8 The Gompertz distribution model

The incubation period distribution can also be modelled as a Gompertz distribution.

Lui et al. (1986) and Wilkie (1989) used this method and its cumulative density function is given as

$$F(t) = 1 - e^{-\alpha(e^{\beta(t-s)} - 1)} \quad \text{for } t \geq s$$



2.4.9 The Staging Models

University of Fort Hare
Together in Excellence

The progression or evolution from HIV infection to AIDS can be divided into stages that are random and unequal in duration. The hazard functions of transition may differ from one stage to the other and this differentiates the stage model from the gamma model where the transition rates are assumed to be equal and constant. Brookmeyer and Liao (1990) used the model as

$$F(t) = \int_0^t \lambda_1 \alpha s^{\alpha-1} e^{-\lambda_1 s^\alpha} \{1 - e^{-\lambda_2 (t-s)}\} ds$$

The equation models the depletion of CD4+ cell. The time from infection to depletion of CD4+ to less than 200 CD4+ T cells is assumed to follow the Weibull distribution with hazard function $\lambda_1 \alpha t^{\alpha-1}$ and the time from less than 200 CD4+ to AIDS follow an

exponential distribution with hazard function λ_2 , then the incubation distribution is given as a convolution of the Weibull and exponential distributions discussed above.

2.4.10 Mixture models

The Mixture models partitions the infected individuals or children into two groups and assign them different incubation distributions: α the proportion with incubation distribution $F_1(t)$ and $(1-\alpha)$ with incubation distribution $F_2(t)$, then the incubation distribution for the entire population of the infected individuals / children as

$$F(t) = \alpha F_1(t) + (1-\alpha) F_2(t)$$

Lui et al (1988) used the mixture model to study incubation period among homosexual men whereas Auger et al (1988) used a mixture of Weibull distributions in a study of the incubation period among maternally infected newborns.

2.5 COMMENTS ON BACK-CALCULATION METHODS

There are some important points that need to be noted about back-calculation methods. Most criticisms of back-calculation models are based on the limitations or uncertainties surrounding the components and assumptions of the models some of which have been discussed above.

The incubation distribution is the most criticized. For instance, most estimates of the incubation period were estimated from data obtained from the study of a few cohorts who may not be representative of the general population in which the distribution is applied. The transfusion-associated AIDS cases study (Lui et al. 1986), the Multicenter AIDS Cohorts Study (MACS) (Munoz et al. 1989), (Kou et al. 1991), the San Francisco Men's Health study (Bacchetti and Jewell 1991) which provided data for the estimation of the incubation period may not have adequately represented the entire population. The infection times of the transfusion-associated AIDS cases were ascertained retrospectively after AIDS had developed, thus leading to right truncation. Also, the periodic testing of a cohort of hemophiliacs (Goedert et al. 1989) and homosexual men in San Francisco (Hessol et al. 1989) until they test positive and the ascertainment of the date of seroconversion from the interval defined by the latest screening test that was negative and the earliest screening that was positive, give rise to doubly censored data. This censoring has been handled by imputing the seroconversion time as the midpoint of the interval (Lui et al. 1988 ; Hessol et al. 1989 ; Biggar 1990). However, it has been shown by (DeGruttula and Lagakos 1989 ; Jewell 1990), that the midpoint approach

will only be optimal if the seroconversion density is uniform over the censoring interval, but the intervals are often sufficiently long to make this assumption void. Other authors have used parametric (Bookmeyer and Goedert 1989) and nonparametric (DeGruttula and Lagakos 1989 ; Bacchetti and Jewell 1991) modelling of both the seroconversion and incubation distribution.

In using back-calculation, the initial assumption was that the incubation distribution is stationary throughout the period and that the probability of progression to AIDS is the same for all infected individuals. Studies have shown that progression to AIDS depends on many variables particularly age (months/ years). Infants and the elderly have shorter incubation periods than other infected individuals. Use of antiretroviral therapy in the treatment of HIV infected individual may also lengthen the incubation period therefore faulting the stationarity assumption. Recent developments have modified the incubation distribution to accommodate changes in incubation period due to therapy. It is also a fact that the accuracy of AIDS incidence data is affected by underreporting and reporting delay. Delays in reporting has a serious effect on the most recent AIDS data series. These two sources of errors may undermine estimates obtained using back-calculation. To overcome these limitations, some underreporting and reporting delay adjustments have been proposed (Brookmeyer and Damiano 1989 ; Centers for Disease Control 1990 ; Harris 1990 ; Lawless and Sun 1992).

Due to the incubation period, the AIDS incidence data contain little information about the most recent infections. The numbers of individuals infected in the last one or two

years are not reliably reflected in the AIDS incidence data. This makes the model imprecise in estimating the recent infection rates.

2.6 IMPROVEMENTS ON BACK-CALCULATION METHODS

A great deal of improvements and advancements in back-calculation methods have occurred over the years. These major developments have arisen mainly in the attempt to overcome and rectify the limitations or uncertainties in back-calculation methods as has been reviewed in the previous section.

The first issue of concern was the problem of data truncation caused by reporting delays of AIDS incidence data. In an effort to compensate for the effects of these delays, several Authors (Harris 1990 ; Brookmeyer and Liao 1990 ; Rosenberg 1990 ; DeGruttula 1992 ; Bacchetti 1994), have suggested various ways of correcting such truncation and incorporating reporting delay distribution in back-calculation methods. Also the concern of whether the information contained in the AIDS incidence data is sufficient to reconstruct the HIV incidence curve led researchers to seek other sources of information to augment or substitute the AIDS data. Some authors have used a combination of AIDS data and information on date of AIDS diagnosis (Tango 1989), date of HIV diagnosis (Aalen et al. 1994 ; Cui and Becker 2000). Mariotto and Verdecchia (2000) substituted AIDS mortality data for AIDS registered cases while Deuffic and Costagliola (2006) considered the use of pre-AIDS mortality data instead of AIDS data. Chau et al. (2003 ; 2004), used data on HIV positive test only in place of

AIDS data. While Bellocco and Marschner (2000) analyzed jointly the HIV and AIDS surveillance data.

The introduction of effective therapy for the treatment of AIDS in both adults and children has resulted in the lengthening of the AIDS incubation period distribution. This led to the adjustment of the incubation distribution to accommodate changes in the distribution due to treatment (Brookmeyer and Liao 1990 ; Gail et al. 1990 ; Rosenberg et al. 1991). Therefore, rather than use a single incubation distribution, a family of distributions indexed by the calendar year of infection such that $F(t)$ becomes $F(t/s)$. This gives the probability that an individual or child infected with HIV at calendar time s develops AIDS within t years/months of infection. Hence, incubation distribution can be obtained for different cohorts.

The recognition that the incubation period distribution may vary with duration of infection led to the development of the stage model which assumes that HIV infected individuals or children pass through stages before developing AIDS. There could be two stages (Brookmeyer and Liao 1990 ; Brookmeyer 1991), or multiple stages defined by clinical criteria (Longini et al. 1989) or by multiple CD4 levels (Longini 1990). Assumptions are made to incorporate treatment into the model which has the effect to reduce the hazard of transition from one stage to the other.

The stage models do not incorporate the effect of changes in AIDS surveillance definition in the hazard model. The time-since-infection (TSI) models (Rosenberg et al. 1991) incorporate both treatment and the effect of redefinition of AIDS through the

hazard model. Unlike the stage models, the TSI models does not assume reduced hazard due to treatment effect until some time has elapsed since infection. Thus, the efficacy function is assumed to vary with time since infection.

In a study to investigate the incubation period of AIDS in patients infected via blood transfusion, Medley et al (1987) found that the mean incubation time for children is much shorter than that of adult patients and the incubation period for patients older than 59 years is less than that of younger adults. Also Goedert and his colleagues (1989) found that hemophiliacs over the age of 30 at infection are at higher risk of progression to AIDS than individuals 19 -30 years old at the time of infection. These findings led to the development of back-calculation models whose incubation distributions depend on age (Rosenberg 1994). Others (Becker and Marschner 1993 ; Marschner et al. 1998 ; Williams et al. 2001), incorporated age as a covariate in back-calculation model because it carries information about incubation period of the individual.

Since it is not easy to identify HIV incidence curve, Becker et al (1991 ; 1993) adopted a nonparametric approach to back-calculation using the maximum likelihood estimation procedure implemented in EM algorithm that incorporates smoothing at each step of iteration. The advantage of this method is that it gives the data greater power to determine the shape of the estimated infection intensity and avoids the assumption of parametric distribution of infection curve. Several authors (Cui and Becker 2000 ; Chau et al. 2003 ; Chau and Yip 2004) used this methods in conjunction with HIV data in back-calculation method.

2.7 PARAMETRIC BACK-PROJECTION

The term parametric back-projection is used to represent all back-calculation approaches where a particular functional form is assumed for the HIV incidence curve or the AIDS incidence curve. This includes the use of step functions, splines or any other form of parametric function. In particular, we shall review the works of Brookmeyer and Gail (1986 ; 1988) and Rosenberg and Gail (1991).

- i) Let $T_0, T_1, T_2, \dots, T_k$ be the time points (for example month, quarter, year) at which counts of AIDS cases are available.
- ii) T_0 is the start of the epidemic and T_k is the latest time at which AIDS data are available.
- iii) Partition the calendar time into $k+1$ intervals as $(T_0, T_1], (T_1, T_2], \dots, (T_{k-1}, T_k], (T_k, T_\infty]$.

Since there is no available data for the last interval $(T_k, T_\infty]$ the total number of individuals infected is not known.

- iv) Let Y_k be the number of AIDS cases diagnosed in the k th interval $(T_k - T_{k-1})$,

$$k = 1, 2, 3, \dots, k$$

$$\text{and } Y = \sum_{k=1}^K Y_k$$

- v) Let Y_{K+1} be the unobserved counts for the last interval $(T_K, T_\infty]$

- vi) The total number of individuals (children) infected with HIV as at T_k is

$$N = Y + Y_{k+1}$$

Let $\nu(s)$ be the infection curve which specifies the expected number of individuals or children infected with HIV in the time interval $(s, s + \delta s)$. And let the general families of all infection curves defined by a basis function $G = \{g_1(s), g_2(s), \dots, g_I(s)\}$ be a basis set.

Then $\nu(s) \in G$ if

$$\nu(s) = \sum_{i=1}^I g_i(s) \beta_i$$

The total number of children infected in an interval $(t_{k-1}, t_k] \subset (T_0, T_k]$ is

$$\int_{t_{k-1}}^{t_k} \nu(s) ds = \sum_{i=1}^I \{G_i(t_k) - G_i(t_{k-1})\} \beta_i \quad \text{where} \quad G_i(t) = \int_0^t g_i(s) ds$$

Hence the total number of children infected before T_k can be computed as

$$N = \int_0^{T_k} \nu(s) ds = \sum_{i=1}^I \{G_i(T_k)\} \beta_i$$

Given the incubation density $f(t)$ and the incubation distribution $F(t)$, we can express the expected number of those infected in the interval (T_{k-1}, T_k) as

$$E(Y_k) = \int_0^{T_k} \nu(s) f(t) dt$$

$$\begin{aligned}
&= \int_0^{T_k} \nu(s) \{F(T_k - s) - F(T_{k-1} - s)\} ds \\
&= \int_0^{T_k} \sum_{i=1}^I g_i(s) \beta_i \{F(T_k - s) - F(T_{k-1} - s)\} ds \\
&= \sum_{i=1}^I \beta_i \int_0^{T_k} g_i(s) \{F(T_k - s) - F(T_{k-1} - s)\} ds \\
&= \sum_{i=1}^I x_{ki} \beta_i \\
&= X\beta
\end{aligned}$$

Where

$$x_{ki} = \int_0^{T_k} g_i(s) \{F(T_k - s) - F(T_{k-1} - s)\} ds$$



$F(t)$ is the incubation distribution and for $t < 0$, $F(t) = 0$. This distribution is assumed known from other studies. Also the general family of infection curve G , is usually assumed to be either a step function (Brookmeyer and Gail 1986 ; 1988) or spline function (Rosenberg and Gail 1991). Having assumed the form of $F(t)$ and G , x_{ki} is computed and β_i is estimated via regression methods. In particular, the multinomial maximum likelihood , the quasi-likelihood and the Poisson likelihood approach can be applied in the estimation of β .

2.7.1 Estimation when G is a basis of indicator function

In this case the general family of infection curve G is defined by a set of base functions $g_i(s) = I\{s \in (t_{i-1}, t_i]\}$, where $I(A)$ is unity in the event of A and zero elsewhere. Hence $\nu(s)$ as defined in the equation in section 2.4 above is a step function.

Let $\mu_k = E(Y_k)$ and $X = [x_{ki}]$ be the $K \times I$ design matrix. $\mu' = (\mu_1, \mu_2, \dots, \mu_k)$ such that

$\mu = X\beta$. Let p_k be the probability that an individual or child infected before T_k is

diagnosed in the k^{th} interval and it is given as $p_k = \mu_k / N$ and $Np = \mu$

To accommodate those who were infected before T_k but not yet diagnosed, we

construct a $(K + 1) \times I$ augmented design matrix X^* whose $(K + 1)^{th}$ row is the vector

$x' = (x_{k+1,1}, x_{k+1,2}, \dots, x_{k+1,I})$. Hence $\mu^* = X^*\beta$ and $p_k^* = \mu_k^* / N$

2.7.2 The Multinomial Maximum Likelihood Estimate

In this case is assumed that the infection times of the N individuals (adults or children)

are independently and identically distributed and that the number of diagnosis within

the intervals follow a multinomial distribution such that

$$Y^* = (y_1, y_2, \dots, y_k, y_{k+1}) \sim Mult(N, p_1, p_2, \dots, p_k, p_{k+1})$$

and the likelihood function is given as

$$L(N, \theta, Y) = \prod_{i=1}^{k+1} \frac{N!}{y_1! y_2! \dots y_k! (N - y)!} p_1^{y_1} p_2^{y_2} \dots p_k^{y_k} p_{k+1}^{y_{k+1}}$$

The log-likelihood is given as

$$\begin{aligned} \ln L(N, \theta, Y) &= \log N! - \sum_{k=1}^K \log y_k! - \log(N - y)! + \sum_{k=1}^{K+1} y_k^* \log p_k^* \\ &= \log N! - \sum_{k=1}^K \log y_k! - \log(N - y)! + \sum_{k=1}^{K+1} y_k^* \log\left(\frac{\mu_k^*}{N}\right) \end{aligned}$$

$$\begin{aligned}
&= \log N! - \sum_{k=1}^K l \log y_k! - l \log(N - y)! + \sum_{k=1}^{K+1} y_k^* l \log \mu_k^* - \left(\sum_{k=1}^{K+1} y_k^* \right) \log N \\
&= l \log N! - \sum_{k=1}^K l \log y_k! - l \log(N - y)! + \sum_{k=1}^{K+1} y_k^* l \log \mu_k^* - N l \log N \\
&\approx l \log N! - l \log(N - y)! - N l \log N + \sum_{k=1}^{K+1} y_k^* l \log \mu_k^*
\end{aligned}$$

We note that N is implicitly a linear function of β , $N = \sum_{i=1}^I G_i(T_K) \beta_i$.

Differentiating the log-likelihood above with respect to β , Rosenberg and Gail (1986) obtained the following results;

$$\frac{\partial l_{MULT}}{\partial \beta} = (\ln \mu_{K+1} - \ln N + \psi(N) - \psi(N - y) - 1) \Delta + (N - y) \frac{x_{k+1}}{\mu_{k+1}} + X' \{diag(\mu)\}^{-1} y$$

Also, they derived the information matrix with respect to β as

$$I_{MULT(\beta)} = \left\{ \frac{1}{N} - \psi'(N) + \psi'(\mu_{K+1}) \right\} \Delta \Delta' + X^* \{diag(\mu^*)\}^{-1} X^*$$

Where $\partial \ln N! / \partial N = \psi(N)$ is the digamma function and $\frac{\partial^2 \ln N!}{\partial N^2} = \psi'(N)$

is the trigamma function (Abramowitz and Stegun 1964). Δ is given as

$$\Delta_i = \sum_{k=1}^{K+1} x_{ki}$$

Using the score and the information given above, Rosenberg and Gail (1991) obtained the multinomial maximum likelihood estimate using the updating scheme of the Fisher's scoring algorithm;

$$\hat{\beta}_{n+1} = \hat{\beta}_n + I_{MULT} \hat{\beta}_n^{-1} \frac{\partial l_{MULT}}{\partial \beta} \Big|_{\hat{\beta}_n}$$

2.7.3 Quasi-likelihood Estimate

From the above equation using the iterative reweighted least squares, the Quasi-likelihood is maximized by;

$$\hat{\beta}_{n+1} = \hat{\beta}_n + (X' \hat{\Sigma}^{-1} X)^{-1} X' \hat{\Sigma}^{-1} (y - \hat{\mu})$$

Where $\hat{\Sigma} = \text{diag}(\hat{\mu}) - \frac{\hat{\mu} \hat{\mu}'}{N}$



2.7.4 Poisson regression model

Some authors have also used Poisson regression where the observed AIDS diagnosis are assumed to be independent Poisson variates with means $\mu_k = x_{ki} \beta_i$. The Poisson likelihood function is given as

$$\ln L = \sum_{k=1}^J y_i \ln \sum_{i=1}^I x_{ki} \beta_i - \sum_{k=1}^K \sum_{i=1}^I x_{ki} \beta_i - \sum_{k=1}^K \ln y_k !$$

and the Poisson regression parameters are given by

$$\hat{\beta}_{n+1} = \hat{\beta}_n + (X' \{\text{diag}(\hat{\mu})\}^{-1} X)^{-1} (X' \{\text{diag}(\hat{\mu})\}^{-1} X)$$

2.7.5 The incubation distribution

The distribution of the waiting time between HIV infection and AIDS diagnosis is assumed to follow the Weibull distribution (Lui et al. 1986 ; 1988 ; Brookmeyer et al. 1989 ; Darby et al. 1990 ;). Its Probability density function is given by

$$f(t, \lambda, \gamma) = \begin{cases} \frac{\gamma}{\lambda} \left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-\left(\frac{t}{\lambda}\right)^\gamma} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

and its Cumulative distribution function is

$$F(t) = \begin{cases} 1 - e^{-\left(\frac{t}{\lambda}\right)^\gamma} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

2.7.6 Estimation when G is a spline function

Here the basis set is assumed to be a quadratic spline. According to Rosenberg and Gail (1991), the flexibility of the step function can be greatly enhanced by using a spline function $g_i(s)$ with knots at $t_l, l = 1, 2, \dots, L$ as a basis for G with the requirements that $v(s)$ be continuous at the knots or that $v(s)$ and its derivative $v'(s)$ be continuous.

Using the '+' function notation, they defined $v(s)$ as

$$v(s) = \sum_{j=0}^n \beta_{0j} s^j + \sum_{j=1}^L \beta_{ln} (s - t_l)_+^n.$$

Applying this to AIDS data from US, Rosenberg and Gail (1991) assumed a single knot

and letting $n = 2$, then $v(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{12}(s - t_1)_+^2$

so $g_1(s) = 1, g_2(s) = s, g_3(s) = s^2$ and

$$g_4(s) = \begin{cases} (s - t_l) & \text{if } t_l \geq T_L \\ 0 & \text{otherwise} \end{cases}$$

Hence,
$$x_{ji} = \int_0^{T_j} g_i(s) = F(T_j - s) - F(T_{j-1} - s) \, ds, \quad i = 1, 2, 3$$

integrating within the intervals of $g_i(s)$ then

$$x_{j4} = \int_{t_1}^{T_j} g_4(s) = F(T_j - s) - F(T_{j-1} - s) \, ds$$

The estimate of the HIV population in the same interval i is given as

$$N_i = \int_{g_i(s)} v(s) ds$$

and the estimate of the HIV population (either for adults or children) through the years in which AIDS diagnosis data is available is

$$N = \int_0^{T_j} v(s) ds$$

2.8 NON-PARAMETRIC BACK-PROJECTION

The ordinary back-calculation defines the yearly, quarterly, or monthly incidence of AIDS as an expression in terms of the average incidence of HIV in the corresponding time period as given by the equation ;

$$\mu_t = \sum_{s=1}^t \lambda_s f_{t-s,s}$$

where μ_t is the mean AIDS incidence at time t , λ_s is the mean of HIV incidence at time s and $f_{t-s,s}$ is the probability density function for someone infected at time s and diagnosed at time t .

One of the major disadvantages of the parametric back-projection is the problem of identifying the functional form of the HIV incidence curve because different forms may be consistent with the observed AIDS incidence. To overcome this limitation, Becker et al (1991) proposed an imposition of a smoothness restriction on a non-parametric form and estimates are obtained using the non-parametric maximum likelihood approach implemented in the EM algorithm, which has the following advantages. That;

- i) it avoids the assumptions of parametric distribution of the infection curve.
- ii) it gives the data greater power to determine the shape of the estimated intensity curve.
- iii) it overcomes the problem of identifying the form of the infection (HIV incidence) curve.
- iv) it ensures that all estimated values of HIV incidence is non-negative.
- v) the estimator of the HIV incidence has an explicit formula.

2.8.1 The EM algorithm

Suppose we let A_t be the number of AIDS cases diagnosed at time t and a_t be the observed diagnoses at time t such that $E(A_t) = \mu_t$. Also if N_{st} is the number of individuals or children infected with HIV at time s and diagnosed at time t and n_{st} is its realized but unknown value. Hence a_t is generated by n_{st} as

$$\sum_{s=1}^t n_{st} = a_t$$

We assume that each infected individual/ child has an incubation duration independent of the incubation duration of others and $f_{d,t}$ is the probability that the incubation period is d given that the individual was diagnosed at time t . Also the HIV incidence $N_1, N_2, \dots, N_T$ are assumed to be independent Poisson variates and $A_1, A_2, \dots, A_T$ be distributed independently as Poisson. Also, the conditional distribution of N_{st} given $A_t = a_t$ is binomial such that

$$E(N_{st} / A_t = a_t) = \frac{a_t \lambda_s f_{t-s}}{\mu_t}$$

The EM algorithm makes use of a more complete data set in maximum likelihood estimation. Here the HIV incidence data N_t is considered more complete than the AIDS incidence data since the AIDS data set depend on the HIV data.

Hence the maximum likelihood for the more complete set is given as

$$\ln L(\lambda; n) = \prod_{t=1}^{\tau} \frac{(\lambda_t f_{d,t})^{n_{st}} e^{-\lambda_t f_{d,t}}}{n_{st} !}$$

$$\begin{aligned} \text{so } \ln L(\lambda; n) &= \sum_{t=1}^{\tau} \sum_{s=1}^t \{n_{st} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t}\} - \sum_{t=1}^{\tau} \ln n_{st} ! \\ &\approx \sum_{t=1}^{\tau} \sum_{s=1}^t \{n_{st} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t}\} \end{aligned}$$

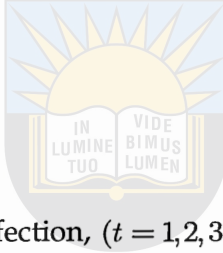
The expectation stage, the E-step, is given as

$$\sum_{t=1}^{\tau} \sum_{s=1}^t \left\{ \frac{a_t \lambda_t f_{d,t}}{\mu_t} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t} \right\}$$

$$= \sum_{t=1}^{\tau} \sum_{s=1}^t \left\{ \frac{a_t \lambda_t f_{d,t}}{\sum_{s=1}^t \lambda_t f_{d,t}} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t} \right\}$$

Becker et al (1991) obtained the M-step (the maximization stage) as

$$\lambda^{[L+1]} = \frac{\lambda_t^{[L]}}{F_{\tau-t,t}} \sum_{d=0}^{\tau-t} \frac{a_{t+d} f_{d,t}}{\sum_{i=1}^{t+d} \lambda_t^{[L]} f_{t+d-i,i}}$$



Where $d = t - s$ is the time since infection, ($t = 1, 2, 3, \dots, \tau$) a_t is the observed AIDS incidence at time t . At time $t < 1$, it is assumed that the disease has not yet emerged in the study population and τ is the last time when reliable data can be obtained and

$$F_{\tau-t,t} = \sum_{d=0}^{\tau-t} f_{d,t}$$

In order to obtain a smooth curve, smoothing step is attached to the EM algorithm

above to obtain $\lambda^{[L+1]}$ as

$$\lambda^{[L+1]} = \sum_{i=0}^k w_i \lambda_{t+i-k/2}^{[L+1]}$$

Where w_i is a symmetrical moving average weight and k is the window width which is usually an even integer.

2.8.2 Convergence

The likelihood is expected to increase at each iteration. Therefore to establish the maximum likelihood estimates, a convergence criterion is incorporated in the EMS iteration. The following criteria which are based on the parameter λ , is normally used.

$$\sum_{t=1}^{\tau} |\lambda_t^{[L+1]} - \lambda_t^{[L]}| < \varepsilon$$

$$\sum_{t=1}^{\tau} \frac{|\lambda_t^{[L+1]} - \lambda_t^{[L]}|}{\lambda_t^{[L]}} < \varepsilon$$

$$\frac{\left| \sum_{t=1}^{\tau} \lambda_t^{[L+1]} - \sum_{t=1}^{\tau} \lambda_t^{[L]} \right|}{\sum_{t=1}^{\tau} \lambda_t^{[L]}} < \varepsilon$$


The last two criteria are used in searching for the convergence in the EM algorithm, while the first criterion is used in the EMS algorithm. In some cases, because of the imprecision of the estimates near the most recent time τ , sometimes $\tau' = \tau - 1$ as the upper limit in the convergence criteria above.

2.9 MODIFICATION OF THE NON-PARAMETRIC BACK-PROJECTION

The ordinary back-calculation method, as considered above, makes use of diagnosed AIDS data in reconstructing the HIV infection curve. This approach has the limitation of not predicting precisely the HIV incidence in the recent past due to the long incubation period between HIV infection and AIDS diagnosis particularly in adults although for children, the period is relatively shorter. In order to overcome this limitation, Cui and

Becker (2000) and Chau et al (2003) suggested the use of HIV data in back-calculation for estimating HIV incidence curve. According to Chau et al (2003), HIV data has the following advantages:

- i) It contains more information than the AIDS data set because not all HIV infected children will develop AIDS by the time of analysis but some of them may have undergone an HIV test.
- ii) It is not affected by the redefinition of AIDS.
- iii) It is not affected by treatment effects as it is unlikely that individuals receive treatment before HIV diagnosis.
- iv) The induction period between HIV infection and HIV diagnosis is shorter than the incubation period which is the time between HIV infection and AIDS diagnosis. This short period implies that HIV data set contains more information than the AIDS data set.
- v) HIV data are more available than the AIDS data.

The back-projection method as given in the last section is modified using HIV diagnosis data instead of AIDS incidence data (Cui and Becker 2000 ; Chau et al. 2003). Hence the modified back-projection is given as

$$\mu'_t = \sum_{s=1}^t \lambda_s f'_{t-s,s}$$

where μ'_t is the mean number of HIV positive diagnosis at time t , λ_s is the mean HIV incidence at time s and $f'_{t-s,s}$ is the probability density function of the induction period for an individual or child infected with HIV at time s and diagnosed with HIV at time t .

The induction period (D') is defined as the period between HIV infection and HIV diagnosis.

The induction probability density function can be derived from HIV diagnosis hazard function which according to Cui and Becker (2000), can be expressed as an additive model through the natural hazard function for AIDS.

The natural hazard function associated with the waiting time between HIV infection and AIDS diagnosis (incubation period) is known and is given by

$$P(x / \mu) = \frac{f(x / \mu)}{1 - F(x / \mu)}$$

The hazard of an infected person/child taking a positive HIV test is assumed to arise from two sources:

i) Routine HIV testing: This refers to tests performed for reasons other than illness.

we assume that the hazard for this test is a constant v .

ii) Symptom-related testing: Are tests performed because of illness and its hazard

is assumed to be proportional to AIDS natural hazard function.

Hence the hazard function for HIV diagnosis is an additive hazard model given by

$$p'(x / \mu) = \begin{cases} v + \gamma p(x / \mu) & \text{if } x + \mu \geq \tau_0 \\ 0 & \text{if } x + \mu < \tau_0 \end{cases}$$

where τ_0 is the time when HIV diagnosis data became available and the unknown parameter γ is the coefficient of proportionality.

In order to estimate the two parameters v and γ , we need to obtain information on HIV diagnosis. If we define D' and D as the induction period for HIV and incubation period for paediatric AIDS respectively, then the probability that an HIV infected child may have HIV test before developing AIDS is given as

$$\Pr(D' < D / \mu) = \int_0^{\infty} f'(x / \mu)[1 - F((x / \mu)]dx = P$$

Where

$$f'(x) = \{v + \gamma p(x)\}e^{-\int_0^x \{v + \gamma p(w)\}dw}$$

and

$$f(x) = p(x)e^{-\int_0^x p(w)dw}$$

University of Fort Hare
Together in Excellence

Simplifying the above we have

$$\Pr(D' < D / \mu) = \int_0^{\infty} \{[v + \gamma p(x / \mu)]e^{-\int_0^x v + (\gamma+1)p(w/\mu)dw}\}dx = P \quad (i)$$

If we now let R be the waiting time from infection to HIV diagnosis if only routine HIV tests were available and S is the waiting time from infection to HIV diagnosis if only test were conducted when symptoms occur, then

$$f_R(x) = ve^{-vx}$$

And
$$f_s(x) = \gamma p(x)e^{-\int_0^x \gamma p(w)dw}$$

are the probability density function for the routine (R) and the Symptom (S) testing respectively, and

$$p(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta} \right)^{\alpha-1}$$

is the incubation period density.

Hence, the proportion of HIV positive tests from routine testing is given as

$$\begin{aligned} \Pr(R < s) &= \int_0^{\infty} f_R(x)[1 - F_s(x)]dx \\ &= v \int_0^{\infty} \left[e^{-\int_0^x (v + \gamma p(w))dw} \right] dx \end{aligned}$$



University of Fort Hare
Together in Excellence

If it is assumed that HIV routine tests and HIV-related symptoms tests contribute proportionately (say by p^*) to make children go for HIV diagnostic test, then,

$$\Pr(R < S / D' < D) = \frac{\Pr(R < S \cap D' < D)}{\Pr(D' < D)} = P^*$$

$$\begin{aligned} \Pr(R < S \cap D' < D) &= \Pr(R < S / D' < D) \Pr(D' < D) \\ &= P^* P \end{aligned}$$

Also

$$\begin{aligned} \Pr(R < D) &= \int_0^x v e^{-vz} [1 - F(x)] dx \\ &= v \int_0^{\infty} e^{-\int_0^x [v + p(w/\mu)]dw} dx \approx PP^* \end{aligned} \quad (ii)$$

Therefore in estimating the two parameters v and γ , we may solve equations (i) and (ii) simultaneously (DeGruttula et al. 1992 ; Cui and Becker 2000).

Since the induction period distribution is not affected by drug therapy, the different drug regimes as defined for the AIDS incubation distribution function does not apply here. Therefore, the induction distribution function is given as

$$F'(x / \mu) = \begin{cases} 1 - e^{-\int_0^{x+0.5} p'(w/u)dw} & \text{if } x + u \geq \tau_0 \\ 0 & \text{otherwise} \end{cases}$$

Hence the HIV incident is updated as

$$\lambda'^{[L+1]} = \frac{\lambda_t^{[L]}}{F'_{\tau-t,t}} \sum_{d=0}^{\tau-t} \frac{h_{t+d} f'_{d,t}}{\sum_{i=1}^{t+d} \lambda_t^{[L]} f'_{t+d-i,i}}$$

where h_t is the number of HIV diagnosis at time t and f' is defined as above.

2.10 CONCLUSION

In the discussions above, several statistical models / methods have been reviewed as well as their applications to analysis of various types of HIV/ AIDS data. We have discussed their strengths, weaknesses and limitations at length. It is clear from the discussions that to get good, accurate and meaningful results depends on the appropriate choice of method(s) and the type of data to which it is applied. In the next chapter, we discuss the statistical methods/ models used in this study and their application to the data collected.

CHAPTER THREE

METHODOLOGY

3.1 INTRODUCTION

This chapter describes the research and sample design for the study. It gives a detailed account of how the research was designed and data collected in line with the stated objectives. It also outlines how ethical issues related to the research were addressed. It discusses some challenges around statistical models and estimations used in studying mother-to-child (vertical transmission) of HIV and incubation period as well as their short comings. Lastly, an improved model to overcome these short comings is proposed for the study.



3.2 RESEARCH DESIGN AND DATA COLLECTION

Data was collected from Cecilia Makiwane (Mdantsane) hospital in East London, Eastern Cape province. A total of 1000 children was sampled from among 5700 children whose data was stored in computer Excel spreadsheet files in the Paediatric department of the hospital. These were some of the children presented to the hospital between 1st January 2003 to 30th June 2010. The Hospital was chosen because it is one of the major provincial referral hospitals, located in a heavily populated township and the patients came from diverse economic and social backgrounds both within the town and the surrounding peri-urban areas thus capturing the population dynamics

of the whole Eastern Cape province and to some extent, many parts of rural South Africa.

The Department of Health under whose jurisdiction the hospital is, keeps fairly good records of surveillance data on paediatric HIV infections for first admissions and follow-ups since 1990 although the details for 2010 were not complete at the time of data collection (June 2010).

The children were given registration/admission numbers which were entered on hard copy files for each patient (child) before being allowed to see a doctor. Clinical diagnoses and other test results prescribed by the doctor were subsequently entered into the file of each patient and all the information later captured accurately in computer excel spreadsheet files. The records in the spreadsheet files gave accurate details and up to date information for references including full clinical diagnoses for the admissions and subsequent visits, tests results and follow-ups thereafter.

A random sample of 1000 children were chosen using computer generated random numbers and information extracted from these files and filled accurately in a prepared questionnaire by trained nurses (see Appendix 2). No names or addresses were filled in the questionnaires for reasons of confidentiality and ethical considerations, but only numerical numbers corresponding to the admission numbers on their admission files for purposes of identification and analysis. The total summary of information from all the questionnaires formed the data base for analysis.

Children born to HIV positive mothers and those presented as inpatients or outpatients with signs and symptoms of WHO clinical definition of Paediatric AIDS were tested for HIV serum antibodies. Those who fulfilled the WHO clinical case definition and confirmed antibody positive were considered to be HIV positive. The following classification was used:

Probable maternal origin if the mother tests antibody positive ; probable transfusion origin if the mother test negative and the child has received at least one blood transfusion before the onset of the symptoms; unknown origin if the mother tests negative and the child has never received any blood transfusion; probable maternal origin if the mother has not been tested for HIV antibodies (mother has died, refused to be tested, or not the child's biological mother). Other tests and results on subsequent follow-up visits were also recorded.

3.3 ETHICAL CONSIDERATIONS

Due to the sensitive nature of the study, certain ethical considerations were put in place;

- i) Ethical clearance was obtained from both the University and the hospital from where the data was collected.
- ii) Questionnaires were prepared in such a way that names of children enrolled in the study from the sampled files were replaced with numerical numbers for purposes of identification and coding in data analysis.
- iii) Confidentiality of the data extracted from the sampled files and filled in the Questionnaires was maintained at all times and was used only for purposes of

data analysis in the study.

3.4 CHALLENGES IN SOME STATISTICAL MODEL ESTIMATION

This section discusses the challenges of estimating the rate of vertical (i.e. mother-to-child) transmission of human immunodeficiency virus (HIV) and, for children who acquire the virus, the age distribution at diagnosis of acquired immune deficiency syndrome (AIDS). This distribution, also known as the paediatric AIDS incubation period, has been estimated using data from AIDS surveillance systems (Auger et al. 1988; Downs, Salamina, and Ancelle-Park 1995), and sometimes supplemented with data from neonatal HIV screening programs (DeGruttola, Tu, and Pagano 1992; MaWhinney and Pagano 1994). However, the focus on this study is on data of children born to known HIV-positive mothers collected from the hospital. The analysis of such studies may sometimes appear straight forward, but complications are introduced by the difficulty in diagnosing HIV infection in young children (Consensus Workshop 1992).

The standard diagnostic test for HIV infection is based on the detection of HIV-specific immunoglobulin class G (IgG) antibody, but because maternal IgG antibody crosses the placenta, all children born to infected mothers initially test antibody positive. Maternal antibody is usually assumed to have cleared by age 15 months or 18 months, although a 12- month and a 24-month threshold have also been noted in some studies (Dabis et al. 1993). In some studies, results of tests that directly measure the presence of the virus

(i.e antigen, virus culture, polymerase chain reaction [PCR]) are also available. A positive virus test result at any age is indicative of infection, but these tests have not generally been regarded as sufficiently sensitive to allow the absence of infection to be inferred from a negative test result. Certain clinical signs and symptoms also provide evidence of infection status, but a definitive diagnosis would usually be made only following the diagnosis of AIDS.

The first stage in a standard analysis is to classify children as infected, uninfected, or of indeterminate infection status, on the basis of clinical, immunological, and virological information (Dabis et al. 1993). Although standard classification systems have been proposed (e.g. Centers for Disease Control 1987a), different studies have used slightly different criteria (Matheson et al. 1995). The vertical transmission rate is then estimated as the ratio of the number of infected children to the number of infected plus uninfected children (Boylan and Stein 1991). Children whose infection status has not been determined because they were lost to follow-up, died from causes unrelated to HIV infection, or simply were born shortly before the date of analysis do not enter the calculation. This results in a biased estimate, because the unconditional probability of vertical transmission generally is not equal to the probability of transmission given that no infection-defining event has been observed.

The pediatric AIDS incubation period is usually estimated by the product-limit method (Kaplan and Meier 1958), in which the analysis is based on known infected children and children who do not develop AIDS are right-censored at their last clinical examination

(Blanche et al. 1994; European Collaborative Study 1994). Exclusion of indeterminate children again gives rise to biased estimates, as these children do not enter the risk set at any age despite the fact that some are truly infected and thus at risk of AIDS.

A technique to reduce these biases is to limit the analysis to children whose infection status should in principle be known. For example, if antibody persistence beyond age 18 months were regarded as proof of infection, then children born within 18 months of the date of analysis would be excluded. A disadvantage of this technique is a loss of information, which may be particularly serious when a comparatively high number of children are born after the inclusion date. Two alternative methods of analysis that are connected with cure models (Laska and Meisner 1992) are described here. The first method used the fact that infected children are persistently HIV antibody positive, whereas uninfected children ultimately lose antibody (European Collaborative Study 1991). Thus the survivor function of the proportion of children who remain antibody positive should, when applied to the entire cohort, asymptote at a level that corresponds to the transmission rate. This method has the advantage of including all antibody test results, irrespective of whether the child had been followed to age 18 months. For children who died due to HIV infection (i.e., non-independent censoring), the censoring age was set to the age they would have been at the date of analysis had they not died.

A different analytic approach was used in the ACTG- 076 trial, which aimed to assess the efficacy of zidovudine in preventing vertical HIV transmission (Connor et al. 1994).

Here transmission rates were estimated from a Kaplan-Meier analysis of time to first positive virus culture. A concern with this analysis is that few infected children were followed beyond the age at which the curves asymptoted (24-36 weeks), and the possibility that later events will be observed cannot be excluded. Moreover, because interval censoring was not taken into account, the estimated distribution functions depend on the frequency of testing, although whether this has an important effect on the asymptote is unclear. Also, no justification was given for censoring children at their latest negative antibody test.



The objective of this study is to describe a new method of analysis, which will be used to estimate the probability of infection for every child in the study, conditional on his or her observed data. This entails iterative and simultaneous estimation of the vertical transmission rate, the sensitivities of the virus tests, the AIDS incubation period, and the distribution function of antibody loss, taking into account interval censoring. The distributions are estimated non-parametrically, and there is no need for an arbitrary threshold age at which antibody persistence is taken as proof of infection. By taking children of indeterminate infection status into account, the method avoids the biases intrinsic to the crude estimators described previously. Furthermore, because it utilizes clinical, immunological, and virological data, it should be more efficient than the methods based on cure models, which essentially rely on a single variable. However, in common with the other estimators, the method does require the assumption of independent censoring (Lagakos 1979).

3.5 THE STATISTICAL MODEL FOR THE STUDY

Suppose that :

β = the probability of vertical transmission

A_t = the age at AIDS diagnosis

S_t = the age when the child first becomes seronegative.

One may consider the use of a continuous time model, but for this study, a discrete time model is more convenient and fitting for purposes of explanation and estimation of the parameters used.

Let : $f_t = \Pr(A_t = t / A_t > t-1, \text{ infected})$ be the AIDS hazard at age t months given that the child is infected.

$h_t = \Pr(S_t = t / \text{uninfected})$ be the probability density function of antibody loss at age t months given that the child is uninfected.

$F_t = \prod_{u=0}^{t-1} (1 - f_u)$ and $H_t = 1 - \sum_{u=0}^{t-1} h_u$ denote the survivor functions for AIDS and antibody loss respectively.

And finally, up to V different types of virus tests may be applied with sensitivities

$s = [s_1, s_2, \dots, s_V]$. The tests are assumed to be perfectly specific, so that a child is

inferred to be infected on the basis of one or more positive tests. A further assumption is that all virus tests results are independent, conditional on infection status.

Suppose that each child in a cohort of size n is classified into one of the four mutually exclusive groups (see Table 3. 1) :

n_A : children are observed to develop AIDS, and the age at diagnosis is assumed to be accurate.

n_s : children become antibody negative (seronegative) at an unknown point between the last positive antibody test and the first negative antibody test.

n_v : children have not developed AIDS but are positive on one or more virus tests. and the remaining ;

n_c : children have not developed AIDS, are never positive on a virus test, and remain antibody positive.

This last group will be referred to as censored, which it should be noted, does not mean the same as indeterminate infection status.

The problem can be considered as a mixture model with two states: infected and uninfected children. A special feature of the data, analogous to the problem considered by Hosmer (1973), is that state membership is known for some observations. Children who develop AIDS or test virus positive are assumed to be infected, and children who become seronegative are assumed to be uninfected.

The contribution to the likelihood of a subject whose state is known is the product of the unconditional probability of belonging to that state and the conditional probability of the observed data. The contribution of a subject whose state is not known is the sum across states of all such terms. Table 3.1 shows each child's contribution to the likelihood function, ignoring terms that do not involve the unknown parameters.

Consider, for example, a censored child who has not developed AIDS by age a_k , remains antibody positive at age l_k , and who has had, by virus test v , a total of y_{kv} negative results $v = (1, 2, \dots, V)$. If the child is infected, he provides no information about the antibody loss distribution H , and if he is uninfected, he provides no information about the AIDS incubation period F or about the sensitivities of the virus tests s . His contribution to the likelihood function is thus the sum of two components.

Table 3.1 : classification of children and information based on the proposed method

Group	HIV infected	Number of children	Required information	Contribution to likelihood
Clinical AIDS (AIDS)	Yes	n_A	Age at AIDS diagnosis a_k Number of positive virus tests x_{kv} Number of negative virus tests y_{kv}	$\beta f_{ak} \prod_{t=0}^{a_k-1} (1 - f_t) \prod_{v=1}^V S_{\nu}^{x_{kv}} (1 - S_{\nu})^{y_{kv}}$
Negative			Age at last positive antibody test l_k	$(1 - \beta) \sum_{t=l_{k+1}}^{r_k} h_t$
Antibody test (SERO)	No	n_S	Age at first negative antibody test r_k	
Positive virus			Age at last clinical assessment a_k	
Test but not AIDS (VIR)	Yes	n_V	Number of positive virus tests $X_{kv} (\geq 1)$ Number of negative virus tests y_{kv}	$\beta \prod_{t=0}^{a_k} (1 - f_t) \prod_{v=1}^V S_{\nu}^{X_{kv}} (1 - S_{\nu})^{y_{kv}}$
None of the above (CENS)	No assumption	N_C	Age at last clinical assessment a_k Age at last antibody test l_k Number of negative virus tests y_{kv}	$\beta \prod_{t=0}^{a_k} (1 - f_t) \prod_{v=1}^V (1 - S_{\nu})^{y_{kv}} + (1 - \beta) \sum_{t=l_{k+1}}^{\infty} h_t$

NB : Subscript k denotes the k th child within a subgroup and subscript v denotes the v th type of virus test.

The first component is the product of the unconditional probability of being infected, β ,

and of the conditional probabilities of not developing AIDS, $\prod_{t=0}^{a_k} (1 - f_t)$, and of always

testing virus negative, $\prod_{\nu=1}^V (1 - s_\nu)^{y_{kv}}$. The second component is the product of the

unconditional probability of not being infected, $(1 - \beta)$, and the conditional probability

of remaining antibody positive, $\left(\sum_{t=l_{k+1}}^{\infty} h_t = 1 - \sum_{t=0}^{l_k} h_t \right)$. The contribution to the likelihood

function of children in the AIDS, VIR, and SERO groups are derived below.

Let the log-likelihood $L(\beta, s, f, h)$ be the sum of the logarithm of the individual contributions given by ;

$$\begin{aligned}
 & \sum_k \left\{ \log(\beta) + \log(f a_k) + \sum_{t=0}^{a_k-1} \log(1 - f_t) + \sum_{\nu=1}^V [x_{kv} \log(s_\nu) + y_{kv} \log(1 - s_\nu)] \right\} \\
 & + \sum_{\substack{\text{SERO} \\ k}} \left\{ \log(1 - \beta) \left[\sum_{t=l_{k+1}}^{r_k} h_t \right] \right\} \\
 & + \sum_{\substack{\text{VIR} \\ k}} \left\{ \log(\beta) + \sum_{t=0}^{a_k} \log(1 - f_t) + \sum_{\nu=1}^V [x_{kv} \log(s_\nu) + y_{kv} \log(1 - s_\nu)] \right\} \\
 & + \sum_{\substack{\text{CENS} \\ k}} \log \left[\beta \prod_{t=0}^{a_k} (1 - f_t) \prod_{\nu=1}^V (1 - s_\nu)^{y_{kv}} + (1 - \beta) \sum_{t=l_{k+1}}^{\infty} h_t \right] \tag{i}
 \end{aligned}$$

Subject to $0 \leq \beta \leq 1$, $0 \leq s_\nu \leq 1$, $0 \leq f_t \leq 1$, $0 \leq h_t \leq 1$, $\sum_{t=0}^{\infty} h_t = 1$.

3.6 IDENTIFIABILITY

In the absence of virus test data, it can be shown that scaling β , $1 - F$, and $1 - H$ by the factors c , $1/c$, and $(1 - \beta) / (1 - c\beta)$ has no effect on the value of the likelihood function. This signifies a lack of identifiability, and the existence of unique maximum likelihood estimates requires that either $\hat{F}$ or $\hat{H}$ be fixed at a point t . However, it suffices to apply the mild constraint that all antibody loss occurs by the oldest age at which an antibody test was performed. In fact, even this constraint is unnecessary if the oldest age at AIDS diagnosis exceeds the oldest censoring age for AIDS, as this removes the indeterminacy in the right tail of $\hat{F}$ (Kaplan and Meier 1958). The issue of identifiability does also not arise if there is at least one child in the VIR category or at least one child in the CENS category who has had a negative virus test result.

3.7 ESTIMATION OF THE PARAMETERS

Authors such as Dempster, Laird, and Rubin (1977) showed that finite-mixture models can be cast in the framework of an EM algorithm by introducing unobserved dummy variables to represent state membership. In the E step, the conditional probabilities of belonging to each of the states are estimated for each observation, given current parameter estimates. The contribution of each observation to the complete-data log-likelihood is the weighted sum of the log-likelihood associated with each of the states, with weights given by the conditional probabilities. The M step involves re-estimating the parameters by maximizing the complete-data log-likelihood.

3.7.1 E Step.

The conditional probabilities of state membership are fixed (0 or 1) for children in the AIDS, SERO, and VIR groups. For children in the CENS group, the conditional probabilities of infection are, using Bayes's theorem is,

$$\hat{\tau}_k = \frac{\hat{\beta} \hat{F}_{a_{k+1}} \prod_{v=1}^V (1 - \hat{s}_v)^{y_{kv}}}{\hat{\beta} \hat{F}_{a_{k+1}} \prod_{v=1}^V (1 - \hat{s}_v)^{y_{kv}} + (1 - \hat{\beta}) \hat{H}_{k+1}}$$

The complete-data log-likelihood L_C is identical to the observed -data log-likelihood L (Eq. i), except that the contribution from the censored children becomes

$$\sum_{\substack{\text{CENS} \\ K}} \hat{\tau}_k \left[\log(\beta) + \sum_{t=0}^{a_k} \log(1 - f_t) + \sum_{v=1}^V y_{kv} \log(1 - s_v) \right] + (1 - \hat{\tau}_k) \left[\log(1 - \beta) + \log \left(\sum_{t=k+1}^{\infty} h_t \right) \right].$$

University of Fort Hare
Together in Excellence

3.7.2 M Step.

This step consists of four parts :

1. Re-estimation of β by the ratio of the estimated number of infected children to

the total cohort size ; $\hat{\beta} = \left(n_A + n_V + \sum_{\substack{\text{CENS} \\ K}} \hat{\tau}_k \right) / n$, which is the solution of the

equation

$$\frac{\partial L_C}{\partial \beta} = \frac{n_A + n_V}{\beta} - \frac{n_S}{1 - \beta} + \sum_{\substack{\text{CENS} \\ K}} \left(\frac{\hat{\tau}_k}{\beta} - \frac{1 - \hat{\tau}_k}{1 - \beta} \right) = 0$$

2. Re-estimation of the components of s by the ratio of the number of positive test results to the estimated total of tests performed on infected children :

$$\hat{s}_v = \frac{\sum_{AIDS_K} xkv + \sum_{VIR_K} xkv}{\sum_{AIDS_K} (xkv + ykv) + \sum_{VIR_K} (xkv + ykv) + \sum_{CENS_K} \hat{\tau}_k ykv}$$

which is the solution of the equation

$$\frac{\partial L_C}{\partial s_v} = \sum_{AIDS_K} \left(\frac{xkv}{s_v} - \frac{ykv}{1-s_v} \right) + \sum_{VIR_K} \left(\frac{xkv}{s_v} - \frac{ykv}{1-s_v} \right) - \sum_{CENS_K} \hat{\tau}_k \frac{ykv}{1-s_v} = 0$$

3. Re-estimate of F by the product-limit method where the discrete hazards are based on the *estimated* number at risk of AIDS (estimated number infected) at each stage:

$$\hat{f}_t = \frac{\sum_{AIDS_K} I(a_k = t)}{\sum_{AIDS_K} I(a_k \geq t) + \sum_{VIR_K} I(a_k \geq t) + \sum_{CENS_K} \hat{\tau}_k I(a_k \geq t)}$$

which is the solution of

$$\frac{\partial L_C}{\partial f_t} = \sum_{AIDS_K} \frac{I(a_k = t)}{f_t} - \sum_{AIDS_K} \frac{I(a_k \geq t+1)}{1-f_t} - \sum_{VIR_K} \frac{I(a_k \geq t)}{1-f_t} - \sum_{CENS_K} \hat{\tau}_k \frac{I(a_k \geq t)}{1-f_t} = 0,$$

where I is the identity function.

4. There is no closed-form solution for $\hat{H}$, but a standard algorithm for estimating a distribution function from interval-censored data (Turnbull 1976) can be used.

Although the original method assumed that all observations have equal weight,

DeGruttola et al. (1992) pointed out that unequal weights are permissible. In this

application, the weights are the current conditional probabilities of being uninfected.

The method iteratively reallocates the weight associated with each child across all times

when antibody loss could have occurred, in proportion to the current estimates of the probabilities of antibody loss for those intervals. Thus the estimate $\hat{h}^{(i+1)}$ at iteration $(i + 1)$ is obtained from the estimate $\hat{h}^{(i)}$ at iteration i by

$$\hat{h}_t^{(i+1)} = \sum_{\substack{SERO \\ k:l_k+1 \leq t \leq r_k}} \frac{\hat{h}_t^{(i)}}{\sum_{u=l_k+1}^{r_k} \hat{h}_u^{(i)}} + \sum_{\substack{CENS \\ k:l_k+1 \leq t}} (1 - \hat{\tau}_k) \frac{\hat{h}_t^{(i)}}{\sum_{u=l_{k+1}}^{\infty} \hat{h}_u^{(i)}} \quad (ii)$$

followed by rescaling to ensure that $\sum_{t=0}^{\infty} \hat{h}_t^{(i+1)} = 1$

The estimation of H is itself an EM algorithm, and when algorithms are nested as here, it is computationally inefficient to drive the inner algorithm to convergence (Healy and Dempster et al. 1977). For data arising from an exponential family where the M step does not have a closed-form solution, it has been proved that convergence to a local maximum is ensured with a single Newton- Raphson iteration in the M step (Rai and Matthews 1993). Although this problem does not fall within this framework, the empirical observation is that convergence is achieved with one iteration of Equation (ii) per cycle of the main EM algorithm.

3.8 PRECISION OF THE ESTIMATES

Some authors have described methods for obtaining the variance-covariance matrix for parameters when the complete-data information matrix is more tractable than the observed-data information matrix (Meng and Rubin 1991). In this application, however, partial double derivatives of the observed-data log-likelihood can be evaluated directly. When calculating pointwise standard errors for the product-limit estimator, only observed event times need be considered. Also, the pointwise standard errors of a survivor function estimated from interval censored data are a function of the nonzero elements of the empirical probability density function (Turnbull 1976). Therefore, what is needed is only the computation of the derivatives of the nonzero elements of g and h . Work on coverage probabilities of confidence intervals for the product-limit estimator suggests the use of a logarithmic scale for analysis (Miller 1981). Thus the AIDS survivor function has been parameterized in terms of $\lambda_t = \log(1 - f_t)$. Then

$$\log F_t = \sum_{u=0}^{t-1} \lambda_u$$

$$\text{and } \text{var}(\log F_t) = \sum_{u=0}^{t-1} \text{var}(\lambda_u) + 2 \sum_{u=0}^{t-1} \sum_{v=u+1}^{t-1} \text{cov}(\lambda_u, \lambda_v).$$

The observed information for β can be expressed as

$$\frac{n_A + n_V}{\hat{\beta}^2} + \frac{n_S}{(1 - \hat{\beta})^2} + \sum_{\substack{\text{CENS} \\ K}} \left[\frac{\hat{\tau}_k - \hat{\beta}}{\hat{\beta}(1 - \hat{\beta})} \right]^2.$$

The last term of this expression indicates the extra information gained by taking children of indeterminate infection status into account. For censored children who

receive no virus tests and who are last observed before the earliest age at AIDS

diagnosis and earliest age at antibody loss, $\hat{\tau}_k = \hat{\beta}$. Thus it does not matter whether or not these children are retained in the analysis.

3.9 COVARIATES

There may be interest in assessing the effects of covariates on the model parameters, particularly the vertical transmission risk and the AIDS incubation period. For a discrete covariate, it is straight forward to introduce category-specific parameters in the relevant steps of the algorithm. For the transmission risk or sensitivities of the virus tests, the significance of the covariate can be assessed by a likelihood ratio test.

A natural way to assess significance with respect to AIDS incubation period is by the

$\sum (O - E)^2 / E$ variant of the log-rank test (Peto and Peto 1972), using the estimated numbers at risk of AIDS (under the null model). This procedure will be somewhat under conservative, as it does not take into account uncertainty in the estimated numbers at risk. In the general case, the log-rank test can be derived as a score test under the proportional hazards model (Breslow 1975). However, in this application it is not possible to regard the underlying hazard as a nuisance distribution that can be conditioned out, because of the contribution to the likelihood of the censored children. Finkelstein (1986) developed a proportional hazards model for interval censored data that could be considered for the antibody loss distribution. It would not be straight-forward to modify the EM algorithm to accommodate this model, and direct numerical

maximization of the likelihood function might need to be considered. Also, an intrinsic feature of this model (unlike the analogous model for exact or right-censored data) is the need to estimate the baseline distribution. This could affect the stability of the estimates of the covariate coefficients, particularly if age was not recorded appropriately.



University of Fort Hare
Together in Excellence

CHAPTER FOUR

DATA ANALYSIS

4.1 INTRODUCTION

As has already been discussed in the three previous Chapters, Mother-to-child transmission of HIV infection is a serious problem in most developing countries, particularly in South Africa where the infection rate of HIV though decreasing, is still quite high. This chapter deals with the analysis and estimate of the rate of human immunodeficiency virus (HIV) transmission in children and the pediatric acquired immune deficiency syndrome (AIDS) incubation period using data collected from Cecilia Makiwane Hospital in the Eastern Cape province of South Africa. Since 1985, doctors and pediatricians working in various hospitals in South Africa have recorded data on children born to mothers known to be HIV infected at the time of delivery (UNICEF, 2010). In principle, children who attend clinics are examined clinically and tested for HIV antibody every 3 months until age 2 years and every 6 months thereafter. Virus tests were done using laboratory expertise and facilities which exist at Cecelia Makiwane hospital. This analysis utilizes the results of PCR and virus culture assays performed after the first month of life. Age restriction was applied to allow for the rapid rise in sensitivity after birth. This method has been used widely (Dunn et al. 1995; McIntosh et al. 1994). The study used the 1987 Centers for Disease Control (CDC) pediatric AIDS definition (Centers for Disease Control 1987b) with two exceptions: It excludes children with asymptomatic lymphoid interstitial pneumonitis

diagnosed by X-ray only and includes children considered to have died as a consequence of HIV infection as per the AIDS case definition. The same was applied in this study.

A total of 1000 children were sampled for the study (June 2010), excluding twins or children with an older siblings in the study. Of these, 42 children had developed AIDS (range age 2-74 months), and 654 children had become seronegative. The earliest negative anti-body test was observed at 2 months (4 children), and the latest positive test at 21 months (2 children). Ignoring tests performed on the seronegative children, there were 605 virus cultures and 234 PCR assays. One or more positive virus test results were observed for 70 children who had not developed AIDS. The remaining 234 children were censored. The proposed method requires that each child be classified into one of the groups as in Table 3.1 above, but in practice inconsistencies arise, even after eliminating obvious laboratory errors. In the data, 6 apparently infected children were transiently HIV antibody negative, 3 children lost HIV antibody after developing AIDS, and 14 children had one or more positive virus tests despite becoming consistently antibody negative. Similar inconsistencies have been observed in other studies and explanations for them have been put forward (Bryson 1995; Simpson and Andiman 1994). To resolve these inconsistencies, the following hierarchical rule was applied :

1. If a child develops AIDS, then classify him or her in the AIDS group.
2. Otherwise, if the last antibody test is negative, then classify him or her in the SERO group.

3. Otherwise, if any virus test is positive, then classify him or her in the VIR group.
4. Otherwise, classify him or her in the CENS group.

Giving precedence to AIDS diagnoses over laboratory findings is consistent with the CDC classification of HIV infection (Centers for Disease Control 1987a), and serological tests are at present judged more reliable than virological tests. This hierarchy could be reordered without changing the general form of the likelihood, since in most studies results will not be sensitive to the precise choice of hierarchy. The method described in Chapter Three was developed in the context in which the clinical diagnosis of AIDS coupled with HIV testing can reasonably be assumed to be 100% specific for detection of HIV infection especially in settings where AIDS infant mortality rate is high such as in South Africa (World Health Statistics, 2009).

4.2 DATA ANALYSIS

One of the methods used in analyzing the data is the Expectation Maximization (EM) algorithm method. This method yielded maximum likelihood (ML) estimates of 0.173 with a standard error (SE) of 0.021 for vertical transmission rate; 0.798 with SE of 0.036 for sensitivity of viral culture; and 0.947 with SE of 0.049 for sensitivity of PCR. Figure 4.1 shows the estimated survivor functions for antibody loss and AIDS up to age 24 months. Figure 4.2 extends the AIDS survivor function up to age 48 months and indicates the precision of pointwise estimates. These distributions should be represented as step functions, but for clarity, the age-specific estimates have been connected with straight lines. An estimated 0.154 (90% confidence interval of 0.099 - 0.204) of infected children were diagnosed with AIDS by age 6 months, with a fairly constant hazard thereafter (approximately 0.091 per year). The median duration of antibody persistence among uninfected children is 11 months, with all children losing antibody by 22 months of age. Figure 4.1 also shows the conditional probability of infection for a child who is antibody positive, who has not developed AIDS, and for whom no virus test results are available. This function remains relatively flat until age 10-11 months before rising rapidly. The data rule as infected all children who test antibody positive after 22 months. The effect of each negative virus test result is to scale the odds of infection by $(1 - \text{sensitivity})$; that is, by 0.238 for a negative virus culture result and by 0.091 for a negative PCR result. These analyses assumed that all parameters remained constant over the duration of the study. Using the methods

described in chapter three section 3.5, the 425 children born in 2003-2006 were compared to the 575 children born in 2007-2010 with respect to the rate of vertical transmission and the AIDS incubation period. Under the assumption of a stationary incubation period, the estimated transmission rate was 0.125 for 2003-2006 and 0.181 for 2007-2010 (likelihood ratio $\chi^2 = 4.92$, $P = 0.03$)

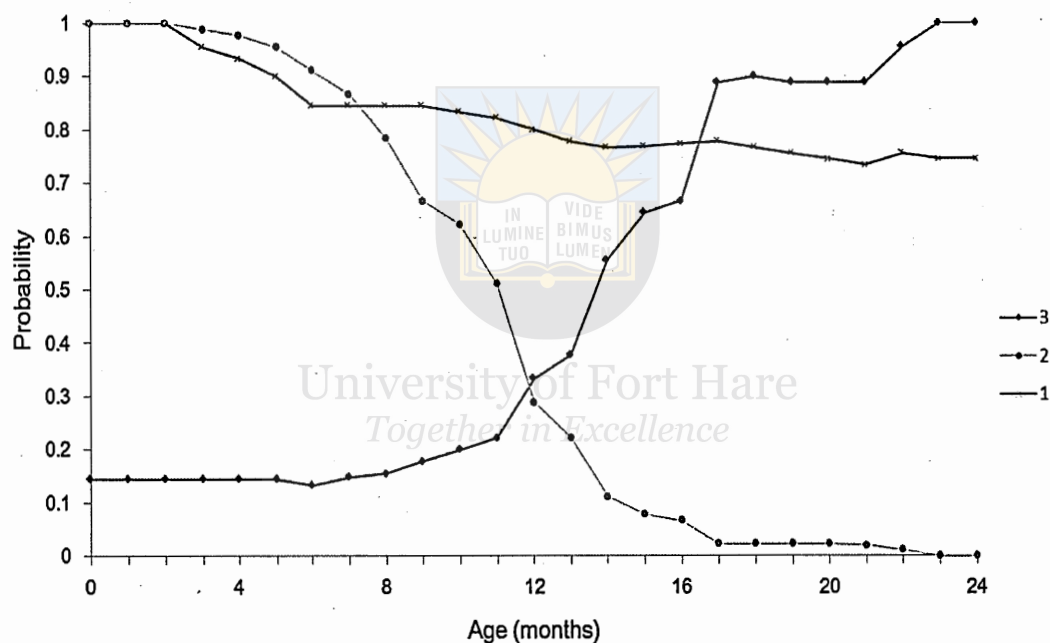


Figure 4.1 Estimates of the survivor functions for AIDS(1). Antibody Loss (2) and the Conditional probability of infection for a Child Age t months who is antibody positive , has not developed AIDS, and has never received a virus test (3).

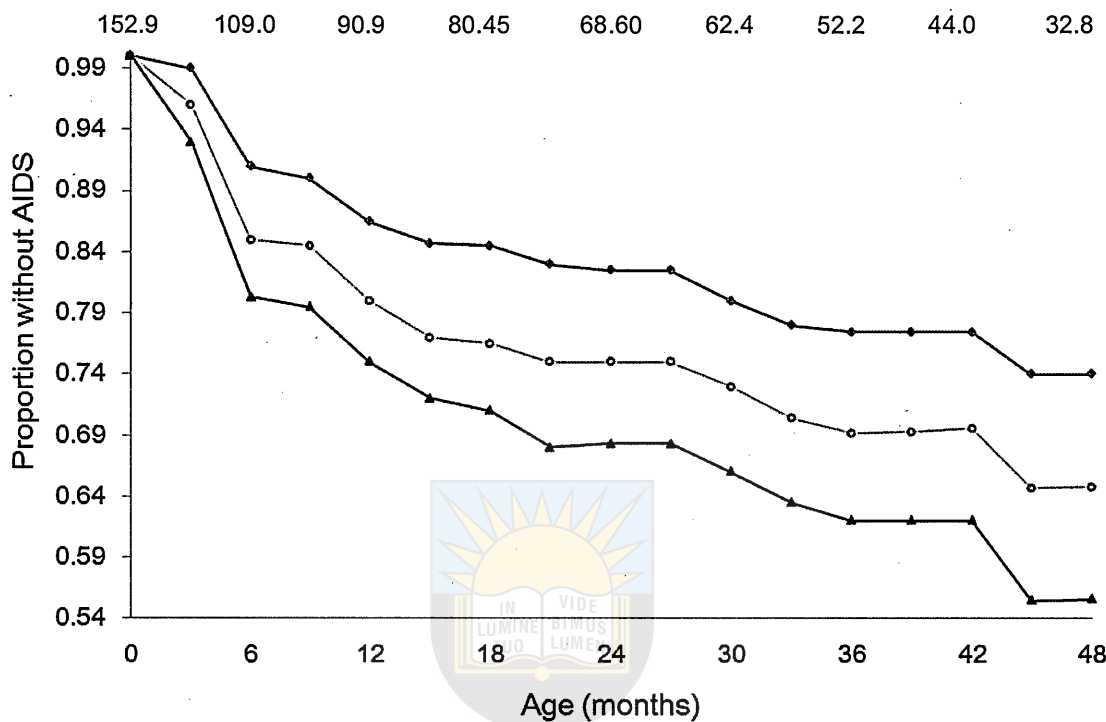


Figure 4.2 Estimates of the survivor functions (Logarithmic scale) for AIDS, with 5 % (Blue) and 95 % (Red) Confidence Limits. Numbers at the top of the figure represent the estimated number of children at risk of AIDS at 0, 6, 12,months.

Figure 4.3 compares the AIDS incubation periods between the two cohorts, allowing time-specific transmission rates.

This appears to indicate that rapid progression to AIDS was more common in the earlier period but that the hazard rates at older ages are fairly similar. An overall comparison provides moderate evidence of a difference between the AIDS incubation periods (log-rank $\chi^2 = 2.85$, $P = 0.09$). A more significant result is obtained by restricting the comparison to the first 6 months of life (log-rank $\chi^2 = 7.14$, $P = 0.008$), although a posteriori analyses should be interpreted cautiously. With time-specific AIDS incubation periods, the estimated trans-mission rates were 0.122 and 0.178.

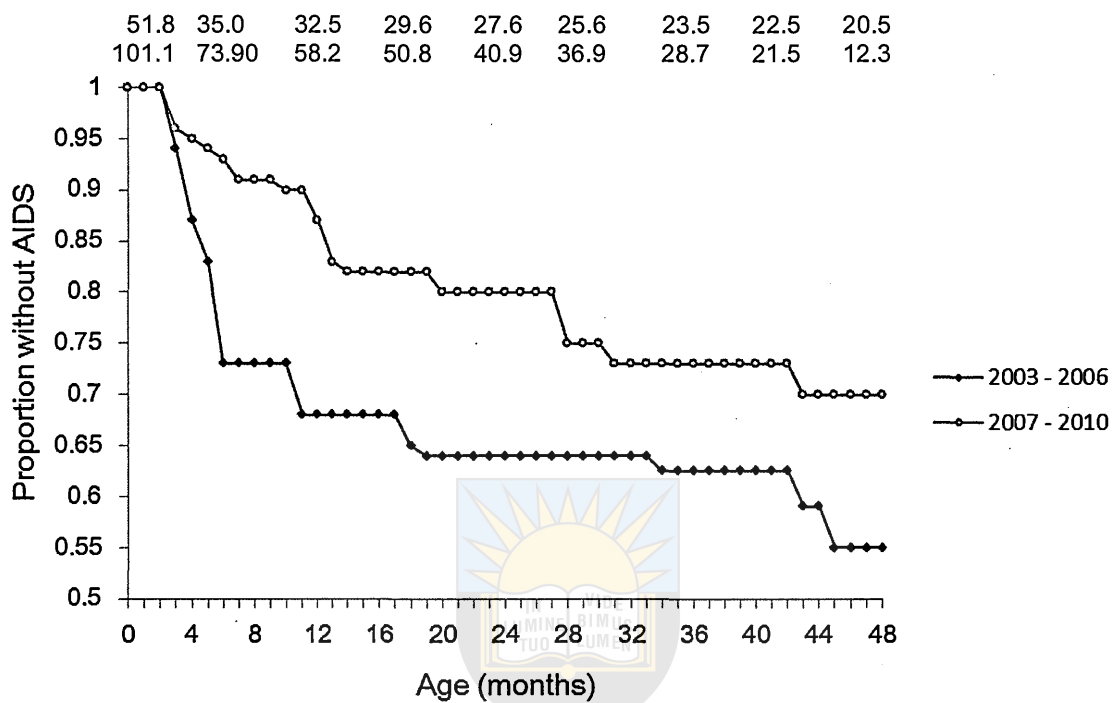


Figure 4.3 Estimates of the survival function (Logarithmic scale) for the AIDS among the 425 children born 2003-2006 and 575 children born 2007-2010. Numbers at the top of the figure represents the estimated number of children at risk of AIDS at 0, 6, 12,months.. The top line refers to 2003-2006 and the bottom line 2007-2010.

4.3 COMPARISON WITH CRUDE ESTIMATORS

In drawing comparisons with the crude estimators described earlier in Chapter three section 3.3, different critical ages at which anti-body persistence is regarded as proof of infection are examined (Table 4.1)

Table 4.1 Vertical Transmission Rate and Proportion of infected children who have Not developed AIDS by age 24 months.

		Including virus test data		Excluding virus test data	
		$\hat{\beta}$	$\hat{F}_{24}$	$\hat{\beta}$	$\hat{F}_{24}$
Maximum likelihood estimates		.156 (.016)	.748 (1.066)	.154 (.018)	.743 (1.068)
Crude estimates					
Critical age 15 months	Restricted*				
	No	.177 (.017)	.753 (1.067)	.153 (.016)	.714 (1.069)
	Yes	.172 (.017)	.749 (1.066)	.152 (.016)	.714 (1.069)
18 months	No	.176 (.017)	.749 (1.066)	.146 (.016)	.698 (1.071)
	Yes	.170 (.016)	.736 (1.066)	.146 (.016)	.698 (1.071)
21 months	No	.172 (.017)	.738 (1.068)	.136 (.016)	.681 (1.075)
	Yes	.161 (.016)	.725 (1.069)	.135 (.016)	.681 (1.075)
24 months	No	.170 (.016)	.735 (1.067)	.134 (.015)	.672 (1.076)
	Yes	.157 (.016)	.722 (1.064)	.136 (.014)	.672 (1.076)

*Analysis restricted to children born at least 15, 18, 21 or 24 months before date of data collection.

Note: Values in parentheses are $SE(\hat{\beta})$ and $\exp[SE(\log(\hat{F}_{24}))]$. An appropriate $(1 - \alpha)$ confidence for F_{24} is derived by dividing and multiplying the point estimator by $\exp[SE(\log(\hat{F}_{24}))]$ raised to the power $\phi^{-1}(\alpha/2)$.

Considering first the analysis based solely on AIDS diagnoses and antibody test results (i.e. excluding virus test results), the crude estimates of the vertical transmission rate are lower than the ML estimate (under the proposed model) of 0.154, the difference being more pronounced the older the critical age. The reason for this is that the crude

estimator ignores children of indeterminate infection status, who as they become older are more and more likely to be infected (Fig 4. 1). However, this can be compensated for by choosing a critical age that is too early, thereby including as infected a number of children who will eventually lose antibody. But it would be difficult to justify using a 15-month or 18-month threshold, as maternal anti-body was detected after these ages. Conversely, the crude estimates of the AIDS survivor function are too low (progression rate too high) because the number of children at risk of the event has been underestimated. Excluding children who had not reached the critical age for antibody persistence by the date of analysis did not alter the estimated AIDS incubation period (no child born within 24 months of the data of analysis had developed AIDS) and had only a marginal effect on the estimated transmission rate. As the standard method of analysis takes account of positive virus test results but ignores negative results, the inclusion of virus test data must increase estimates of the transmission rate and decrease estimates of the rate of progression to AIDS. However, the estimates of the transmission rate are now high and the rate of progression to AIDS is still overestimated, although the bias in the latter is now less severe. The exclusion of recently born children had a greater effect than in the analysis in which the virus test data were not considered, bringing estimates of the transmission rates closer to the ML estimates but increasing the bias of the estimated AIDS incubation period.

4.4 COMPUTING

For the main analysis in Section 4.2, the log-likelihood changed by less than 1.0×10^{-10} after 650 iterations done through SAS. No attempt was made to optimize the starting values, which were obtained simply by setting the conditional probability of infection for all censored children, $\hat{\tau}_k = 0.5$, and proceeding with the M step of the algorithm.



University of Fort Hare
Together in Excellence

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

5.1 INTRODUCTION

This chapter discusses the summary and conclusions of the study. It highlights some of the important results from the study and compares these results with other findings from other similar studies done elsewhere. Some suggestions and recommendations based on the results of the study are also given.

5.2 Summary and Conclusions

Considering some similar studies carried out elsewhere, the European Collaborative Study has been notable for its low rate of vertical transmission. This analysis however, provides evidence of a considerable rise in the rate of vertical transmission, which is higher than some recent estimates from other studies in the developed world. Since the risk of vertical transmission depends on the degree of maternal immune suppression (European Collaborative Study 1992), it would be surprising if the transmission rates are not higher especially in South Africa where the HIV infection rate among women is still high. These results based on data from Cecilia Makiwane Provincial Hospital in the Eastern Cape serving both rural, urban and peri-urban communities around it may be generalizable for similar areas in the country with high incidences of HIV infection rate among women although there may be slight variations as some Hospitals may have different types of data depending on the rate of infections where the Hospitals are located. Another reason is that there are many pregnant women with an unrecognized

HIV infection and their children as well, who were not part of the study. However, the study gives a fairly accurate estimations of vertical transmission rate of HIV and incubation period in similar areas in the country. In some instances of large cohort studies, an insufficient number of children have been observed to progress to AIDS to allow reliable characterization of the incubation period. Although comparatively precise estimates have been derived (at least for the shape of the distribution) using data from AIDS surveillance systems, these data have a number of limitations (DeGruttola et al. 1992). Among these are the periodic revisions in the pediatric AIDS case definition, and comparisons between studies must take account of differences in definitions. For instance, of the 41 AIDS cases according to the criteria used in European Collaborative Study, 3 died before fulfilling the 1987 CDC definition and presumably would not have been noticed. Conversely, this analysis excluded children with asymptomatic lymphoid interstitial pneumonitis on the grounds that AIDS is supposed to reflect a clinically significant manifestation of HIV infection. This condition was responsible for about 26% of initial pediatric AIDS-defining diagnoses in the United Kingdom up to April 1995 (Fiona Holland, personal communication). The finding of an inflexion point in the AIDS hazard rate at around age 6 months is consistent with other reports (Auger et al. 1988; DeGruttola et al. 1992), although this effect was much less striking among children born in the later part of the study. The reasons for this are unclear, although one possible explanation is an improvement in the clinical management of young infants. Analyses of routine AIDS data have generally assumed that the incubation period is stationary. The analysis in this study points to a need for

critical examination of this assumption, although with this type of data it is only possible to compare incubation periods conditional on developing AIDS by a given age (DeGruttola et al. 1992).

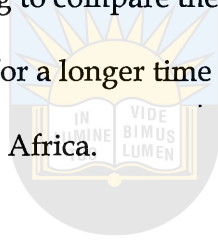
The algorithm is not guaranteed to converge to estimates that are biologically reasonable. For example, unsatisfactory estimates of the antibody loss distribution may be obtained if either the number of seronegative children is small relative to the number of censored children or there are children who are antibody positive beyond, say, age 2 years who have repeatedly tested virus negative. This gives rise to a "conflict" between the antibody test results and virus test results, and if the latter predominate, then the antibody loss distribution may have a too-long tail. Although these problems are unlikely when modeling extensive data, modifications to the method may need to be considered for smaller data sets, such as estimating the antibody loss distribution parametrically or imposing a smoothness penalty. With the European Collaborative Study data, fitting a normal distribution with a power transformation of age made virtually no difference to estimates of the transmission rate or of the incubation period. In common with other estimators, the method used in this study assumes that the censoring mechanism is independent of infection status, time to AIDS (if infected), and time to antibody loss (if uninfected). But it is not possible to directly examine these assumptions (Lagakos 1979). Children can be censored either because they were born shortly before the date of analysis or because they were genuinely lost to follow-up. For

the former group the assumptions would appear reasonable, but for the latter group there is concern that ill children, including a disproportionate number of HIV-infected children, may be more likely to remain in contact with the study pediatrician. If this is true, then the vertical transmission rate and the rate of progression to AIDS would both be overestimated.

Survival analysis was used for virus culture data in the ACTG-076 trial (Connor et al. 1994). This implicitly assumes a transition age for each infected child, before which he or she always tests negative and after which he or she always tests positive. In contrast, our model assumes that each virus test has a fixed sensitivity that is independent of age (after the neonatal period). This is obviously a simplification, but it would be extremely difficult to model all the technical facets of the assays that influence sensitivity. A more productive development would be to include test results on neonates and to model the effects of age. It is also plausible that some infected children, averaging over a series of virus test assessments, are more likely to test positive than others. Failure to model "frailty" would result in attaching too much diagnostic significance to multiple negative test results, thereby underestimating the transmission rate. Furthermore, estimates of marginal sensitivity would be affected if virus tests were selectively performed on infected children who were more, or less, likely to test positive. We found differences between the crude estimates and the ML estimates under the proposed model that were of moderate size (Sec. 4.3). This partly reflects the long duration of the European

Collaborative Study (10 years), and this allowed the infection status of a large majority of the enrolled children to be established. The method described in this study should be particularly useful in studies of shorter duration, particularly in the efficient analysis (possibly interim) of randomized controlled trials to reduce vertical transmission.

Finally, although the method has been developed in the context of HIV transmission rate, it may also be applicable to other vertically transmissible infections, such as hepatitis C virus. It would be interesting to compare the results of this study if it was carried out on a larger population and for a longer time in a developing country with a high HIV prevalence rate such as South Africa.



University of Fort Hare
Together in Excellence

5.3 RECOMMENDATIONS

- i) The study was carried out in the Eastern Cape Province of South Africa, one of the nine Provinces in the Country with a very high HIV prevalence rate. It would be interesting if a similar study was done regionally in all the nine provinces covering the whole country. This would give a complete picture of the impact of both the rate of vertical transmission of HIV and incubation period not only by province but also nationally among the infected children.
- ii) Such a study would be of immense benefit to both the Provincial and national governments. This is because it would provide the Department of Health with interventions to help in reducing vertical transmission rate of HIV infection among

children as well as giving the already infected children an opportunity to live longer.

iii) The method used in this study has been developed in the context of vertical transmission of HIV infection in children but with slight modifications, it can also be useful in studies in the efficient analysis of randomized controlled trials as well as in application to reduce other vertically transmissible infections, such as hepatitis.

iv) The study also indicates that more women should be encouraged and incentivized to attend antenatal clinics during their early pregnancy period and be tested so that infected mothers can start early ART treatment to reduce the rate of mother-to-child transmission.

REFERENCES

- [1] Aalen O. O., Farewell V. T., Angelis, D. D., and Day, N. E. (1994), The use of human Immunodeficiency Virus diagnosis information in Monitoring AIDS epidemic. *Journal of Royal Statistical Society, Series A*, 157(1):3-16.
- [2] Abramowitz, M., and Stegun, I. A. (1964), Handbook of Mathematical functions with Formulas, Graphs and Mathematical Tables. *Washington DC: National Bureau of Standards*.
- [3] American congress committee of obstetricians and Gynecologists (ACOG). (2010), Scheduled Caesarean Delivery and the Prevention of Vertical Transmission of HIV Infection, 234.
- [4] Ibid
- [5] AIDS Information, HIV and Women. (2013), Preventing Mother-to-child Transmission of HIV during Childbirth, Education material, Fact sheet ,2013.
- [6] Alkema, L., Raftery, A.E., and Brown, T. (2008), Bayesian melding for estimating certainty in national HIV prevalence estimates. *Sexually Transmitted Infections*, 84(1): i11- i16.
- [7] Auger et al. (1988), Incubation Periods for Paediatric AIDS Patients. *Nature*, 336: 575-577.
- [8] Ibid

- [9] Baachetti, P., Segal, M. R., and Jewell, N. P. (1993), Backcalculation of HIV Infection Rates. *Statistical Science*, 8(2):82-101.
- [10] Bacchetti, P., and Jewell, N. P. (1991), Nonparametric estimation of the incubation period of AIDS based on a prevalence cohort with unknown infection times. *Biometrics*, 47:947- 960.
- [11] Baachetti, P. (1994), The impact of lengthening AIDS reporting delays and uncertainty about underreporting on incidence trends and projections. *Journal of AIDS*, 7:860-865.
- [12] Becker, N.G., and Marschner, I. C. (1993), A method for estimating the age-specific relative risk of HIV infection from AIDS incident data. *Biometrika*, 80:165-178.
- [13] Becker, N.G., Watson, L. F., and Carlin, J. B. (1991), The method of nonparametric back-projection and its application to AIDS data. *Statistics in Medicine*, 10:1527-1542.
- [14] Bellocco, R., and Marschner, I. C. (2000), Joint analysis of HIV and AIDS surveillance data in back-calculation . *Statistics in Medicine*, 19:297-311.
- [15] Biggar, J. (1990), AIDS incubation in 1890 HIV seroconverts from different exposure group. *AIDS*, 4:1059-1066.
- [16] Blanche et al., and French Pediatric HIV Infection Study Group. (1994), "Relation of the Course of HIV Infection in Children to the Severity of the Disease in Their Mothers at Delivery." *New England Journal of Medicine*, 330, 808-812.

- [17] Blythe, S. P., and Anderson, R. M. (1988), Distributed incubation and infection periods in models of the transmission dynamics of human immunodeficiency virus (HIV) . *IMA Journal of Mathematics Applied in Medicine and Biology*, 5:1-19.
- [18] Boldson, J. L. et al. (1988), Incubation time distribution and the Danish AIDS data. *Journal of Royal Statistical Society A*, 151:42-43.
- [19] Bonnel, R. (2000), 'HIV/ AIDS and Economic Growth: A Global Perspective'. *South African Journal of Economics*, 65(5).
- [20] Boylan, L., and Stein, Z. A. (1991), "The Epidemiology of HIV-1 Infection in Children and Their Mothers-Vertical Transmission," *Epidemiologic Reviews*, 13, 143-177.
- [21] Breslow, N. E. (1975), "Analysis of Survival Data Under the Proportional Hazards Model," *International Statistical Review*, 43, 45-48.
- [22] Brookmeyer, R., and Gail M. H. (1986), Minimum Size of the acquired Immunodeficiency Syndrome (AIDS) epidemic in the United States . *The Lancet*, ii:1320-1322.
- [23] Brookmeyer, R., and Gail M. H. (1988), A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of American Statistical Association*, 83(402):301-308.
- [24] Brookmeyer, R., and Goedert, J. J. (1989), Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*, 45:325-335.

- [25] Brookmeyer, R., and Gail, M. H. (1994), *AIDS Epidemiology. A Quantitative Approach*. Oxford University Press.
- [26] Brookmeyer, R., and Damiano, A. (1989), Statistical Methods for short-term projections of AIDS incidence . *Statistics in Medicine*, 8:23-34.
- [27] Brookmeyer, R., and Liao, J. (1990), The analysis of delays in disease reporting: Methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, 132:355-365.
- [28] Brookmeyer, R., and Gail, M. H. (1987), Biases in prevalence cohorts. *Biometrics*, 43:739-749.
- [29] Brookmeyer, R., and Liao, J. (1990), Statistical modelling of AIDS epidemic for forecasting healthcare needs. *Biometry*, 46:1151-1163.
- [30] Brookmeyer, R. (1991), Reconstruction and Future trends of the AIDS epidemic in the United States . *Science*, 253:37-42.
- [31] Brown, T., Salomon J.A, Alkema, L., Raftery, A.E., and Gouws, E. (2008), Progress and challenges in modelling country-level HIV/ AIDS epidemics: the UNAIDS Estimation and Projection Packages. *Sexually Transmitted Infections*, 84(1): i5- i10.
- [32] Bryson, Y. J. (1995), "HIV Clearance in Infants-A Continuing Saga," *AIDS*, 9, 1373-1375.
- [33] Caldwell et al. (1998), Projection of AIDS and HIV incidence among children born infected with HIV *Statistics in Medicine*, 17: 169-181.

- [34] Centers for Disease Control. (1990), Estimates of HIV prevalence and projected AIDS cases: Summary of a workshop Oct 31 - Nov 1, 1989. *Morbidity and Mortality Weekly Report*, 39:110-119.
- [35] Centers for Disease Control (1987a), "Classification System for Human Immunodeficiency Virus (HIV) Infection in Children Under 13 Years of Age," *Morbidity and Mortality Weekly Report*, 36, 225-235.
- [36] Centers for Disease Control (1987b), "Revision of the CDC Surveillance Case Definition for Acquired Immune Deficiency Syndrome," *Morbidity and Mortality Weekly Report*, 36, 3S-15S.
- [37] Chau, P. H., Yip P. S.F., and Cui J. S. (2003), Reconstructing the incidence of Human immunodeficiency virus (HIV) in Hong Kong by using data from positive tests and diagnoses of acquired immune deficiency syndrome. *Applied Statistics*, 52(2).
- [38] Chau, P. H., and Yip P. S.F. (2004), Non-parametric back-projection of HIV positive tests using multinomial and poisson settings. *Applied Statistics*, 31(5).
- [40] Consensus Workshop (1992), "Early Diagnosis of HIV Infection in Infants: Report of a Consensus Workshop Held in Siena, Italy, January 1992," *Journal of the Acquired Immune Deficiency Syndrome*, 5, 1169-1178.
- [41] Conner et al. (1994), Reduction of Maternal-Infant Transmission of HIV Type I with Zidovudine Treatment, AIDS Clinical Trials Group Protocol 076 Study Group. *New England Journal of Medicine*, 331 : 1173- 1180.

- [42] Ibid
- [43] Cui, J. S., and Becker, N. G. (2000), Estimating HIV incidence using dates of both HIV and AIDS diagnosis . *Statistics in Medicine*, 19:1165-1177.
- [44] Ibid
- [45] Dabis et al. and Working Group on Mother-to-Child Transmission of HIV (1993), "Estimating the Rate of Mother-to-Child Transmission of HIV. Report of a Workshop on Methodological Issues, Ghent (Belgium), 17-20 February 1992," *AIDS*, 7, 1139-1148.
- [46] Ibid
- [47] Darby, S. C., Doll, R., Thakrar, B., and et al. (1990), Time from infection with HIV to onset of AIDS in patients with hemophilia in the UK . *Statistics in Medicine*, 9:681-689.
- [48] Day, N. E., Gore, S. M., McGee, M. A., and South, M. (1989), Prediction of the AIDS epidemic in the UK: the use of the back projection methods. *Philosophical Transactions Royal Society, London B*, 325:123-134.
- [49] DeGruttola, V., and Lagakos, S.W. (1989), The value of AIDS incidence data in assessing the spread of HIV infection . *Statistics in Medicine*, 8:35-44.
- [50] DeGruttola, V., and Lagakos, S. W. (1989), Analysis of doubly censored survival data, with application to AIDS . *Biometrics*, 45:1-11.

- [51] DeGruttula V., Tu, X. M., and Pagano, M. (1992), Pediatric AIDS in New York City : Estimating the distribution of Infection, latency and reporting delays and projecting future incidence. *Journal of American Statistical Association*, 87:633-640.
- [52] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- [53] Downs, A. M., Salamina, G., and Ancelle-Park, R. A. (1995), "Incubation Period of Vertically Acquired AIDS in Europe Prior to the Widespread Use of Prophylactic Therapies," *Journal of Acquired Immune Deficiency Syndromes and Human Retro-virology*, 9, 297-304.
- [54] Department of Health. (2006), South African National HIV and Syphilis prevalence survey 2006, www.doh.gov.za/docs/reports/2005/hiv.pdf
- [55] Department of Health. (2009), South African National Antenatal Sentinel HIV and Syphilis Prevalence Survey 2008, www.doh.gov.za/docs/reports/2008/hiv.pdf
- [56] Deuffic, B. S., and Costagliola, D. (2006), Including pre-AIDS mortality in back-calculation model to estimate HIV prevalence in France, 2000. *European Journal of Epidemiology*, 21:389-396.
- [57] De Ruiter et al. (2008), *British HIV Association and Children's HIV Association guidelines for the management of HIV infection in pregnant women*. HIV Med 9: 452-502, 2008, www.bhiva.org

- [58] Dunn et al. (1995), "The Sensitivity of HIV- 1 DNA Polymerase Chain Reaction in the Neonatal Period and the Relative Contributions of Intra-Uterine and Intra-Partum Transmission," *AIDS*, 9, F7- 11.
- [59] Eyster, M. E., Gail, M. H., Ballard, J. O., and et al. (1987), Natural history of human immunodeficiency virus in haemophiliacs : effects of T-cell subsets, platelet counts and age. . *Annals of Internal Medicine*, 107:1-6.
- [60] European Collaborative Study (1994), "Natural History of Vertically Acquired Human immuno-deficiency Virus-1 Infection," *Paediatrics*, 94, 815-819.
- [61] European Collaborative Study (1991), "Children Born to Women With HIV-1 Infection: Natural History and Risk of Transmission," *Lancet*, 337, 253-260.
- [62] Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval- Censored Failure Time Data," *Biometrics*, 42, 845-854.
- [63] Fowler, T.B., Gouws, E., and Mishra, V. (2008), Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalized epidemics: implications for calibrating surveillance data. *Sexually Transmitted Infections*, 84(1): i17- i23.
- [64] Freund, H. P., and Book, D. L. (1990), Determination of the spread of HIV from the AIDS incidence history. *Math. Biosciences*, 98:227-241.
- [65] Gaffeo, E. (2003), The Economics of HIV/ AIDS: A survey. *Development Policy Review*,(1) 27-49.

- [66] Gail, M. H., and Rosenberg, P. S. (1992), Perspectives on using backcalculation to estimate HIV prevalence and project AIDS incidence. In Jewell N P, Dietz K, and Farewell V T eds., *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston.
- [67] Gail, M. H., Rosenberg, P. S., and Goedert, J. J. (1990), Therapy may explain recent deficits in AIDS incidence . *Journal of AIDS*, 3:296-306.
- [68] Gelman, A., and Rubin, D.B.(1992), Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457-511.
- [69] Gelman, A., and Rubin, D.B. (1992), *A single series from the Gibbs sampler provides a false sense of security*. In Bayesian Statistics : Bernado, J.M., Berger, J.O., Dawid, A.P., and Smith, eds. Oxford University Press.
- [70] Ghys, P. D., Brown, T., Grassly, N.C., Garnett, G., Stanecki, K.A., Stover, J., and Walker, N. (2004), The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics. *Sexually Transmitted Infections*, 80(1): i5
- [71] Ghys, P.D., Walker, N., McFarland, W., Miller, R., and Garnett, G.P. (2008), Improved data, methods and tools for the 2007 HIV and AIDS estimates and projections. *Sexually Transmitted Infections*, 84(1): i1- i4.
- [72] Gigli, A., and Verdecchia, A. (2000), Uncertainty of AIDS incubation time and its effects on backcalculation estimates. *Statistics in Medicine*, 19:175-189.
- [73] Global Health Policy. (2014), The global HIV / AIDS epidemic : Kaiser Family Foundation ; *KFF.org*

- [74] Goedert, J. J., Biggar, R. J., Weiss S H, and et al. (1986), Three year incidence of AIDS in five cohorts of HTLV-III-infected risk group members . *Science*, 231:992-995.
- [75] Goedert, J. J., Kessler, C. M., Aledort L M, and et al. (1989), A perspective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia. *New England Journal of Medicine*, 321:1141-1148.
- [76] Govender et al. (2004) , Quantifying effects of illness and death on Education at school level: Implications for HIV/ AIDS response , Johannesburg: Health and development in Africa.
- [77] Haris, J. E. (1987), Delay in reporting acquired immunodeficiency syndrome (AIDS). *Working Paper No. 2278. National Bureau of Economic Research*.
- [78] Haris, J. E. (1990), Reporting delays and the incidence of AIDS. *Journal of American Statistical Association*, 85:915-924.
- [79] Healy, M. J. R., and Tillett, H. E., (1988), Short-term Extrapolation of AIDS Epidemic. *Journal of Royal Statistical Association A*, 151(1):50-61.
- [80] Hessol et al. (1987), The natural history of human immunodeficiency virus infection in cohorts of homosexual and bisexual men. *Third international Conference on AIDS, Washington DC, Department of Health and Human Services and the World Health Organisation*.
- [81] Hessol, N. A., Lifson, O'Malley P. M., and et al. (1989), Prevalence, incidence and progression of human immunodeficiency virus infection in homosexual

- and bisexual men in hepatitis B vaccine trials 1978-1988. *American Journal of Epidemiology*, 130:1167-1175.
- [82] Hollister, A.I.T., and Deanery, S. (2010), Questioning AIDS. Under-reporting of HIC in Rural South Africa. *Comeytrowe Centre, Tauton, US*.
- [83] Hosmer, D. W. (1973), "A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Sample," *Biometrics*, 29, 761-770.
- [84] Jewel N. P. (1990), Some statistical issues in studies of epidemiology of AIDS. *Statistics in Medicine*, 9:1387-1416.
- [85] Kalbfleisch, J. D., and Lawless, J. F. (1989), Inference based on retrospective ascertainment: An analysis of data on transfusion related AIDS. *Journal of American Statistical Association*, 84:360-372.
- [86] Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457-481.
- [87] Kou, J. M., Taylor, J. M. G., and Detels, R. (1991), Estimating the AIDS incubation period from a prevalence cohort. *American Journal of Epidemiology*, 133:1050-1057.
- [88] Lagakos, S. W. (1979), "General Right Censoring and Its Impact on the Analysis of Survival Data," *Biometrics*, 35, 139-156.
- [89] Laska, E. M., and Meisner, M. J. (1992), "Nonparametric Estimation and Testing in a Cure Model," *Biometrics*, 48, 1223-1234.

- [90] Lawless, J. and Sun, J. A. (1992), comprehensive Back-calculation framework for the estimation and prediction of cases. In: AIDS Epidemiology: Methodological issues, eds., Jewel N.P., Dietz K., and Farewell V. T. *Birkhauser Basel*, pages 81-104.
- [91] Longini, I. M., Clark W. S., Byers R. H., and Ward J. W. and et al (1989), Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, 8:831- 843.
- [92] Longini, I. M. Jr. (1990), Modelling the decline of CD4+ T-lymphocyte counts in HIV infected individuals. . *Journal of Acquired Immune Deficiency Syndromes*, 3:930- 931.
- [93] Love life. (2001), The National Survey of South African Youth, 2001: Africa Strategic Research Corporation/Kaiser Family Foundation.
- [94] Lui, K. J., Lawrence, D.N., Morgan, W.M., Peterman, T. A., Haverkos, H. W., and Bregman, D. J. (1986), Model-based approach for estimating the mean incubation period of transfusion associated acquired immunodeficiency syndrome. *Proc. atl. Acad. Sci. USA: Applied Mathematics*, 83:3051-3055.
- [95] Lui, K. J., Darrow, W. W., and Rutherford, G. W. (1988), Model-based estimate of the mean incubation period for AIDS in homosexual men. *Science*, 240:1333-1335.
- [96] Map of South Africa. (2014)
www.saexplorer.co.za/south-africa/map.asp

- [97] Mariotto, A. B., and Verdecchia, A. (2000), Using AIDS data to reconstruct HIV/ AIDS epidemics. *Statistics in Medicine*, 19:161-174.
- [98] Marschner, I. C., and Bosch, R. J. (1998), Flexible assessment of trend in age-specific HIV incidence using two-dimensional penalized likelihood . *Statistics in Medicine*, 17:1017-1031.
- [99] Matheson et al., and New York City Perinatal HIV Transmission Collaborative Study Group (1995), "Comparison of Methods of Estimating the Mother-to-Child Transmission Rate of Human Immunodeficiency Virus Type 1 (HIV-1)," *American Journal of Epidemiology*, 142, 714-718.
- [100] MaWhinney, S., and Pagano, M. (1994), "Distribution of the Latency Period for Perinatally Acquired AIDS," *Statistics in Medicine*, 13, 2031- 2042.
- [101] McEvoy, M., and Tillett, H. E. (1985), Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the UK. *Lancet*, ii: 541-542.
- [102] McIntosh et al. (1994), "Blood Culture in the First 6 Months of Life for the Diagnosis of Vertically Transmitted Human Immunodeficiency Virus Infection," *Journal of Infectious Diseases*, 170, 996-1000.
- [103] Medley, G. F., Anderson, R. M., Cox, D. R., and Billard, L. (1987), Incubation period of AIDS patients infected via blood transfusion. *Nature*, 328:719-721.
- [104] Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899-909.

- [105] Miller, R. G. (1981), *Survival Analysis*, New York: John Wiley.
- [106] Moodley D. (1997), The relationship between maternal-infant antibody levels and vertical transmission of HIV-1 infection. *J Trop Paediatric Journal*, 43 :75-79.
- [107] Montana, L.S., Mishra, V., and Hong, R. (2008), Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sexually Transmitted Infections*, 84(1): i78- i84.
- [108] Morgan, W. M., and Curran, J. W. (1986), Acquired immunodeficiency syndrome: current and future trends. *Public Health Report*, 101:459- 465.
- [109] Munoz A., Wang M. C., Bass, S., Taylor, J. M. G., Kingsley, L.A., Chmiel, J.S., and Polk, B. F. (1989), Acquired immunodeficiency syndrome (AIDS)-free time after Human Immuno deficiency Virus Type 1 (HIV-1) seroconversion in homosexual men . *Journal of American Epidemiology*, 130:530-539.
- [110] Nastouli, E et al. (2007), *False-positive HIV antibody results with ultrasensitive serological assays in HIV uninfected infants born to mothers with HIV*. *AIDS* 21 : 1222-1223.
- [111] National Institute of Allergy and Infectious Diseases (NIH). (2008), *HIV infection in Infants and Children*. US, Sept 2008.
- [112] National Department of Health. (2008), South African National Antenatal Sentinel HIV and Syphilis Prevalence Survey, 2008, Pretoria, SA

- [113] National Department of Health. (2011), South African National HIV, prevalence, incidence and Behaviour survey 2012, Pretoria, SA.
- [114] National Department of Health. (2013), The South African National Antenatal Sentinel HIV and Prevalence Survey 2012, Pretoria, SA.
- [115] Newell et al. (2004), Mortality of Infected and Un-infected Infants born to HIV-infected Mothers in Africa: A pooled Study. *Lancet* 2004, 364 : 1236-43.
- [116] Nicolay, N. (2008), AIDS and Demographic change.
www.actuarialsociety.org.za
- [117] Nojilana B. et al. (2009), Quality of Cause of Death Certification at an academic hospital; Cape Town, South Africa. *South African Medical Journal*, vol 99 (9).
- [118] Peto, R., and Peto, J. (1972), "Asymptotically Efficient Rank-Invariant Test Procedures" , *Journal of the Royal Statistical Society*, Ser. A, 135, 185-206.
- [119] Rai, S. N., and Matthews, D. E. (1993), "Improving the EM Algorithm," *Biometrics*, 49,587-591.
- [120] Rees, M. (1987), The sombre view of AIDS. *Nature*, 326:343-345.
- [121] Rosenberg, P. S., and Gail, M. H. (1991), Backcalculation of Flexible Linear Models of the HIV Infection Curve . *Applied Statistics*, 40 : 269-282.
- [122] Rosenberg, P. S., Gail, M. H., and Pee, D. (1991), Mean square error of estimates of HIV prevalence and short-term AIDS projections derived from backcalculation. *Statistics in Medicine*, 10:1167-1180.

- [123] Rosenberg, P. S., and Gail, M. H. (1990), Uncertainty in estimates of HIV prevalence derived backcalculation. *Annals of Epidemiology*, 1:105-115.
- [124] Rosenberg, P. S., Biggar, R. J., Goedert, J. J., and Gail, M. H. (1991), Backcalculation of the number with human immunodeficiency virus infection in the United States. *American Journal of Epidemiology*, 133:276-285.
- [125] Rosenberg, P. S. (1990), A simple correction of AIDS surveillance data for reporting delays. *Journal of AIDS*, 3:49-54.
- [126] Rosenberg, P. S., Gail, M. H., and Carrol, R. J. (1991), Estimating HIV prevalence and AIDS incidence in the United States: a model that accounts for therapy and changes in the Surveillance definition of AIDS . *Statistics in Medicine*, 11:1633-1655.
- [127] Rosenberg, P. S., Gail, M.H., Schrager, L. K., Vermund, S. H., and et al. (1991), National AIDS incidence trend and the extent of zidovudine therapy in selected demographic and transmission groups . *Journal of AIDS*, 4:392-401.
- [128] Rosenberg, P. S. (1994), Back-calculation models of age-specific HIV incidence rates. *Statistics in Medicine*, 13:1975-1990.
- [129] Saphonn, V., Hor, L.B., Ly, S.P., and Chhuon, S. (2002), How well do antenatal clinic (ANC) attendees represent the general population? *International Journal of Epidemiology*, 31(2):449 - 455.

- [130] Shisana, O., Rehle, T., Simbayi, L.C., Zuma, K., Jooste, S., Zungu N., Labadarios, D., Onoya, D. et al. (2014). *South African National HIV Prevalence, Incidence and Behaviour Survey, 2012*. Cape Town, HSRC Press.
- [131] Simpson, B. J., and Andiman, W. A. (1994), "Difficulties in Assigning HumanImmuno deficiency Virus-I Infection and Seroreversion Status in a Cohort of HIV-Exposed Children Using Serologic Criteria Established by the Centers for Disease Control and Prevention," *Pediatrics*, 93, 840- 842.
- [132] Simon et al. (2006), HIV/ AIDS epidemiology , pathogenesis, prevention and treatment. *Lancet* ; 368 : 489-504
- [133] Snowflake. (2007), The Newborn Baby of an HIV positive woman, [www.Yahoo](http://www.Yahoo.com) site, Jan 2007.
- [134] South Africa, Country profile. (2014).
www.nationsonline.org/oneworld/south_africa.htm
- [135] South Africa Health Review/Surevey. (2008).
www.hst.org.za/publications/841
- [136] South Africa HIV & AIDS Survey. (2008), Stats SA.
www.avert.org/south-africa-hiv-aids-statistics.htm
- [137] South Africa HIV & AIDS Statistics.(2012), Stats SA.
www.avert.org/south-africa-hiv-aids-statistics.htm

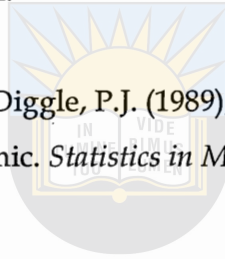
- [138] South African National Health Survey. (2006), Information sheet, SA Medical Research council.
www.mrc.ac.za.bod/edudatasuplmt.pdf
- [139] South African National HIV, Behaviour and Health survey. (2012), HSRC, Pretoria, SA.
www.hsrc.ac.za/en/research_areas-HAST
- [140] Stover, J., Brown, T., and Marston, M. (2012), Updates to the Spectrum/Estimation and Projection Package (EPP) Model to estimate HIV trends for Adults and Children. *Sexually Transmitted Infections*, 88(2): i11- i16.
- [141] Tango, T. (1989), Estimation of Haemophilia-associated AIDS incidence in Japan using individual dates of diagnosis. *Statistics in Medicine*, 8:1509-1514.
- [142] Taylor, J. M. G. (1989), Models for HIV infection and AIDS Epidemic in the United States. *Statistics in Medicine*, 8:45-58.
- [143] Turnbull, B. W. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society, Ser. B*, 38, 290-295.
- [144] UNAIDS.(2001), Factsheet: The global HIV and AIDS Epidemic.
www.umich.edu/globalfactsheets.pdf
- [145] UNAIDS. (2001), Epidemiology Reference Group. Recommended methodology for the estimation and projection of HIV prevalence and AIDS mortality in the short-term. *Meeting held at La Mainaz Hotel, Gex France, 2001.*

- [146] UNAIDS. (2002), Report on AIDS Epidemic update, 2002.
www.unaids.org/update2002_en.pdf.
- [147] UNAIDS. (2003), Report on AIDS Epidemic update, 2002.
www.unaids.org/pub/report/2003_epiupdate_en.pdf.
- [148] UNAIDS. (2004), Report on the Global AIDS Epidemic update, 2004.
www.unaids.org/pub/report/2004_epiupdate_en.pdf.
- [149] UNAIDS. (2005), Reports on the Global AIDS Epidemic update, 2005.
www.unaids.org/pub/publications/IRC_pub061_epi_updates2005.en.pdf
- [150] UNAIDS. (2006), Report on Global AIDS Epidemic update, 2006.
www.unaids.org/pub/report/2006_gr_en.pdf.
- [151] UNAIDS. (2006), Epidemiology Reference Group. Technical report and recommendations: Improving parameter estimation, projection methods, uncertainty estimation and epidemic classification. *Meeting held in Prague, Czech Republic*, 2006.
- [152] UNAIDS/WHO. (2007), Annual Report on Global AIDS Epidemic update, 2007.
www.unaids.org/pub/report/2007_epiupdate_en.pdf.
- [153] UNAIDS. (2008), Annual Report on Global AIDS Epidemic update, 2008.
www.unaids.org/globalreport/2008/jc150_2008globalreport

- [154] UNAIDS. (2008), Epidemiology Reference Group. Technical report and recommendations: Improving EPP and spectrum estimation tools for the 2008-9 round of national estimates with specific attention to prevalence fits, uncertainty, etc. *Meeting held in London, 2008.*
- [155] UNAIDS/UNICEF. (2009), Report on Epidemic update on HIV prevalence in South Africa.
<http://www.unaids.org/HIVdata/Epiupdate.pdf>, 2009.
- [156] UNAIDS. (2009), AIDS Epidemic update, 2009.
www.unaids.org/Report,2009
- [157] UNAIDS/UNICEF. (2010) , *Report on child and maternal health in South Africa*, Annual Report, 2010. www.unicef.org/southafrica
- [158] UNAIDS. (2010), Global report on AIDS Epidemic, 2010.
www.unaids.org/globalreport/2010/southafrica
- [159] UNAIDS. (2011), World AIDS Day Report, 2011.
www.unaids/worldaidsday-report_2011_en.pdf/Southafrica
- [160] UNAIDS/WHO. (2011), Progress report, Global HIV/ AIDS response, 2011.
www.unaids.org/2011 and www.who.int/hiv/pub/report/2011/en
- [161] UNAIDS. (2012), Global report on AIDS Epidemic, 2012.
www.unaids.org/globalreport/2012/southafrica
- [162] UNAIDS/WHO. (2013), Report on the Global AIDS Epidemic, 2013.
www.unaids.org/2013/global_report,2013

- [163] UNAIDS. (2013), Factsheet on AIDS, South Africa, 2013.
www.unaids.org/countries/2013/southafrica
- [164] UNAIDS. (2014), Global report on AIDS Epidemic, 2014.
www.unaids.org/en/resources/2014/southafrica
- [165] UNAIDS. (2014), The Gap Report, Factsheet 2014.
www.unaids.org/2014/uniads_Gap_report
- [166] Velayati, A., Bakayev, V., Bahadori, M., Tabatabaei, S., Alaei, A., Farahboud, A., and Masjedi, M. (2007), Religious and Cultural Traits in HIV/ AIDS Epidemics in Sub-Saharan Africa. *Archive of Iranian Medicine*, 10(4):486 – 497.
- [167] Wai yaun, T. (2000), *Stochastic Modeling of AIDS epidemiology and HIV pathogenesis*. World Scientific Publishing Co. Pte. Ltd.
- [168] Whiteside, A. (2002), Poverty and HIV/ AIDS in Africa. *Third World Quarterly*, 23(2):313 - 332.
- [169] WHO. (2003), Estimation/ projection of HIV infection and AIDS cases/ deaths. *HIV/AIDS in Asia and Pacific Region. Annex 4*, : 102-109.
- [170] WHO. (2009), World Health Statistics, 2009.
www.who.int/world_health_statistics/en
- [171] WHO (2012) Treatment of children living with HIV. 2012.
<http://www.who.int/hiv/topics/paediatric/en/index.html>.

- [172] Wilkie, A. D. (1989), Population projections for AIDS using an Actuarial model. *Phil. Trans. Royal Soc., London B*, 625:61-74.
- [173] Williams, B., Gouws, E., Wilkinson, D., and Karim, S. K. (2001), Estimating HIV incidence rates from age prevalence data in epidemic situation. *Statistics in Medicine*, 20:2003-2016.
- [174] Yudkin, P. et al. (2009), Deaths Caused by HIV Disease Under-reporting in South.AIDS, (23)12 : 1600-12.
- [175] Zeger, S. L., See, L. C., and Diggle, P.J. (1989), Statistical Methods for monitoring the AIDS epidemic. *Statistics in Medicine*, 8:3-21.



University of Fort Hare
Together in Excellence

APPENDIX A

LIST OF ABBREVIATIONS

AIDS	:	Acquired immune deficiency syndrome
ANC	:	Antenatal care
ART	:	Antiretroviral treatment
ARV	:	Antiretroviral
EPP	:	Estimation and Projection Package
HIV	:	Human immunodeficiency virus
HSRC	:	Human Sciences Research Council (SA)
MACS	:	Multicenter AIDS cohort study
MTCT	:	Mother-to-child transmission
PMCT	:	Prevention of Mother-to-child transmission
SA	:	South Africa
SAS	:	Statistical Analysis System
TSI	:	Time since infection
UNAIDS	:	United Nations programme on HIV and AIDS
UNICEF	:	United Nations International Children's Fund
USA	:	United States of America
WHO	:	World Health Organisation

APPENDIX B



University of Fort Hare
Together in Excellence

DEPARTMENT OF STATISTICS

HOSPITAL MANAGER,
CECILIA MAKIWANE HOSPITAL,
NU 4, MDANTSANE,
5219.

Dear Sir,

12th February 2009

**RE : ACCESSING SECONDARY DATA AT CECILIA MAKIWANE HOSPITAL
BY MR JAMES ONYANGO NDEGE**

The above subject refers concerning my visit and Mr. Ndege to your Hospital in January this year. I am making it official as you suggested.

The bearer of this letter (Mr. James Onyango Ndege) is a member of staff at University of Fort Hare in the Department of Statistics and is currently doing his PhD in Biostatistics.

His Topic is entitled : STATISTICAL MODELLING OF MOTHER-TO-CHILD
TRANSMISSION RATE OF HIV INFECTION AND
INCUBATION PERIOD OF PAEDIATRIC AIDS.

He would like to use some of the data you already have in your data bank. This is secondary data and will not involve contact with patients at all. He will explain to you the type of data he needs.

The data is only for purposes of the PhD study and ethical considerations will be adhered to. Please accord him all the necessary help.

Yours sincerely,

Prof J C Tyler
HOD STATISTICS DEPARTMENT

APPENDIX C

HIV / AIDS SURVEILLANCE QUESTIONNAIRE

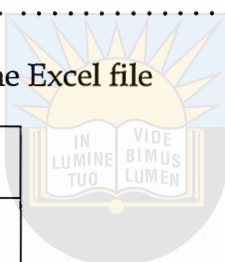
**ALL THE INFORMATION FILLED IN THE QUESTIONNAIRE ARE
EXTRACETED FROM THE SAMPLED FILES BEARING THE CHILD'S
ADMISSION NUMBER.**

001 Questionnaire Number :

002 Child's card number (only) :

003 Date of data extraction from the Excel file

Day	Month	Year



University of Fort Hare
Together in Excellence

004 Date of Birth

Day	Month	Year

005 Sex of the child

Male	Female

006 Was the child born on or after 1st January 2003 ?

If Yes continue recording

☐

If No stop the recording

☐

007 When did the child first become sick ?

Day	Month	Year

008 When did the child first visit the clinic / doctor ?

Day	Month	Year

009 What was the child suffering from ? (List all the observed diagnosis)

.....

.....

.....

.....

.....



University of Fort Hare
Together in Excellence

010 What was the child’s age at the time ?

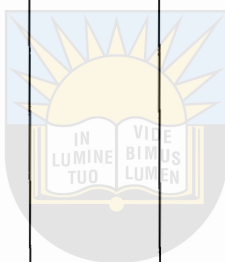
Day	Month	Year

011 Has the child ever had a HIV test ?

Yes	No

012 What type of test was done? (List all the types of tests done)

Type of test	Day	Month	Year	Result	
				Positive	Negative



University of Fort Hare
Together in Excellence

013 Has the child ever received any blood transfusion(s) ?

Yes	No

014 If yes, when was this ? (List all the dates)

Dates			Blood screened before Transfusion		
Day	Month	Year	Yes	No	Not recorded

015 Is the parent the biological mother of the child?

Yes	No

016 Has the mother ever been tested for HIV ?

Yes	No

017 If yes, when was this ? (List all the dates and give the results)

Dates			Results	
Day	Month	Year	+ ve	- ve

018 Was the child ever been given any anti-retroviral treatment ?

Yes	No

019 If yes, when did the child start receiving this anti-retroviral treatment ?

Day	Month	Year



University of Fort Hare
Together in Excellence

020 Is the child still alive ?

Yes	No

020 If yes, is the child still receiving anti-retroviral treatment ?

Yes	No

[If yes, go to Question 027]

021 If no, when did the child die ?

Day	Month	Year

023 If the child is dead, did he/she have any symptoms of HIV/ AID ?

Yes	No

024 List all the diseases/symptoms observed and when they were first observed by the doctor ?

Diseases / Symptoms	When first observed		
	Day	Month	Year

025 How long did the doctor treat the child ?

Months	Year(s)

026 If dead , what was registered as the cause of death ?

.....

.....

.....

.....

027 If the child is still alive, does he/she have any symptoms of HIV/AID ?

Yes	No



028 If yes, list all the diseases/symptoms observed and when they were first observed by the doctor ?

Diseases / Symptoms	When first observed		
	Day	Month	Year