ESTIMATING THE INCIDENCE OF ACUTE HIV INFECTION FROM A SINGLE CROSS-SECTIONAL SAMPLE

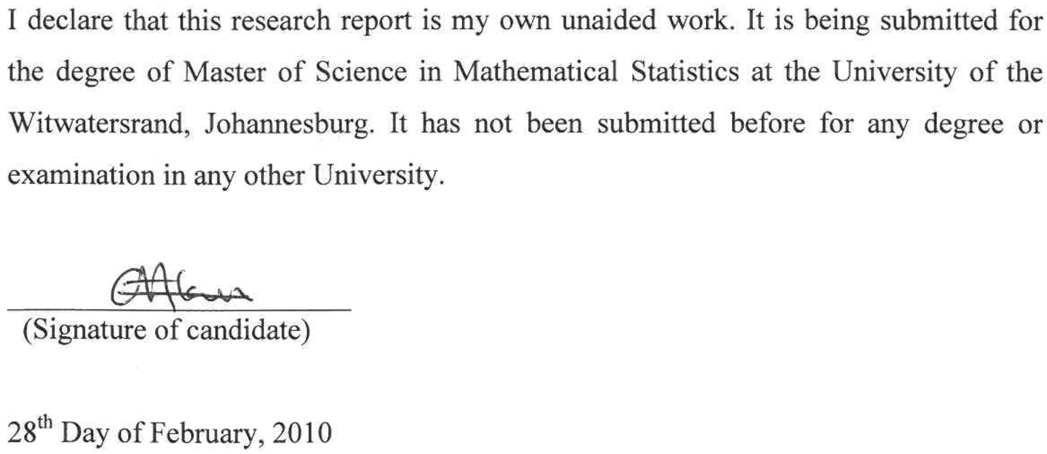
Omotola Omokunbi Akindolani

Supervisor: Prof. J. Galpin
School of Statistics and Actuarial Science

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

February, 2010.

DECLARATION



I declare that this research report is my own unaided work. It is being submitted for the degree of Master of Science in Mathematical Statistics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

(Signature of candidate)

28th Day of February, 2010

ABSTRACT

The Human Immunodeficiency Virus (HIV) epidemic is currently one of the greatest challenges and most important health issues in the world. South Africa has one of the fastest growing epidemics in the world therefore reliable estimates of prevalence and incidence are required for understanding the magnitude of the epidemic and improving the methods of prevention.

This study examines the estimation of HIV incidence from a cross-section of people, using one of the laboratory methods that discover recent HIV infection in blood samples. The incidence estimate is obtained at a single point in time, thereby saving time and cost expended in following a cohort over a period of time. It also examines incidence from pooled blood samples, and evaluates the assumptions of the different methods of estimating HIV incidence, comparing each of them; and checking the sensitivity of the estimates to the assumptions.

Results from the simulation study shows that accurate estimates of incidence can be obtained by pooling blood samples; and these estimates are obtained at a fraction of the cost of individual testing.

DEDICATION

To my parents
Mr and Mrs Akindolani
whose love and encouragement has made me who I am today.

ACKNOWLEDGEMENTS

To God almighty, without whom I am nothing.

My sincere thanks goes to my supervisor Prof J.S Galpin who was a constant source of inspiration and encouragement to me throughout the process of completing this report.

To my awesome family, for cheering me on, always being there and believing in me, even when I did not believe in myself: I love you all!!

A special word of gratitude goes to my colleague and course mate, Mr. S. Salau for his invaluable contributions and suggestions especially during the periods I had “writers’ block”.

Last but not least, special thanks also go to friends who have encouraged me and strengthened me with their words when I needed a shoulder to lean on. I am indeed very grateful and blessed to have them around me.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 : INTRODUCTION

1.1 BACKGROUND

The Human Immunodeficiency Virus (HIV) epidemic is currently one of the greatest challenges and most important health issues in the world. The virus needs no introduction, as it is the one of the deadliest diseases in the world, with no known cure as yet. It is spread most commonly by having unprotected sex with an infected partner, having contact with infected blood, sharing of needles or syringes with someone infected with the virus and from mothers to their babies during pregnancy, birth or breast feeding.

Over the years, knowledge of the spread and prevention of HIV infection has increased considerably (Pettifor, Macphail, Rees and Cohen, 2008; UNAIDS, 2008). However, despite this increase in knowledge, the epidemic is still spreading at an alarming rate in Africa, especially in Southern Africa. UNAIDS (2008) estimates that by the end of 2007, there were 5.7 million South Africans who were infected with the virus, and this figure makes this the country with the largest HIV epidemic in the world.

1.2 THE HIV EPIDEMIC IN SOUTH AFRICA

In addition, South Africa has one of the fastest growing epidemics in the world (UNAIDS, 2008), thus reliable statistics are required for understanding the magnitude of the epidemic and improving the methods of prevention. In order to measure how fast the epidemic is growing, estimates of prevalence and incidence are needed to determine the magnitude and rate of new infections.

The first cases of the epidemic are reported to have occurred around the early 80's, and were limited to homosexual men. However the pattern soon changed and by the late 1980's, the epidemic had spread to the heterosexual population, and even the children, through mother-to-child transmission. The main mode of transmission of HIV is through sexual intercourse (Gilbert and Walker, 2002). The spread of HIV in

South Africa is mainly due to heterosexual contact, with subtype C being the predominant subtype (Abdool Karim and Abdool Karim, 1999).

Like most African countries, South Africa is plagued with poor quality of data on HIV (Williams and Gouws, 2001). However, the country and its government continue to try to find ways to improve on this deficiency. Data on HIV is collected from Voluntary Counseling and Testing (VCT) centers and Sexually Transmitted Diseases (STD) clinics, Family Planning Clinics and hospitals. Williams, Gouws, Wilkinson, and Abdool Karim (2001) acknowledge that incidence estimates are not available in South Africa because of the difficulties and cost associated with obtaining these estimates.

A number of studies have been undertaken by different bodies or research organisations, for the purpose of collecting data on HIV. Examples of these include: The Nelson Mandela/HSRC study of HIV/AIDS (Shisana and Simbayi, 2002); the Medical Research Council (MRC) research among sex workers in KwaZulu Natal (Ramjee, Abdool Karim and Sturm, 1998) and the Carletonville survey carried out on a sample of mineworkers living in informal settlements in Johannesburg (Gilgen, Campbell, Williams, Taljaard and Macphail, 2000). However, these studies only focus on prevalence rates.

In order to plan HIV prevention and treatment efforts, the number of infections that are present within a population has to be determined or estimated (Kral, Lorvick, Gee, Bacchetti, Rawal, Busch, and Edlin, 2003) and this can be achieved through the estimation of prevalence (the number of HIV positive people in the population) and incidence (the number of new infections in the past year or other time period).

1.3 HIV PREVALENCE

The UNAIDS 2008 report defines HIV prevalence as the total number of people living with HIV irrespective of when they were infected. HIV prevalence estimates are usually obtained using one of two methods:

1) By monitoring sentinel populations: Sentinel surveillance is a system in which specific sites and population groups are selected, and a predetermined number of

people are routinely tested in a regular and consistent way (Nelson, 2005). Examples of Sentinel Populations include:

- Pregnant women in antenatal clinics
- STD clinic attendees
- Blood donors
- Tuberculosis patients and
- Commercial sex workers.

This is the most common method of estimation, especially in the less developed countries. In South Africa for instance, a national survey of HIV prevalence among women attending public antenatal clinics has been conducted by the Department of Health annually since 1990 (Department of Health, 2007). The information from these surveys is then used to estimate prevalence in other groups through modelling. The limitations of the estimates obtained from the surveys are that we cannot draw conclusions about certain parts of the population, such as men, non pregnant women and the sexually inactive. They also do not give estimates of incidence, unless the same people are re-tested and the person identifier is known; and this is not usually the case.

2) By population based surveys. This method uses a nationally representative sample of households to determine HIV prevalence. This is done by screening a representative sample using an antibody test such as Enzyme Linked Immunosorbent Assay (ELISA), and estimating the proportion of people infected with HIV from the known or estimated population total (Salomon and Murray, 2001). Examples of population–based surveys in South Africa are those carried out by the Human Sciences Research Council (HSRC) in conjunction with the Nelson Mandela Foundation and by the MRC.

Salomon and Murray (2001) note that although these surveys provide information on trends in the epidemic, and are more representative of the population; they have been undertaken in only a few developing countries. They are also very expensive to conduct and suffer from non-response bias due to respondents’ refusal.

Although HIV prevalence estimates are useful in determining the state of the epidemic, they only give us a ‘historical’ perspective on the rate of infection, giving us information only on what occurred in the past (Gouws, Williams, Sheppard, Barryett and Abdool Karim, 2002). With the high increase in HIV/AIDS related deaths, it is not clear whether a slowing down of the prevalence rate can be related to a decrease in the incidence rate, a high increase in death rate or both.

1.4 HIV INCIDENCE

HIV incidence is defined as the number of people who became infected with the HIV virus over a period of 1 year (UNAIDS, 2008).

Incidence can be expressed in two ways:

- Incidence rate, which is the represented by the number of incident cases divided by the population at risk within a specific time period in the course of a time period. It is also known as Incidence density.
- Incidence risk, which is the ratio of incident cases to the population at risk at the beginning of the observation period. Philippe (2001) notes that it is also known as the cumulative incidence.

HIV incidence data are less straightforward and more complex to obtain; and are usually estimated rather than measured directly. Research has shown that the HIV virus can be present in an individual without any clinical symptoms for a long time (Parekh and McDougal, 2005). Sometime after an individual becomes exposed to the HIV virus, there is a window period of 130-170 days during which HIV antibodies cannot be detected in the blood using standard HIV antibody tests. According to Rosenberg (2002), this period after infection with HIV, but before the development of detectable antibodies is known as Acute HIV Infection or Primary HIV infection. During this window period, people are usually unaware of the fact that they are infected, because when they go for tests, the tests are negative, yet they actually have the virus.

Thus the challenge in obtaining incidence estimates is to establish when an individual first contracted the virus, and this is no easy feat. Nevertheless, some indirect methods of obtaining incidence estimates exist, and they include the back-calculation method, cohort studies and cross-sectional studies. Since this study is focused more on incidence than prevalence, these methods will be described in detail in the literature review section. It should be noted that while the different methods are outlined, not all of them will be used in this study, as this study focuses on the technique based on pooling samples and using more than one test for HIV. This technique involves using portions of blood samples submitted to a laboratory, with the person identifiers removed, and then testing samples obtained by pooling the blood of several patients.

Laboratory methods and tests to identify recent HIV infection have improved over time (Parekh and McDougal, 2005). However, one main issue with using these tests is the cost associated with the administering of the tests. The pooling of blood samples is one of the mechanisms used in the reduction of the costs associated with HIV testing.

Estimates of HIV incidence are used to formulate treatment plans and also help to examine and evaluate programs aimed at reducing the spread of the virus. When new cases of HIV arise in any country, this shows that the public programs to eradicate the epidemic are not working and more ways should be devised to educate people on the need for prevention (Rutherford, Schwarcz, and McFarland, 2000; Kral *et al.*, 2003).

1.5 PROBLEM STATEMENT

HIV incidence estimates in South Africa, and indeed the whole of Africa, are scarce and mostly unreliable. Most of these estimates are obtained from cohort studies, which are based on studies of pregnant women (Gilbert and Walker, 2002; Williams *et al.*, 2001). Thus they are non-representative of the whole population since they do not include men and non-pregnant women. The question now is: Are there other cost effective methods of estimating incidence, which do not have to depend on only a select group of the population? Can these methods be used without having to organise a cohort study? Will any of these methods provide accurate estimates, be as precise or even more precise as the widely used cohort method?

This study examines the estimation of HIV incidence from a cross-section of people, using one of the laboratory methods that discover recent HIV infection in blood samples. Thus the incidence estimate is obtained at a single point in time, thereby saving time and cost expended in following a cohort over a period of time. It also examines the estimation of incidence from pooled blood samples, which provide accurate estimates of incidence at a lower cost than the common antenatal surveys.

1.6 BACKGROUND TO THE STUDY

In 2004, medical researchers from the University of the Witwatersrand, Johannesburg carried out a cross-sectional study to assess the prevalence of Acute HIV infection among South African men and women attending a primary health care clinic in downtown Johannesburg.

Consultations between the researchers and representatives of the School of Statistics and Actuarial Science (University of the Witwatersrand) were set up in order to assist with the various concerns raised during the study. One of the concerns was the determination of which confidence intervals to use for the research. During the course of these consultations, several questions were raised with regards to the effect of sensitivity and specificity of the tests used, and the effect of prevalence and the window period on estimates of incidence.

1.7 OBJECTIVES

The primary objective of this research is to investigate the statistical properties of estimates of incidence from a cross-sectional sample, using a pooling technique. The other objectives are to:

1. Evaluate the assumptions in the different methods of estimating HIV incidence, comparing them; and checking the sensitivity of the estimates to the assumptions.
2. Examine the strengths, weaknesses and appropriateness of the current approaches to estimating the incidence of HIV.
3. Examine the accuracy of pooling blood samples for the detection of HIV infection.
4. Compare incidence estimates derived from individual testing and pooled testing of different grouping strategies, discussing the different results.

In addition, an estimate of HIV incidence based on a sample from the Esselen clinic will be provided, as well as applying the findings from this report to the dataset from the clinic.

CHAPTER 2 : LITERATURE REVIEW

This chapter provides a review of the literature around the different methods of estimating HIV incidence, and the advantages and disadvantages associated with each method. Due to the importance of cost in group testing, there will also be a detailed discussion on the pooling method and its application.

2.1 METHODS OF ESTIMATING HIV INCIDENCE

HIV incidence rates help us understand the epidemiology of HIV, and also help in the formulation of sound heath policies, the appropriate allocation of resources, and the planning of programs for the primary and secondary prevention of infection. Cleghorn, Jack, Murphy, Edwards, Mahabir, Paul, O'Brien, Greenberg, Weinhold, Bartholomew, Brookmeyer, and Blattner (1998) note that HIV incidence is the best measure of the current trends of the epidemic, and very important in determining Acquired Immune Deficiency Syndrome (AIDS) incidence. Despite the rise of infection rates all over the world, HIV incidence estimates are still difficult to obtain. There are different types of studies which provide estimates of HIV incidence, and they are discussed below.

2.1.1 COHORT STUDIES

A cohort study is a study in which patients who presently have exposure to a certain risk are followed over time (in days, weeks, months, years) through periodic and repeated testing, and are compared with another group who are not exposed to the risk under investigation (Grimes and Schulz, 2002). These studies involve repeated sample collection and testing of individuals at set intervals.

In a typical cohort study, groups of people called cohorts are chosen because they have an exposure of interest. The respondents are free of the disease of interest at the beginning of the study and then they are followed up through time to determine who develops the disease. Incidence is then estimated by dividing the number of new cases discovered by the number of person years of exposure (Kaplan and Brookmeyer, 1999; Satten, Janssen, Busch, and Datta, 1999). Cohort studies are also known as longitudinal studies or follow up studies, and they are considered to be the best method for determining the incidence and natural history of a condition because it enables calculation of true incidence rates and relative risks (Grimes and Schulz, 2002).

Cohort studies are of different types: In a closed cohort study, subjects or respondents are recruited into the study, and followed up for a specific period of time. During this time, no new subjects may be recruited into the study, even if some subjects may be lost through death, withdrawal or migration (Cleghorn *et al.*, 1998). In an open cohort, subjects are continually enrolled into the study, thus making the study long and undefined.

Although cohort studies are regarded as the common approach to estimating incidence, they have several disadvantages and limitations. The results from the study may be non-representative of the whole population if they involve a select group of such as women attending antenatal care clinics.

Brookmeyer, Quinn, Shepherd, Mehendale, Rodrigues and Bollinger (1995) note that this type of study is often plagued by a ‘follow up bias’. Bias in cohort studies occurs when individuals who return for follow up have different incidence rates from those who do not return. This is usually because during follow up, they may have received counselling, and have become aware of the dangers of the disease, and therefore engaged in lower-risk behaviour. Follow up bias may also occur when individuals who are at higher risk of HIV infection do not return for follow up. In addition, they are time consuming because the researchers have to wait for the conditions of interest to develop before arriving at a conclusion (Kaplan, Kedem, and Pollack, 1998). Other disadvantages are the fact that cohort studies are expensive, and difficult to undertake, because they require assembling and following up a large number of people. (Brookmeyer, 1997; Gouws *et al.*, 2002; Parekh and McDougal, 2005).

An example of a cohort study is the one conducted by Beyrer, Brookmeyer, Natpratan, Kunawararak, Niraroot, Palapunya, Khamboonruang, Celentano and Nelson (1996). The cohort study, which was conducted in Thailand, involved using 2 different populations: commercial male sex workers and commercial female sex workers. The study was conducted between 1989 and 1994.

The results show high incidence rates of 23.8 per 100 person-years among female sex workers, and 11.9 per 100 among the males. However, follow up among the respondents was incomplete because of the high migration rates and mobility of the sex workers. Thus the estimates obtained from this study were based only on those who could be followed up.

Few cohort studies have been carried out in Africa because of the outlined disadvantages.

2.1.2 BACK-CALCULATION

Estimates of HIV incidence can be obtained from AIDS incidence data through the method of back-calculation, or back-projection. Back-calculation is a method of estimation that makes inferences about the HIV incidence through the use of AIDS prevalence, and the distribution of the incubation period over time (Kaplan and Brookmeyer, 1999; Brookmeyer, 1991). The concept of back-calculation is to use the incubation period to work backwards and reconstruct how many people must have been infected to give rise to the observed pattern of AIDS cases (Brookmeyer, 1996).

Since most developed countries have some form of AIDS reporting, this method of estimation is favoured because it makes use of readily available data on AIDS. According to Bacchetti, Segal, and Jewell (1993) back-calculation depends on 3 major factors:

1. The incubation period: this is the time between HIV infection, and AIDS diagnosis.
2. The number of AIDS diagnoses over time (obtained from national surveillance systems).
3. A model for the distribution of infections.

The basic equation of back-calculation is given by:

$$a(t) = \int_0^t I(s)F(t-s \mid s)ds \qquad (2.1)$$

where: $a(t)$ is the cumulative number of cases of AIDS diagnosed by the year *t*

$I(s)$ is the infection rate in year *s*, and

$F(t \mid s)$ is the incubation period distribution among individuals infected in year *s*.

Brookmeyer (1996) gives a simple explanation for this equation: an individual, who has been infected in a previous year *s*, would have an incubation period shorter than *t* – *s* if that individual is diagnosed with AIDS in year *t*. The value of $a(t)$ can be obtained from national surveillance data on AIDS and $F(t)$ can be derived from various epidemiological studies. These can then be used to obtain information about the incidence of new infection $I(s)$.

In using this method, it is assumed that:
1. There is sufficient data available for accurate estimation
2. Incubation does not vary across populations (Bacchetti *et al*., 1993).

Earlier works using back-projection include studies done by Bacchetti *et al*. (1993) and Brookmeyer (1996).

Estimates from back-calculation are not reliable, and produce large statistical uncertainty due to limited information on the incubation period distribution and because the time of HIV infection is usually unknown (Kaplan and Brookmeyer, 1999). In most African countries, there are no accurate data and statistics on the AIDS epidemic due to incomplete reporting of AIDS cases. For those countries with data on AIDS, there are errors in AIDS incidence data due to underreporting of AIDS cases, reporting delays, and non-response bias.

Williams *et al.* (2001) note that methods of back-calculation have not been applied to South African data as reporting of AIDS cases is voluntary and the data are too inconsistent and incomplete to be of use.

2.1.3 CROSS-SECTIONAL STUDIES

A cross-sectional study is a study which examines the presence or absence of a disease and its exposure at a particular point in time. The respondents who take part in a cross-sectional study usually do not have previous knowledge of their disease or exposure. Although cross-sectional studies are flexible because they can either study whole populations, or a representative sample; they are prone to selection bias in choosing the respondents. Also, the characteristics of the population surveyed may change with time due to deaths from the HIV virus. Despite these disadvantages however, cross-sectional studies are more cost effective and time saving than the cohort study. Because cross-sectional studies are more flexible in terms of cost and ease of implementation, most researchers have resorted to using the concept to propose different methods of estimating incidence.

2.1.4 TRENDS IN HIV PREVALENCE

Various statistical models have been developed to estimate incidence from cross-sectional surveys. These estimates can be obtained by using trends in HIV prevalence obtained from serial cross-sectional surveys. Incidence is indirectly estimated by the slope of prevalence over time assuming the population remains constant over time. However, some other models have been described in literature (Brookmeyer *et al.*, 1995; Williams *et al.*, 2001).

Cleghorn *et al.* (1998) used one of these models to estimate HIV incidence in patients attending a Sexually Transmitted Disease clinic in Trinidad and Tobago between 1987 and 1991. The annual incidence rates were calculated as the difference in prevalence at two time points divided by the time period between the two surveys. Although estimates obtained from serial cross-sectional surveys are less expensive to obtain, one major disadvantage is that there are high levels of non response. Also, the population surveyed could be changing because of unknown selection bias. (Cleghorn *et al.*, 1998).

In South Africa, Williams *et al.* (2001) presented a method of estimating the incidence of HIV infection by combining data from a single cross-sectional survey with information on the change in the overall prevalence with time. They state that while most models work under the assumption that prevalence is unchanging, this is not applicable to the South African scenario where the epidemic is still on the increase. To address this issue, the authors propose a model that links the prevalence of infection among women of age *a* at time *t*, to the incidence with assumptions about the growth rate of the epidemic and AIDS mortality rate. Their model was based on observed age-specific sero-prevalence data on 3,163 women, aged between 15 and 49 years, attending antenatal clinics in a rural district in Kwazulu-Natal province.

They presented 2 methods of obtaining age-specific incidence from age-prevalence curves (Williams *et al.*, 2001).

In the first method, a smoothed prevalence curve, estimates of the AIDS related mortality and the current epidemic growth rate were used. The authors suggest using the slope of the age-prevalence curve = $P(a,t) - P(a-1, t-1)$,

where $P(a,t)$ = Prevalence at age *a* at time *t* and

$P(a-1, t-1)$ = The prevalence at age $a-1$ at time $t-1$.

Therefore to obtain incidence, the prevalence at any age $a-1$ must be reduced by the amount by which the prevalence has increased and by the proportion of people who have died due to AIDS related diseases in the last year. The incidence at age *a* at time *t*; $I(a,t)$) is then given as:

$$I(a,t) = P(a,t) - P(a-1,t)\frac{\overline{P}(t-1)}{\overline{P}(t)}e^{-\mu}$$

Where $P(a,t)$ is the age-specific prevalence at age *a* at time *t*, $\overline{P}(t)$ is the average adult prevalence at time t, and μ is the AIDS related mortality per year.

Williams *et al.* (2001) also note that although this model is simple and easy to use, it assumes that the survivorship function of those with HIV infection is exponential and does not take into consideration the fact that infections in those of age *a* will have been accumulated over the years.

In the second method, $P(a,t)$ is a function of the proportion of infected people; $c(a,t)$, and the proportion of susceptible people of age *a* at time t; $s(a,t)$.

i.e.

$$P(a,t) = \frac{c(a,t)}{c(a,t) + s(a,t)}$$

This proportion of people at risk $s(a,t)$ is the probability that a person who has reached age *a* remains uninfected by the virus. This method assumes any form for the survivorship function, and permits changes in prevalence over time. It however requires that the data be detailed in order to give correct estimates.

Using these models, the average annual incidence per susceptible persons in the population aged 15-49 was 11.4%, with a confidence interval of (10.0-13.1). The annual incidence of infection per susceptible increased from 5.4% (3.3–8.5%) at age 15 years to 24.5% (20.6–29.1 %) at age 22 years and declined to 1.3% (0.5–2.9%) at age 50 years (Williams *et al.,* 2001).

Parekh and McDougal (2005) however, pointed out that although this approach was useful, the assumptions about the growth rate and mortality rate would have a significant impact on the outcome.

2.1.5 LABORATORY METHODS

It is important to note that none of the preceding methods above provide information on exactly when an individual became infected with HIV, and this is the most important consideration in measuring new occurrences or incidence. It has been suggested that dependence on cohort studies and estimates based on statistical models of AIDS cases and prevalence data can be eliminated if there was a reliable means of distinguishing individuals with recent infection within a population. (McDougal, Pilcher, Parekh, Gershy-Damet, Branson, Marsh and Wiktor, 2005).

Some laboratory tests have recently been developed that detect new HIV infection in the time between HIV infection and detection of HIV antibodies. These test methods differentiate recent HIV infection from established infection and estimate incidence from cross-sectional surveys (Parekh and McDougal, 2005; McDougal *et al.*, 2005). Incidence estimates are subsequently obtained by using the relationship between prevalence, incidence and the duration of the window period.

One of such tests is the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS) test, detailed by Janssen, Satten, Stramer, Rawal, O'Brien, Weiblen, Hecht, Jack, Cleghorn, Kahn, Chesney and Busch (1998). This test is also known as the 'sensitive-less sensitive' test and it is a testing algorithm which involves testing a single blood specimen from a person with HIV infection using two separate ELISA tests, with one of the tests being more sensitive to antibodies than the other. In the study by Rutherford *et al.* (2000) for example, patients who are positive on the sensitive test, and negative on the less sensitive test were estimated to have become infected within 129 days, with a 95% confidence interval of 109–149 days (Rutherford *et al.*, 2000). The estimate of incidence was given as:

$$I = \left(\frac{n}{N}\right) x \left(\frac{365}{T}\right) x (100) \tag{2.2}$$

where I is the incidence per year,
n is the number of persons who were positive on the sensitive test and negative on the less sensitive test
N is the sum of n and those who are HIV negative on both tests,
T is the estimated number of days between seroconversion on both tests, given as 129 days in the above study, which was based on ELISA testing.

A study carried out by Gouws *et al.* (2002) in South Africa estimated the incidence of HIV using the STARHS data for women attending antenatal clinics in Hlabisa, Kwazulu-Natal. Blood serum was collected from 2623 women aged 15 to 50, and this was done anonymously.

The authors also used the mathematical model developed by Williams *et al.* (2001) to compare estimates to those obtained from the STARHS method. The results were very similar-except for higher estimates of incidence in women aged 40 to 44 years using the STARHS method. They state that more research is required to confirm the accuracy of these results.

Brookmeyer and Quinn (1995) also propose a method of estimating HIV incidence rates based on a single cross-sectional survey. In this approach, a group of individuals are tested for HIV antibodies. Those who test negative or indeterminate for HIV are tested again via the p24 antigen test, which recognizes early HIV infection better than the standard HIV antibody test. The expected proportion of p24 antigen positive is given as:

$$p \approx I * \mu \tag{2.3}$$

where ‘I’ is the incidence rate and μ is the mean duration of the p24 antigen seroconversion period. This estimated mean duration was given as 22.5 days with a

95% confidence interval of 13-42 days, based on a Markov model of transitions between states, fitted to their data.

One source of uncertainty with this method is concerned with μ, the mean duration. Le Vu, Pillonel, Semaille, Bernillon, Le Strat, Meyer, and Desenclos (2008) explain that the number of days during which the p24 antigen is detectable is too short, and has even become shorter because of the development of new antibody tests that are more effective. Due to this uncertainty in the mean duration, incidence estimates obtained should be cautiously recommended. Another limitation with this method is that large sample sizes are required, especially in low incidence areas (Brookmeyer and Quinn, 1995).

Few of these laboratory methods have been introduced in Africa, despite frequently being used in the more developed countries like the United States.

2.2 ISSUES IN HIV TESTING

Although current HIV antibody testing methods detect infections, they do not distinguish between recent (incident) and long-term (prevalent) infections (Parekh and McDougal, 2001). Therefore, to accurately estimate an individual's status, it is important that the type of HIV test used is one that easily makes this distinction between what type of infection it is.

Tests to diagnose HIV infection are of different types: those that detect antibodies, tests that identify antigens such as the p24 antigen test, tests that detect /monitor viral nucleic acids and those that provide an estimate of T lymphocyte numbers (Constantine, 2001).

When choosing a test, considerations should be given to the cost of the test, its sensitivity and its specificity. The sensitivity of a test is the probability that a diseased person will be correctly classified by a test; while the specificity of a test is the probability that a non-diseased person will be correctly classified by a test (Litvak, Tu, and Pagano, 1994).

A perfect test has 100% specificity and 100% sensitivity but currently, there is no perfect HIV test. Studies have shown that the p24 test has a specificity of 99%, and a sensitivity of about 79%. Conversely, the HIV RNA test has a sensitivity of 100% but a specificity of 97%. The ELISA test has a sensitivity of 77% and specificity of 97% (Hecht, Busch, Rawal, Webb, Rosenberg, Swanson, Chesney, Anderson, Levy, and Kahn, 2002). As new tests evolve, the costs, sensitivity and specificity of the tests are constantly changing. The table below summarizes the details of the three tests from Hecht *et al.* (2002) and it is important to note that these numbers have changed over the years.

Table 2.1: Sensitivity, specificity & cost of different HIV tests

TEST	SENSITIVITY	SPECIFICITY	COST
p24 Antigen	79	99	$25
HIV RNA	100	97	$119
ELISA	77	97	$20

Litvak *et al.* (1994) provided an overview of the relationship between sensitivity and specificity. If sensitivity is improved by retesting all negative samples, specificity will decrease; and if the specificity is improved by retesting all positive samples, sensitivity will decrease. Both approaches involve an increase in cost because either way, testing is duplicated. According to Litvak *et al.* (1994), the only way of increasing the quality of the test kits and reducing cost simultaneously is by pooling the samples together.

2.3 POOLED TESTING

In a standard HIV screening test, patients are tested one at a time. Patients who test positive are retested with a confirmatory test, while those who test negative are not retested. When estimating incidence, standard screening is not cost efficient because we have to test each individual. Various strategies have been developed recently to overcome this inefficiency in cost. One of these developments is the grouping of samples together and testing them as if they were one sample. This is known as pooling.

Pooled testing is also known as group testing and this procedure was developed by Dorfman (1943). The method was first used in the United States to detect syphilis in

U.S soldiers going off to World War II. Under this program, a sample of blood was first drawn from each man, and each blood sample was then tested for the presence of the syphilis antigen. However, Dorfman (1943) proposed that instead of testing each individual sample, the blood samples should be pooled into a certain number of groups, and these groups would now be tested instead of the individual samples. If the pool tests negative, it means that none of the individual samples in that pool has the disease, hence, it will not be necessary to test the individual samples making up that pool. If the pool tests positive, the individuals making up the positive pool would be tested one by one to determine which particular ones are positive (Dorfman, 1943; Brookmeyer, 1999). A diagrammatic representation of the Dorfman pooling algorithm is given below in Figure 2.1.

Chen and Swallow (1990) outlined the goals behind pooled testing which are mainly two-fold: Firstly, to detect all individuals in the population as being infected with a particular disease or not, while reducing the number of tests at the same time. This is known as the "Classification problem". The second goal is to estimate the proportion of individuals who have a particular disease in a population (prevalence), without identifying the infected individuals. In statistical literature, this is known as the "Estimation problem" (Kim, Hudgens, Dreyfuss, Westreich, and Pilcher, 2007).

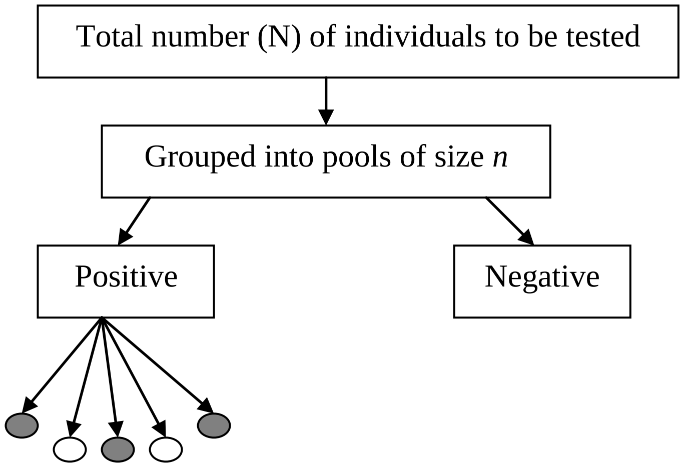

Total number (N) of individuals to be tested
Grouped into pools of size *n*
Positive
Negative

Figure 2.1: **Diagram of dual stage pooling algorithm as proposed by Dorfman (1943). Individual blood sera pooled into groups of n and these groups are tested. The individuals making up the positive pools are retested to determine which particular individuals are positive.**

The classification problem aims at reducing the expected number of tests involved in the grouping procedure, while the estimation problem aims to minimize the mean squared error of the proportion of positives for a fixed number of tests (Hung and Swallow, 2000).

2.3.1 VARIANTS OF POOLING

The original pooling algorithm as proposed by Dorfman (1943) is quite simple and has been used as a foundation for studies of infectious diseases. Kline, Brothers, Brookmeyer, Zeger and Quinn (1989) for instance undertook a study to determine the accuracy and cost effectiveness of estimating HIV sero-prevalence in a population survey using pooled testing. In their procedure small groups of at most 15 samples were created. If the test result is negative, all 15 samples are declared negative. If the test result is positive, then each of the 15 samples is tested individually. They observed a cost reduction of up to 80% as compared to individual testing of all samples.

Some other pooling procedures have been proposed as an extension of the above pooling strategy. Litvak *et al.* (1994) discussed two different group testing methods in which positive groups are split into several subgroups of almost equal size, denoted as T_2 and a subsequent extension as T_2^+. For the first test (T_2), if the pooled sample example above tests negative, then all 15 samples are declared negative. If it tests positive, then it is divided into 2 different subgroups of size 7 and size 8. The testing procedure is then repeated within each of the subgroups. For the second test (T_2^+), if the pooled sample tests negative, it is retested.

Each group that produces a negative outcome is retested at most r-1 times (where r is the maximum number of times a group will be retested before being classified or split further into smaller groups). If all r-1 tests are negative, the group is classified as negative; otherwise, it is classified as positive. They concluded that both methods were better than Dorfman's procedure.

Johnson and Gastwirth (2000) describe a two stage procedure which is similar to Dorfman's algorithm. In the first stage every individual is randomly pooled into groups of size $k \geq 1$. Due to the fact that some errors may occur in this stage, a second stage is adopted. In the second stage, a random sample of units from the pools/groups that tested negative in the first stage are pooled into new pools and are retested. If a pool or group tests positive in the second stage, then all the individual samples that make up that pool are tested by a gold standard test to identify the samples with the specific characteristic in question. The idea behind re-pooling a portion of individuals from negative pools is that accurate inferences can be made about the accuracies of the test used; and the probability of having a false negative result pass through the system is limited.

A simple extension of Dorfman's procedure is known as Hierarchical group testing or multistage testing. In this method, a master pool (*n*) comprising all samples is first tested. If the result is positive, the pool is then divided into non-overlapping sub-pools. As the name implies, the re-grouping of such samples can be divided into as many stages as possible, until individual samples are tested (Kim *et al.*, 2007). Brookmeyer (1999: 609) describes the procedure for the single stage pooling as follows:

"N randomly selected individuals are allocated to one of n_1 pools that each consist of c_1 individuals. Suppose r_1 of the n$_1$ pools test positive (i.e., at least one of the individuals in each of these pools has the disease) and s_1 of the pools test negative (i.e., all individuals in each of these pool (sic) do not have disease), where $r_1 + s_1 = n_1$. The likelihood function is given as:

$$\mathrm{L}_1 = \binom{n_1}{r_1}\left\{(1-p)^{c_1}\right\}^{s_1}\left\{1-(1-p)^{c_1}\right\}^{r_1} \qquad (2.4)$$

The maximum likelihood estimator of the overall prevalence p (Thomson, 1962) is:

$$\hat{p} = 1-\left(\frac{s_1}{n_1}\right)^{1/c_1} \text{"} \qquad (2.5)$$

Brookmeyer (1999: 609) further states that:

"In a multistage pooling study, the positive pools identified at each stage are subdivided and tested."

“For example, in stage 2, each of the r_1 positive pools identified at stage 1 is divided into pools consisting of c_2 individuals. Suppose the *jth* positive pool from stage 1 yields r_{2j} positive and s_{2j} negative pools in stage 2 (*j*=1, …, r_1). The r_{2j} positive pools are then subdivided and tested in stage 3. More generally, at the *ith* stage, pools of size c_i are constructed from the positive pools identified at stage *i*-1. The numbers of positive and negative pools that originate from the *jth* positive pool of stage *i*-1 are called r_{ij} and s_{ij} respectively. The total number of positive pools at stage *i* is: $r_i = \sum_{j=1}^{r_{i-1}} r_{ij}$ (for i≥2) and the total number of negative pools at stage *i* is $s_i = \sum_{j=1}^{r_{i-1}} s_{ij}$. The total number of stages is called k.

The likelihood function of the general multistage pooling study can be expressed as a product of conditional probabilities of the pooling results from stage *i* given the results from stage *i* – 1. At stage *i,* the conditional probability of observing r_{ij} positive pools and s_{ij} negative pools originating from the *j*th positive pool at stage *i*-1, conditional on the event that at least one of the c_{i-1} individuals from this pool has the disease is:

$$L_{ij} = \frac{\binom{r_{ij} + s_{ij}}{r_{ij}} \left\{(1-p)^{ci}\right\}^{sij} \left\{1-(1-p)^{ci}\right\}^{rij}}{1-(1-p)^{c_{i-1}}}. \tag{2.6}$$

The likelihood function for k stages is:

$$L = L_1 * \prod_{i=2}^{k} \prod_{j=1}^{r_{i-1}} \mathrm{L_{ij}}” \tag{2.7}$$

Substituting for L_{ij} “and simplifying, we obtain:

$$\mathrm{L} \propto (1-p)^{N_1} \left\{1-(1-p)^{c_k}\right\}^{r_k} \tag{2.8}$$

where $N_1 = \sum_{i=1}^{k} c_i s_i$ is simply the total number of individuals in negative pools, r_k is the final number of positive pools identified at *kth* stage and c_k is the number of individuals in each of these positive pools.” (Brookmeyer, 1999: 609).

“Each individual in the negative pool contributes the factor $\left(1-p\right)$ and each positive pool contributes the factor $\left(1-\left(1-p\right)^{c_k}\right)^{r_k}$. The likelihood depends only on the final size of the positive pools, c_k, and not on the pool sizes at earlier stages (c_i, i<k). After setting $\partial \log L \big/ \partial p$ =0, we obtain

$\hat{p} = 1 - \left(\frac{N_1}{N}\right)^{\frac{1}{c_k}}$.” (Brookmeyer, 1999: 609-10). (2.9)

Brookmeyer (1999) explains that if there is an assay to detect early onset of the disease, we can estimate incidence from a multistage pooling study by substituting $\hat{p}$ for p in the incidence equation: $p = I * \mu$. Thus an estimate of incidence $\hat{I}$ is given as:

$$\hat{I} = \frac{1}{\mu}\left\{1 - \left(\frac{N_1}{N}\right)^{\frac{1}{ck}}\right\}. \qquad (2.10)$$

A practical example of multistage pooling is outlined in the NC STAT program by Pilcher, McPherson, Leone, Smurzynski, Owen-O'Dowd, Peace-Brewer, Harris, Hicks, Eron, and Fiscus, (2002) using a three stage hierarchical algorithm as follows: First, disjoint master pools of 90 specimens are tested. Second, positive master pools are divided into sub pools of 10 specimens each and these sub pools are tested. Third, specimens from the positive sub pools are individually tested.

Multistage testing has been used in the detection of acute HIV infection (Pilcher, Price, Hoffman, Galvin, Martinson, Kazembe, Eron, Miller, Fiscus, and Cohen, 2004; Pilcher, Fiscus, Nguyen, Foust, Wolf, Williams, Ashby, O'Dowd, McPherson, Stalzer, Hightow, Miller, Eron, and Cohen, 2005; Quinn, Brookmeyer, Kline, Shepherd, Paranjape, Mehendale, Gadkari, and Bollinger, 2000; Brookmeyer, 1999) and will be the focus of this research report.

Johnson, Kotz, and Wu (1991) examined hierarchical algorithms in the presence of test errors. They derived the expected number of tests, pooling sensitivity and pooling specificity for a hierarchical algorithm. They also allow sensitivity and specificity to be dependent on the number of pools in each stage.

An alternative to Hierarchical group testing is a method called array based specimen pooling. Kim *et al.* (2007) note that this approach is frequently employed in genetics but remains underutilized in the infectious disease setting. To begin with, n^2 specimens are usually arranged in an *n x n* matrix format, and pools of size *n* are created from all samples in the same row or in the same column. Using the assumption that there are no false negatives, the pools are tested, and all positive specimens will lie at the intersection of a positive row and a positive column pool. Kim *et al.* (2007) note that this method is not effective for estimating infectious diseases mainly because it requires an extremely large number of specimens, and it does not consider test errors i.e. false negatives and false positives. It should be noted that this variant is mentioned for the sake of completeness, but not used in the study.

2.3.2 ASSUMPTIONS IN POOLING

The feasibility of pooling blood samples have been studied by various authors such as Quinn *et al.* (2000), and Brookmeyer (1999). Sergeant and Toribio (2004) report that pooled testing is based on a number of assumptions:

1. The status of an individual is independent of the status of others both within and outside the pool.
2. The test used is sensitive and specific enough to detect when one of the pooled samples is positive or negative.
3. The sensitivity of the test for a pool is approximately the same as it is for an individual sample. (Litvak *et al.*, 1994).
4. The dilution of individual samples into the group does not affect the accuracy of the tests.

2.3.3 THE DILUTION EFFECT

Gupta and Malina (1999) define the dilution effect as the failure of tests to detect an infected sample when it is diluted with a large number of negative samples. Zenios and Wein (1998) note that if the size of a pool is large, some of the positive sera may be excessively diluted by negative sera and consequently become undetectable in the pool. This can compromise the sensitivity of a test thereby resulting in the underestimation of prevalence and may also lead to a loss of accuracy.

To examine the effect of dilution, Wein and Zenios (1996) develop a hierarchical statistical model which links the HIV test output with the antibody concentration in the pool, thus capturing the effect of pooling different samples together. Their approach simultaneously captures the dilution effect, while estimating prevalence directly from optical density (the attribute that is observed and measured in an ELISA test) readings. They assumed that the optical density of the pool is the average of the optical densities of the individual samples. Their results show that although the dilution effect may lead to a slight loss of accuracy during pooled testing; this can be overcome by the choice of the pool/group size. In spite of this concept, the analysis has one main disadvantage- consideration is not given to the effect of the window period donors in the model.

2.3.4 OPTIMAL POOL SIZE

One advantage of pooled testing is the reduction of the number of tests used, and the subsequent reduction in the cost of testing. The optimal pool size is the one that provides accurate results whilst minimizing the loss in sensitivity and specificity of the tests. In dual group and multistage testing, the optimum choice of the size of pools depends on the prevalence of the disease in the population, and the accuracy and cost of the tests used (Johnson and Gastwirth, 2000). Generally, group sizes are predetermined by the laboratory personnel in charge of the procedure.

2.3.5 CONFIDENCE INTERVALS

Measures of the uncertainty due to random error are typically presented in epidemiologic research in the form of a confidence interval (Cole, Chu and Brookmeyer, 2006). Kline *et al.* (1989) conducted a study to evaluate the feasibility and accuracy of pooled samples in estimating HIV prevalence in large samples. Point and interval confidence estimates were obtained from the pooled sera. In their study, blood samples from two large population surveys were pooled separately. Prevalence was estimated by maximum likelihood and given as:

$$\hat{p} = 1 - (1 - [s/n])^{\frac{1}{c}} \tag{2.11}$$

and the asymptotic variance of $\hat{p}$ is estimated as :

$$v = \frac{\{s/n[1-(s/n)]^{[2/c]-1}}{c^2 n} \tag{2.12}$$

where n represents the number of groups/pools, c is the number of blood samples making up n and s is the total number of positive pools.

Generally the approximate 95% confidence interval was given by Kline *et al.* (1989) as:

$$\hat{p} \pm 2\sqrt{v} \tag{2.13}$$

The confidence intervals to be used and compared in this report are based on what has been used in similar literature (Cowling, Gardner, and Wesley, 1999; McDougal *et al.*, 2005; Hauck, 1991 and Brookmeyer *et al*, 1995). The confidence intervals are defined as:

Normal Approximation Method of the Binomial Confidence Interval:

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{2.14}$$

where p = proportion of interest and n = sample size.

Exact Binomial Confidence Interval:

Upper limit: $\sum_{i=x}^{m} \binom{m}{i} P_U{}^i (1 - P_U{}^i)^{m-i}$ (2.15)

Lower limit: $\sum_{i=0}^{x} \binom{m}{i} P_L{}^i (1 - P_L{}^i)^{m-i}$ (2.16)

where m is the number of pools and x is the number of positive samples detected.

Exact Poisson Confidence Interval:

Upper limit: $\frac{1}{2}\chi^2\left(\alpha/2; 2_x + 2\right)$ (2.17)

Lower limit: $\frac{1}{2}\chi^2\left(1-\alpha/2; 2_x\right)$ (2.18) where x is the observed count and $\chi^2(P;a)$ is the P^{th} quantile of the chi-square distribution with a degrees of freedom.

Another confidence interval defined in McDougal *et al.*, 2005 is given as:

$$\hat{p} \pm 1.96 \frac{I}{\sqrt{N}} \qquad (2.19)$$

where $\hat{p}$ is the proportion of interest, 'I' is the annualized incidence rate, and N is the number of infected individuals classified as positive on the assay or test used (McDougal *et al.*, 2005). This will be termed the McDougal approach confidence interval.

Cowling *et al.* (1999) note that in low prevalence populations, the lower confidence limit may be negative due to the "inappropriateness of the large-sample normal theory in extreme-low-prevalence situations". To resolve this problem, Hauck (1991) proposed an alternative confidence interval that will not give values less than zero, and does not depend on the asymptotic variance (v). The two methods recommended were: 1) to obtain an asymmetric interval that uses the normal approximation with a continuity correction and 2) to obtain exact confidence intervals by assuming a binomial distribution for the number of positive pools (Hauck, 1991). Due to the high prevalence of HIV in South Africa, it is safe to assume that we are unlikely to obtain a negative confidence limit; therefore, the method of Kline *et al.* (1989) is used to obtain the confidence intervals in this report.

2.3.6 OTHER AREAS OF APPLICATION

Pooled testing has been successfully used to screen blood donors in less-developed countries for the HIV virus (Emmanuel, Bassett, Smith, and Jacobs, 1988). However, the validity of this has been questioned because of the possible loss in sensitivity during this process (Johnson and Gastwirth, 2000; Kennedy, 2004a).

Pooling has been applied to the study of other infectious diseases such as Hepatitis B, Hepatitis C, and West Nile Virus (Westreich, Hudgens, Fiscus and Pilcher, 2008). The extension of this method into the study of HIV has also proved useful especially in the area of reducing the costs of HIV tests (Litvak *et al.*, 1994; Saraniti, 2006; Kline *et al.*, 1989). Chen and Swallow (1990) pointed out that it is used to quantify resistance factors in plants. It has also been used to estimate disease prevalence in animals and infection rates in insect vectors (Sergeant and Toribio, 2004; Munoz-Zanzi, Johnson, Thurmond, and Hietala, 2000; Maherchandani, Munoz-Zanzi, Patnayak, Malik, and Goyal, 2004). Other applications are in the area of genetics and the demonstration of cost efficiency and reduction in pharmaceutical companies (Westreich *et al.*, 2008; Xie, Tatsuoka, Sacks, and Young, 2001).

Kennedy (2004b) notes that in populations with low prevalence rates, pooling can result in more accurate tests. Although it is a cost reducing method, cost savings are only going to be substantial where there are a large number of samples.

In summary, the various methods of estimating incidence have been examined. Although all the methods outlined above have their disadvantages, they are still in use in different countries all over the world today. Pooled testing provides a way to group samples together, test them, and still be able to estimate prevalence and incidence from these samples-except of course, at a lower cost. One issue that has not been addressed in the literature is the sensitivity of incidence estimates to the different assumptions in pooling and this will be discussed in upcoming chapters.

CHAPTER 3 : METHODOLOGY

3.1 STUDY POPULATION

The sample that initiated this study was made up of consenting attendees at the Hillbrow Esselen Street clinic in Johannesburg. This clinic is a primary care facility, providing HIV testing services, Sexually Transmitted Infections (STI) services, and Voluntary Counselling and Testing (VCT). The respondents were recruited into the study between April and October 2004. Data was collected for each individual on age, sex, and the arm of the clinic where the test took place: either the STI or the VCT arm. No physical examination or behavioural data were collected in this survey. A total number of 1906 individuals were anonymously tested, of which 970 (50.9%) were male and 936 (49.1%) were female. The study was performed on anonymous and unlinked samples as specified by the ethics committee of the University of the Witwatersrand (Stevens, Akkers, Myers, Motloung, Pilcher and Venter, 2005).

3.2 THE POOLING ALGORITHM (ESSELEN CLINIC DATA)

From the 1223 respondents who tested negative on the ELISA test, a portion of the blood samples taken from each respondent was then used to create 12 pools of 100 samples. This was done specifically to have an equal number of specimens in each pool. The balance of the samples (23) was later pooled, with negative results.

Applying the RNA test on these 12 pools, 3 pools were found to be positive, indicating that there were positive individual samples within the pools; and 9 pools were negative-no further testing was required on these 9 pools. Therefore 300 samples would have to be retested. The 3 positive pools were split into 6 pools of 50. After testing, 3 of the 6 pools were found positive and 3 pools were negative. Thus 150 samples were retested, while the other 150 needed no further testing. The 150 positive samples were again grouped into 15 pools of 10 from which 7 pools were found positive, and 8 pools were found negative. The 70 samples were then tested individually, and 8 of these were positive. The remaining 80 samples were negative, and needed no further testing.

Of the 23 other samples 2 pools of 10 were created, which tested negative on the RNA test. The remaining 3 samples were tested individually and also tested negative on the RNA test.

3.3 POOLING STRATEGIES

Two main algorithms are considered: The first is based on Dorfman's (1943) dual stage approach where a number of samples are combined into pools of a certain size. If the pooled samples produce a value equal to or greater than the lower limit of interest for the pool then all individual samples in the affected pools are tested in the second stage. If the sample values are below the lower limit of interest, they are considered negative and no further tests would be performed on these pools.

The second algorithm follows the example of a pooling strategy set by Quinn *et al.* (2000). The strategy is a 3- stage Hierarchical pooling algorithm with the following general structure:

a) Firstly, a "Master Pool" containing all blood samples is tested.
b) If the master pool is positive, we group samples into non-overlapping sub-pools and test these sub-pools.
c) Individual specimens from the positive sub-pools are tested.

The pooling algorithm describing the method used is found below and is a modification of the algorithm used by Quinn *et al.* (2000).

3.4 OTHER MODEL ASSUMPTIONS

- The outcome of both the antibody test and the HIV RNA test is either positive or negative: there are no indeterminate results.
- There are no errors in testing.
- Constant sensitivity and specificity which is independent of pool size.

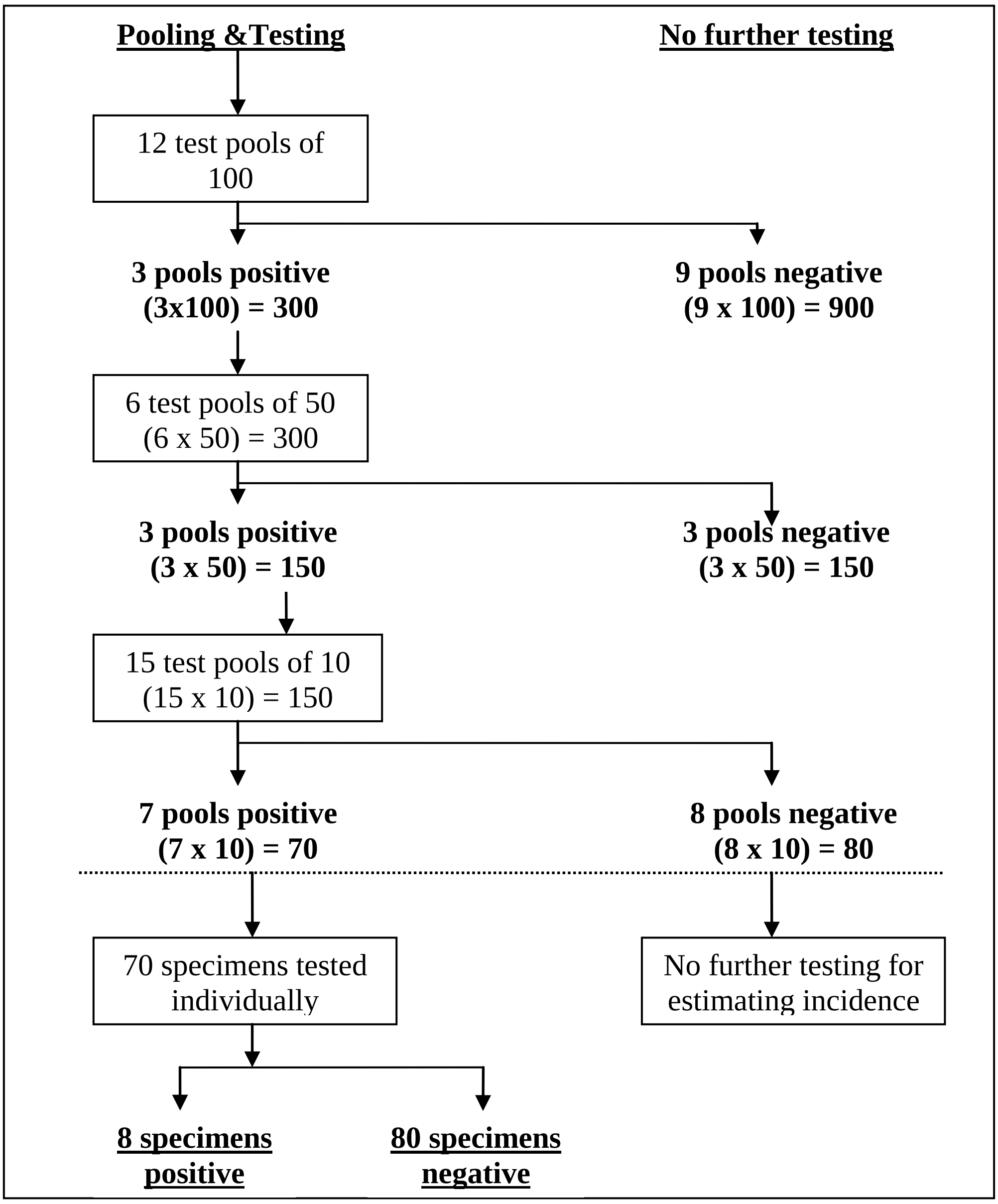

Pooling &Testing
No further testing
12 test pools of 100
3 pools positive (3x100) = 300
9 pools negative (9 x 100) = 900
6 test pools of 50 (6 x 50) = 300
3 pools positive (3 x 50) = 150
3 pools negative (3 x 50) = 150
15 test pools of 10 (15 x 10) = 150
7 pools positive (7 x 10) = 70
8 pools negative (8 x 10) = 80
70 specimens tested individually
No further testing for estimating incidence
8 specimens positive
80 specimens negative

Figure 3.1: Pooling Strategy used in HIV testing for 1200 antibody negative respondents attending the Esselen Clinic in Johannesburg, South Africa. The results for the remaining 23 samples have been omitted in this figure.

3.5 OPERATING CHARACTERISTICS OF POOLING ALGORITHMS

The operating characteristics of pooled algorithms have been defined in group testing literature (Kim *et al*, 2007 and Westreich *et al*, 2008). Denoting the dual stage algorithm defined by Dorfman (1943) by S2 and the three stage algorithm defined by Quinn *et al*. (2000) by S3, the characteristics of the algorithms are:

Efficiency: The efficiency of a pooling algorithm is the expected number of tests per specimen required to identify all positive specimens (Kim *et al*, 2007). The efficiency of individual testing is 1, and an efficiency of less than 1 indicates that the pooling algorithm will require fewer tests on average than individual testing. Dorfman (1943: 438) states that "the extent of the savings attainable by the use of the group method depends on the group size and the prevalence rate".

The efficiencies of the above algorithms are given as:

$$\text{E (S2)} = \frac{1}{n} + (1 - q^n) \tag{3.1}$$

$$\text{E (S3)} = \frac{1}{n} + \frac{1 - q^n}{\sqrt{n}} + 1 - q^{\sqrt{n}} \tag{3.2}$$

Where n =pool size

q = 1- p

p =Prevalence

Pooling Specificity $S_p(A)$: This is defined as the probability that an individual is categorized as negative by a particular pooling algorithm (A) given that the individual is truly negative.

Pooling Sensitivity $S_e(A)$: The probability that an individual is categorized as positive by a particular pooling algorithm (A), given that the individual is truly positive.

Positive Predictive Value (PPV): The probability that an individual is truly positive given that s(he) is categorized as positive by A and is given as:

$$\text{PPV} = \frac{pS_e(A)}{(1-p)\{1 - S_p(A)\} + pS_e(A)} \tag{3.3}$$

Negative Predictive Value (NPV): The probability an individual is truly negative given that s(he) is categorized as negative by A and is given as

$$\text{NPV} = \frac{(1-p)S_p(A)}{(1-p)S_p(A) + p\{1 - S_e(A)\}} \tag{3.4}$$

3.6 SIMULATION

Incidence estimation will be further investigated using a simulation study, which examines the effects of pooling on different algorithms. The laboratory results obtained from the Esselen Street data were analysed and will be integrated into a simulation model aimed at evaluating the efficiency of two main pooling algorithms: The Dorfman (1943) 2-stage procedure and the Quinn *et al*. (2000) 3-stage pooling procedure. It should be noted that one can go beyond 3 stages in multistage pooling; however, most of the literature on this subject is limited to three stages.

Firstly, a dataset with N subjects is generated to mimic a scenario where the true prevalence is p. This is done by simulating each subject's initial infection probability using a random number from a uniform distribution U [0, 1]. If the simulated infection probability is less than the selected prevalence level, the data point is labelled as positive and if it's greater than or equal to the selected prevalence level, it is labelled as negative. This typically corresponds to the 1st test in which negative samples have antibody values below 500copies/ml and positive samples have antibody values equal to and greater than 500copies/ml.

At this stage, each observation is classified as ELISA positive or ELISA negative. For each ELISA negative value, the 'new infection' status is set by generating another uniform [0, 1] number, and classifying the observation as an old infection if this is >=specified annual incidence level (incidence * 18/52), otherwise it is classified as a new infection. The factor 18/52 is used to annualize the incidence estimate based on the window period estimate given by Janssen *et al.* (1998): 129 days (18 weeks) which is the mean time of seroconversion between the sensitive and the non-sensitive tests divided by 365 days (52 weeks), which is the number of days in a year. The RNA value for the observation is then generated.

Because the Esselen Street study employed the use of HIV RNA test as the 2nd assay, the characteristics of the RNA test were used for detecting recent infections in the simulated samples. Previous studies (Le Corfec, Le Pont, Tuckwell, Rouzioux, and Costagliola, 1999) indicate that threshold assay value used to identify positive samples based on the HIV RNA test is 750,000 copies per ml, with a value less than or equal to 400 being classified as RNA undetectable, and all other values are classified as negative. The RNA values follow a log normal distribution (Wein and Zeinos, 1998). In order to take viral load into account, the values are generated as a mixture of normal distributions as follows:

- For new infections: exp{N (0, 1)*0.1+log (10000} i.e. exp {N (4, 0.1)}
- For old infections, another U [0, 1] number is generated.
    - If this is less than 0.5, the RNA value is generated from exp {N (0, 1)*0.1+log (100000} i.e. exp {N (5, 1)}.
    - Otherwise it is generated as exp {N (0, 1)*0.1+log (1000000} i.e. exp {N (6, 0.05)}.

Viral load values were not available for the Esselen Street clinic data and very few South African studies report mean RNA values. A study done by Shisana, Rehle, Simbayi, Parker, Zuma, Bhana, Connolly, Jooste and Pillay (2005) gives the median HIV RNA viral load as 3.9. In order to have infected individuals with varied viral load values, a mixture of distributions was used and a mean HIV RNA range of 4-6 was considered.

The simulated samples are divided into predetermined pools of size *k* to imitate the random allocation of samples to pools in an actual laboratory testing. For each group of negative samples, 2 pooling algorithms are examined:

a) Dual or 2-stage pooling: This is based on the Dorfman (1943) algorithm where the negative samples are combined into one pool. Here, pool samples with a value less than the set threshold will be considered as negative for HIV and no further grouping or testing will be performed. Otherwise, all the samples in that pool would be tested individually because it shows that there is at least one positive sample in that pool.

b) Multistage (3 stage pooling).: This follows the same structure as the Dorfman algorithm, except that we further re-pool positive samples and only test individually in the third and final stage.

PARAMETERS:

To set up the simulation described above, the general parameters required are:

- p =Prevalence: Prevalence is set between 25%-35% following the current HIV prevalence rates in South Africa (Welz, Hosegood, Jaffarc, Batzing-Feigenbaum, Herbst and Newell, 2007).
- I = Incidence range: The results from the Hillbrow study by Stevens *et al.* (2005) suggest that incidence is 12.9% (95% CI, 11.01%-14.8%). The range for incidence in the model will be set between 12% -15% as recommended by Beyrer *et al* (1996).
- μ =Mean duration of the RNA test before seroconversion. Quinn *et al*. (2000) estimates the period of HIV antigen positivity before seroconversion to be either 22.5 days (for the p24 test) or 28 days (for the HIV RNA test). The values of the window period used in the simulation will vary around these two figures.
- Pool sizes: Pools of size 1 (all samples), 5, 10, 25, 50, 100 and 1000 are selected arbitrarily for the ease of analysis.

A number of combinations are possible from the parameter set above but only a selected number of scenarios will be evaluated in Chapter 4. Estimates of the incidence rate, confidence intervals and number of tests needed are obtained for each of generated data sets and pooling scenario. For a given prevalence range, the relative efficiency and accuracy of both pooling strategies are examined and compared to individual testing. Discussions on the findings are presented in Chapter 4.

3.7 STUDY LIMITATIONS

The result of the data from the Esselen Street clinic is available for the direct estimation of incidence. However in order to discuss other characteristics of pooling algorithms, a simulation study is required to examine the theoretical aspects of this procedure. The method of choosing the sizes of pools is one of the limitations of this study - pooled sizes have been chosen depending on which sizes would make sub-pooling easier. A major assumption of the setup is that there is no error in testing at any of the stages. The study also does not assess the sensitivity to the assumption of no dilution effect, or the cost thereof. This is because the pool size chosen by the lab experts is assumed to take account of this.

CHAPTER 4 : DATA ANALYSIS AND RESULTS

This chapter evaluates and discusses the results and examines issues around the different pooling algorithms, size and cost. The issues of optimal pool sizes and cost saving are also explored. All computations were done using SAS version 9.1 (SAS® Institute Inc., Cary NC, USA).

4.1 SUMMARY RESULTS FROM ESSELEN STREET DATA

The main objective of study conducted at the Esselen street clinic was to determine the prevalence of acute primary HIV infection among South African men and women attending a primary care clinic. The study further generates an estimate of HIV incidence based on a multistage pooling method.

Despite the laboratory testing which was carried out, data made available from the study only contained summary level information which included demographic frequencies of the respondents, and pooled results. Individual results of the antibody tests were not included. The summary of the demographic and test characteristics of patients attending the Esselen Clinic in Hillbrow are given in Table 4.1 below:

Table 4.1: **Frequency table of respondents attending the Esselen Clinic. A weakly reactive result means that the antibody test results were equivocal or indeterminate and needed confirmatory testing.**

	Male	**Female**	**Total**
Negative	670	553	1223
Weakly Reactive	5	6	11
Positive	295	377	672
Total	**970**	**936**	**1906**

Of the 1906 individuals tested at the Esselen Street Clinic, 672 respondents were HIV antibody positive using the ELISA test and 11 were antibody equivocal or indeterminate. A total of 4 out of the 11 antibody indeterminate respondents tested positive for HIV RNA. As it was not clear whether these 11 specimens had been re-tested using the ELISA test before they were tested using the RNA test, they were discarded for the purposes of the simulation study.

The 1,223 respondents who were antibody negative were then tested with HIV RNA to detect acute infections using the pooling algorithm described in Figure 3.1. The results show that 8 respondents tested positive for HIV RNA. The prevalence of acute HIV infection in the overall study population was then calculated as $\frac{12}{1223}$ or 0.98%. This is consistent with findings from a previous study by Kamanga, Thumbi, Nkhoma, Manamela, Bogoshi, Latka, Martinson, Karim, Kumwenda, Rees, Churchyard, McCauley, Gay and Cohen (2008). In their study, 6674 people attending two Sexually Transmitted Diseases (STD) clinics in Malawi and one research facility in South Africa were screened for HIV, with an overall prevalence of 1.2%.

Based on the multistage pooling method, incidence is then estimated using the relationship between prevalence and the test window period as described in equation (2.2):

$$I = \left(\frac{12}{1223}\right) x \left(\frac{365}{28}\right) x (100)$$

Therefore, the estimate of acute HIV incidence of the Esselen Street data is given as: 12.79% (95% CI, 11.01%-14.8%).

The major limitation of the data provided was the lack of adequate pooling information. However, the data was useful for investigating the sensitivity of the final conclusion to the assumptions. As a result, the simulation steps introduced in Chapter 3.6 were performed and results are discussed in the following sections.

4.2 SIMULATION RESULTS

The subsequent sections show and discuss the direct estimates of incidence obtained from the different methods discussed in preceding chapters.

The first approach examines incidence estimates derived from a single cross-sectional survey using the relationship between prevalence and the estimated window period: $p \approx I * \mu$. Therefore calculations of incidence= $I = \frac{p}{\mu}$ per day.

Table 4.2: Estimated incidence rates (% per year) as a function of p, the expected proportion of positives, and mean duration μ ranging from 15 days to 40 days

Prevalence (p)	Window period				
	15	22.5	28	30	40
0.1	2.4	1.6	1.3	1.2	0.9
0.12	2.9	1.9	1.6	1.5	1.1
0.15	3.7	2.4	2.0	1.8	1.4
0.2	4.9	3.2	2.6	2.4	1.8
0.25	6.1	4.1	3.3	3.0	2.3
0.3	7.3	4.9	3.9	3.7	2.7
0.35	8.5	5.7	4.6	4.3	3.2
0.4	9.7	6.5	5.2	4.9	3.7

For a given set of values of prevalence (proportion of positives), Table 4.2 shows the incidence rates that would be estimated from a cross-sectional survey and shows the sensitivity of these rates to the different assumptions about the mean duration (μ).

As indicated in equation (2.5), incidence rates are inversely proportional to the mean duration of the assumed test. When prevalence is given at 30% with a 40 day mean duration of an assumed test, incidence is estimated to be 2.7%. A window period of 28 days with a given prevalence of 15% results in a lower estimate of 2%.

Brookmeyer and Quinn (1995) state that disadvantages of this method are that large sample sizes are required; and there are uncertainties associated with the estimation of the mean duration or window period. For the purpose of this study, a mean duration of 28 days is used throughout the analysis as suggested in the literature (Le Corfec *et al*, 1999).

Measures of uncertainty in epidemiologic research are typically obtained by computing confidence intervals for calculated estimates (Cole *et al, 2006*). Table 4.3 below presents the various confidence intervals for individual-incidence estimates based on the confidence intervals defined in equations (2.14) to (2.17).

Table 4.3: Confidence intervals of incidence estimates with varying prevalence and sample size and μ =28 days

	N=1000			N=2000			N=5000		
	p=25%	p=30%	p=35%	p=25%	p=30%	p=35%	p=25%	p=30%	p=35%
Incidence	19.40	30.49	40.48	21.17	26.09	34.56	17.60	22.95	31.22
Confidence Intervals:									
'McDougal Approach'	18.53-20.28	29.36-31.63	39.14-41.82	20.52-21.81	25.36-26.82	33.7-35.43	17.23-17.97	22.52-23.38	30.69-31.74
Normal Approximation	8.02-30.78	15.73-45.26	23.02-57.95	12.77-29.57	16.52-35.66	23.27-45.86	12.8-22.39	17.34-28.57	24.42-38.02
Exact Binomial	9.72-34.51	17.5-49.16	24.86-61.99	13.6-31.37	17.39-37.54	24.16-47.82	13.12-23.09	17.67-29.29	24.77-38.79
Poisson	9.69-34.72	17.43-49.52	24.73-62.52	13.56-31.5	17.34-37.71	24.08-48.07	13.1-23.14	17.64-29.37	24.71-38.9

Using the simulation procedure described in Chapter 3.6, data points are generated from a uniform distribution, U (0, 1) with the parameter p ranging from 25% to 35% and sample size ranging from 1,000 to 5,000. When prevalence (p) is set to 25%, and the sample size is set to 1,000, the first simulation results indicate that 261 samples are ELISA positive while 739 are positive. From these negative results, 11 individual samples were positive for HIV RNA thus giving a direct incidence estimate of 19.4% (Kline *et al*, 1989, CI: 18.53-20.28). The normal approximation gives a wider interval of (CI: 8.02-30.78), however the Binomial and Poisson intervals are very similar (CI: 9.72-34.51; 9.69-34.72). This similarity between the latter intervals is acknowledged and discussed by Lehmann (1999), for cases where n is large (n> 30). When prevalence is assumed to be 35% in N=2000 samples, the incidence is calculated to be 34.56% (CI: 33.7%-35.43%) using the McDougal approach as given in equation (2.17); and a wider interval of 23.27%-45.86% using the normal approximation approach. Results from Table 4.3 above show that there are similarities in the results of the confidence intervals obtained from the Binomial and Poisson approach.

Although the McDougal approach to the confidence intervals yields the shortest confidence interval; Cowling *et al*, (1999) note that these confidence intervals provide poor coverage of the limits and can sometimes give a misleading confidence in the estimate. Hauck (1991) suggests that confidence intervals should be based on exact binomial probability calculations instead.

4.3 POOLING RESULTS

For the pooling section of the study, samples were generated according to the simulation process in Chapter 3.6. For each generated sample, incidence is calculated using 2 main pooling algorithms:

1. Dual (2) stage pooling
2. Multistage (3 stage pooling).

Incidence estimates for groups of 1 means that all samples are tested individually, and this is taken to be our "single stage" pooling; results from which will be compared to the 2 and 3-stage algorithms.

4.4 DUAL STAGE ALGORITHM (2-STAGE POOLING):

A 2 stage algorithm is introduced to examine the efficiency of pooling on incidence estimates and to compare the confidence intervals from this scenario to the individual level (single stage) procedure.

For each sample size generated, a 2 stage pooling scheme is applied to that set. Two broad set-ups are considered, with prevalence varying from 25%-35%, while setting the true incidence to be 15% and 12 % respectively.

4.4.1 ASSUMING INCIDENCE OF 0.15

Table 4.4 shows the calculated estimates of incidence, its confidence interval and the number of tests required for 3 different sample sizes: N=1,000; N=2,000 and N=5,000.

Considering Table 4.4: With an assumed prevalence of 25% and a sample size of 1,000; incidence is calculated as the number of new infections divided by the number of antibody negative individuals. The antibody negative samples were obtained from our simulation as: number of antibody positive samples (258) subtracted from total number of samples (1000) which gives us 742 negatives. These 742 samples are then grouped into 7 pools of 100, with the 8th pool containing only 42 samples; as stipulated in the simulation procedure. The number of tests required at this stage is 8, since this is the number of pools available. Only one pool of 100 was found to be negative, thus 7 pools had to be tested individually to identify how many samples were positive. 6 pools containing 100 (600) were tested individually, with 1 incomplete pool of 42. Therefore, the total amount of tests required for this simulation was 8+600+42; which gives 650 tests in total. 14 individual samples were found to be RNA positive, given the parameters. Incidence for the 2-stage procedure is calculated as given in equation (2.5); with an estimate of 26.83 (CI: 24.40%-29.26%).

Table 4.4: Estimated 2-stage incidence estimates when true incidence=0.15

μ =28; I =0.15

Parameters	True Prevalence	Incidence	(CI)	Tests required
Pool size =100 N=1000	25	26.83	(24.40-29.26)	650
	30	25.12	(22.78-27.46)	595
	35	25.12	(22.71-27.54)	607
Pool size =100 N=2000	25	34.83	(32.41-37.25)	1396
	30	20.81	(19.47-22.15)	1215
	35	25.12	(23.44-26.81)	1214
Pool size =100 N=5000	25	26.17	(25.13-27.22)	3338
	30	28.33	(27.15-29.51)	3236
	35	31.27	(29.87-32.66)	3048
Pool size =20 N=1000	25	27.01	(23.70-30.32)	298
	30	42.63	(38.05-47.21)	375
	35	42.52	(37.81-47.24)	353
Pool size =50 N=1000	25	28.33	(25.73-30.93)	507
	30	32.26	(29.25-35.26)	502
	35	37.67	(34.11-41.24)	513

It should be noted that the confidence intervals obtained for the dual stage algorithm has been ‘corrected’ for pooling as suggested in equation 2.13.

We observe that when sample size is held constant at 1,000 and pool size varied, the differences in incidence estimates are very similar when the true prevalence is at 25%. For instance, when the pool size is 100, 20 and 50 the incidence estimates are 26.83% (95% C.I 24.40-29.26), 27.01% (95% C.I 23.70-30.32) and 28.33% (95% C.I 25.73-30.93) respectively.

Whereas, given the same specification and the true prevalence of either 30% or 35 %, results indicate that the variation in the incident estimates become more prominent. In general, incidence estimates tend towards the true value with an increase in pool sizes.

With sample sizes varied between 1000 and 5000, μ =28; I =0.15 and pool size =100, it can be seen from Table 4.4 that a moderate sample size of 2000 and a true prevalence value of 30 % gave the lowest incidence value and the highest value was obtained for the case where prevalence was 25% with the same sample size. Otherwise the incidence values are fairly consistent even with sample size varied. Confidence intervals tend to be relatively close to the true incidence value unlike the estimates in Table 4.3.

4.4.2 ASSUMING INCIDENCE OF 0.12

The estimates in Table 4.5 show a considerable difference when true incidence is set at 0.15 compared to 0.12.

Table 4.5: Estimated 2-stage incidence estimates when true incidence=0.12

μ =28; I =0.12

Parameters	True Prevalence	Incidence	CI	Tests required
Pool size =100 N=1000	25	17.95	(16.34-19.55)	547
	30	25.12	(22.77-27.47)	591
	35	25.12	(22.70-27.54)	607
Pool size =100 N=2000	25	26.00	(24.34-27.66)	1293
	30	25.12	(23.48-26.76)	1213
	35	25.12	(23.43-26.81)	1214
Pool size =100 N=5000	25	21.87	(21.01-22.73)	3138
	30	25.48	(24.44-26.52)	3136
	35	30.89	(29.51-32.26)	3033
Pool size =20 N=1000	25	22.79	(19.79-25.80)	257
	30	32.87	(28.99-36.76)	315
	35	38.91	(34.46-43.37)	333
Pool size =50 N=1000	25	23.67	(21.41-25.93)	454
	30	32.26	(29.24-35.27)	498
	35	30.37	(27.42-33.32)	463

When the true incidence is set at 12%, sample size held constant at 1000, the lowest estimate of incidence is observed when the true prevalence is 25% and pool size is 100. Estimates of incidence when samples were pooled in groups of 20 and 50 were very similar compared to when they were grouped into pools of size 100 for the same sample size (1000). Again we observe that the differences in incidence estimates are minimal when the prevalence is at 25% with varying pool sizes at a sample size of 1000.

From Table 4.5, incidence estimates are the same when the true prevalence is 30% or 35%, with a varying sample size of 1000 and 2000. However, when the sample size increases to 5000, there are differences between the estimates. In proportion to sample size, the total numbers of tests required are least for pool sizes of 20 and 50 when sample size is constant at 1000. The values of these estimates are lower than when the incidence was set at 0.15.

4.5 MULTI STAGE ALGORITHM (3-STAGE POOLING)

The differences in estimates for between the 1 stage (individual samples), 2 stage and 3 stage algorithms are examined and discussed. Similarities in their confidence intervals are also mentioned and highlighted.

Table 4.6: Estimated 3 Stage incidence estimates when pool size=100:20:5

	True Prevalence	Incidence	CI	No of tests required Multi Stage
Sample size=1000; I=0.12	25%	20.01	(8.27-31.74)	130
	30%	32.03	(16.52-47.54)	169
	35%	36.39	(19.32-53.45)	174
Sample size=1000,I=0.13	25%	21.86	(9.59-34.13)	135
	30%	34.09	(18.09-50.09)	178
	35%	38.60	(21.02-56.18)	183
Sample size=1000, I=0.15	25%	25.58	(12.31-38.85)	155
	30%	40.31	(22.90-57.72)	197
	35%	43.12	(24.51-61.72)	197

Table 4.6 displays the results of the multistage algorithm for estimating incidence. Taking the scenario of assumed prevalence of 25% and a sample size of 1,000; the number of positive samples on the simulated ELISA test was 261, while the negative samples were 739. The negative sample (739) samples are then grouped into 7 pools of 100, with the 8th pool containing only 39 samples; since it was incomplete. This step will be regarded as stage 1 where 7 pools of 100 samples each are formed. Results show that 6 of the 7 pools are positive. Thus in the next stage, the 600 samples are pooled into groups of 20, giving 20 pools of 30 samples each. It should be noted that the 39 extra samples from stage 1 are not included in the next stage. They will be retested alongside the other (if any) incomplete samples at the end of the procedure. Results from the 2nd stage show that 10 out of the 20 pools are positive and it is these 200 samples that will be tested again in the 3rd and final stage. In stage 3,

200 samples are divided into 40 groups of 5 and each of the 40 groups is tested. 11 of these pools are positive, thus 55 samples were each tested individually. Therefore, the total amount of tests required for this simulation was 8+27+40+55; which gives 130 tests in total. Incidence is calculated as given in equation (2.10); with an estimate of 20.01 (CI: 8.27%-31.74%).

From the results, we observe that the lower the true prevalence, the lower the number of tests required in estimating incidence. Even though the incidence estimates fluctuate slightly, the lower bounds of the confidence interval for low prevalence (25%) captures the value of the true incidence.

4.6 EFFICIENCY OF POOLING ALGORITHMS (2 STAGE AND 3 STAGE)

For given values of prevalence (p), the optimally efficient 2-stage and 3-stage procedure is determined by the value of n that minimizes equation (3.1) and (3.2) above. The relative efficiency and accuracy of the pooling strategies for assumed prevalence rates are demonstrated in Table 4.7. The efficiency of a pooling algorithm is defined as the expected number of tests required per individual specimen evaluated, ignoring confirmatory retesting of individual positive specimens (Westreich *et al*, 2008). Therefore an efficiency of less than '1' indicates that the pooling algorithm will require a lesser amount of tests on the average, than individual tests.

Figure 4.1 shows the expected number of tests per individual sample at various group sizes ranging from 10 to 500. A group size of '1' represents individual testing. The shapes of the relative cost curve vary for the different prevalence rates above. For a prevalence rate of 0.01% and a group size of 50, we will only require 41% as many tests as individual testing for a 2 stage algorithm and 14% of tests for a 3 stage algorithm. When the prevalence is at 0.1%, the 2-stage algorithm shows a 6% decrease for group sizes of 20.

For group sizes of 200, with the same prevalence rates of 0.01% and 0.1%, S2 shows 19 % efficiency compared to individual testing while S3 shows 3% efficiency, respectively. When the prevalence is higher, at 25% for instance, both the 2-stage and

3-stage algorithms show 100% efficiency for groups of 10 and 20. We also note that as prevalence increases, efficiency between both algorithms show substantial differences.

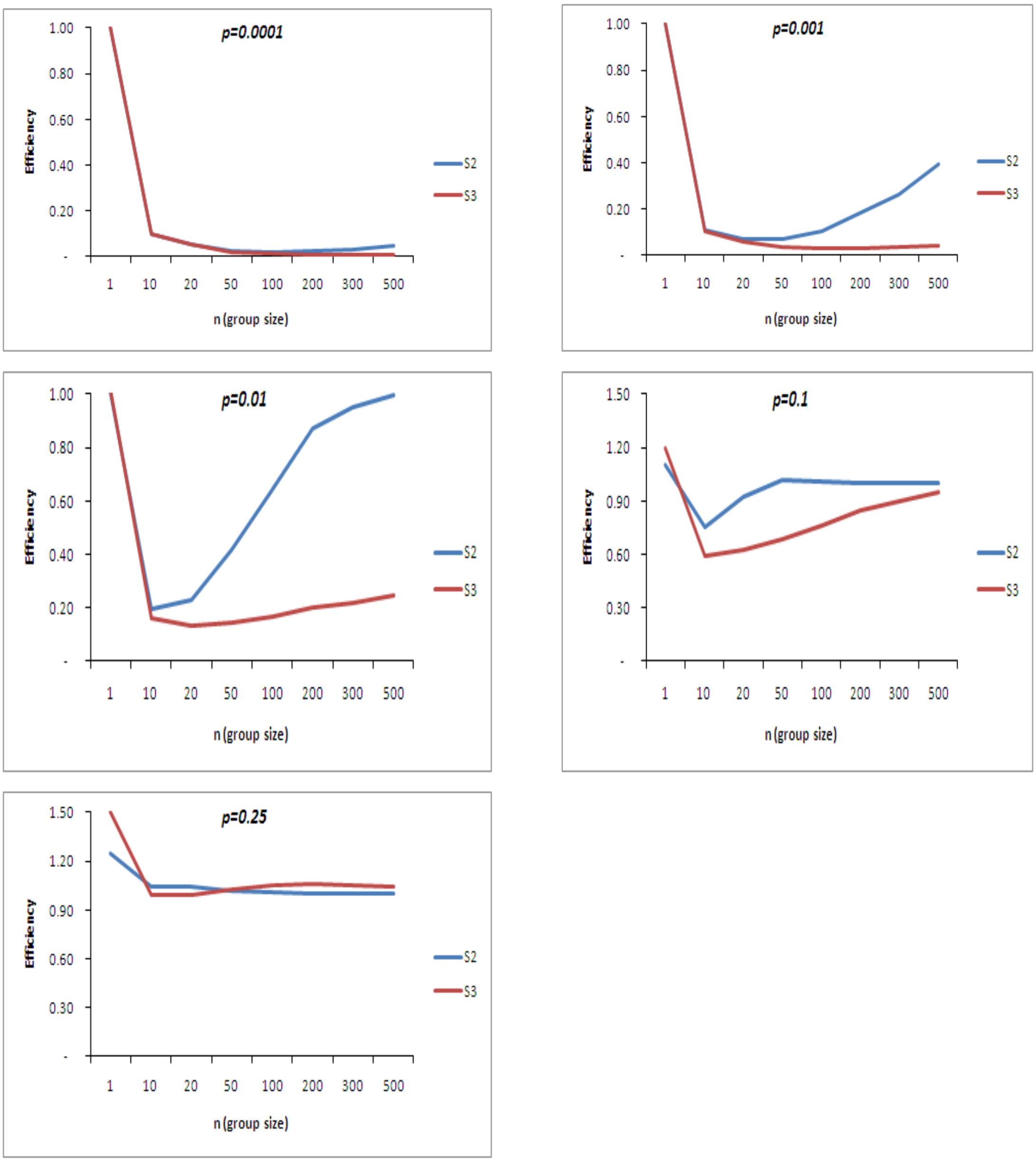


Figure 4.1: **Expected number of tests for 2 stage and 3 stage algorithms at given prevalence rates. S2 is a 2-stage "Dorfman Algorithm" while S3 is the 3-stage /Multistage Algorithm.**

The 3-stage algorithm consistently shows a lower efficiency across all the given prevalence rates showing that the more stages involved the more significant the results. As group sizes increase, the expected numbers of tests decrease.

The asymptotic relative efficiency (ARE) of $\hat{p}$ after k stages relative to individual testing of each specimen is computed. The asymptotic efficiency is defined as the ratio of the variance with individual testing $\frac{p(1-p)}{N}$ divided by the variance with multistage and is given by:

$$\frac{c_k p(1-p)^{c_k -1}}{(1-(1-p)^{c_k})} \qquad (4.1)$$

Table 4.7 illustrates the relative efficiency for an arbitrary range of prevalence values after each stage of a multistage pooling study, given a sample of N=1000 and initially sub divided groups of 100.

Table 4.7: Asymptotic efficiency of $\hat{p}$ after 3 stages relative to individual testing

		Prevalence			
Stage	Pool size	0.0005	0.01	0.02	0.04
1	100	0.98	0.58	0.31	0.07
2	10	1	0.96	0.91	0.83
3	1	1	1	1	1

At the end of the stage 2, if p=0.0005 $\hat{p}$ is nearly as efficient as individual testing as shown in the table (ARE=0.98). When p =0.01 the multistage estimator is about half as efficient as individual testing (ARE=0.53).

4.7 COST COMPARISON OF POOLING ALGORITHMS

One major advantage of pooling studies over individual testing is cost savings. The cost of administering an RNA test to an individual is $119 (R952 at the exchange rate of R8 to $1). Cost is calculated directly and does not take into account other costs like the actual cost of pooling each blood sample, value of the cost kit, labour and cost of retesting the pool if positive.

Tables 4.8, 4.9 and 4.10 show the cost savings associated with each pooling scheme when compared to individual testing of all samples. The comparison is done for different sample sizes, holding the pool sizes constant.

Table 4.8: Cost savings when sample size =1000

N=1000, 2 stage k=100, 3 stage k=100:20:5			
Prevalence	25%	30%	35%
Cost to test all specimens	R 703,528	R 651,168	R 613,088
Cost to test 2-stage	R 520,744	R 562,632	R 577,864
Cost to test 3-stage	R 123,760	R 160,888	R 165,648
Cost reduction-2stage vs ALL	26%	14%	6%
Cost reduction-3stage vs ALL	82%	75%	73%
Cost reduction-3stage vs 2 Stage	76%	71%	71%

Table 4.8 shows the results of cost savings of direct estimation compared to two pooled scenarios, when sample size is 1000. As expected, multistage pooling shows a considerable amount of savings in cost compared to the 2 stage pooling, and compared to individual testing of all specimens. At a true prevalence of 25%, there is an 82% savings on cost compared to individual testing, and a 76% savings when comparing multistage to 2-stage pooling. For the same parameters, there is only a 26% reduction in cost when comparing the 2 stage pooling to individual testing of all specimens.

Table 4.9: Cost savings when sample size=2000

N=2000, 2 stage k=100, 3 stage k=100:20:5			
Prevalence	25%	30%	35%
Cost to test all specimens	R 1,407,056	R 1,331,848	R 1,256,640
Cost to test 2-stage	R 1,230,936	R 1,154,776	R 1,155,728
Cost to test 3-stage	R 277,032	R 295,120	R 317,016
Cost reduction-2stage vs ALL	13%	13%	8%
Cost reduction-3stage vs ALL	80%	78%	75%
Cost reduction-3stage vs 2 Stage	77%	74%	73%

Table 4.10: Cost savings when sample size =5000

N=5000, 2 stage k=100, 3 stage k=100:20:5			
Prevalence	25%	30%	35%
Cost to test all specimens	R 3,596,656	R 3,406,256	R 3,140,648
Cost to test 2-stage	R 2,987,376	R 2,985,472	R 2,887,416
Cost to test 3-stage	R 592,144	R 295,120	R 317,016
Cost reduction-2stage vs ALL	17%	12%	8%
Cost reduction-3stage vs ALL	84%	91%	90%
Cost reduction-3stage vs 2 Stage	80%	90%	89%

CHAPTER 5 : SUMMARY AND CONCLUSIONS

5.1 SUMMARY

Estimates of HIV incidence are traditionally obtained from cohort studies which are time consuming and cost prohibitive. This report focused on an alternative to estimating HIV incidence from a cross-sectional sample without the disadvantages associated with the cohort study.

Chapter 1 gives an introduction of the epidemic in South Africa and an overview of the incidence and prevalence which are measures of the magnitude and rate of new infections of the disease. In Chapter 2, a literature review of the methods of estimating incidence is provided, along with an introduction to the pooling technique. Chapter 3 discusses the operating characteristics, pooling assumptions, the simulation algorithm used and the parameters used in the simulation model. In Chapter 4, estimates of incidence are calculated by direct estimation and using varying pool sizes and sample sizes. The accuracy of pooling blood samples was discussed along with the cost savings associated with different pooling strategies.

Incidence estimates calculated using the relationship between prevalence and mean duration of recent infection are compared to estimates derived from group testing procedures.

The results show that estimates from the pooled testing procedures are similar to, but more precise than the direct testing approach. In Table 4.3 for instance, with a sample of 1000, and a given prevalence of 25%; incidence was calculated as 19.4% (CI: 9.72%-34.51%). Using the same criteria the 2-stage algorithm yielded a 17.95% estimate (CI: 16.34%-19.55%); while the 3-stage algorithm resulted in a 20.01% incidence estimate (CI: 8.27%-31.74%).

For pooled testing, the findings suggest that for same number of samples, pooled testing provides a decrease in the average number of tests required per specimen, compared to individual testing. This is reflected in the charts in Figure 4.1, which shows gains in efficiency for both 2-stage and 3-stage algorithms, compared to individual testing. The findings also reveal that the 3-stage algorithm consistently offers a lower efficiency compared to its 2-stage counterpart; implying that as group sizes increase, the expected numbers of tests decreases, thereby saving time and money.

To further emphasize the benefits of pooling, a cost comparison of the different strategies is highlighted in the results. Tables 4.7 to 4.9 show that in terms of cost, there is up to 80% reduction in the cost of testing going from individual testing to multiple stage testing; and up to 70% reduction in cost going from single stage testing to multiple stage testing. These results just further show that there is substantial cost savings involved in pooled testing, regardless of the number of stages involved.

5.2 CONCLUSION

Estimating HIV incidence from methods like cohort studies and back calculation is currently difficult to perform. The findings from this study show that laboratory methods like the pooled testing procedure holds great advantages over either of the above mentioned methods. The precision of the estimates obtained with both pooled testing algorithms is shown to be better than individual testing, as well as being closer to the true incidence. This conclusion is further reinforced by the narrow width of the confidence intervals in both pooling algorithms compared to individual testing. Findings suggest that incidence estimates derived from pooled testing are precise, cost saving and time saving compared to individual testing

5.3 RECOMMENDATIONS

In the current study, one of the assumptions applied was that the sensitivity of the test for a pool is approximately the same as it is for an individual sample. It would be of interest to introduce the implication and effect of test errors into the study like the modeling of sensitivity and specificity as a function of pool size. This will give an indication as to whether more precise estimates of incidence can be derived, and will examine the issue of dilution effects more closely.

Further investigations or research into the application of the pooling method to estimate incidence should include a matrix based approach, mentioned briefly in Chapter 2.3.1. The comparison of this method to other multistage pooling strategies can be undertaken to determine which is more efficient and optimal.

Data on the incidence of HIV in South Africa, and Africa in general is still limited and insufficient. However, with an introduction of pooled testing strategies, the disadvantages can be minimized significantly. The integration of pooled studies into the estimation of HIV incidence in South Africa will be cost effective thus allowing money saved to be spent in other areas of researching the HIV Virus. Pooled testing will also enable the monitoring of the virus to be feasible and to limit the development and transmission of HIV in resource constrained settings.

REFERENCES:

Abdool Karim, Q. and Abdool Karim, S.S. (1999), South Africa: Host to a new and emerging HIV epidemic, *Sexually Transmitted Infections*, 75, pp. 139-147.

Bacchetti, P., Segal, M.R., and Jewell, N.P. (1993), Back calculation of HIV Infection Rates, *Statistical Science*, 8 (2), pp. 82-101.

Beyrer, C., Brookmeyer, R., Natpratan, C., Kunawararak, P., Niraroot, V., Palapunya, P., Khamboonruang, C., Celentano D. and Nelson K. (1996), Measuring HIV-1 Incidence in Northern Thailand: Prospective Cohort Results and Estimates Based on Early Diagnostic Tests. *Journal of Acquired Immune Deficiency Syndromes*, 12 (5), pp. 95-499.

Brookmeyer, R. (1991), Reconstruction and Future Trends of the AIDS Epidemic in the United States, *Science*, 253, pp. 37-42.

Brookmeyer, R. (1996), AIDS, Epidemics and Statistics, *Biometrics*, 52, pp. 781-796.

Brookmeyer, R. (1997), Accounting for follow-up bias in estimation of Human Immunodeficiency Virus Incidence Rates, *Journal of the Royal Statistical Society*, 160 (1), pp. 127-140.

Brookmeyer, R. (1999), Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence, *Biometrics*, 55, pp. 608-612.

Brookmeyer, R. and Quinn, T. (1995), Estimation of Current Human Immunodeficiency Virus Rates from a Cross-Sectional Survey Using Early Diagnostic Tests, *American Journal of Epidemiology*, 141(2), pp. 166-172.

Brookmeyer, R., Quinn, T., Shepherd M., Mehendale S., Rodrigues J. and Bollinger R. (1995), The AIDS Epidemic in India: A New Method for Estimating Current Human Immunodeficiency Virus (HIV) incidence rates, *American Journal of Epidemiology*, 142 (7), pp. 709-713.

Cleghorn, F.R., Jack, N., Murphy, J.R., Edwards, J., Mahabir, B., Paul, R., O'Brien, T., Greenberg, M., Weinhold, K., Bartholomew, C., Brookmeyer, R., and Blattner, W.A. (1998), Direct And Indirect Estimates of HIV-1 incidence in a High Prevalence Population, *American Journal of Epidemiology*, 147, pp. 834-839.

Chen, C.L. and Swallow, W.H. (1990), Using Group Testing to Estimate a Proportion, and To Test the Binomial Model, *Biometrics*, 46 (4), pp. 1035-1046.

Cole, S.R., Chu, H., Brookmeyer, R. (2006), Confidence Intervals for Biomarker-based Human Immunodeficiency Virus Incidence Estimates and Differences using Prevalent Data, *American Journal of Epidemiology*, 165, pp. 94-100.

Constantine, N. (2001), HIV Antibody Assays. HIV InSite Knowledge Base Chapter. http://hivinsite.ucsf.edu/InSite?page=kb-02&doc=kb-02-02-01.

Cowling, D.W., Gardner, I.A. and Wesley, J.O. (1999), Comparison of Methods for Estimation of Individual Level Prevalence based on Pooled Samples, *Preventive Veterinary Medicine*, 39, pp. 211-225.

Department of Health (2007), *The National HIV and Syphilis Prevalence Survey in South Africa, 2007.* Pretoria: South African Department of Health; 2008.

Dorfman, R. (1943), The Detection of Defective Members of Large Populations, *The Annals of Mathematical Statistics*, 14 (4), pp. 436-440.

Emmanuel, J. C., Bassett, M. T., Smith, H. J., Jacobs, J. A. (1988), Pooling of sera for human immunodeficiency virus (HIV) testing: An economic method for use in developing countries, *Journal of Clinical Pathology*, 41, pp. 582–585.

Gilbert, L. and Walker, L. (2002), HIV/AIDS in South Africa: An Overview, *Cad. Saúde Pública*, 18(3), pp. 651-660.

Gilgen, D., Campbell, C., Williams, B., Taljaard, D. and Macphail, C. (2000), *The Natural History of HIV/AIDS in South Africa: A Biomedical and Social Survey in Carletonville.* Johannesburg: Council for Scientific and Industrial Research.

Gouws, E., Williams, B.G., Sheppard, H.W., Barryett, E., and Abdool Karim, S. (2002), High Incidence of HIV-1 in South Africa Using a Standardised Algorithm for Recent Seroconversion, *Journal of Acquired Immune Deficiency Syndromes*, 29(5), pp. 531-535.

Grimes, D.A., and Schulz, K.F. (2002), An overview of clinical research: the lay of the land, *The Lancet*, 359, pp. 57-61.

Gupta, D. and Malina, R. (1999), Group Testing in Presence of Classification Errors, *Statistics in Medicine*, 18, pp. 1049-1068.

Hauck, W.W. (1991), *Confidence intervals for Seroprevalence Determined from Pooled Sera*, Biostatistics Technical Report #5, pp. 1-11.

Hecht, M.F., Busch, M.P., Rawal, B., Webb, M., Rosenberg, E., Swanson, M., Chesney, M., Anderson, J., Levy, J., and Kahn, J.O. (2002), Use of Laboratory Tests and Clinical Symptoms for Identification of Primary HIV Infection, *Journal of Acquired Immune Deficiency Syndromes*, 16, pp. 1119-1129.

Hung, M. and Swallow, W.H. (2000), Use of Binomial Group Testing in Tests of Hypotheses for Classification or Quantitative Covariables, *Biometrics*, 56(1), pp. 204-212.

Janssen, R.S., Satten, G.A., Stramer, S.L., Rawal, B.D., O'Brien, T.R., Weiblen, B.J, Hecht, F.M., Jack, N., Cleghorn, F.R., Kahn, J.O., Chesney, M.A., Busch, M.P. (1998), New Testing Strategy to Detect Early HIV-1 Infection for Use in Incidence

Estimates and for Clinical and Prevention Purposes, *Journal of the American Medical Association*, 280(1), pp. 42-48.

Johnson, N. L., Kotz, S. and Wu, X. (1991), *Inspection Errors for Attributes in Quality Control*. New York: Chapman and Hall Ltd.

Johnson, W.O., and Gastwirth, J.L. (2000), Dual Group Screening, *Journal of Statistical Planning and Inference*, 83, pp. 449-473.

Kamanga, G., Thumbi, P., Nkhoma, M., Manamela, P., Bogoshi, M., Latka, M., Martinson, F., Karim, S.S.A., Kumwenda, J., Rees, H., Churchyard, G., McCauley, M., Gay, C., Cohen, M.S. (2008), High prevalence of acute HIV infection in Sub-Saharan Africa: a cross-sectional, multi-centre screening study, *Abstract XVII International AIDS Conference, Mexico City*.

Kaplan, E.H., and Brookmeyer, R. (1999), Snapshot Estimators of Recent HIV incidence Rates, *Operations Research*, 47(1), pp. 29-37.

Kaplan, E.H., Kedem, E., and Pollack, S. (1998), HIV Incidence in Ethiopian Immigrants to Israel, *Journal of Acquired Immune Deficiency Syndrome*, 17, pp. 465-469.

Kennedy, N.L. (2004a), Multistage Group Testing Procedure (Group Screening), *Communications in Statistics*, 33 (3), pp. 621-637.

Kennedy, N.L. (2004b), Testing For the Presence of Disease by Pooling Samples, *Australian and New Zealand Journal of Statistics*, 46(3), pp. 383–390.

Kim, H., Hudgens, M.G., Dreyfuss, J.M., Westreich, D.J., Pilcher, C.D. (2007), Comparison of Group Testing Algorithms for Case Identification in the Presence of Test Error, *Biometrics*, 63, pp. 1152-63.

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger. S., and Quinn, T.C. (1989), Evaluation of Human Immunodeficiency Virus Seroprevalence in Population Surveys Using Pooled Sera, *Journal of Clinical Microbiology*, 27(7), pp. 1449-1452.

Kral, A.H., Lorvick, J., Gee, L., Bacchetti, P., Rawal, B., Busch, M., and Edlin, B.R. (2003), Trends in Human Immunodeficiency Virus Seroincidence among Street-Recruited Injection Drug Users in San Francisco, 1987-1998, *American Journal of Epidemiology*, 157, pp. 915-922.

Le Corfec, E., Le Pont, F., Tuckwell, H.C., Rouzioux, C., and Costagliola, D. (1999), Direct HIV Testing in Blood Donations: Variation of the Yield with Detection Threshold and Pool Size, *Transfusion*, 39, pp. 1141-1144.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Le Vu, S., Pillonel, J., Semaille, C., Bernillon, P., Le Strat, Y., Meyer,L., and Desenclos, J.C. (2008), Principles and uses of HIV incidence estimation from recent infection testing- *A review. Eurosurveillance,* 13 (7–9), pp. 11-16.

Litvak, E., Tu, X.M., and Pagano, M. (1994), Screening for the Presence of a Disease by Pooling Sera Samples, *Journal of the American Statistical Association*, 89(426), pp. 424-431.

Maherchandani, S., Munoz-Zanzi, C.A., Patnayak, D.P., Malik, Y.S. and Goyal, S.M. (2004), The effect of pooling sera on the detection of avian pneumovirus antibodies using an enzyme-linked immunosorbent assay test, *Journal of Veterinary Diagnostic Investigation*, 16, pp. 497–502.

McDougal, J.S., Pilcher, C.D., Parekh, B.S., Gershy-Damet, G., Branson, B.M., Marsh, K. and Wiktor, S.Z. (2005), Surveillance for HIV-1 incidence using tests for recent infection in resource–constrained countries, *Journal of Acquired Immune Deficiency Syndromes*, 19, S25-S30.

Munoz-Zanzi, C.A., Johnson, W.O., Thurmond, M.C., and Hietala S.K. (2000), Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle, *Journal of Veterinary Diagnostic Investigation*, 12, pp. 195–203.

Nelson, L.J. (2005), *Current Status and Future of HIV Surveillance*. Joint working group meeting, HIV and Drug Resistance Surveillance and Testing, Versailles, October 2005.

Parekh, B.S. and McDougal, J. S (2001), New Approaches for Detecting Recent HIV-1 Infection, *AIDS Review*, 3, pp. 183-193.

Parekh, B.S. and McDougal, J.S. (2005), Application of laboratory methods for estimation of HIV-1 incidence, *Indian Journal of Medical Research*, 121, pp. 510-518.

Pettifor, A., Macphail, C., Rees, H. And Cohen, M. (2008), HIV and Sexual Behaviour among Young People: The South African Paradox, *Sexually Transmitted Diseases*, 35(10), pp. 843–844.

Philippe, P. (2001), Density Incidence and Cumulative Incidence: A Fundamental Difference. *The Internet Journal of Internal Medicine*, 2, pp. 1-3.

Pilcher, C. D., McPherson, J. T., Leone, P. A., Smurzynski, M., Owen-O'Dowd, J., Peace-Brewer, A. L., Harris, J., Hicks, C. B., Eron, J. J. and Fiscus, S. A. (2002), Real-time, universal screening for acute HIV infection in a routine HIV counselling and testing population, *Journal of the American Medical Association*, 288, pp. 216-221.

Pilcher, C. D., Price, M. A., Hoffman, I. F., Galvin, S., Martinson, F. E., Kazembe, P. N., Eron, J. J., Miller, W. C., Fiscus, S. A., and Cohen, M. S. (2004), Frequent detection of acute primary HIV infection in men in Malawi, *Journal of Acquired Immune Deficiency Syndromes*, 18, pp. 517–524.

Pilcher, C. D., Fiscus, S. A., Nguyen, T. Q., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J. O., McPherson, J. T., Stalzer, B., Hightow, L., Miller, W. C., Eron, J. J. and Cohen, M.S. (2005), Detection of acute infections during HIV testing in North Carolina, *New England Journal of Medicine*, 352, pp. 1873–1883.

Quinn, T.C., Brookmeyer, R., Kline, R., Shepherd, M., Paranjape, R., Mehendale, S., Gadkari, D.A., and Bollinger, R. (2000), Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute Primary HIV-1 infection and estimate HIV incidence, *Journal of Acquired Immune Deficiency Syndromes*, 14, pp. 2751-2757.

Ramjee G., Abdool Karim S. S. and Sturm A. W. (1998), Sexually transmitted infections among sex workers in KwaZulu-Natal, South Africa. *Journal of Sexually Transmitted Diseases*, 25, pp. 346-349.

Rosenberg, E. (2002), Backgrounder: Recognizing and diagnosing primary HIV infection. *Research Initiative Treatment Action (RITA)*, 7(2), pp. 5-10.

Rutherford, G.W., Schwarcz, S.K., and McFarland, W. (2000), Surveillance for Incident HIV Infection: New Technology and New Opportunities, *Journal of Acquired Immune Deficiency Syndromes*, vol. 25, pp. S115-S119.

Salomon, J.A. and Murray, CJL. (2001), Modelling HIV/AIDS epidemics in sub-Saharan Africa using seroprevalence data from antenatal clinics, *Bulletin of the World Health Organization*, 79(7) pp. 596-607.

SAS® Institute Inc. (2004), *SAS User's Guide, Version 9.1*, Cary, NC: SAS Institute Inc.

Saraniti B.A. (2006), Optimal pooled testing, *Health Care Management Science*, 9, pp. 143–149.

Satten, G.A., Janssen, R., Busch, M.P., and Datta, S. (1999), Validating Marker-Based Incidence Estimates in Repeatedly Screened Populations, *Biometrics*, 55, pp. 1224-1227.

Sergeant, E. and Toribio, J. (2004), *Estimation of Animal-Level Prevalence from Testing of Pooled Samples*, A report prepared for The Australian Bio security Cooperative Research Centre for emerging infectious diseases.

Shisana, O. and Simbayi, L. (2002), *Nelson Mandela/HSRC Study of HIV/AIDS: South African National HIV Prevalence Behavioural Risks and Mass Media Household Survey Cape Town*, South Africa: Human Sciences Research Council.

Shisana, O., Rehle, T., Simbayi, L.C., Parker, W., Zuma, K., Bhana, A., Connolly, C., Jooste, S. and Pillay, V. (2005), *South African National HIV Prevalence, HIV Incidence,Behaviour and Communication Survey Cape Town,* South Africa: Human Sciences Research Council Press.

Stevens, W., Akkers, E., Myers, M., Motloung, T., Pilcher, C. and Venter, F. (2005), High prevalence of undetected, acute HIV infection in a South African primary care clinic, abstract MoOa0108. *The Third IAS Conference on HIV Pathogenesis and Treatment. International AIDS Society,* Rio de Janeiro, Brazil.

Thompson, K. H. (1962), Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, 18, pp. 568-578.

UNAIDS (2008), *Report on the Global AIDS Epidemic*.

Wein, L.M. and Zenios, S.A. (1996). Pooled testing for HIV screening: Capturing the Dilution Effect, *Operations Research,* 44, pp.543–569.

Welz, T., Hosegood, V., Jaffar, S., Bätzing-Feigenbaum, J., Herbst, K. and Newell, M. (2007). Continued very high prevalence of HIV infection in rural KwaZulu-Natal, South Africa: A population based longitudinal study, *AIDS*, 21, pp 1467–1472.

Westreich, D.J., Hudgens, M.G., Fiscus, S.A., and Pilcher, C.D. (2008). Optimizing Screening for Acute Human Immunodeficiency Virus Infection with Pooled Nucleic Acid Amplification Tests, *Journal of Clinical Microbiology,* vol.46, No.5, pp1785-1792.

Williams, B.G., and Gouws, E. (2001). The Epidemiology of Human Immunodeficiency Virus in South Africa, *Philosophical Transactions of the Royal Society of London*, B., 356, pp. 1077-1086.

Williams, B.G., Gouws, E., Wilkinson, D. and Abdool Karim, S. (2001), Estimating HIV incidence rates from age prevalence data in epidemic situations, *Statistics in Medicine*, 20, pp. 2003-2016.

Xie, M., Tatsuoka, K., Sacks, J. and Young, S.S. (2001). Group testing with blockers and synergism, *Journal of the American Statistical Association* 96, pp92-102.

Zenios, S. A. and Wein, L. M. (1998). Pooled testing for HIV prevalence estimation: Exploiting the dilution effect, *Statistics in Medicine* 17, pp.1447–1467.