Incidence estimation and calibration from cross-sectional data of acute infection HIV-1 seroconvertors

A Research Report Presented

by

Eustasius Musenge

Submitted to the School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Masters of Science in Medicine in the Field of Biostatistics and Epidemiology.

June 2007

Johannesburg, South Africa

DECLARATION

I, Eustasius Musenge declare that this research report is my own work. It is being submitted for the degree of Masters of Science in Medicine in the Field of Biostatistics and Epidemiology in the University of Witwatersrand , Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

Signature:………………………………

This 25th Day of June 2007

Incidence estimation and calibration from cross-sectional data of acute infection HIV-1 seroconvertors

A Research Report Presented

by

Eustasius Musenge

Approved as to style and content by:

____
Supervisor: Edmore Marinda

____
Supervisor: Alexander Welte

____
Head School of Public Health : Sharon Fonn

DEDICATION

I dedicate this research report to my former Statistics students from the University of Botswana and all the people who encouraged me throughout the Msc Medicine (Biostatistics and Epidemiology) course. Sometimes I felt like giving up, but you stood with me spiritually, I just got more and more strength to go on and before I knew it was over.

ACKNOWLEDGMENTS

Many thanks go to my supervisors Edmore Marinda and Alexander Welte for all the academic and personal motivation and inspiration. I also would like to thank SACEMA for the funding that enabled the writing and full completion this research report. I am also thankful to the CAPRISA study team for allowing me to use their data in the analysis. It is not by might nor by power but by God’s Holy Spirit, I praise you Lord Jesus Christ for divine enablement.

ABSTRACT

Incidence estimation and calibration from cross-sectional data of acute infection HIV-1 seroconvertors.

May 2007
Eustasius Musenge

Masters in Medicine in the Field of Biostatistics and Epidemiology

Supervised by: Mr E Marinda and Dr A Welte

Background: The HIV-1 incidence (a very important measure used as a proxy for disease burden) can be estimated from a cross-sectional study. This incidence estimate has the advantage of reducing on costs and time, thus enabling more timely intervention; it is also ideal for developing nations. A common procedure used in making this estimate utilizes two antibody tests (Sensitive/Less sensitive tests). Due to the long window period of such tests (at least three months), persons classified as recently infected would have been infected more than three months prior to the test date. Detecting acute HIV-1 infection is very important since this is the most infectious stage of the disease. This research report explores a method of estimating incidence using an antibody test and a virological test, Polymerase Chain Reaction Ribonucleic Acid (PCR-RNA).The cross-sectional data used are from the Centre for the AIDS Programme of Research in South Africa (CAPRISA).

Methods: Actual follow-up cohort data from CAPRISA acute infection cohort (AIC), comprised of 245 sex workers, were used to estimate the incidence of HIV-1 using a PCR-RNA ,virology test based, incidence formula. The result obtained was compared to the incidence estimate obtained by the classical method of estimating incidence (prospective cohort follow-up). As a measure to reduce costs inherent in virological tests (PCR-RNA), multistage pooling was discussed and several pooling strategies simulations were proposed with their uncertainties. Point estimates and interval estimates of the window period, window period prevalence and incidence from cross-sectional study of the AIC cohort were computed.

Findings: The mean window period was 6.6 days 95% CI: (2.7 – 13.0). The monthly window period prevalence was 0.09423 percent 95 % CI: (0.0193 – 0.1865)%. The incidence from the prospective cohort follow-up was 5.43 percent 95% CI: (3.9 – 9.2) %. The incidence estimate from cross-sectional formulae was 5.21 percent 95% CI: (4.1– 4.6). It was also shown by use of simulations that an optimum pool sample size is obtained when at least half the samples are removed on every run.

Interpretation and recommendations: The PCR-RNA test is very sensitive at detecting acute HIV-1 infected persons. The incidence estimate from the cross-sectional study formulae was very similar to that obtained from a follow-up study. The number of tests needed can be reduced and a good estimate of the incidence can still be obtained. The calibration was not accurate since the samples used were small and the window period duration was too short, hence, it was difficult to extrapolate to the whole population. Further work still needs to be done on the calibration of the proposed incidence formulae as it could be a very useful public health tool.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

NOMENCLATURE

STARHS	Serologic Testing Algorithm for Recent HIV-1 Seroconversion
WHO	World Health Organization
HIV	Human Immunodeficiency Virus
AIDS	Acquired Immune Deficiency Syndrome
MLE	Maximum Likelihood Estimator.
RNA	Ribonucleic Acid
ELISA	Enzyme Linked Immunosorbent Assay
CAPRISA	Centre for the AIDS Programme of Research in South Africa
AIC	Acute HIV Infection Cohort
PCR	Polymerase Chain Reaction
CDC	Centre for Disease Control
env	Envelope
gag	Group associated antigen

CHAPTER 1 : INTRODUCTION

1.1 Introduction

About 922 million people (11 percent of the world's population) reside in Africa of which more than 25 million people are living with HIV/AIDS[1]. More than a fifth (5.5 million) of these people are from South Africa.[2] South Africa as a nation has the highest absolute number of HIV-infected persons in the world and a prevalence of 30.2%.[3] The nation is divided into nine provinces, one of which is KwaZulu-Natal, which has the highest population and HIV prevalence of 39.1% for antenatal attendees.

Several studies on HIV prevention are underway country-wide including studies evaluating the effect of male circumcision on HIV transmission and also studies evaluating the impact of microbicide gel use by women on HIV transmission. With millions of rands being spent towards HIV-related endeavours, the need for accurate statistics and greater understanding of the dynamics of the virus is critical. Prevalence (which is the number of cases of a disease that are present in a particular population at a given time) and incidence (which is the number of newly diagnosed cases during a specific time period) are two very important indicators of disease burden and spread, respectively. Another useful measure is the test-specific window period duration, which when accurately measured, will enable intervention early after infection (which is the most infectious period).[4] Scientific evidence gathered from South Africa, the region with the highest absolute number of HIV-infected persons in the world, will be very useful if reliably collected and well validated.

1.2 Statement of the problem and rationale

Acute HIV infection is the stage of disease progression during which HIV viral replication and shedding occur before detectable antibodies occur.[5] This is also the time when the viral load peaks in the blood and genitals,[6] which is also the most infectious period.[3] Estimation of acute HIV-1 incidence is essential in HIV prevention. The public health benefits are that persons with acute HIV infection can be counselled about risk reduction behaviours such as abstinence and safer sex to reduce secondary infections.[7] This has great benefits such as reduction in transmission to uninfected sexual partners, channelling of resources towards this most infectious group and early treatment of the acute HIV infected.[8]

The incidence estimates are important for the purposes of planning vaccine trials and disease monitoring and evaluation. The classical method of estimating the incidence within a research setting is based upon following an uninfected cohort over time until some are infected. This has several limitations, the most common are it is costly, the loss to follow-up and also difficulties with respect to distinguishing between those who were recently infected (HIV) and those long infected (AIDS). In order to strengthen the fight against HIV, there is need to detect places and persons with the highest levels of infectiousness in the right time period in order to implement public health interventions. A useful epidemiological indicator best suited for this purpose is the acute HIV-1 incidence.

1.3 Study Objectives

The aim of this study is to create a calibrating tool useful in estimating HIV incidence from recently infected persons in cross-sectional studies. The tool development will assess the estimation of the window period (duration between infection and testing antibody positive), window period prevalence and incidence estimation. The study will provide calibrated combinations of parameters (incidence, window period prevalence and window period). The 95% confidence intervals will also be given. The specific objectives are:

1) Estimation of window period for the acute infection cohort.
2) Computation of incidence from cross-sectional acute infection cohort (AIC) data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA).
3) Validation of optimal multistage pooling strategy using data simulations.

1.4 Literature review of methods for measuring incidence

Prevalence rates which are the most commonly reported statistic for HIV/AIDS include full-blown AIDS and recently infected individuals. This may not be very useful in tracking the progression of the disease or in providing an instantaneous state of the epidemic. Incidence rates which are integral in the design of vaccine efficacy studies, calculating sample sizes and allocation interventions are routinely measured from prospective follow-up cohort studies.[9] The following are several ways in which the incidence rates are estimated.

1.4.1 Observing seroconversions in a prospective follow-up study

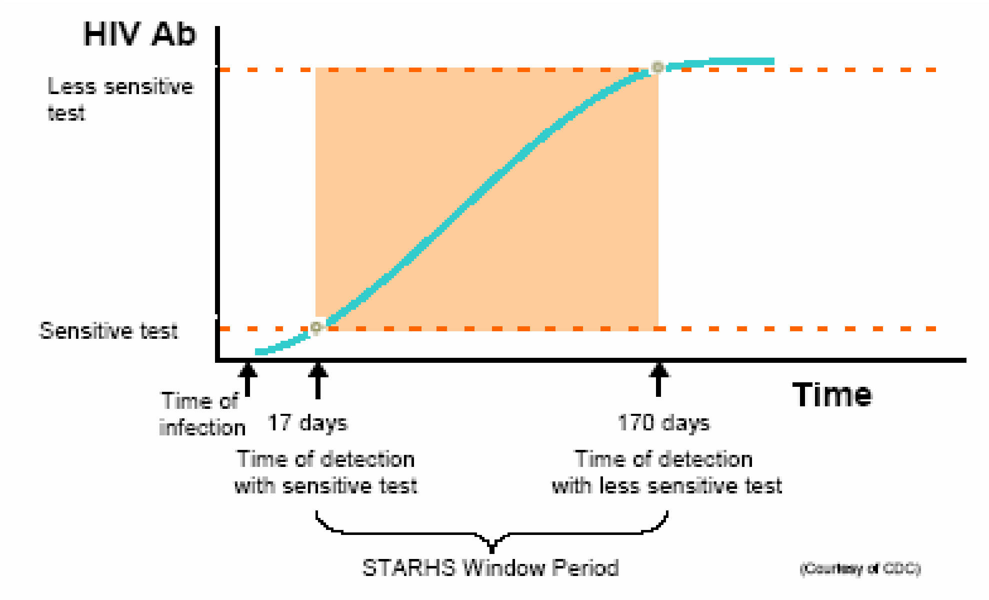
Incidence can be defined as the proportion of seronegative individuals who seroconvert during a defined period of observation. Estimating incidence is done by enrolling an HIV-negative population in a longitudinal or prospective cohort study and testing the participants at regular intervals for new HIV infections, thereby deriving an incidence rate (number of new infections per total number of person-years of follow-up).[10] Also, the proportion of positives identified in a cross-sectional study that has markers of recent infection is used for estimating incidence.

1.4.2 Identifying recent seroconverters from a cross-sectional sample using two HIV antibody tests of differing sensitivity for HIV antibodies

This method is described by the Centre for Disease Control (CDC) as the Serological Testing Algorithm for Recent HIV Seroconversion (STARHS).[11] It is also known as the "detuned assay" or the "Sensitive Assay/Less Sensitive Assay". This is a comparison of two tests on a single diagnostic specimen. The regular HIV antibody test that is used to diagnose HIV infection and a less sensitive version of the same test that only detects high levels of HIV antibodies.

The first test indicates whether the person is infected with HIV. If infected, the second, less-sensitive test can indicate whether or not the patient has a high level of HIV antibodies. Since a person's level of antibodies gradually increases in the early stages after infection, the result of the second test suggests whether they have been infected within a shorter (approximately 6 months or less) or longer time. However the probability of identifying an infected individual within six months of infection is a function of how often they go for testing.[12]

Figure 1-1. STARHS method to estimate stage of infection using a single diagnostic specimen.[6]



The STARHS approach uses the following formula to calculate the annual incidence rates (equation 1.2 is the tailor form of equation 1.1):

$$i = \frac{R(365.25/T_w)}{N_{neg} + R(365.25/T_w)} \times 100$$ ....Equation 1.1 [6]

$$i = \frac{365.25R}{N_{neg}T_w}\left\{1 + \frac{365.25}{N_{neg}T_w}\right\}^{-1} \times 100$$ .... Equation 1.2

Where R: recent infection (antibody negative), T_w: window period, N_{neg}: number of HIV seronegative. This approach was used by Parekh *et al.*[13]

In South Africa, another version of the formula was used in the annual survey for South African national HIV incidence, commissioned by the Nelson Mandela Foundation:[2]

$$i = \frac{R(365.25/T_w)}{N_{neg} + R/2(365.25/T_w)} \times 100 \text{ .... Equation 1.3}^{2}$$

$$i = \frac{365.25R}{N_{neg}T_w}\left\{1 + \frac{365.25}{2N_{neg}T_w}\right\}^{-1} \times 100 \text{ .... Equation 1.4}$$

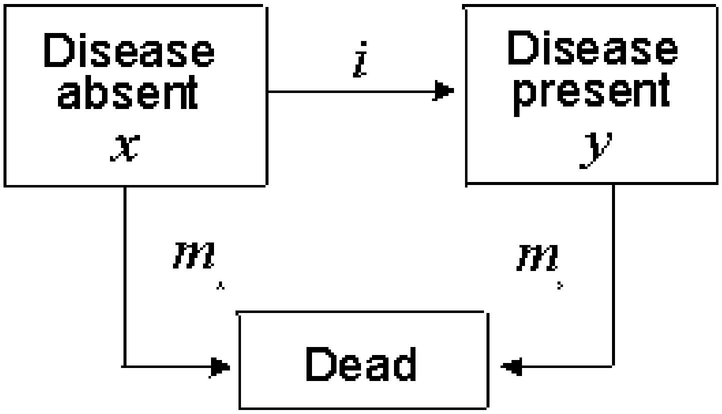
For the algebraic expansion ($(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + ...$ where *n* is negative or a fraction exponent and valid if and only if $-1 < x < 1$) to be possible, the above equations 1.2 and 1.4 $\frac{365.25R}{N_{neg}T_w} < 1$ and $\frac{365.25R}{2N_{neg}T_w} < 1$ (respectively) are needed, but this is not always possible in these formulae, e.g., when the window period is very small (less than 7 days). The assumption made on application of these formulae is that the incidence is constant throughout the year preceding the calculation. The formulae are globally calibrated to estimate incidence only when antibody-based assays are used to identify recently infected HIV-1 persons.[2]

1.4.3 Inferring incidence from serial cross-sectional surveys

With this method, incidence is indirectly estimated by the slope of the seroprevalence against time, assuming the population being surveyed remains representative over time.[14]

The figure below shows a two-state model of disease within a cohort. At a given age, *a,* *x(a)* is the number of people without the disease, *y(a)* is the number of people with the disease, *i(a)* is the incidence rate, and $m_x(a)$ and $m_y(a)$ are the mortality rates among those with and without the disease, and where *i* is the actual incidence rate based on the susceptible population.

Figure 1-2. Two state-deterministics model, source[15] *.*



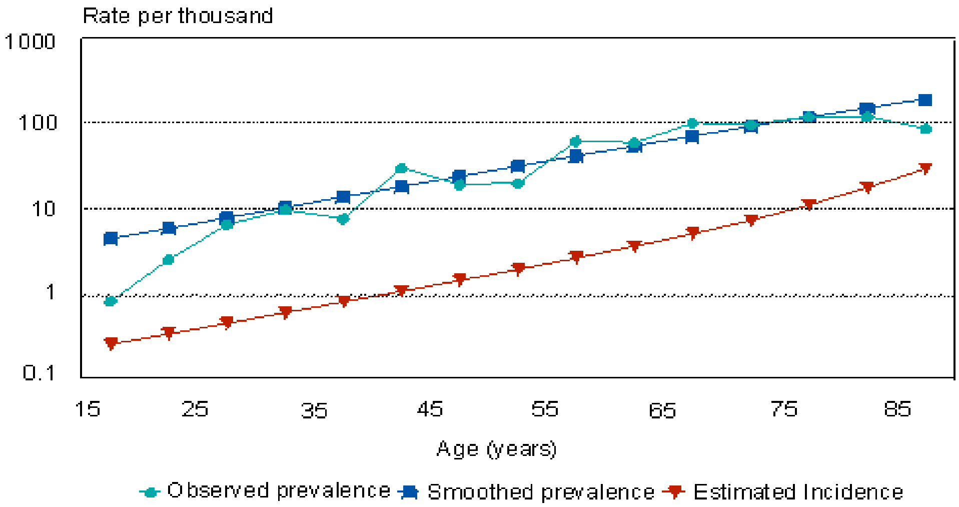
The age-specific prevalence of a disease can be obtained from population surveys, either by interview or by examination. The mortality of people with the disease can be obtained from following up the survey subjects, demographic surveillance surveys or from cohort studies.[15]

$$i(a) = \frac{p(a)N(a)}{\{1 - p(a)\} + \{[m_y(a) - m_x(a)]p(a)\}}$$ ....Equation 1.5 [15]

Where p(a) is the prevalence and N(a)is the total of the susceptible population.

The estimates of age-specific prevalence are usually ‘noisy’, it is necessary to smooth them, taking into account that *p(a)* must lie between 0 and 1 and, for most diseases, increases with age. To reduce the noise a suitable smoothing function is the logistic *ln*[*p(a)*/{1 - *p(a)*}]. Figure 1.4 shows the graphs obtained when this approach was used for a group of diabetic women in Canada.

Figure 1-3. Prevalence and incidence of diabetes in Canadian women, source[16].



The graph shows the effect of the smoothing technique and an increase in the estimated incidence with increasing age.

1.4.4 Use of surrogate marker for recent (age specific) infection

In this approach to estimate HIV incidence, the number of reported AIDS cases in the youngest age range of adult cases, ages 13-25, is used as a surrogate for recent trends in incidence. The justification for this approach is that the onset of sexual and drug-using risk behaviour in the teenage years (or later) leads to the inference of AIDS cases in this age group. Predominately those with a short incubation time from infection to AIDS reflect relatively recent infections (less than 5 years on average).[17] Also the AIDS-related mortality would be less significant in this younger cohort and however the incidence among the 18 year olds may differ from that of those in the fifties thus generalisability becomes an issue.

1.4.5 Back-calculation from reported AIDS cases

This approach estimates incidence or prevalence by use of a mathematical model called 'back calculation', which combines the available data on the number of reported AIDS cases and the incubation period distribution of AIDS (the mathematical function that estimates the probability of developing AIDS for each year following HIV infection) to derive how many HIV infections occurred during years past.[18] Back-calculation is done by use of the convolution equation:

$$a(t) = \int_{-\infty}^{t} I(s)F(t-s \mid s)ds \quad \text{...Equation 1.6}^{18}$$

Where *a(t)* is the expected cumulative number of AIDS cases diagnosed by time *t* , *I(s)* is the HIV infection rate at time *s* and *F(t|s)* is the probability of developing AIDS within *t* years of infection for those who were infected at time *s*. This approach uses information on *a(t)* and *F(t|s)* to estimate the infection rate *I(s)*.[19]

1.4.6 Using capture-recapture methods in serial surveys

The sixth method is a variant of the cross-sectional survey approach that uses 'capture-recapture', a method long used by biologists to study wildlife populations. It requires a unique identifier, of individuals included in repeated surveys, so that the seroconverters among those repeatedly tested can be identified.[20]

1.5 Estimating the window period for HIV primary infection

The window period is the interval during which an infected individual tests negative to an antibody test. Recently infected individuals may however be detected by virological assays such as Polymerase Chain Reaction Ribonucleic Acid (PCR-RNA) which also has a window period. A procedure known as incidence window period (IWP) is used among blood donors. This is derived from an epidemiological relation, **Prevalence=Incidence x Window** period under a steady state assumption on the infection dynamics. The IWP estimates the window period by[21]

$$\hat{\pi}_0 = \hat{I} \times \overline{w} = \frac{N}{\sum_{i=1}^{N} (t_{i(n_i+1)} - t_{i1})} \times \overline{w} \text{ ....Equation 1.7}$$

Where $\hat{\pi}_0$ is the prevalence, $\hat{I}$ is the incidence, $\overline{w}$ is the average length (duration) of HIV window period for the blood bank of interest, N is the total number of repeat donors and $(t_{i(n_i+1)} - t_{i1})$ is time between the donors consecutive donations.

1.6 Optimization of lab work in 'recent infection' prevalence studies

Pooling is a strategy used on biological specimens. Individual serum samples are grouped and randomly selected specimens are tested in each group (pool). Based on the outcome, the whole pool is classified as positive or negative. Two objectives of pooling biological specimens are to identify infected individuals and to estimate the prevalence (when it is low) of infection in the population at a lower cost than testing individual samples.[22] When the process is done several times, it is known as multistage pooling.

In multistage pooling the cohort is divided into equal-numbered pools which are then tested. Each positive pool identified in each stage is subdivided to smaller pools in the following stage and this is repeated up to the last pool of size one.

For the multistage pooling study after z stages of pooling algorithm, the incidence rate can be estimated by the following relation:

$$i=\frac{1}{T_w}\left[\left(1-\left(\frac{N_1}{N}\right)^{\frac{1}{S_Z}}\right)\right]$$ ……………………………………………..Equation 1.8 [24]

Where *N* is the sample size, S_1 is the size of the initial pool sample (thus S_Z is the size of the z th pool sample) and N_1 the number in the first stage, that is, $N=N_1.S_1$.

This equation uses concepts discussed by Brookmeyer et al. (2000) in obtaining the optimal pooling size. The major aim of pooling blood samples is to identify the number of positive infections *R* in a group of *N* samples using as few tests as possible (PCR runs in the case of the HIV-RNA test). For this to be possible, at least half of the samples on every run (stage) must be eliminated and an optimal initial pool size s_1. must be used (see appendix C).

1.7 Forthcoming discussions

Having discussed the different approaches of estimating incidence rates, there is still a gap to be filled for ideal methods applicable to developing countries and high-risk populations in which it is difficult to follow cohorts to identify seroconverters. This research report discusses a systematic approach to estimating the incidence and calibration, using results from an antibody test (ELISA) and a virological test (PCR-RNA). The following section discusses the methods employed in the study, design and analysis.

CHAPTER 2 : MATERIALS AND METHODS

2.1 Study design and sample description

The data which were used for this study were from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) in their phase II Acute Infection Cohort (AIC). The study design was a prospective observational cohort study conducted at the Doris Duke Medical Research Institute (MRI) at the Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, and the CAPRISA Vulindlela Research Facility. Study

participants with acute HIV infection were identified among female sex workers in KwaZulu-Natal; from participants in a Phase II/IIb microbicide trial in Durban; and from research cohorts in Vulindlela. A diagrammatic illustration of the screening, recruitment and enrolment process is shown in Appendix E.[23]

A total of 775 participants were assessed from which a cohort of 245 uninfected high-risk women was selected, where the prevalence among them was greater than 10% or incidence greater than 3%. The average age was 34.3 years (range 18-58) and the majority (78.8%, n=193) identified themselves as sex-workers. The loss-to-follow-up rate was 15.1%.[23]

The 245 female sex workers were then assessed monthly between September 2004 and July 2006. In this particular study the participants who became HIV infected were detected soon after the infection, which made the data suitable for greater in-depth observation of early viral and CD4+ T cell bio-dynamics. During the follow-up, two tests were administered monthly: a virological test (HIV-RNA PCR) and the sensitive antibody test 'ELISA'. This enabled the estimation of a hypothetical window period, the actual duration between the last ELISA negative and first ELISA positive, which has an unknown distribution.

2.2 Testing protocol

The two tests administered were an antibody test (ELISA) and a virological test (RNA-PCR) in individual samples and pools. The ELISA assay assessed for serum binding antibodies in blood reacting with purified HIV proteins or peptides from *env* and *gag*.[23]

The COBAS AMPLICOR™ HIV-1 MONITOR Test, v1.5 (Standard or the Ultrasensitive, Roche Diagnostics) was used to measure viral loads. The test quantifies HIV-1 RNA over the range of 50-750 000 copies/ml and has a specificity greater than 99.85% for quantification of HIV-1 Groups M subtypes A-G. A series of runs was performed to ensure reliability and reproducibility including inter-laboratory reproducibility.[23]

For detection of HIV-1 RNA in pooled samples, the AMPLISCREEN™ HIV-1 Tests v1.5 (Roche Diagnostics) were used. A primary pool contained 24 or less samples and when found positive, was disaggregated into smaller secondary pools.[23]

2.3 Data analysis concepts

The incidence estimate from the CAPRISA longitudinal prospective follow-up cohort was computed and compared with the incidence estimate from a cross-sectional prevalence. Due to the expenses incurred in running PCR-RNA tests on individual samples, multistage pooling was used to reduce the number of laboratory tests and costs. Three pooling options will be discussed for the multistage pooling strategy. The statistical analysis was done using STATA 9.0 and Excel. The programming of different algorithms was carried out using MATLAB. The following section discusses the concepts backing the methods utilized for each of the objectives.

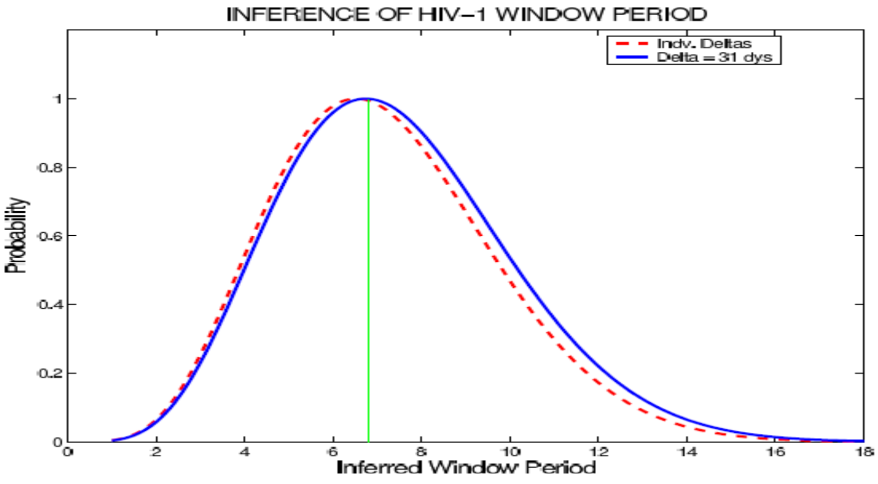
2.3.1 Inference of the window period duration

The estimation of the PCR-RNA window period was done using the interval between the first RNA positive and last RNA negative (which we will call delta Δ), multiplied by the probability of seeing an individual in the window period.

$T_w = \overline{\Delta} \times \hat{p}$ ……………………………………………………………Equation 2.1

Where T_w is the estimate of the window period, $\overline{\Delta}$ is the average of the interval between visits and $\hat{p}$ is the probability of seeing an individual in the window period. The window period can be estimated using either the individual deltas or average delta (as shown above). The latter approach is less involved and yields a very similar distribution to the former. Mwanga (2006)[24] obtained the following distributions for the individual and the average delta and a window period of 6.8 days.

Figure 2-1. Posterior density window period for individual and average delta.[23]



2.3.2 Modelling of relation between incidence and prevalence of 'recent infections'

In the previous chapter different approaches of estimating the incidence were discussed. This section focuses on an approach of estimating incidence from the window period prevalence proposed by Mwanga (2006) [25], which will be validated in this research report.

Consider an uninfected individual who has had an infectious contact (sexual or other), the following stages occur in the prognosis of HIV, namely:

- Stage 1: infection, this occurs immediately (if contact with infected person was infectious).
- Stage 2: entry into the 'window state' (infection is now detectable by a virological assay such as PCR-RNA). This occurs at the individual's own time say t_1 (which varies) and globally at say T_1 (which is fixed), after infection. In fact this test detects acute infection prior to seroconversion. HIV-PCR stands for HIV- Polymerase Chain Reaction and is also known as viral load testing because it detects the presence of the immuno-deficiency virus in blood
- Stage 3: exit from the 'window state' (infection is now detectable by both assays- PCR and ELISA), when antibodies can now be detected. This happens at the individual's time t_2 (globally T_2) after infection

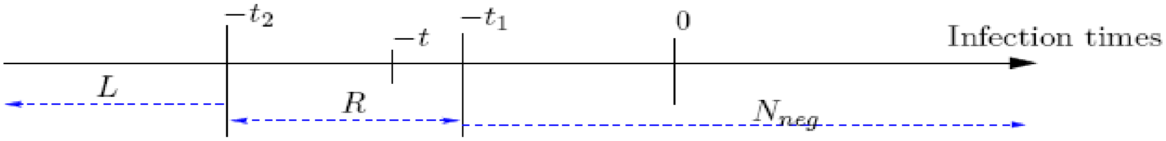
The biological distribution of the times t_1 and t_2 over the population is given by $\rho(t_1, t_2)$ which is unknown. Infected individuals can be grouped based on the outcome of the two tests as a proxy to indicate how far back from the day of testing they contracted the virus, that is t_0.

Figure 2-2. Schematic illustration of the PCR and ELISA test result.[8]

PCR	*Negative*	*Positive*	*Positive*	*Negative*
ELISA	*Negative*	*Negative*	*Positive*	*Positive*
Classification	Negative (N_{neg})	Recently Infected (R)	Long infected (L)	Indeterminate

Figure 2-3 below shows how individuals can be classified from a point in time $t_0=0$ on the day when they come for testing. There are long-infected (L), recently infected (R) and susceptible individuals (N_{neg}). These classifications are relative to the time from the day of testing and when they contracted the HIV, for distant past (long-infected), recent past (recently infected) and less than a week (susceptible individuals).

Figure 2-3. Classification of infected persons at time t_0=0.



The "window period" is the time it takes for a person who has been infected with HIV to *seroconvert* (test positive) for HIV antibodies. For simplicity the window period $T_w=|T_2-T_1|$ when individuals enter the window period at a fixed time since infection T_1 and leave at fixed time T_2. Hence $\rho(t_1,t_2)$, assuming both times to be independent, can be written as:

$\rho(t_1,t_2)=\delta(t_1-T_1)\delta(t_2-T_2)$ …………………………………………...Equation 2.2

In an infinitesimal cohort the probability of anyone being infected in the window period around time t in a period dt is given by:

$P(t_1 < t_0 - t \text{ and } t_2 > t_0 - t) = \int_0^{t_0 - t} \int_{t_0 - t}^{t_2^{\max}} \rho(t_1, t_2) dt_2 dt_1$ ....Equation 2.3

$P(t_1 < t_0 - t \text{ and } t_2 > t_0 - t) = \int_0^{t_0 - t} \int_{t_0 - t}^{t_2^{\max}} \delta(t_1 - T_1)\delta(t_2 - T_2) dt_2 dt_1$ ....Equation 2.4

Where $t_2^{\max}$ is the individual's maximum time from the test time t_0, when he or she was PCR positive and ELISA negative. The time $t_i = t_0 - t_2^{\max}$ is the earliest time that is considered, since all people infected before this time would have left the window period by the time they are observed. Consider a gross incidence *I* from a population (N_s) which gets infected at an incidence rate *i*. Thus, in a period time *dt*, the number of new cases is given by $I(t)dt = i(t).N_s.dt$ for a susceptible population N_s.

The number of persons infected between t_i and t_0 is given by:

$N_i = \int_{t_i}^{t_0} I(t)dt = \int_{t_i}^{t_0} i(t) N_s(t) dt$ ....Equation 2.5

Thus the expected number of persons in the window period is:

$N_w = \int_{t_i}^{0} \int_0^{-t} \int_{-t}^{t_2^{\max}} i(t) N_s(t) \rho(t_1, t_2) dt_2 dt_1 dt$ ....Equation 2.6

$N_w = \int_{t_i}^{0} \int_0^{-t} \int_{-t}^{t_2^{\max}} i(t) N_s(t) \delta(t_1 - T_1)\delta(t_2 - T_2) dt_2 dt_1 dt$ ....Equation 2.7

Since the expected number (N_W) of the recently infected is also given by the experimental value *R*, the above equations can be rewritten in terms of *R*:

$R = \int_{t_i}^{0} \int_0^{-t} \int_{-t}^{t_2^{\max}} i(t) N_s(t) \delta(t_1 - T_1)\delta(t_2 - T_2) dt_2 dt_1 dt$ ....Equation 2.8

In general, the incidence rate of the susceptible population can be modelled by a Taylor series, that is :

$i(t) = i(0) + i'(0)t + i''(0)t^2 / 2 + $ ....Equation 2.9

And:

$$N(t) = N(0) + N'(0)t + N''(0)t^2 / 2 + \ldots.. \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{..Equation 2.10}$$

The incidence was estimated using the formula below whose derivation is discussed in appendix A :

$$i = \frac{R}{N_{neg} T_w} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{..Equation 2.11}$$

Where *i* is the incidence rate, *R* the number of recently infected persons, N_{neg} are the susceptible persons and *Tw* is the window period. This incidence relation can also be written, in relation to window period prevalence and window period, where P_w is the window period prevalence (i.e., the number of people seen in the window period over susceptible P_w=R/N_{neg}) and T_w is the window period duration estimated from delta (Δ), the difference between the two dates for RNA negative and positive:

$$i = \frac{P_w}{T_w} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{Equation 2.12}$$

2.3.3 Bayesian estimation

Bayesian estimation draws inference about unobservable parameters or hypotheses by combining two sources of information:

a) (Prior) beliefs about the parameters formed from past evidence, such as pilot studies or similar studies.

b) Sample data that the study generates.

The Bayesian approach estimates the probability of the hypothesis *H* conditional on the observed data, i.e., Prob(H | data). This probability is called the posterior.

Bayes theorem calculates the posterior probability of *H* as follows:

$$P(H \mid data) = \frac{P(data \mid H)P(H)}{P(data)} \text{ ....Equation 2.13}$$

Where *P(data | H)* is the likelihood, that is, the probability of observing the data conditional on the hypothesis. *P(H)* is called the prior probability, which is a function which quantifies our prior knowledge or beliefs about *H*. *P(data)* is the normalizing constant, that is, the factor that makes the total probability equal to one. [25]

2.3.4 Computation of the incidence confidence interval

Two main procedures were used to compute the confidence intervals for the estimates obtained from the data: the Wald interval estimation of binomial proportions and Bayesian likelihood estimation.

The Wald procedure is an approximate confidence interval estimation of a binomial proportion in finite population sampling, computed by:

$$CI_{Wald}(1-\alpha)\times 100\% = \hat{p} \pm z_{\alpha/2}\sqrt{1-f}\sqrt{\frac{\hat{p}(1-\hat{p})}{\hat{n}+1}} \text{ ....Equation 2.14}$$

Where p is the prevalence, f is the sample fraction and $\hat{n} = n + z^2_{\alpha/2}$. This procedure was used in estimating the confidence interval of the ‘window period prevalence’, which was done using STATA 9.0.

The Bayesian likelihood procedure was used to estimate the incidence confidence interval. The likelihood function of obtaining the number of recent infections for a particular incidence with an associated window period is given as:

$$L(R \mid i) = \frac{\int L(R \mid <t_w>, i)\cdot prob(<t_w>)d<t_w>}{\int prob(<t_w>)d<t_w>} \text{ ....Equation 2.15}$$

Where $prob(< t_w >) = L(callibrated\ data\ |< t_w >)$. L is the likelihood and R are the recently infected.

The denominator in equation 2.15 is for normalizing the likelihood. This procedure was performed in Excel by making use of small discrete intervals of time. It was used to estimate the confidence interval for the 'window period' as well as that for the 'incidence'. The procedure did not require any distributional assumptions of any distribution. The Bayesian approach also has the advantage of providing an the required estimate and interval band(credibility interval), distribution (posterior) and graphical output.

2.3.5 Multistage pooling incidence rate

For the multistage pooling study after, e.g., *z* stages of pooling algorithm, the incidence rate will be estimated by the following relation after replacing P_w with $\hat{p}$ from equation B6 in appendix B.

$$i = \frac{1}{T_w}\left[1-\left(\frac{N_1}{N}\right)^{\frac{1}{S_Z}}\right] \text{...Equation 2.16}^{24}$$

The error introduced by lack of sensitivity and specificity is known.[26] The selection of the initial pool size is integral to the effect of minimizing the impact of false positives and false negatives.

2.3.6 Optimal pooling strategy

The major aim of pooling blood samples is to identify the number of positive infections *R* in a group of *N* samples using as few tests as possible (PCR runs in the case of HIV-RNA test). For this to be possible, at least half of the samples on every run (stage) needs to be eliminated (PCR runs in the case of HIV-RNA test), and optimal initial pool size s_1 must be used.

$$s_1 = \frac{-\ln 2}{\ln(1-p)} = \frac{\ln 2}{|\ln(1-p)|} = \frac{\ln 2}{f} \quad \text{..Equation 2.17}^{8}$$

The equation 2.17 derived in appendix C gives the optimal initial pool size where, *f* is the expected frequency of individual being in the window period. When, $p = 1\text{-}\varepsilon$ (very high prevalence) where ε is a small nonnegative number, individual testing is preferable.

The true value for disease prevalence (*p*) is not known in advance, and poor choice of *p* may lead to an imprecise estimate of initial pool size. The multistage pooling study allows one to adapt the pool size. This is done by observing the cost (variance) of reducing or adding of the pool size after each pooling stage. In practice, adapting the pool size after each stage is not pleasant and clinicians prefer having the pooling algorithm before starting to run the PCR in order to reduce time wastage. It has been observed that once one has a good estimate of the initial pool, positive pools that are broken into sub-pools of half the size provide an optimal pooling algorithm.

2.4 Ethical Clearance

The study was granted ethical clearance from three universities including the University of the Witwatersrand, clearance number **MM040202.** The research report received ethical clearance from the University of the Witwatersrand, clearance number **W-CJ-070504-1.**

CHAPTER 3 : RESULTS

3.1 Incidence from direct follow-up

Two hundred and forty-five sex workers were followed up and 19 seroconverted observed in 350 person years (127 353 person-days). This yields an incidence of 5.43% per year with a 95 % confidence interval of (3.9 - 9.2) % per year.

3.2 The results of systematic estimation of incidence

Table 3.1 below shows the interval between observations delta (Δ**)**, which is calculated by differencing the last PCR negative and first PCR positive. A participant was observed in the window period if the PCR status switched from negative to positive and antibody negative on the same day as the first PCR positive. The data had a total of 21 participants who seroconverted and of these 2 did not have any follow-up information, and were thus removed from the analysis. The results show that out of the 19 persons who seroconverted, 4 of them were seen in the window period.

Table 3.1. Seroconvertors and those seen in the window period.

Participant id	Δ (in days)	seen in window
100136	23	No
100040	26	No
100200	27	No
100239	28	No
100174	28	No
100225	28	No
100137	28	No
100222	28	No
100069	28	Yes
100221	28	Yes
100177	28	Yes
100229	29	No
100065	29	No
100037	29	No
100129	33	No
100045	34	Yes
100008	35	No
100085	53	No
100206	55	No
Mean Δ	**31.42105**	

The probability of seeing an individual in the window period was 4/19 and this was used to fit different models to estimate the likelihood probability of finding a seroconverter in the window period, namely Binomial and Poisson. The following were the corresponding probabilities together with their respective confidence intervals.

Table 3.2. Fitted likelihood functions.

Model	**Likelihood L**	**Standard Error**	**Confidence Interval (95%)**
Binomial	0.2105263	0.0935288	(0.0605245 , 0.4556531)
Poisson	0.2105263	0.1052632	(0.0573613 , 0.539031)

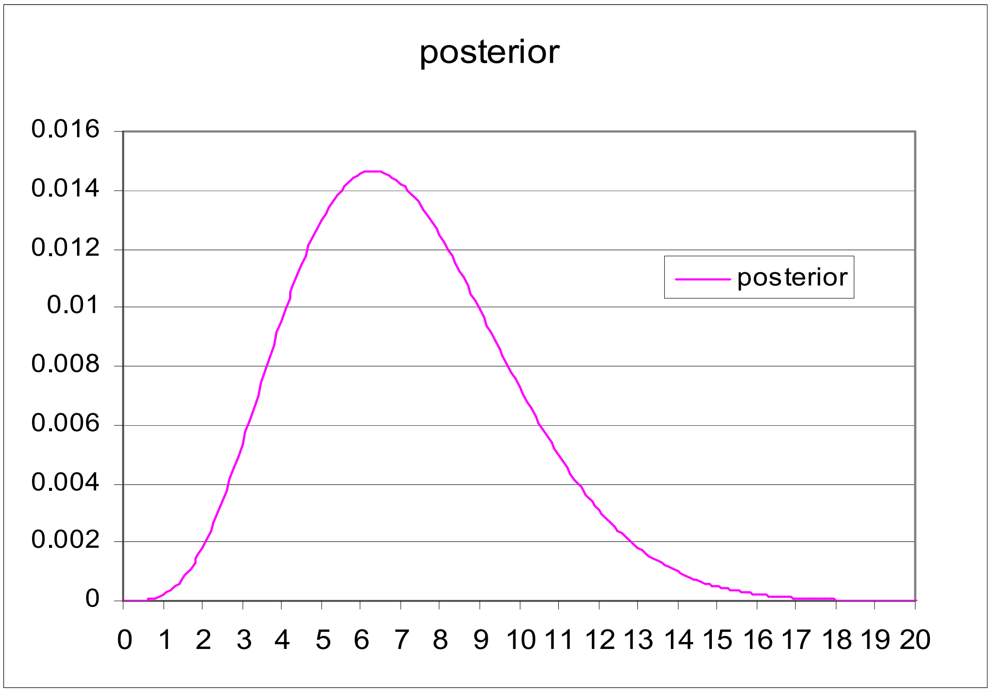
The window period is estimated by using the delta function, the likelihood function and the following:

- Four participants were observed in the window period.

- There were 21 infections (seroconversions) of which 19 had enough information for the computation of the interval (Δ) between consecutive observation times.
- The average interval (Δ) between observation times is 31.42 days.

Thus $T_w = 31.42105 \times 0.2105263 = 6.61496 days \approx 0.0181 years$ days with a 95% confidence interval of (1.90–14.32) days assuming a binomial likelihood using the Wald's approximation. This was also computed directly (exact) using the cumulative 95% confidence intervals and linear interpolation techniques yielding (2.69 – 13.01)%, which is narrower. The latter approach was more preferred than the former since the normal approximation is not ideal for expected mean less than 5 (in this instance mean *np=4*).

Figure 3-1. Bayesian posterior likelihood for 'window period'.



The figure above gives the posterior likelihood function for the window period of 6.6 days with a 95% confidence interval of (2.69 – 13.01) days using Excel.

3.3 Incidence and mean window period prevalence estimation

The window period prevalence was computed using:

$$P_w = \frac{R}{N_{neg}}$$

From the four people seen in the window period over the 4245 person-months (350 person-years), the window period prevalence estimate is 0.09423 % and a 95% confidence interval of (0.0193 – 0.1865)% using the Wald's confidence interval approximation for finite populations.

This yields an incidence: $i = \frac{0.09423}{0.0181} = 5.21\%$ using the equation:

$$i = \frac{P_w}{T_w} \text{ ………………………………………………….. Equation 2.12.}$$

This has a 95% confidence interval of (4.144 – 14.564)%.

3.4 Example of 'ideal' pooling strategy versus 'practical' pooling strategy

The available data were not sufficient to test for the optimum pooling strategy, since this required a large sample to be divided into different pools. Mwanga (2006) [25] discussed three pooling algorithms. The first has pool size of (100; 50; 10; 1), which was used in detection of acute HIV-1 infection in North Carolina[27] and in South Africa.[28] The second (Strategy 2) was computed with a prevalence $p = 0.0109$ and halving the positive pools at each stage pooling algorithm with pool a size of (64; 32;

16; 8; 4; 2; 1). The last pooling algorithm (25; 5; 1), given as strategy 3, is proposed as a practical pooling strategy proposed by local experts.

Incidence estimates using the three pooling algorithms are shown in table 3.3 where $\hat{i}$ is the estimate of incidence using a multistage pooling procedure (equation 2.16) and i is the exact incidence from prospective cohort follow-up. In various studies [20, 21] as well as the simulations in this study, the window period duration was 28 days which was made from a sample of size N=6400.

Table 3.3. Incidence estimates using three pooling algorithms.

Strategy								
1.	**Pool size**	***100***	***50***	***10***	***1***			
	$\hat{i}$ % per year	8.67	8.61	8.56	8.56			
	$\lvert (i)-\hat{i} \rvert \times 10^{-6}$	14.22	0.54	7.69	0			
2.	**Pool size**	***64***	***32***	***16***	***8***	***4***	***2***	***1***
	$\hat{i}$ % per year	8.62	8.58	8.57	8.56	8.56	8.56	8.56
	$\lvert (i)-\hat{i} \rvert \times 10^{-6}$	8.46	19.72	8.95	0.2	1.71	0.57	0
3.	**Pool size**	***25***	***5***	***1***				
	$\hat{i}$ % per year	8.57	8.56	8.56				
	$\lvert (i)-\hat{i} \rvert \times 10^{-6}$	8.8	2.29	0				

Table 3.3 above shows by use of simulations that an optimum pool sample size is obtained when at least half of the samples are eliminated at every run. The results also show that in screening a large population using a multistage pooling algorithm for the purpose of estimating HIV incidence, it is not necessary to stop at pools of size one. Strategy 2 shows it suffices to stop at the pools of size 8, which will still give a good approximation of the incidence. Similarly, strategies 1 and 3 show that one can stop at the pools of size 10 and 5, respectively, and still produce a robust estimate of incidence.

Thus, applying a multistage pooling algorithm for estimating HIV incidence rate has the potential of dramatically reducing the cost of testing.

3.5 Summary of results

The mean window period was 6.6 days 95% CI: (2.7 – 13.0). The monthly window period prevalence was 0.09423 % 95 % CI: (0.0193 – 0.1865)%. The incidence from direct follow-up was 5.43 percent 95% CI: (3.9 – 9.2)%. The incidence estimate from cross-sectional formulae was 5.21 percent 95% CI: (4.1 – 14.6). The cross-sectional incidence is about 4.05% lower than that obtained by direct follow-up of the cohort. It was also shown by use of simulations that an optimum pool sample size is obtained when at least half of the samples are eliminated at every run. The simulated data also showed that multistage pooling has great potential of reducing the cost of estimating incidence in a cross-sectional study.

CHAPTER 4

CONCLUSIONS, LIMITATIONS & PROPOSAL FOR FURTHER WORK

Epidemiological principles that govern a quality programme for accurate and representative prevalence estimates also apply to cross-sectional incident infections. These include the selection of the appropriate target population, the validation of it being representative, the identification of any selection, accrual or testing bias that may distort the representation, as well as programmatic issues such as enrolment, specimen handling, transport, testing, data management and quality assurance.[29]

The estimation of incidence is the result of a calculation requiring three measurements: the number classified as incident; the window period; and the number of seronegative (at-risk) members of the population.[29] The accuracy of the estimate is thus dependent on the accuracy of all three measurements.

The window period for the HIV-PCR in this study was found to be 6.6 days for sensitivity of at least 50copies/ml. Fiebig et al. (2003) proposed six stages of primary (acute) HIV-1 infection with regards to blood samples being detected by RNA, p24 antigen, enzyme immunoassays (EIA) and the Western Blot. Table 4.1 shows the results obtained together with the 95% confidence intervals.

Table 4.1. Laboratory stages of primary HIV infection, source[30].

Stage	RNA	p24-Antigen	EIA - not sensitive	EIA - sensitive	Western Blot	Duration in days (95% CI)
I	+	-	-	-	-	5.0 (3.1 , 8.1)
II	+	+	-	-	-	5.3 (3.7 , 7.7)
III	+	+	-	+	-	3.2 (2.1 , 4.8)
IV	+	+/-	-	+	|	5.6 (3.8 , 8.1)
V	+	+/-	+/-	+	+	69.5 (39.7 , 121.7)
VI	+	+/-	+	+	+	Open ended

This study showed that the PCR-RNA test was very sensitive and yielded a very similar result to that obtained with the lab results in table 4.1. The window period of 6.6 days 95% CI: (2.69 – 13.01) is similar to the window period obtained from the same cohort a year before, i.e., 6.8 days 95% CI: (3 – 13). The latter yielded better point and incident estimates since more persons were tested (23, including others from other studies) compared to this year's 19 (AIC only). The table also shows that HIV-RNA testing may still increase the sensitivity for HIV detection. It is also a worthwhile addition to HIV prevention efforts and useful in identifying persons with acute HIV infection, which should remain a public health priority. Hence the sensitivity also shows that the use of HIV RNA testing can readily identify persons with acute HIV infection (who are the most infectious and more likely to transmit the virus)[3] and other acute sexually transmitted infections. It may be more useful to use a longer window period by detuning the antibody assay, in order to mimic the real life situation more practically.

There are also issues regarding the sample size estimation (McDougal et al. 2005). If the prevalence (P) is known, the total population size needed is N_{neg} = (no. at risk)/(1 - P). A Sensitive/Less sensitive assay with a 150-day window period and an anticipated incidence of 5% would require that 4642, 1168 or 519 seronegative specimens accrue throughout the follow-up to achieve a ±1%, 2%, or 3% CI, respectively. An assay with a 10-fold shorter window period, such as the viral assays, would require 10-fold higher numbers.[29]

The monthly window period prevalence for the study was found to be 0.09423 % and the 95% CI: (0.0193 – 0.1865)%, which is comparable to that found a year before 0.15% and 95% CI: (0.06 – 0.34) %. The incidence from direct follow-up was 5.43 percent 95% CI: (3.9 – 9.2)% .The incidence estimate from cross-sectional formulae was 5.21 percent 95% CI: (4.1 – 14.6). This interval is wider and other procedures such as Bootstrap estimations may also be used in further studies.

It was also shown by use of simulations that an optimum pool sample size is obtained when at least half the samples are eliminated at every run. In screening a large population using a multistage pooling algorithm for the purpose of estimating HIV incidence, it is not mandatory to stop at pools of size one. Strategy 2 shows it suffices to stop at the pools of size 8, which will still give a good approximation of the incidence. Similarly, strategies 1 and 3 show that one can stop at the pools of size 10 and 5, respectively, and still produce a robust estimate of incidence. Thus, applying a multistage pooling algorithm for estimating HIV incidence rate can dramatically reduce the cost of testing.

For the multistage pooling strategy, it was shown that an optimal pooling algorithm can be achieved by halving the pool size at each stage of executing the pooling algorithm.

This was done using strategy 2, yet this runs more tests until one pool is left. This showed less error (obtained by use of the mean and standard deviation of PCR runs) compared to the other two strategies (see appendix D). Where *E[Npositive]* is the expected number of positive specimens in the sample of 6400 specimens. Multistage pooling when used to run the PCR test in combination with a very sensitive assay can reduce the cost of estimating the incidence remarkably and the number of tests by over 75% compared to p24 antigen testing.[31]

In general, the issues surrounding the generalization or extrapolation of results from a sentinel population to the larger epidemic are the same for prevalence and incidence data. A unique feature of incidence that is relevant to extrapolation is a point estimate, whereas prevalence reflects cumulative experience with HIV-1. Prevalence reasonably reflects HIV-1 exposure before, during and after pregnancy for antenatal clinic data. Incidence reflects recent exposure shortly before or during pregnancy in the case of antenatal attendees. If there is a difference in risky behaviour and incidence before and during pregnancy, the extrapolation to non-pregnant women of the same age may not be valid.[29]

A more informative incidence estimate, a pooled estimate based on estimates from different provinces, may be more useful to generalize the incidence to the whole population. The actual number of persons in the window period is also very low and large data may need to be used to validate the formulae. The PCR-RNA test is very sensitive at detecting acute HIV-1 infected persons. The incidence estimate from the cross-sectional study formulae was very similar to that obtained from a follow-up study.

The number of tests needed can be reduced and a good estimate of the incidence can still be calculated. The calibration was not accurate since the samples used were small and the window period duration too short, hence it was difficult to extrapolate to the whole population. Further work still has to be done on the calibration of these incidence formulae as it can serve as a very useful public health tool.

APPENDIX A: DERIVATION FOR ACUTE HIV-1 FORMULAE

This is the derivation of an incidence for cross-sectional studies based on results obtained from and antibody test (ELISA) and virological test (RNA-PCR). This formula was proposed by Mwanga (2006) [25].

If a linearly decreasing closed population (at rate λ) is considered, the number susceptible can be written as:

$N_s(t) = N_s(0) - N_s(0)\lambda t$ ………………………………………………..Equation A1

At time *t₀*, when participants appear for testing. Shifting the reference point to the date of the first PCR positive test, that is, time $-t_1$, the susceptible population can be given as $N_s(t) = N_s(-t_1) - N_s(-t_1)\lambda t$ and replacing $\lambda' = N_s(-t_1)\lambda$:

$N_s(t) = N_s(-t_1) - \lambda' t$ …………………………..……………………...Equation A2

Then the number of recently infected *R* (section 2.3.2) is given by:

$R = \int_{t_i}^{0}\int_{0}^{-t}\int_{-t}^{t_2^{\max}} i(t)\{N_s(-t_1) - \lambda' t\}\rho(t_1,t_2)dt_2dt_1dt$ ……………………...….Equation A3

$R = \int_{t_i}^{0}\int_{0}^{-t}\int_{-t}^{t_2^{\max}} i(t)\{N_s(-t_1) - \lambda' t\}\delta(t_1 - T_1)\delta(t_2 - T_2)dt_2dt_1dt$ ……………...Equation A4

Assuming a constant yearly incidence, that is, *i(t) = i*.

$R = i\left(N_s(-t_1)T_w + \lambda'\frac{T_w^2}{2}\right)$ …..…………………………………..….Equation A5

Replacing $N_s(-t_1)$ with N_{neg}:

$i = \dfrac{R/T_w}{(N_{neg} + \frac{\lambda'}{2}T_w)}$ ……………………………………………………...Equation A6

$i = \dfrac{R}{N_{neg}T_w}\left\{1 + \dfrac{\lambda'}{2N_{neg}}T_w\right\}^{-1}$ ……………………………..……..….Equation A7

Series expansion of equation:

$$i = \frac{R}{T_w N_{neg}} - \frac{\lambda' R}{2N_{neg}^2} . + \frac{\lambda'^2 R T_w}{4N_{neg}^3} - \frac{\lambda'^3 R T_w^2}{8N_{neg}^4} + ... \quad \text{.........................Equation A8}$$

Assuming the population does not vary much over time, hence keeping only the first order term we obtain:

$$i = \frac{R}{N_{neg} T_w} \quad \text{...Equation 2.11}$$

APPENDIX B : MULTISTAGE POOLING INCIDENCE

In a one-stage pooling study, if there is a sample of *N* seronegative specimens (for HIV-RNA test), the sample can be divided into n_1 pools of size s_1, thus $N=n_1.s_1$ given x_1 of these pools are positive and y_1 are negative such that $n_1 = x_1+y_1$. In order to systematically determine the probability *p* that a person has the disease without error (assuming perfect sensitivity and specificity of the test), the maximum likelihood estimate (MLE) must be determined. The individual tests can be taken as Bernoulli trials, thus the likelihood would be a binomial distribution.

$$L_1 = \binom{n_1}{x_1} \left\{ (1-p)^{s^1} \right\}^{y_1} \left\{ 1-(1-p)^{s_1} \right\}^{x_1} \quad \text{...Equation B1}$$

Finding the natural logarithm of both sides.

$$InL_1 = c + s_1 y_1 In\left\{ (1-p) \right\} + x_1 In\left\{ 1-(1-p)^{s1} \right\} \text{..........................Equation B2}$$

To find the maximum likelihood, $\partial(InL_1)/\partial p = 0$

$$\frac{\partial InL_1}{\partial p} = -\frac{s_1 y_1}{1-\hat{p}} + \frac{x_1 s_1 (1-\hat{p})^{s_1 - 1}}{1-(1-\hat{p})^{s_1}} = 0 \quad \text{..Equation B3}$$

Thus after some algebra, this yields the MLE of *p*:

$$\hat{p} = 1 - \left(\frac{x_1}{n_1} \right)^{\frac{1}{s_1}} \text{..Equation B4}$$

For large n_1 the variance of the MLE of p is estimated by [10].

$$\operatorname{var}(\hat{p}) = \frac{1 - (1 - \hat{p})^{s_1}}{n_1 s_1^{\ 2} (1 - \hat{p})^{s_1 - 2}} \text{...Equation B5}$$

This is only applicable to the one stage pooling case, but one pooling is usually not sufficient, thus, there is a need for multistage pooling. It can similarly be shown that *p* yields a good estimate for a multistage study pooling study, which is the generalization of a single stage pooling study.

$$\hat{p} = 1 - \left(\frac{N_1}{N} \right)^{\frac{1}{s_z}} \text{..Equation B6}$$

And the corresponding estimate of the variance.[10]

$$\operatorname{var}(\hat{p}) = \frac{1 - (1 - \hat{p})^{s_1}}{N s_z^{\ 2} (1 - \hat{p})^{s_1 - 2}} \text{...Equation B7}$$

In order to estimate the incidence from cross-sectional prevalence, it should be assumed that *R* individuals in the sample of *N* seronegative individuals are found to be recently HIV infected. If the mean window period duration (*Tw*) of these individuals is known:

$$i = \frac{P_w}{T_w} = \frac{R}{T_w N} \text{..Equation 2.11}$$

Where P_w=R/N is the proportion of individuals seen in the window period namely the window period prevalence.

For the multistage pooling study after, e.g., *z* stages of pooling algorithm, the incidence rate will be estimated by the following relation after replacing P_w with $\hat{p}$ from equation B6.

$$i = \frac{1}{T_w}\left[\left(1 - \left(\frac{N_1}{N}\right)^{\frac{1}{S_Z}}\right)\right] \text{...Equation 2.16}$$

The error introduced by lack of sensitivity and specificity is known.[24] The selection of the initial pool size is integral to the effect of minimizing the effect of false positives and false negatives.

APPENDIX C : DERIVATION OPTIMAL POOLING STRATEGY

If there is a sample of *N* specimens that is pooled into n_p pools of size s_1 (s_1 being the initial pool size) and *p* is the prevalence of infected individuals in this sample, then the probability that in this pool of size s_1 there are exactly *k* infected individuals can be modelled by the binomial distribution.

$$P(X = k) = \binom{s_1}{k} p^{k_1} (1-p)^{s_1 - k} \text{ ……………………………….……………..Equation C1}$$

Where, *X* is a random variable and $k \le R$. Thus, the probability of a pool testing positive is $P[X \ge 1]$, i.e., at least one individual in the pool has infection. This can be written as:

P(at least 1 positive) = 1- P(none positive) i.e. $P(X \ge 1) = 1 - P(X < 1) = 1 - P(X = 0)$

$$P(X \ge 1) = 1 - (1-p)^{s_1}$$

To minimize the number of PCR runs, at least half of the initial pools should test negative. If

$P(X = 0) = \frac{1}{2} = (1-p)^{s_1}$ and making $s_1$the subject, the following equation is obtained:

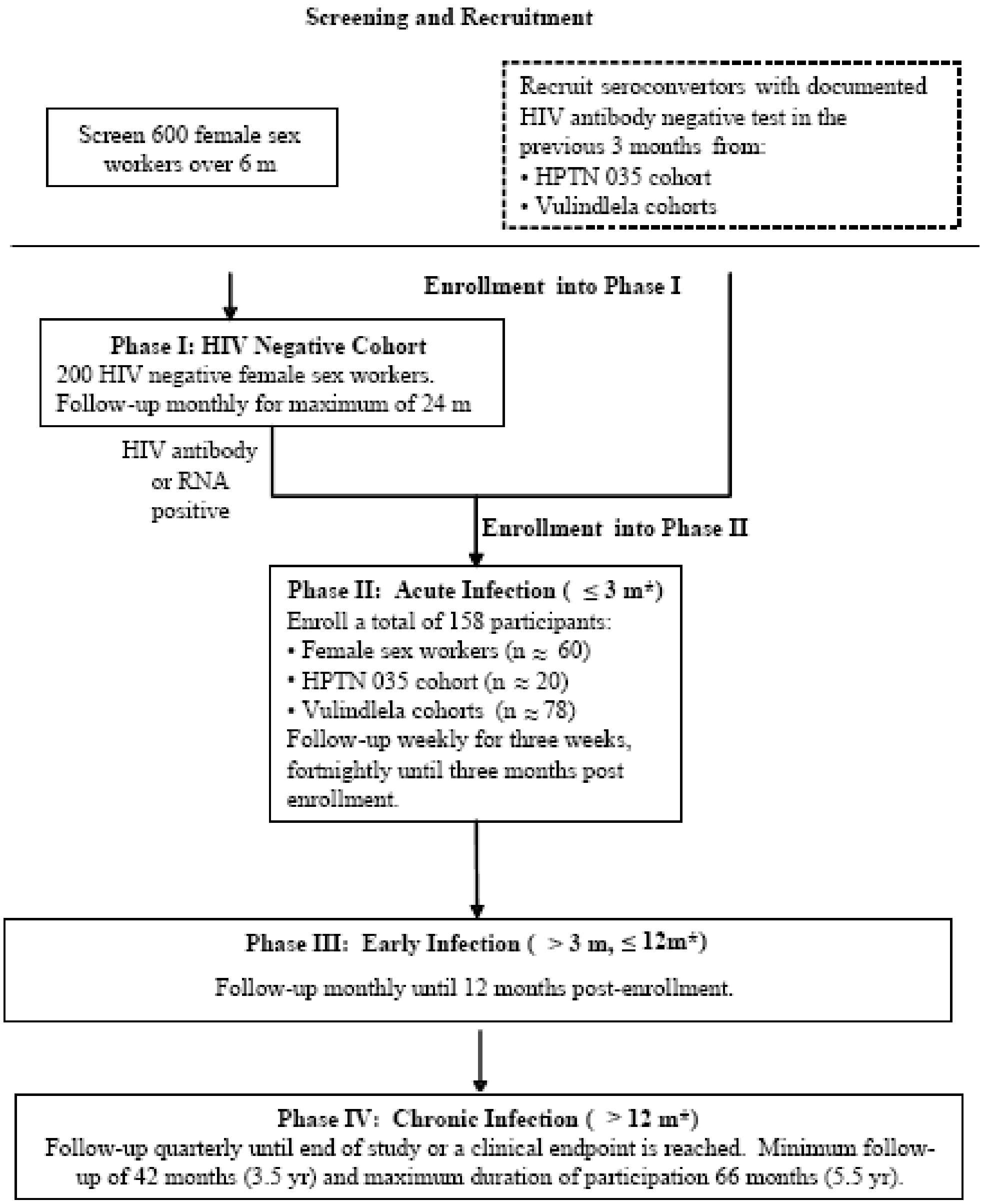
$$s_1 = \frac{-In2}{In(1-p)} = \frac{In2}{|In(1-p)|} = \frac{In2}{f} \text{ ……………………………………………..............Equation 2.17 }^{8}$$

Equation 2.17 gives the optimal initial pool size where *f* is the expected frequency of individual being in the window period. When $p = 1\text{-}\varepsilon$ (very high prevalence), where ε is a small nonnegative number, individual testing is preferable.

APPENDIX D: STRATEGIES MEASURES OF SPREAD AND LOCATION

E[Npositive]	Mean and stdev	Strategy 1	Strategy 2	Strategy 3
2	mean	97.93	123.96	275.98
	stdev	0.74	0.47	0.36
4	mean	131.5	147.8	295.86
	stdev	2.15	1.05	0.94
6	mean	164.77	171.46	315.66
	stdev	3.29	1.7	1.49
8	mean	197.77	194.97	335.39
	stdev	4.27	2.3	2
10	mean	230.26	218.32	355.04
	stdev	5.66	2.9	2.43
12	mean	262.83	241.67	374.58
	stdev	6.53	3.38	2.96
14	mean	294.67	264.76	394.12
	stdev	7.56	3.86	3.35
16	mean	326.42	287.74	413.39
	stdev	8.65	4.53	3.96
18	mean	357.87	310.45	432.67
	stdev	9.65	5.08	4.51
20	mean	389.16	333.06	451.96
	stdev	10.59	5.62	4.94
22	mean	419.77	355.84	471.07
	stdev	11.68	6.08	5.37
24	mean	450.58	378.23	489.98
	stdev	12.38	6.68	5.96
26	mean	480.87	400.48	509.15
	stdev	13.23	7.21	6.32
28	mean	511.09	422.66	527.94
	stdev	14.02	7.71	6.92
30	mean	540.95	444.7	546.8
	stdev	15.34	8.2	7.31

APPENDIX E: STUDY DESIGN



Screening and Recruitment

Screen 600 female sex workers over 6 m

Recruit seroconvertors with documented HIV antibody negative test in the previous 3 months from:
- HPTN 035 cohort
- Vulindlela cohorts

Enrollment into Phase I

Phase I: HIV Negative Cohort
200 HIV negative female sex workers.
Follow-up monthly for maximum of 24 m

HIV antibody or RNA positive

Enrollment into Phase II

Phase II: Acute Infection (≤ 3 m*)
Enroll a total of 158 participants:
- Female sex workers (n ≈ 60)
- HPTN 035 cohort (n ≈ 20)
- Vulindlela cohorts (n ≈ 78)

Follow-up weekly for three weeks, fortnightly until three months post enrollment.

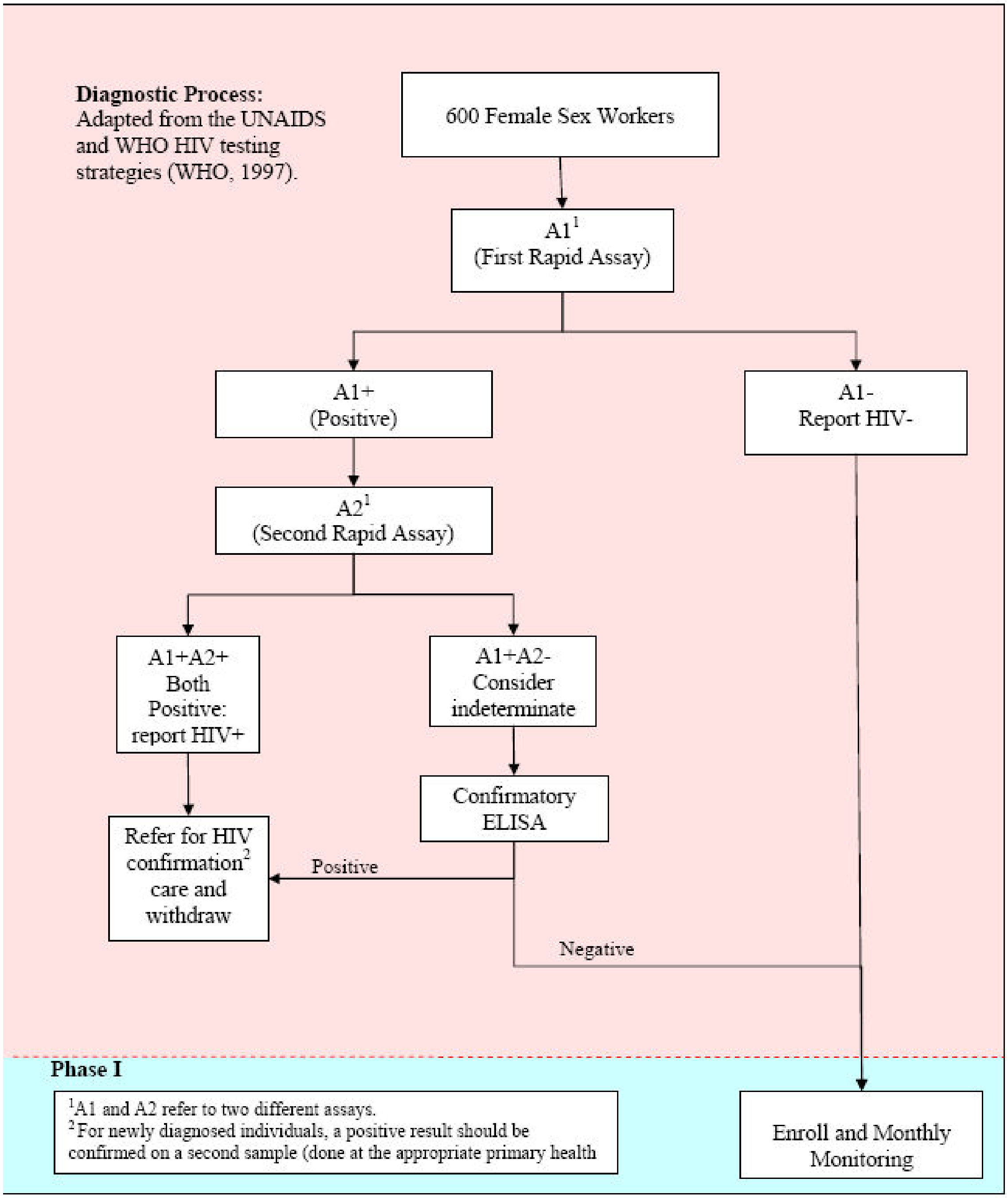
Phase III: Early Infection (> 3 m, ≤ 12m*)

Follow-up monthly until 12 months post-enrollment.

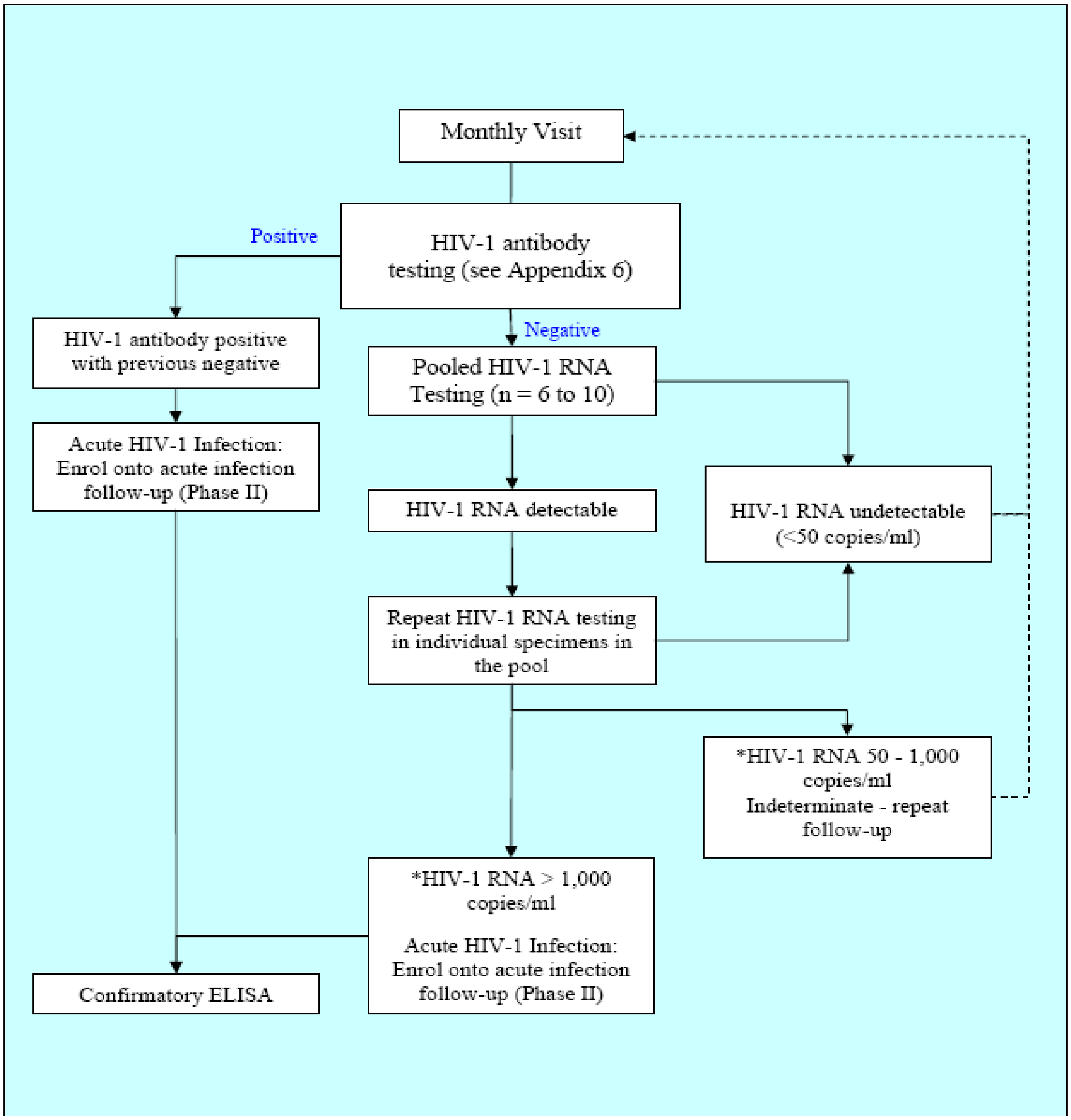
Phase IV: Chronic Infection (> 12 m*)
Follow-up quarterly until end of study or a clinical endpoint is reached. Minimum follow-up of 42 months (3.5 yr) and maximum duration of participation 66 months (5.5 yr).

* Indicates months post enrollment into Phase II.

APPENDIX F : SCREENING AND DIAGNOSTIC ALGORITHM


Diagnostic Process:
Adapted from the UNAIDS and WHO HIV testing strategies (WHO, 1997).
600 Female Sex Workers
A1[1] (First Rapid Assay)
A1+ (Positive)
A1- Report HIV-
A2[1] (Second Rapid Assay)
A1+A2+ Both Positive: report HIV+
A1+A2- Consider indeterminate
Confirmatory ELISA
Refer for HIV confirmation[2] care and withdraw
Positive
Negative
Phase I
[1]A1 and A2 refer to two different assays.
[2]For newly diagnosed individuals, a positive result should be confirmed on a second sample (done at the appropriate primary health
Enroll and Monthly Monitoring

APPENDIX G : HIV DIAGNOSTIC ALGORITHM FOR PHASE I (HIV NEGATIVE FSW COHORT)


Monthly Visit
HIV-1 antibody testing (see Appendix 6)
Positive
Negative
HIV-1 antibody positive with previous negative
Acute HIV-1 Infection: Enrol onto acute infection follow-up (Phase II)
Pooled HIV-1 RNA Testing (n = 6 to 10)
HIV-1 RNA detectable
HIV-1 RNA undetectable (<50 copies/ml)
Repeat HIV-1 RNA testing in individual specimens in the pool
*HIV-1 RNA 50 - 1,000 copies/ml Indeterminate - repeat follow-up
*HIV-1 RNA > 1,000 copies/ml
Acute HIV-1 Infection: Enrol onto acute infection follow-up (Phase II)
Confirmatory ELISA

* HIV RNA >5,000 copies/ml in the absence of a positive HIV antibody test, diagnosis of acute HIV infection. (Hecht *et al.* 2002, Walker and Altfeld, 2003). Follow up testing to confirm subsequent antibody seroconversion will be done to provide final confirmation of the diagnosis (adapted from Marcus Altfeld & Bruce D. Walker, in Hoff & Kamps (Eds), 2003, p.50).

REFERENCES

[1] Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, *World Population Prospects: The 2006 Revision* and *World Urbanization Prospects: The 2005 Revision*, http://esa.un.org/unpp, Tuesday, October 30, 2007; 8:51:33 AM.

[2] Steinbrook R. The AIDS pandemic in 2004. Global Health.

[3] National HIV and Syphilis prevalence survey 2005 South Africa. Department of Health 2006.

[4] Pilcher C.D., Shugars D.C., Fiscus S.A.,Miller W.C., Menezes P., Giner J. *et al*. HIV in body fluids during primary HIV infection: implications for pathogenesis, treatment, and public health. AIDS. 2001;15:837–845.

[5] Quinn T.C. Acute primary HIV infection. Journal American Medical Association*,*
(JAMA). 1997;287:58–62.

[6] Daar E.S., Moudgil T., Meyer R.D.,Ho D.D. Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. New England Journal of Medicine. 1991;324:961–964.

[7] Colfax G.N., Buchbinder S.P., Cornelisse P.G.A.,Vittinghoff E., Mayer K., Celum C. Sexual risk behaviors and implications for secondary HIV transmission during and after HIV seroconversion. AIDS. 2002;16:1529–1535.

[8] Patel P., Klausner J., Bacon O., Liska S., Taylor M., Gonzalez A., *et al*. Detection of acute HIV infections in high-risk patients in California *Journal of AIDS* 2005.

[9] Brookmeyer R., Quinn T.C. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. 1995 American Journal of Epidemiology.

[10] Winkelstein W. Jr., Samuel M., Padian N. S., Wiley J. A., Lang W., Anderson R. E.,*et a*l. The San Francisco Men's Health Study: III. Reduction in human immunodeficiency virus transmission among homosexual/bisexual men, 1982-86. American Journal of Public Health. 1987 Jun;77(6):685-9.

[11] Schwarcz S., Kellogg T., McFarland W., Louie B., Kohn R., Busch M.,*et al.*Differences in the temporal trends of HIV seroincidence and seroprevalence among sexually transmitted disease clinic patients, 1989-1998: application of the serologic testing algorithm for recent HIV seroconversion. American Journal of Epidemiology. 2001 May 15;153(10):925-34.

[12] Freidan T. R. HIV incidence in New York , 2001.New York's Community Health vol1, No s1 2003

[13] Parekh B.S., Kennedy M.S., Dobbs T., Pau C., Byers R., Green T.,*et al*. Quantitative detection of increasing HIV type 1 antibodies after seroconversion: A simple assay for detecting recent HIV infection and estimating incidence. AIDS Research and Human Retroviruses, 18(4):295 307, 2002.

[14] Wong K., Tsai W., Kuhn L. Estimating HIV hazard rates from cross-sectional HIV prevalence data. Statistics in Medicine. 2006; 25:2441–2449.

[15] Hill G. B., Forbes W. F., Kozak J. A Simple Method for Estimating Incidence from Prevalence. Chronic diseases in Canada Volume 20, No.4 – 2000.

[16] Statistics Canada (Health Statistics Division). National Population Health Survey (NPHS): public use microdata files, 1994-95. Ottawa, 1995.

[17] Denning P. H., Jones J. L., Ward J.W. Recent trends in the HIV epidemic in adolescent and young adult gay and bisexual men. Journal of AIDS Human Retrovirology. 1997 Dec 15;16(5):374-9.

[18] Janssen R.S., Satten G.A., Stramer S.L., Rawal B.D., O'Brien T.R., Weiblen B.J., *et al*. 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 (JAMA). 1998 Jul 1;280(1):42-8.

[19] Liao J., Brookmeyer R. An Empirical Bayes Approach to Smoothing in Backcalculation of HIV Infection RatesBiometrics, Vol. 51, No. 2. (Jun., 1995), pp. 579-588..

[20] Moss A.R., Vranizan K., Gorter R., Bacchetti P., Watters J., Osmond D. HIV seroconversion in intravenous drug users in San Francisco, 1985-1990. AIDS. 1994 Feb;8(2):223-31.

[21] Kleinman S., Busch M.P., Korelitz J. J., Schreiber G.B. The incidence /window period model and its use to assess the risk of transfusion - transmitted HIV and hepatitis C virus infection. Transfusion Medicine Reviews 11 , 155- 172 ,1997.

[22] Quinn T.C, Brookmeyer R., Kline R., Shepherd M., Paranjape R., Mehendale S,*et a*l. Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. AIDS, 14:2751-2757, 2000.

[23] Karim S.A. Viral set point and clinical progression in HIV-1 subtype c
infection: The role of immunological and viral factors during acute and early infection. June 2004.Version 2.0.

[24] Welte A,Mwanga G.G. Dynamical Models of Acute HIV Infection: Application to Interpretation of Seronegativity/RNA Positivity Survey Data to Estimate HIV Incidence. Masters dissertation, University of Witwatersrand 2006 (unpublished).

[25] Gelman A., Carlin J. B., Stern H.S., Rubin D.B. Bayesian Data Analysis.Chapman and Hall,(2003).2nd Edition .

[26] Brookmeyer R. Analysis of Multistage pooling studies of Biological specimens for estimating disease incidence and prevalence. Biometrics , 55(2):606-612,1999.

[27]Pilcher C.D., Fiscus S.A., Nguyen T.Q., Foust E., Wolf L., Williams D,*et al.* Detection of acute infection during HIV testing in North Carolina. New. England. Journal of. Medicine,352(18):1873-1883, 2005.

[28] Stevens W., Akkers E., Myers M., Motloung T., Pilcher C., Venter F. High prevalence of undetected, acute HIV infection in a South African primary care clinic. Third International AIDS Society Conference on HIV Pathogenesis and Treatment, Rio de Janeiro.Wits, 2005.

[29] Mcdougal J.S., Pilcher C.D., Parekh B.S., Gershy-Damet G., Branson B.M., Marsh K.,et.al. Surveillance for HIV-1 incidence using tests for recent infection in resource-constrained countries. AIDS:volume 19 Supplement 2May 2005pp 25-30.

[30]Fiebig E.W.,Wright D.J., Rawal B.D. ,Garrett P.E., Schumacher R.T., Peddada L. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS:Volume 17(13)5 September 2003pp 1871-1879.

[31] Quinn T.C., Brookmeyer R., Kline R., Shepherd M., Paranjape R., Mehendale S., *et al*. Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. AIDS 2000; 14:2751-2757.