Evaluation of Factors Associated with HIV Prevalence among 15- to 49-Year-Olds in Chimanimani District of Zimbabwe: Divergence and Similarity of the Two Statistical Methods

By

Elliot Chikaka

Dissertation Submitted in Partial Fulfilment of Master of Science in Biostatistics Degree

University of Zimbabwe

FACULTY OF MEDICINE

DEPARTMENT OF COMMUNITY MEDICINE

UNIVERSITY OF ZIMBABWE

HARARE

AUGUST 2009

Declaration

This dissertation is the original work of Elliot Chikaka. It has been prepared in accordance with the guidelines for Masters in Biostatistics dissertations in the University of Zimbabwe. It has not been submitted elsewhere for another degree at this or any other university.

Name of student: ____

Signature: ___________________________ Date: __________________

Name of Academic Supervisor: ____________________________________

Signature______________________________ Date: _________________

Chairman of the Department of Community Medicine, University of Zimbabwe

Signature______________________ Date: ______________________________

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof S Rusakaniko for his guidance and support. I acknowledge Biomedical Research and Training Institute for the opportunity they offered me to work on this project. Their support will never be forgotten.

Special thanks to my wife, family and colleagues for the support that you have always given me.

" This project was made possible through a fellowship received by Elliot Chikaka from the Forgarty International Centre, National Institute of Health (NIH-USA) through the International Clinical, Operational and Health Services and Training Award (ICOHRTA) Programme"

Elliot Chikaka

University of Zimbabwe

List of Abbreviations

AIDS	ACQUIRED IMMUNE DEFICIENCY SYNDROME
BSS	BEHAVIOURAL RISKS and HIV SEROSTATUS SURVEY
CDC	CENTRE FOR DISEASE CONTROL
CSTE	COUNCIL for STATE and TERRITORIAL EPIDEMIOLOGISTS
HIV	HUMMAN IMMUNO-DEFICIENCY VIRUS
LDA	LINEAR DISCRIMINANT ANALYSIS
LR	LOGISTIC REGRESSION
OMT	ORAL MUCOSAL TRANSUDATE
OVC	OPHANED AND VULNERABLE CHILDREN
PCA	PRINCIPAL COMPONET ANALYSIS
ROC	RECEIVER OPERATING CHARACTERISTIC CURVE
SSA	SUB-SAHARAN AFRICA
STI	SEXUALLY TRANSMITTED INFECTIONS
WHO	WORLD HEALTH ORGANISATION

Definition of Terms

Convergence refers to the two statistical methods giving the same or similar results in the analysis of factors associated with HIV prevalence among the 15-49 year olds in Chimanimani district as measured by (i) classification error (ii) receiver operating characteristic curve (iii) magnitude of the beta and unstandardized coefficients (iv) relative importance of the predictors using the Z-statistic and the standardized coefficients (v) specificity and sensitivity at different cut-off points and (vi) kappa

Divergence refers to the two statistical methods giving the different results in the analysis of factors associated with HIV prevalence among the 15-49 year olds in Chimanimani district.

The **Eigen Value** is also called the *characteristic root of each discriminant function*, reflects the ratio of importance of the dimensions which classify cases of the dependent variable.

Unstandardized discriminant Coefficients are used in the formula for making the classifications in DA, much as b-coefficients are used in regression in making predictions. The constant plus the sum of products of the unstandardized coefficients with the observations yields the discriminant scores.

Standardized Discriminant Coefficients, also termed the *standardized canonical discriminant function coefficients*, are used to compare the relative importance of the independent variables, much as beta weights are used in regression.

Sensitivity is defined as the proportion of subjects predicted to have the outcome who really have it (true-positive test) and calculated as [true positives/ (true positives + false negatives)] x 100.

Specificity is defined as the proportion of subjects predicted not to have the outcome who do not have it (true-negative test) and is calculated as [true negatives/ (true negatives + false positives)] x 100.

Positive Predictive Value (PPV) is defined as the percentage of individuals with a positive test result who actually have the disease and is calculated as [true positives/ (true positives + false positives)] x 100.

Negative Predictive Value (NPV) is defined as the percentage of individuals with a negative test result who actually do not have the disease and is calculated as [true negatives/ (true negatives + false negatives)] x 100.

Misclassification (error) Rate is defined as the percentage of individuals who have been misclassified and is calculated as [(false positives + false negatives)/ sample size] x 100.

Accuracy is a measure of the degree of conformity of a measured or calculated quantity to the actual value. It is calculated as the proportion of the true results of a binary classification test (true positive and true negative) among all possible results i.e. [(true positives + true negatives)/ sample size]

Abstract

Introduction

AIDS surveillance has been the cornerstone of national efforts to monitor the spread of HIV infection in the world and to target HIV-prevention programs and health-care services. HIV case surveillance provides data to better characterize populations in which HIV infection has been newly diagnosed, including persons with evidence of recent HIV infection such as adolescents and young adults. The aim of this work after summarizing the properties of the two discriminating methods is to explore the convergence (give same results) and divergence (give different results) of the two analytical methods when they are used to classify participants as “HIV positive” or “HIV negative” using the symptoms of HIV.

Methods

The study was conducted in Chimanimani district, Manicaland Province of Zimbabwe in 2005.This was secondary data analysis of data from baseline study that utilised the Household survey of HIV-prevalence and behaviour in Chimanimani District, Zimbabwe which sort to quantify the magnitude of HIV and AIDS problem among children and adults; determine the knowledge ,attitudes, behaviour and practice of the general population; identify prevention and care programmes and human rights issues concerning HIV and AIDS among the general public and provide evidence-based information to policy makers on HIV and AIDS preventive mitigatory needs.

Results

The data was analysed using discriminant analysis (DA) and logistic regression (LR). The results of the DA showed that 87.4% of the cases were correctly classified as either HIV-

positive or HIV-negative whilst LR managed to classify 89.1% of the same cases. LR identified 11 variables that include swollen lymph nodes, burning urine, clothes too large which were not picked by DA as significant. Comparing the results obtained from logistic regression and discriminant analysis indicate that the two techniques gave almost the same percentage of correct classification, different error rates and kappa coefficients but have a very high overall kappa coefficient of 0.98. The Logistic Regression model had positive coefficients whilst the Discriminant Analysis model had negative coefficients but the two can correctly identify the almost the same number of patients who are HIV-positive or HIV- negative. This means that the clinician can apply any model to classify a patient.

Overall HIV prevalence among the 15-49 year age group was 15.1%, 95% CI = [13.1-16.9].

Conclusions

The logistic model has proved to be an efficient tool for classifying patients as HIV-positive or negative. It has shown that it can correctly classify 89.1% of the patients who come presenting with HIV symptoms whilst the discriminant mode can correctly classify 87.4% of the patients. LR has a lower error rate (10.9) and a higher Predictive Value Positive (79.6) compared to DA's 12.6 error rate and 30.5 Predictive Value Positive. The results of Logistic regression model were closer to those of discriminant analysis model and any model can be used.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1:

Background

1.0 Background

AIDS surveillance has been the cornerstone of national efforts to monitor the spread of HIV infection in the world and to target HIV-prevention programs and health-care services. Although AIDS is the end-stage of the natural history of HIV infection, in the past, monitoring AIDS-defining conditions provided population-based data that reflected changes in the incidence of HIV infection. However, recent advances in HIV treatment have slowed the progression of HIV disease for infected persons on treatment and contributed to a decline in AIDS incidence. These advances in treatment have diminished the ability of HIV/AIDS surveillance data to represent trends in the incidence of HIV infection or the impact of the epidemic on the health-care system. As a consequence, the capacity of local, state, and public health agencies to monitor the HIV epidemic has been compromised.[18]
In response to these changes in HIV treatment practices and the information needs of public health and other policymakers, Centre for Disease Control (CDC) and Council of State and Territorial Epidemiologists (CSTE) have recommended that all states and territories extend their AIDS case surveillance activities to include HIV case surveillance and the reporting of HIV-exposed infants.

The global HIV /AIDS epidemic is far more extensive than initially anticipated particularly in Sub-Saharan Africa (SSA), where the tragedy is of epic proportions and has altered the region’s demographic outlook. It is reducing life expectancy, raising mortality, lowering fertility, creating an excess of men over women and leaving millions of orphans in its wake. Africa remains the global epicentre of the AIDS pandemic with SSA being the worst affected.

Human immunodeficiency virus (HIV) infections in Sub-Sahara Africa, in particular Central and Southern Africa, are a major health problem. Sub-Sahara is known to be world`s region that is most affected by the epidemic. It is home to 10% of the world`s but has 70% of the global total HIV infected persons.

The first reported case of AIDS in Zimbabwe occurred in 1985. By the end of the 1980s, around 10% of the adult population were thought to be infected with HIV. This figure rose dramatically in the first half of the 1990s, peaking and stabilising at 29% between 1995 and 1997. But since this point the HIV prevalence is thought to have declined, making Zimbabwe one of the first African nations to witness such a trend. According to government figures, the adult prevalence was 24.6% in 2003 and fell to 15.6% in 2007. [23]

Yet although survey results indicate a fall in Zimbabwe's adult HIV prevalence, caution should be taken when interpreting the data available; it is not yet known whether the trend is a sign of long-term change or merely a temporary movement. In Chimanimani district, which is predominantly rural and lies 155km south-east of the provincial capital Mutare in Manicaland Province, has an overall HIV prevalence of 9.7% which increased with age. It was highest in the urban sector (14.6%) followed by large-scale commercial sector at 12.2% and was lowest in communal and small-scale sectors both at 8%. The symptoms that were associated with HIV prevalence in this district include weight loss, fever, sores on the mouth and skin, swollen lymph nodes, recurrent headaches, persistent cough, shingles/herpes zoster and Tuberculosis among others. These symptoms and others can be modelled to predict the HIV status of a patient using statistical methods like logistic regression and linear discriminant analysis. These models can be used by clinicians in resource limited areas like Chimanimani to screen patients as HIV positive or HIV negative and can determine the World Health Organisation (WHO) or Centre for Disease Control (CDC) stage of the patient and can take appropriate action to serve the patient.

1.1 Research Question

Are there any differences between Logistic Regression and Discriminant Analysis in evaluating the symptoms associated with HIV prevalence?

1.2 Problem statement

The 15-49 year group is an important group for monitoring reduction of HIV incidence in the population as elucidated in the United Nations General Assembly Special Session (UNGASS) on HIV and AIDS whose primary objective is to reduce the transmission rate in this age group. An understanding of the extent and magnitude of HIV infection among the 15-49 is important in the prioritisation, design, implementation and evaluation of programmes to support the HIV-infected people and provides a more realistic and useful estimate of resources needed for patient care. Reliable data on HIV incidence are essential for monitoring the spread of HIV and for impact assessment of sexual health interventions. Unfortunately, incidence data are difficult and costly to collect in Zimbabwe. This age group (15-49) changes sexual partners frequently and expose themselves to the risk of infection with HIV and other sexually transmitted infections (STIs).

In Zimbabwe the health centres are receiving an increasing number of patients presenting with varying HIV symptoms. However, the services for HIV testing are sometimes not available especially in rural areas and the clinician cannot put the patient on HIV/AIDS treatment before conclusive test have been done, usually in urban areas or district hospitals where most patients cannot afford to go. The HIV status of a patient can be predicted by applying mathematical modelling on the symptoms presented by a patient to the clinicians. The association between these symptoms and HIV prevalence has not been explored and is making it difficult for service providers to give appropriate intervention especially in resource limited areas. Therefore we need to put a system in place that will be used by all health centres to capture everyone who is HIV-positive using the symptoms of HIV.

Chapter 2: Literature Review

In their paper, Brenn and Arnesen [7], compared the ability of discriminant analysis, logistic regression, and Cox model when applied in a dataset of 6595 men aged 20–49, who were followed for 9 years for total and coronary deaths, in order to select possible risk factors. People in the population sampled were divided into two groups, one with mortality 5 per 1000 and one with 93 per 1000. Logistic regression and Cox model derived the same set of variables, and discriminant analysis set of variables had only minor differences. The researchers also noticed that a time-saving option, offered for both the logistic and Cox selection, showed no advantage compared with discriminant analysis, since by analyzing more than 3800 subjects, the logistic and Cox methods consumed, respectively, 80 and 10 times more computer time than discriminant analysis. Thus, the researchers reached to the conclusion that discriminant analysis is preferred for preliminary or stepwise analysis, otherwise Cox method should be used to select possible risk factors.

In a study by Montgomery et al.[8], who compared the two methods in veterinary data using stepwise linear discriminant analysis and logistic regression in a first dataset and comparing the selected variables, the order of selection and the sign and the magnitude of the estimated coefficients of the discriminating models in a second dataset, resulted that although both methods converged logistic regression is preferable to discriminant analysis particularly when the assumptions of normality and equal variance are not met.

Pohar et al.[6], in their study which used several simulated datasets and discrimination indexes, found that the two methods converge when the linear discriminant assumptions for normality of the distribution of explanatory variables are met. The authors concluded that linear discriminant analysis is a more appropriate method when the explanatory variables are normally distributed. For categorized predictor variables, linear discriminant analysis remains preferable, and logistic regression overcomes discriminant analysis only when the number of categories is small (2 or 3). When the assumptions of linear discriminant analysis are not met, the usage of it is not justified, while logistic regression gives good results regardless of the distribution of the predictors.

According to George Antonogeorgos et al [16], linear discriminant function is a better method than logistic regression when the normality assumptions are met, the differences between them become negligible when the sample size is large enough (50 observations or more). Symptoms of HIV vary according to what stage of the infection you are in. According to M. Cichocki et al [22], the symptoms of HIV and AIDS include diarrhoea, thrush, weight loss, fatigue, burning of feet and hands among others.

Many people who have been infected with HIV have few or no HIV symptoms initially; testing is the only way to know for sure if you are infected with HIV. There are, however, assortments of HIV symptoms that can be associated with a new infection.

HIV disease staging and classification systems are critical tools for tracking and monitoring the HIV epidemic and for providing clinicians and patients with important information about HIV disease stage and clinical management. Two major classification systems currently are in use: the U.S. Centers for Disease Control and Prevention (CDC) classification system and the World Health Organization (WHO) Clinical Staging and Disease Classification System.[20]

The CDC disease staging system (last revised in 1993) assesses the severity of HIV disease by CD4 cell counts and by the presence of specific HIV-related conditions. The definition of AIDS includes all HIV-infected individuals with CD4 counts of <200 cells/µL (or CD4 percentage <14%) as well as those with certain HIV-related conditions and symptoms. Although the fine points of the classification system rarely are used in the routine clinical management of HIV-infected patients, a working knowledge of the staging criteria (in particular the definition of AIDS) is useful in patient care. In addition, the CDC system is used in clinical and epidemiologic research. [17]

In contrast to the CDC system, the WHO Clinical Staging and Disease Classification System (revised in 2005) can be used readily in resource-constrained settings without access to CD4 cell count measurements or other diagnostic and laboratory testing methods. The WHO system classifies HIV disease on the basis of clinical manifestations that can be recognized and treated by clinicians in diverse settings, including resource-constrained settings like Zimbabwe, and by clinicians with varying levels of HIV expertise and training.

In emergency departments and family practice offices, people come in with symptoms like fever, headache, muscle and joint pain, sore throat, rash and diarrhoea. In response to these symptoms, physicians diagnose the flu and send the patient on their way. In the majority of cases, their diagnosis proves correct. But unfortunately, a number of people with these vague, indistinct symptoms have a more serious illness than the flu; these symptoms may signal the acute stages of HIV infection.

The immune system is the body's natural defense against invading foreign agents such as viruses and bacteria. HIV attacks the immune system, weakening it and making the body more susceptible to infection. After infection, some symptoms affect the immune system. Symptoms include:

1. Swollen lymph nodes in the neck, axilla, or groin (lymphadenopathy)

ii) Fever / night sweats

HIV can have a profound effect on the gastrointestinal system and nutrition. Like any virus, HIV can cause symptoms ranging from poor appetite to diarrhoea. Here are some of the most common symptoms:

1. Rapid weight loss
2. Diarrhoea
3. Poor appetite
4. Profound fatigue

Some of the most serious symptoms of HIV involve the respiratory system. Symptoms associated with breathing and respiration usually occur later in course of the disease. Unfortunately, many people's first symptom of HIV involves the respiratory system, illustrating why HIV testing is so important. Any cough or shortness of breath could be a sign of bacterial pneumonia or pneumocystis (carinii) jiroveci pneumonia (PCP). Symptoms include:

1. Shortness of breath
1. Dry cough / productive cough

Many of the symptoms of HIV are a result of infection by other viruses, fungi or bacteria. Often, these infections involve the skin or mucous membranes (for example, inside the mouth).

1. Skin rash
2. Red, brown, pink or purple lesion on the skin - suggestive of Kaposi's sarcoma
3. White patches on the tounge, inside of the mouth, or gums - could be signs of thrush, candida, or hairy oral leukoplakia
4. Canker sores / apthous ulcers

Very early on in the epidemic, experts realized that HIV affected both the physical and emotional self. Neurological and emotional symptoms of HIV can include the following:

1. Depression
2. Numbness, tingling, or burning in the feet, hands, or face. (peripheral neuropathy)
3. Confusion, weakness, or changes in level of consciousness

2.1 Justification

In many African countries, the availability of HIV testing is limited in clinical and community settings; diagnosis and management are often based on clinical presentation and use of the WHO clinical staging system. According to National Aids Council, Zimbabwe has made great progress in prevention of AIDS as seen by the decline in the HIV/AIDS prevalence from 33.7% in 2001 to 15.6% in 2007. This figure remains too high and there is need to identify the high risk subpopulation with potential of spreading HIV, establish their status and put interventions in place to curb the spread of the pandemic. While surveys typically have national coverage and generate data for women and men in urban and rural areas, they also have limitations. The main limitations are the potential for bias introduced by non-response and the exclusion from the sampling frame of population groups at high risk of HIV infection [7]. Also population-based testing can only be undertaken every 3 to 5 years in most countries, because of the size and expense of the surveys. Sentinel surveillance testing is often reported annually, and provides a good benchmark for measuring progress over short time periods.

In addition, population-based testing is dependent on the population's willingness to be voluntarily tested for HIV. In cases where the characteristics of those who agreed to be tested are different than those who refused testing, bias may result. Hence there is need to develop a predictive or discriminant model that can be used to predict someone's HIV status or discriminate people whether they are HIV positive or negative. We have decided to use two statistical methods; Logistic Regression and Discriminant Analysis to analyse the factors associated with HIV prevalence in addition to finding out where the two statistical methods converge and diverge and control for confounding. This approach may allow health care providers in low resource settings to gather the data at much faster rate and use to predict/discriminate HIV positive patients and provide care and offer HIV/AIDS prevention programmes the strategic opportunity to reach subpopulations that are at risk of HIV infection.

2.2 Purpose of Study

The aim of this work is to look at the predictive and discriminant methods and explore where they converge (agree) and diverge (disagree) when the two analytical methods (logistic regression and discriminant analysis) are used to classify participants as “HIV positive” or “HIV negative”.

Logistic regression and linear discriminant analyses are multivariate statistical methods which have been used for the evaluation of the associations between various covariates and categorical outcomes in HIV prevalence data. Both methodologies have been extensively applied in research, especially in medical and sociological sciences.

Although the theoretical properties have been studied extensively throughout the literature, the choice of the proper method in data analysis is still a question for the researcher. This study aims to summarize the properties of the two discriminating methods and explore the convergence and divergence of the two analytical methods when they are used to classify participants as “HIV positive” or “HIV negative”. The revised HIV classification system provides uniform and simple criteria for categorizing conditions among adolescents and adults with HIV infection and should facilitate efforts to evaluate current and future health-care and referral needs for persons with HIV infection. The addition of a measure of severe immune-suppression, as defined by a CD4+ T-lymphocyte count of less than 200 cells/uL or a CD4+ percentage of less than 14, reflects the standard of immunologic monitoring for HIV-infected persons and will enable AIDS surveillance data to more accurately represent those who are recognized as being immune-suppressed, who are in greatest need of close medical follow-up, and who are at greatest risk for the full spectrum of severe HIV-related morbidity. The addition of three clinical conditions; pulmonary TB, recurrent pneumonia, and invasive cervical cancer; to AIDS surveillance criteria reflects the documented or potential importance of these diseases in the HIV epidemic. Two of these conditions (pulmonary TB and cervical cancer) are preventable if appropriate screening tests are linked with proper follow-up. The third, recurrent pneumonia reflects the importance of pulmonary infections not included in the 1987 definition as leading causes of HIV-related morbidity and mortality. Successful implementation of expanded surveillance criteria will require the extension of existing safeguards to protect the security and confidentiality of AIDS surveillance information.

The expanded AIDS surveillance case definition is expected to have a substantial impact on the number of reported cases. The immediate increase in case reporting will be largely attributable to the addition of severe immune-suppression to the definition; a smaller impact is expected from the addition of pulmonary TB, recurrent pneumonia, and invasive cervical cancer, since many persons with these diseases will also have CD4+ T-lymphocyte counts of less than 200 cells/uL. However, not all persons are aware of their HIV infection and of those who know their HIV infection status, not all have had an immunologic evaluation; thus, the immediate impact on the number of AIDS cases will be considerably less.

Based on current levels of HIV and CD4+ testing, CDC estimates that the expanded definition could increase cases reported in 1993 by approximately 75%. Early effects of expanded surveillance will be greater than long-term effects because prevalent as well as incident cases of immune-suppression will be reported following implementation of the expanded surveillance case definition. In subsequent years, the effect on the number of reported cases is expected to be much smaller.

2.4 Study Objectives

Broad Objective

To evaluate the divergence and convergence of Logistic Regression and Discriminant Analysis in assessing factors associated with HIV prevalence using the HIV symptoms

Specific Objectives

To determine the HIV prevalence by demographic characteristics

To provide a predictive model for HIV prevalence using Logistic Regression

To provide a predictive model for HIV prevalence using Discriminant Analysis

To compare the two models for convergence and divergence

Chapter 3.0: Methodology

3.1 Design

This was secondary data analysis from a baseline study that utilised the Household survey of HIV-prevalence and behaviour in Chimanimani District, Zimbabwe which sort to quantify the magnitude of HIV and AIDS problem among orphans and vulnerable children; determine the knowledge ,attitudes, behaviour and practice of the general population; identify prevention and care programmes and human rights issues concerning HIV and AIDS among the general public and provide evidence-based information to policy makers on HIV and AIDS preventive mitigatory needs.

3.2 Data Sources

The Behavioural Risks and HIV Serostatus (BSS) survey employed the second-generation behavioural surveillance model that involves combining both biological (i.e. HIV testing) and behavioural surveillance surveys in the same study in order to determine the link between the two. In this cross-sectional survey, 13 wards were randomly selected from a total of 23 wards of Chimanimani district, located in the Manicaland Province of Zimbabwe. In each of the 23 wards, individuals in four age strata of 2-11 and 12-14 years for children, 15-24 years for youths and 25+ years for adults were randomly selected based on proportional sampling using the total populations of each age group in each ward. The participants were identified in randomly selected households in the 13 wards. Age specific questionnaires on demographic, socioeconomic and child care practices and HIV-related behavioural aspects were administered and blood samples collected for HIV testing.

3.3 Laboratory Data

Dried Blood Spots (DBS) samples were collected on labelled Whatman No. 3 filter paper, which was cut into three strips, from participants in all the age groups. Oral Mucosal Transudate (OMT) was collected from children aged 2-14 years only using the Orasure device which was then placed in the labelled (same bar code as on DBS) plastic vial containing preservative fluid. The samples were checked for quality and quantity before storing at room temperature and then transported to the laboratory within 14 days. At the lab, the samples were checked for state, quality and quantity and information on the form and on the specimen. The specimens were stored in a freezer at -20^{0}C until tested for HIV using the Vironostika Uniform II commercial ELISA kits.

3.4 Sample size

Using an STI prevalence of 35% among HIV positive young adults from the Zimbabwe Young Adult Survey (YAS) a sample size was calculated viz:

$$n = \frac{z^2\, p(1-p)}{\Delta^2}$$

where: z = 1.96, Δ (the precision) = 0.05 and p = 35%

$$n = \frac{1.96^2\,(0.35)(0.65)}{0.05^2}$$

Sample size = 350

A minimum sample size of 350 participants was needed for the study to be valid. For this study, we focused on demographic characteristics and Health status (including history of hospitalisation) so that we can assess for confounding and develop predictive/discriminant models for HIV prevalence using HIV symptoms. All participants aged between 15 and 49 years who met the inclusion/exclusion criteria and took part in the baseline survey were included in the study. (see Appendix 1 for details on how the participants were selected into the primary study)

3.5 Selection Criteria

Inclusion criteria

Males and females aged between 15 and 49 years who participated in the baseline household census were considered eligible for the study.

Exclusion Criteria

Males and females aged below 15 years and over 49 years who participated in the baseline survey will not be considered eligible for this study.

3.6 Statistical Methods

Data set from 'The household survey of HIV-prevalence and behaviour in Chimanimani District' and the 'Laboratory Results' were merged and data for those aged between 15 and 49 was used for analysis. Also the HIV/AIDS Behavioural Risks and Surveillance Survey (BSS) Questionnaire was used as source of data

The section looks at the statistical methods used. Descriptive statistics and graphical statistics were performed on the study variables and means and standard deviations were presented on continuous data whilst frequencies were presented on categorical variables. Principal Component Analysis (PCA) was used to select the most significant components that determine the HIV status of a person from the HIV symptoms the participants were asked on their health status. These components were fed into the Logistic Regression (LR) and

Discriminant analysis (DA) models to come up with a predictive and discriminant models to determine if one is HIV positive or negative. The two models were compared to see where they converge and diverge in predicting the HIV status of a person.

3.6.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. It is an approach to factor analysis that considers the total variance in the data, which is unlike common factor analysis. It is recommended when the researcher's primary concern is to determine the minimum number of factors that will account for the maximum variance in the data in use in the particular multivariate analysis. While conducting principal component analysis (PCA), the researcher can get well versed with standard deviation, covariance, eigenvectors and eigenvalues. The eigenvalues in principal component analysis (PCA) refers to the total variance explained by each factor. The standard deviation in principal component analysis (PCA) measures the variability of the data. The task of principal component analysis (PCA) is to identify the patterns in the data and to direct the data by highlighting their similarities and differences. In this case, PCA will be used to identify independent variables (HIV symptoms) that will account for the total variance in the data and will inform the logistic regression and discriminant analysis models. (see Appendix 4 for details)

3.6.2 Logistic Regression

Logistic regression is a form of regression which is used when the dependent variable is dichotomous, discrete, or categorical, and the explanatory variables are of any kind. In medical sciences, the outcome is usually the presence or absence of a stated situation or a disease. Using the logit transformation, logistic regression predicts always the probability of group membership in relation to several variables independent of their distribution. The logistic regression analysis is based on calculating the odds of the outcome as the ratio of the probability of having the outcome divided by the probability of not having it. (see Appendix 3 for detailed analysis)

3.6.3 Discriminant Analysis

Discriminant analysis is a similar classification method that is used to determine which set of variables discriminate between two or more naturally occurring groups and to classify an observation into these known groups. In order to achieve that discriminant analysis is based on the estimation of the orthogonal discriminant functions, the linear combination of the standardized independent predictor variables gives the greatest means differences between the existing groups. Thus, it can be proposed that both discriminant analysis and logistic regression can be used to predict the probability of a specified outcome using all or a subset of available variables. (see Appendix 2 for detailed analysis)

3.7 Statistical Analysis

Data on demographic characteristics, health status and HIV status was analysed using STATA version 10. Testing for presence of HIV antibody at baseline was performed on dried blood spots. In order to evaluate HIV symptoms in the study sample, the participants completed questions on their health status. The independent variables that were associated with the dependent variable, "presence of any HIV symptoms", at a significance level $\alpha = 0.05$ were entered in a principal components analysis (PCA). Kaiser's criterion (eigen value > 1), was applied so that we retain factors that are mutually independent.

The retained components were the predictor variables that were entered in both discriminant and logistic regression models (step-wise). The assumptions for the two models were tested for multivariate normal distribution and mutual independence, absence of multi-collinearity, equivalence of variance-covariance matrices of the groups using Box's M test (p-value > 0.05). We used the standardized canonical discriminant function coefficients and the unstandardized function coefficients for discriminant analysis and Z-statistic (squared Wald statistic) for logistic regression, to evaluate how much each one of the variables contributed to the discrimination between two groups. The contribution of the respective variables to the discrimination depended on how large the coefficients were. We also compared the signs and magnitudes of coefficients.

For each model, we plotted the corresponding response operating characteristics (ROC) curve. An ROC curve graphically displays sensitivity and 100% minus specificity (false positive rate) at several cut-off points. By plotting the ROC curves for two models on the same axes, one is able to determine which test is better for classification, namely, that test whose curve encloses the larger area beneath it. All analyses were performed using the Stata version 10 software.

3.8 Variables

The dependent variable was the 'HIV status' which was assessed by the following independent variables: (i) weight loss, (ii) frequent burning urine, (iii) regular lower abdominal pains, (iv) breast abscess, (v) diagnosed with STIs in the last three months, (vi) genital sores or discharge or ulcers in the last three months, (vii) genital warts, (viii) diarrhoea for more than three days in the last three months, (ix) fever in the last three months, (x) white sores in the mouth in the last three months, (xi) swollen lymph nodes (glands), (xii) persistent coughs for one month or more, (xiii) herpes zoster/shingles in the last 12 months, (xiv) sores on the skin in the last three months, (xv) TB treatment in the last year, (xvi) treated for pneumonia in the last year and (xvii) weak or dizzy in the last three months.

The following demographic independent variables were considered: (i) Age, (ii) Sex, (iii) Educational level, (iv) Religion, (v) Sector and (vi) Orphanhood status.

Table 1: Variables

VARIABLE	DESCRIPTION
Response:	
HIV Status	HIV status of the patient- positive or negative
Explanatory:	
Age	Age of respondent in years (completed)
Sex	Identifies the sex of respondent
	Codes: 1 – Male 2 – Female
Education	Highest educational level of respondent
	Codes: 1-No schooling 2-Primary school 3-Secondary school 4-Tertiary
Sector	Identifies the setting were the interview was conducted
	Codes:1-Communal 2-Resettlement 3-Large scale commercial 4-Small scale official 5-Urban

Religion	Identifies the faith or religious grouping to which the respondent belongs. Codes: 1-Catholic 2-Protestant 3-Pentecostal 4-Apostolic sects 5-Moslem 6-Traditional African 7-Atheist 8-Other
Marital status	Identifies the current marital status of respondent Codes: 1-Married 2-Single/Never married 3-Divorced/Separated 4-Living together/Cohabiting 5-Widowed
Weight change	Identifies whether the recipient lost/gained weight in the last 6 months Codes: 1 – Yes 2 – No
Burning urine	Identifies whether respondent was frequently troubled by burning urine Codes: 1 – Yes 2 – No
Abdominal pains	Identifies whether respondent was having severe lower abdominal pains on a regular basis Codes: 1 – Yes 2 – No
Breast Abscess	Identifies whether a woman respondent had a breast abscess Codes: 1 – Yes 2 – No
STI	Shows whether a respondent was diagnosed with sti during the last three months Codes: 1 – Yes 2 – No
Sores/Ulcers on genitals	Shows whether a respondent had had sores/ulcers on genital organs over the last three months Codes: 1 – Yes 2 – No
Abnormal discharge	Shows whether a respondent had any abnormal discharge from genitals in the last three months Codes: 1 – Yes 2 – No
Genital warts	Identifies whether a respondent had genital warts Codes: 1 – Yes 2 – No
Diarrhoea	Shows whether the respondent had diarrhea that lasted for more than three days in the last three months Codes: 1 – Yes 2 – No
Fever	Identifies whether the respondent had fever for more than one month on end for the last three months
White sores	Identifies whether the respondent had had white sores in the mouth or skin over the last three months Codes: 1 – Yes 2 – No
Swollen lymph nodes	Identifies whether a respondent had swollen lymph nodes in the neck, under the arms or groin Codes: 1 – Yes 2 – No
Pneumonia	Shows whether a respondent was treated for pneumonia more than once during the last year
Difficulty swallowing	Identifies whether a respondent had difficulty swallowing solid foods compared to liquids Codes: 1 – Yes 2 – No
Recurrent headaches	Identifies whether a respondent had recurrent headaches through day and night Codes: 1 – Yes 2 – No
Shingles/Herpes zoster	Identifies whether a respondent had had shingles/herpes zoster over the last 12 months Codes: 1 – Yes 2 – No
TB	Identifies whether a respondent is currently on treatment for TB or have

	been on treatment during the last year Codes: 1 – Yes 2 – No
Persistent cough	Identifies whether a respondent had had a persistent cough for one month or more Codes: 1 – Yes 2 – No
Felt weak/dizzy	Identifies whether a female respondent felt weak/dizzy during the last three months Codes: 1 – Yes 2 – No
Circumcision	Identifies whether a male respondent had had circumcision Codes: 1 – Yes 2 – No
Stistatus	Identifies whether a respondent had sores or ulcers on genital organs, diagnosed with sti, abnormal discharge from genitals and genital warts in the last 3 months Codes: 1 – Yes 2 – No
Skin	Identifies whether a respondent had white sores on the mouth, sores on the skin or shingles/herpes zoster in the last 3 months Codes: 1 – Yes 2 – No

3.9 Data cleaning and coding

Data cleaning and coding was done in Stata. The data for categorical variables were coded and a coding manual is shown in Table 2 below. Frequency tables to check for unusual values were done for all the variables. Outliers for continuous variables and out of range values for discrete were verified and corrected.

Table 2: Codes for Categorical Data

VARIABLE	CATEGORY	CODE
Gender	Female	0
	Male	1
Marital status	Married	1
	Single/never married	2
	Divorced/separated	3
	Widowed	4
Agegroup	15-19	1
	20-24	2
	25-29	3
	30-34	4
	35-39	5
	40-44	6
	45-49	7
HIV status	Negative	0
	Positive	1

3.9.1 Data Analysis issues

The study was a population-based survey based on a random sample and may be considered to be representative enough of the population of Zimbabwe. However, bias could have been introduced during data collection by research assistants when counting households using a given interval since there was no systematic way of counting in the communal areas where the households are also randomly placed. Also bias could have been introduced by the selection of respondents which was based on availability and need to balance sexes. Usually women are at home most of the time compared to men and whether the men found at home had the same characteristics as those not at home could not be verified. The other source of bias was the number of children included in the study, which was small, because the study was done during school days and it could not be ascertained whether those interviewed had the same characteristics as those not interviewed. Another weakness noted was that there was no control for confounding; hence the interpretation of associations between HIV and various variables was very limited. In this study we looked at multivariable analysis to control for confounding.

Chapter 4.0: Results

Preamble

This chapter looks at the demographic characteristics (gender, age group, marital status, area of residence, educational level, religion, HIV status and circumcision) of participants and the prevalence of HIV by the selected variables. It also looks at the predictive variables using logistic regression and the discriminating variables using discriminant analysis including their corresponding Receiver Operating Curves (ROC) to find the model that is better in classifying patients visiting any health centre as HIV-positive or HIV-negative. The two models are then assessed for convergence and divergence when used to predict the HIV status of patients

4.1 Demographic characteristics

The table below shows the distribution of participants by selected demographic characteristics. From a total of 934 participants, 521 (55.8%) were females and 413(44.2%) were males. In this study, 51.6% of the participants were married, 40.0% were single and the rest were divorced, widowed or co-habiting. The participants were put into 7 age categories and most of them (27.6%) were in the 20-24 age group and the least number of participants (38) was in the age group 44-49. Their median age was 24 years, with the upper quartile being 31years and lower quartile 19 years showing that there were many participants in the first three age groups. This also means age distribution was skewed to the right. Education wise, 65.5% of them had had some secondary level of tuition while 2.5% and 1.6% had had no schooling and tertiary tuition respectively. 40.6% belonged to the apostolic sect, 25.4% attended Pentecostal churches, while 23.8% were Protestants and only 2 and 4 participants were Moslems and Atheist respectively. From the blood test done, 141 tested positive for HIV, giving a prevalence rate of 15.1% for the 15-49 age group and only 4.4% of the male respondents were circumcised whilst 24.2% of the females underwent the coming of age/initiation ceremony or ritual.

Table 3: Distribution of Participants by Demographic Characteristics

Variable	Frequency	Percentage
Gender (N=934)		
Male	413	44.2
Female	521	55.8
Marital status (N=934)		
Married	482	51.6
Single/never married	374	40.0
Divorced/separated	41	4.4
Widowed	12	1.3
Living together/cohabiting	25	2.7
Age group (N=934)		
15-19	251	26.7
20-24	258	27.6
25-29	149	16.0
30-34	109	11.7
35-39	83	8.9
40-44	46	4.9
45-49	38	4.1
Median age 24 (Q1= 19, Q3=31)		
Area of residence (N=934)		
Communal	386	41.3
Resettlement	230	24.6
Large scale farms	200	21.4
Small scale farms	36	3.9
Urban	82	8.8
Educational level (N=934)		
No schooling	23	2.5
Primary	284	30.4
Secondary	612	65.5
Tertiary	15	1.6
Religion (N=934)		
Catholic	47	5.9
Protestant	189	23.8
Pentecostal	202	25.4
Apostolic	322	40.6
Moslem	2	0.3
Traditional/African	28	3.5

Atheist	4	0.5
HIV status (N=934)		
Positive	141	15.1
Negative	793	84.9
Circumcision for Males (N=413)		
Yes	18	4.4
No	395	95.6
Initiation for women (N=281)		
Yes	68	24.2
No	213	75.8

The table below shows the prevalence of HIV by gender and agegroup. Overall HIV prevalence among the 15-49 year age group was 15.1%. The prevalence is highest in the 25-29 age group (38.0%) where 41(16 males and 25 females) where HIV positive. This is followed by 40-44 age group at 28.3% with 4 males and 9 females being HIV positive. The 15-49 age group had 4 males and 7 females who were HIV positive and thus giving it the lowest HIV prevalence (4.4%).

Table 4: HIV Prevalence by Gender for Variable Age group N=934

			HIV Status		
Variable			**Negative**	**Positive**	**P-Value**
Agegroup 15-19			**(n=240)**	**(n=11)**	0.000
	Gender	**Male**	115	4	0.000
		Female	125	7	0.000
Agegroup 20-24			**(n=224)**	**(n=34)**	0.000
	Gender	**Male**	100	9	0.000
		Female	124	25	0.000
Agegroup 25-29			**(n=108)**	**(n=41)**	0.000
	Gender	**Male**	47	16	0.000
		Female	61	25	0.000
Agegroup 30-34			**(n=87)**	**(n=22)**	0.000
	Gender	**Male**	43	4	0.000
		Female	44	18	0.001
Agegroup 35-39			**(n=68)**	**(n=15)**	0.000
	Gender	**Male**	26	4	0.000
		Female	42	11	0.000

Agegroup 40-44			**(n=33)**	**(n=13)**	0.003
	Gender	**Male**	20	4	0.001
		Female	13	9	0.394
			(n=33)	**(n=5)**	0.000
Agegroup 45-49					
	Gender	**Male**	20	1	
		Female	13	4	0.029

The table below shows the prevalence of HIV in Chimanimani district for the 15-49 year olds by gender and marital status. 19 out of 25 females living together/or cohabiting were found to be HIV positive giving the highest prevalence rate of 76.0%. This is followed by the widowed females at 50.0 % prevalence. 16 males and 11 females who were either single or never married had the lowest prevalence rate of 7.20%.

Table 5: HIV Prevalence by Gender for Variable Marital Status N=934

Variable			**Negative**	**Positive**	**P-Value**
Marital status					
Married			**(n = 403)**	**(n =79)**	0.000
	Gender	**Male**	153	21	0.000
		Female	250	58	0.000
Single/never married			**(n = 347)**	**(n = 27)**	0.000
	Gender	**Male**	216	16	0.000
		Female	131	11	0.000
Divorced/separated			**(n = 31)**	**(n = 10)**	0.001
	Gender	**Male**	2	5	0.257
		Female	29	5	0.000
Widowed			**(n = 6)**	**(n = 6)**	1.000
	Gender	**Male**	0	0	
		Female	6	6	1.000
Living together/cohabiting			**(n = 6)**	**(n = 19)**	0.009
	Gender	**Male**	0	0	
		Female	6	19	0.009

The table below shows the prevalence of HIV among the 15-49 year olds in Chimanimani district by educational level and gender. 10 males and 49 females with a primary level of education were found to be HIV positive giving an HIV prevalence rate of 20.8%. only 1 female with a tertiary level of education was found to be HIV positive and thus giving the lowest prevalence of 6.7%.

Table 6: HIV Prevalence by Gender for Variable Educational Level N=934

Variable			Negative	Positive	P-Value
Educational level					
No schooling			**(n = 19)**	**(n = 4)**	0.002
	Gender	**Male**	1	0	0.317
		Female	18	4	0.003
Primary			**(n = 225)**	**(n = 59)**	0.000
	Gender	**Male**	88	10	0.000
		Female	137	49	0.000
Secondary			**(n = 535)**	**(n = 55)**	0.000
	Gender	**Male**	274	10	0.000
		Female	261	45	0.000
Tertiary			**(n = 14)**	**(n = 1)**	0.001
	Gender	**Male**	8	0	0.005
		Female	6	1	0.059

The table below shows the HIV status of males who were asked whether they were circumcised or not. Only one out of eighteen males who were circumcised was found to be HIV positive and 10.0% of the uncircumcised males were HIV positive. The Chi2 value of 0.4386 is less than the critical value of 3.814; hence there is no relationship between being circumcised and getting HIV. This is also supported by the p-value of 0.508 which is far much greater than level of significance of 0.05.

Table 7: Circumcision and HIV Status N=413

	HIV Status of Respondent		
Have you been circumcised?	**Positive**	**Negative**	**Total**
Yes	1(5.56)	17(94.44)	18(100.0)
No	41(10.38)	354(89.62)	395(100.0)
Total	42(10.17)	371(89.83)	413(100.0)
	Chi2(1) = 0.4386		p-value=0.508

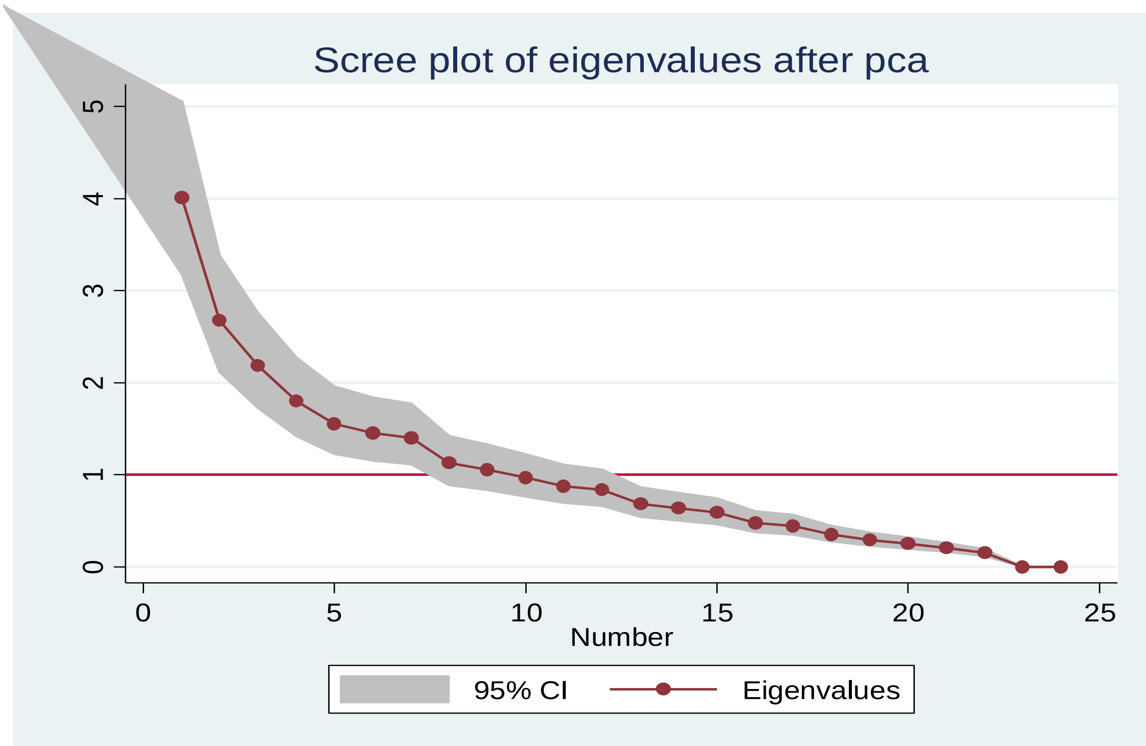
The table below shows the women who were asked if they underwent the initiation ceremony and their corresponding HIV status. Of the 68 female participants who underwent a coming of age/initiation ceremony or ritual, 11(16.2%) were found to be HIV positive and there is no statistically significant relationship between initiation and getting HIV because the p-value is greater than the significance level of 0.05

Table 8: Initiation and HIV Status N=281

	HIV Status of Respondent		
Did you undergo a coming of age/initiation ceremony or ritual?	**Positive**	**Negative**	**Total**
Yes	11(16.18)	57(83082)	68(100.0)
No	21(9.86)	192(90.14)	213(100.0)
Total	32(11.39)	249(88.61)	281(100.0)
	Chi2(1) = 2.0385		p-value=0.153

The univariate and multivariate correspondence analysis yielded 18 symptoms at the cut-off point of 0.25 so that they can be included in the logistic and discriminant models. These include: severe lower abdominal pains (0.026); STI (0.04); sores/ulcers on genitals (0.057); abnormal discharge from genitals (0.017); genital warts (0.016); diarrhoea (0.106); fever (0.000); white sores in the mouth (0.048); sores on the skin (0.180); swollen lymph nodes (0.052); difficulty swallowing solid foods (0.004); shingles/herpes zoster (0.007); currently on TB treatment (0.000) and persistent cough (0.009). A new variable called stistatus was created to encompass all the Sexually Transmitted Infections (STIs) i.e. sores/ulcers on genitals, genital warts, abnormal discharge from genitals and STI. Another variable called skin was also created to cover white sores in the mouth in the last 3 months, sores on your skin over the last 3 months and shingles/herpes zoster over the last 3 months. Using PCA and applying Kaiser's criterion, 9 patterns of our original data were extracted as shown in the scree plot below. These variables were used in both discriminant and logistic regression analyses, and both techniques revealed that the following symptoms: severe lower abdominal pains, fever, TB, difficulty swallowing solids, pneumonia and sores/ulcers on genitals were significant in determining the HIV status of an individual.

Figure 1: Scree plot of Eigen Values after PCA



The table below shows the prevalence of HIV among the 15-49 year olds by selected HIV symptoms. We observed that the following symptoms were statistically significantly associated with HIV-prevalence in Chimanimani district: feeling weak/dizzy (p=0.010), severe lower abdominal pain (p=0.008), genital warts (p=0.002), discharge from genitals (0.020), fever (p=0.026), swollen lymph nodes (p<0.001), difficulty in swallowing solids (p<0.001), shingle/herpes zoster (0.017), TB (p<0.001) and diarrhoea (p=0.018) in the last three months. Other symptoms like sores on genitals (p=0.402), sores on the skin (p=0.388), white sores on the mouth (p=0.614) and weight change (p=0.476) were found not to be statistically significantly associated with HIV-prevalence among the 15-49 year olds.

Table 9: HIV Prevalence by Selected Symptoms N=934

Symptom	**HIV Status**		**P-Value**
	Positive (n=141)	**Negative (n=793)**	
Severe Lower Abdominal Pain	29(20.6)	97(12.2)	0.008

Discharge from Genitals	9(6.4)	21(2.7)	0.020
Genital Warts	8(5.7)	12(1.5)	0.002
Fever	17(12.1)	53(6.7)	0.026
Swollen Lymph Nodes	14(9.9)	16(2.0)	<0.001
Difficulty in Swallowing solids	10(7.1)	16(2.0)	<0.001
Shingles/Herpes Zoster	3(2.1)	3(0.4)	0.017
TB	8(5.7)	11(1.4)	<0.001
Diarrhoea	17(12.1)	51(6.4)	0.018
Weak and Dizzy	31(22.0)	108(13.6)	0.010

Table 10: Predictive Variables for HIVSTATUS Using Logistic Regression Model

Variable	**OR**	**p-value**	**95% CI**
Fever	1.8	0.032	1.1 – 4.0
TB	5.5	0.000	2.3 – 19.5
Agegroup (35-39)	3.2	0.014	1.3 – 8.2
Agegroup (40-44)	8.2	0.018	1.2 – 8.7
Single/never married)	0.4	0.002	0.2 – 0.7
Divorced/separated)	2.1	0.032	1.1 – 4.1
Living together/cohabiting)	3.8	0.035	1.1 – 13.2
Widowed)	14.0	0.000	5.7 – 33.9
Shingles/Herpes zoster	4.8	0.059	0.9 – 24.9
Ststatus	1.3	0.666	0.4 – 3.8
Sex	1.3	0.271	0.8 – 2.1
Diarrhoea	0.7	0.274	0.3 – 1.4
Difficulty Swallowing	2.4	0.075	0.9 – 6.3
Genital warts	1.7	0.467	0.4 – 6.4
No education	2.7	0.141	0.7 – 10.1
Primary education	3.1	0.101	0.8 – 11.7
Secondary education	3.6	0.164	0.6 – 21.3
Tertiary education	3.2	0.400	0.2 – 48.0

From the analysis, patients who presented with TB or have been on TB treatment for the past year were 5.5 times more likely to be HIV-positive compared to those who did not present with TB, (p-value < 0.000 and Confidence Interval [2.3 – 19.5]). Also patients that presented with fever were 1.8 times more likely to be HIV-positive (p-value = 0.032 and Confidence Interval [1.1 – 4.0]) compared to those without fever. Those who presented with shingles/herpes zoster (p-value = 0.059) were 4.8 times more likely to be positive in comparison to those who did not although it was not statistically significant. Those in the age groups 35-39 and 40-44 were respectively 3.2 and 8.2 times more likely to be HIV-positive compared to those in the other age groups (p-values 0.014 and 0.018 respectively). The widowed were 14.0 times more likely to be HIV-positive (p-value < 0.000 and Confidence Interval [5.7 – 33.9]) whilst the single/never married were 60% less likely to be HIV-positive (p-value = 0.002 and Confidence Interval [0.2 – 0.7]). However education was statistically significantly not associated with HIV prevalence (as the p-values were > 0.05 and confidence intervals contained 0). The same applied to Stistatus, gender, genital warts, difficulty in swallowing and diarrhoea.

The table below presents the output from truncated logistic regression. The model predicting HIVSTATUS from age group, gender, marital status, religion, burning urine, genital warts and other symptoms was statistically significant (chi-squared = 125.75, p<0.001). The predictors in the table were statistically significantly associated with the prevalence of HIV at the 0.05 level of significance as the p=values are less than 0.05 and their corresponding confidence intervals do not contain a zero (0). The squared correlation between the observed and predicted HIVSTATUS values was 0.28 indicating that these predictors acconted for 28% of the variability in the outcome variable. The ancillary statistic (sigma) is equivalent to the standard error of estimate in OLS regression. The value of 0.3 can be compared to the standard deviation of HIVSTATUS which was 0.4. this shows a modest reduction. The model is given by:

HIVSTATUS=0.2 + 0.04Agegroup + 0.01Marital status + 0.09Sex + 0.03Religion + 0.1Burning Urine + 0.2 Genital warts + 0.3 Swollen Lymph Nodes + 0.3Difficulty Swallowing Solids + 0.3TB

Table 11: Truncated Logistic Regression Model for HIVSTATUS

HIVSTATUS	Coefficient	P-Value	[95% Confidence Interval]
Age group	0.04	<0.001	0.03 – 0.06
Marital status	0.01	<0.001	0.01 – 0.02
Sex	0.09	<0.001	0.01 – 0.14
Religion	0.03	0.016	0.01 – 0.05
Burning_urine	0.1	0.003	0.002 – 0.40
Genital Warts	0.2	0.008	0.06 – 0.40
Swollen Lymph Nodes	0.3	<0.001	0.20 – 0.50
Difficulty swallowing solids	0.3	<0.001	0.10 - 0.40
TB	0.3	<0.001	0.10 – 0.40
Constant	0.2	0.072	0.04 – 0.20
Sigma	0.3	<0.001	0.30 – 0.34

The table below shows the significant variables and the class functions that were used to classify an individual as HIV positive or negative using discriminant analysis. HIVSTATUS1 and HIVSTATUS2 were the two functions used to classify participants into two categories. If HIVSTATUS1 > HIVSTATUS2, then the participant was classified as HIV negative and HIV positive otherwise.

Table 12: Standardised Canonical Discriminant Function Coefficients and the Significant variables for the Discriminant Model

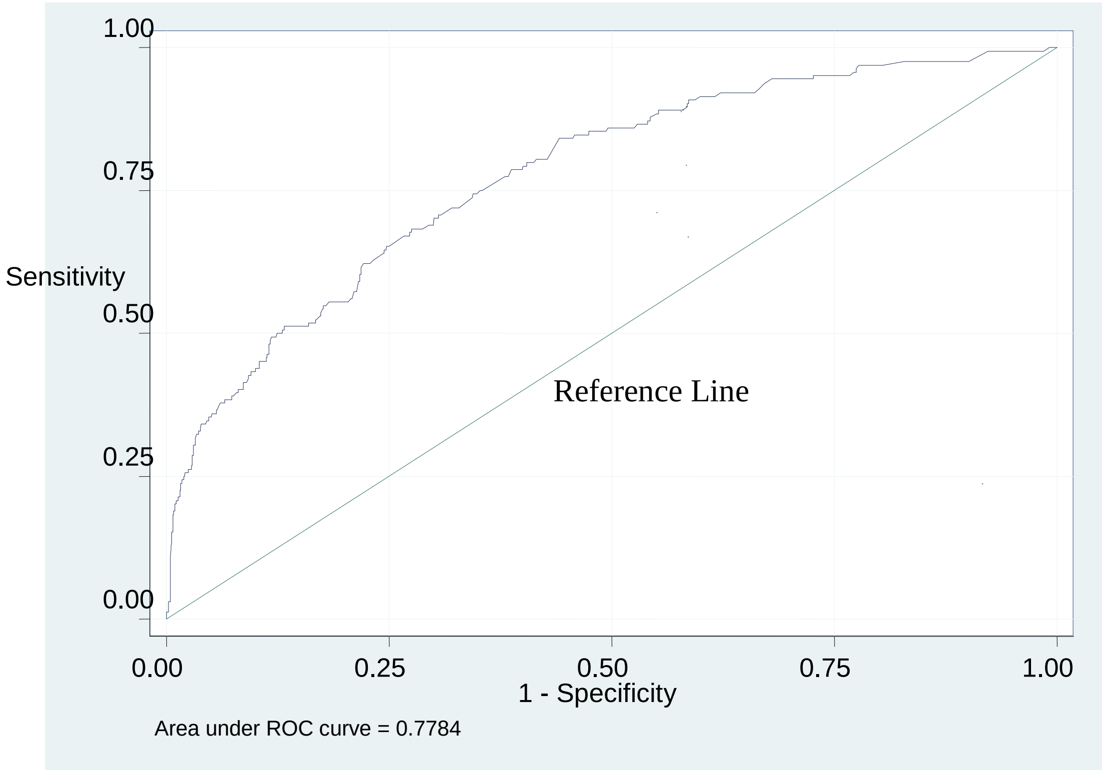
VARIABLE	Standardised Canonical Discriminant Function Coefficients	
	HIVSTATUS 1	HIVSTATUS 2
Marital Status	0.9	0.9
Age Group	2.1	2.3
Sex	8.5	8.9
Education	12.4	11.9
Severe Lower Abdominal Pain	1.1	2.6
STI	-6.6	-5.5
Sore/Ulcers on genitals	2.7	1.8
Abnormal Discharge from Genitals	5.9	8.7
Genital Warts	1.0	2.5
Diarrhoea	6.4	6.2
Fever	2.2	3.6
Pneumonia	3.6	4.7
Difficulty Swallowing Solid Foods	-6.0	-5.3
TB	2.1	5.3
Constant	**-31.7**	**-34.4**

Table 13: Standardised Canonical Discriminant Function Coefficients for the final Model

VARIABLE	Canonical Discriminant Function Coefficients
Marital Status	0.01
Age Group	0.2
Sex	0.3
Education	-0.3
Severe Lower Abdominal Pain	1.1
STI	-0.8
Sore/Ulcers on genitals	-0.7
Abnormal Discharge from Genitals	1.9
Genital Warts	1.0
Diarrhoea	-0.1
Fever	1.0
Pneumonia	0.8
Difficulty Swallowing Solid Foods	0.6
TB	-2.3
Constant	**0.5**

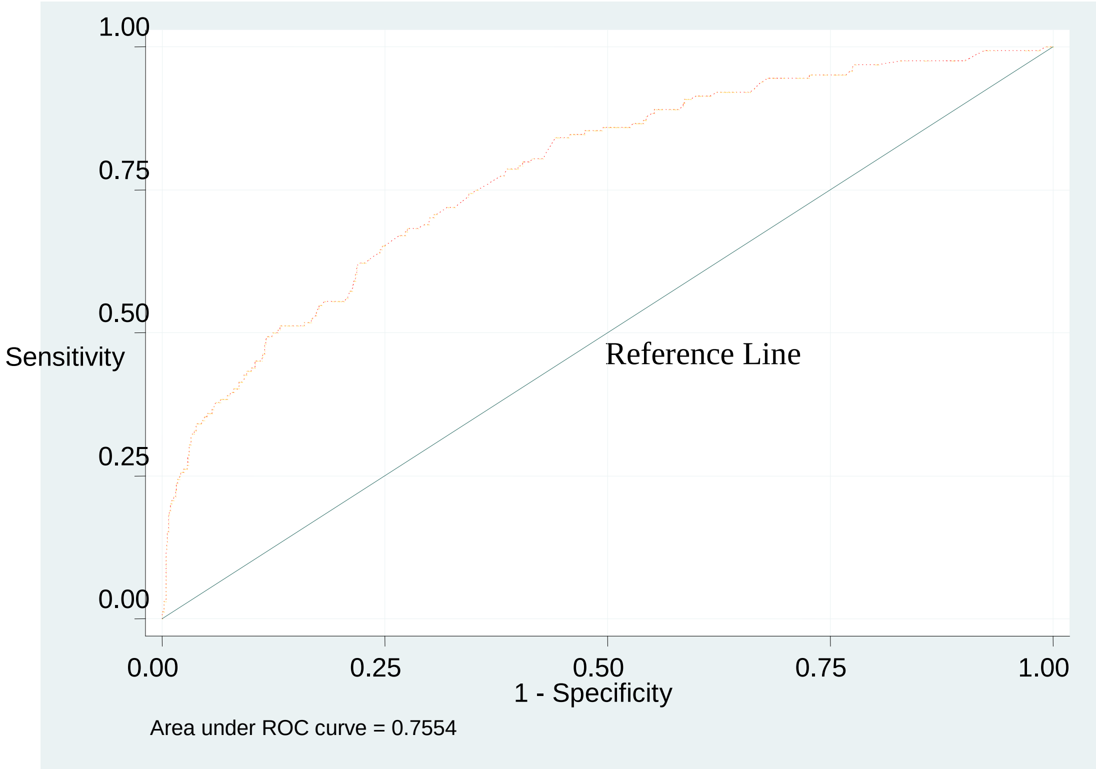
Although some differences are observed between the methods, as we can see in Figure 1 and Figure 2, the ROC curves of the models clearly indicate that the logistic model is similar to the discriminant analysis model (i.e., no difference in the area under the curve (AUC), 77.8% versus 75.5%). We also observed that the direction of the relationships was not the same, and there were extreme differences in the magnitude of the coefficients. This indicates a divergence in the two models in classifying a patient as HIV positive or HIV negative. However, the overall correct classification rate was 87.4% for discriminant analysis and 89.1% for logistic regression analysis showing that they converge.

Figure 2: Receiver operating characteristics (ROC) curve for logistic regression model.



The above curve explains 77.84 % of the variation in predicting HIV status of a patient.

Figure 3:Receiver operating characteristics (ROC) curve for discriminant analysis model.



The above curve explains 75.54 % of the variation in predicting HIV status of a patient.

Tables 10 and 11 below look at the classification tables of the logistic regression and discriminant analysis models for HIVSTATUS. They give the number of participants who were correctly classified as HIV-positive or HIV-negative including the false positive and false negative. The error rate and overall correct classification of each model, together with the agreement coefficient, Kappa, which provides a measure to which the observed and the predicted values concur in each model and between models was calculated.

Table 14: Classification using the Logistic Model

	Observed			
	Classified	**HIV Positive**	**HIV Negative**	**Total**
Predicted	**HIV Positive**	43	11	54
	HIV Negative	75	660	735
	Total	118	671	789

The logistic model can correctly classify 89.1% [((true positives + true negatives) / sample size) x 100] of the participants as HIV-positive or HIV-negative.

Table 15: Classification using the Discriminant Model

	Observed			
	Classified	**HIV Positive**	**HIV Negative**	**Total**
Predicted	**HIV Positive**	36	82	118
	HIV Negative	18	658	676
	Total	54	740	794

The model can correctly classify 87.4% [((true positives + true negatives) / sample size) x 100] of the participants as HIV-positive or HIV-negative.

Symptoms associated with HIV prevalence in Chimanimani District

The following symptoms were significantly associated with HIV prevalence in Chimanimani district: fever, TB, STI, Diarrhoea, Burning urine, severe lower abdominal pains, sores/ulcers on genitals, pneumonia, abnormal discharge from genitals and difficulty swallowing solid foods. The relative importance as measured by Z- statistic for logistic regression model and standardised coefficient for the discriminant model, show that TB, swollen lymph nodes, difficulty swallowing solid foods, living together/cohabiting, and being in the age groups (25-

29 or 40-44) were important in determining the HIV status of an individual although the same variables were weak for the discriminant model with negative coefficients. We also observed that there is a difference in the direction of the relationships with the logistic regression taking the positive and the discriminant taking the negative.

Table 16: : Predictors, standardized, and unstandardized coefficients for the discriminant analysis model and logistic regression model.

Predictors	Logistic regression		Discriminant analysis	
	b coefficients	**Z statistic**	**Unstandardized coefficients**	**Standardized coefficients**
Age group (20-24)	1.2	2.7	-0.3	-0.5
Age group (25-29)	1.6	3.0	-0.3	-0.5
Age group (30-34)	1.3	2.2	-0.3	-0.5
Age group (40-44)	2.1	3.1	-0.3	-0.5
Widowed	2.3	2.8	-0.1	-0.3
Living together/Cohabitating	3.2	5.1	-0.1	-0.3
Resettlement Sector	-0.8	-2.3	0.5	0.1
Urban Sector	-1.8	-2.8	0.5	0.1
Atheist	3.2	2.4	-0.2	-0.2
Burning Urine	-1.4	-2.6	1.0	0.3
Severe Lower Abdominal pain	-0.1	-0.1	-0.2	-0.1
Genital Warts	1.5	1.6	-1.7	-0.3
Fever	0.6	1.3	-0.6	-0.1
Swollen Lymph Nodes	2.0	3.4	-2.4	-0.4
Pneumonia	-0.4	-0.7	0.4	0.1
Difficulty swallowing solid food	2.2	3.2	-2.0	-0.3
TB	1.7	2.6	-2.0	-0.3
Constant	**-4.9**		**2.5**	

LOGISTIC REGRESSION MODEL

Hivstatus = -4.9 +1.2 Agegrp(20-24) + 1.1 Agegrp(25-29) + 1.3 Agegrp(40-44) + 2.3 Widowed + 3.2living together/cohabiting – 0.8 Resettlement Sector – 1.8 Urban + 3.2Atheist-1.4Burning Urine -0.1Severe lower abdominal pain + 1.5Genital Warts +0.6 Fever +2.0Swollen legs -0.4Pneumonia +1.7TB

DISCRIMINANT MODEL

Hivstatus = 2.9 - 1.0 Fever -3.5 TB – 0.4 Agegrp5 (35-39) - 0.4 Agegrp6 (40-44) – 0.1 Mstatus7 (single/never married) – 0.1 Mstatus8 (divorced/separated) – 0.1 Mstatus9 (living together/cohabiting) – 0.1 Mstatus10 (widowed) – 0.5 Skin – 0.8 Stistatus

Chapter 5: Discussion

Many studies compared these two methods but the comparisons remain inconclusive. The only universal criterion for selecting logistic regression over discriminant analysis relates to the predictor data type limitations of discriminant analysis. Because both methods build linear boundaries between the classes they have the same functional form (a sum of weighted predictors). Logistic regression (maximum likelihood estimation) and discriminant analysis (least squares technique) use different methods to estimate the regression coefficients. Because of the difference, discriminant analysis should outperform logistic regression, when its assumptions are valid. However, in other circumstances, logistic regression should be better. Empirical comparisons of these methods do not always support the data type restrictions. As long as the predictors are not categorical there appears to be little difference between the performance of the two methods when sample sizes are reasonably large (>50). If there are outliers it is likely that logistic regression will be better because the outliers could distort the variance – covariance matrix used in the estimation of predictor weights by discriminant analysis.

In this study, both logistic regression and discriminant analyses converged in similar results. Both methods did not estimate different statistical significant coefficients, with different effect size and direction, although logistic regression estimated larger coefficients overall. The overall classification rate for both was good, and either can be helpful in predicting the possibility of a patient being HIV-positive or HIV-negative. Logistic regression slightly exceeds discriminant function in the correct classification rate but the difference in the area under the curve (AUC) is negligible, thus indicating no discriminating difference between the models. Receiver operating characteristics (ROC) curve between sensitivity and 1-specificity is a useful method to evaluate the performance of a diagnostic test in classification of subjects into two categories (say) HIV positive and HIV negative. ROC curve may be used to judge how well the test performs. If area under the curve is near 1 it has higher chance of correct classification and when it is near 0, higher chance of incorrectly classifying in opposite group. The value 0.5 shows the test is no better than just tossing a coin for classification into positive or negative. Hence the logistic model has a 77.84% chance of correctly classifying patients as HIV positive or HIV negative whilst the discriminant model has a 75.54% probability. Although logistic regression has a higher value compared to discriminant analysis, the difference is very negligible. This shows that the two have almost the same predictive/discriminant power of classifying patients into two categories viz HIV-positive and HIV-negative, thus they agree (converge).

Unstandardized discriminant coefficients are simply like the regression beta coefficients, which are used to predict the discriminate score. Standardized discriminant coefficients compare the relative importance of the independent variables. The Logistic model shows that having been treated for TB in the last three months or currently on TB treatment, aged between 40 - 44, cohabiting/living together with a partner and/or being a widow, having swollen lymph nodes, difficulty swallowing solid foods and being an atheist contributed significantly to the HIV status of a patient as shown by the coefficients whilst fever, being divorced/separated contribute less in predicting the HIV status. The discriminant model shows that having been aged between 20 and 44, residing in a resettlement or urban area and having sores/ulcers on genitals contributed significantly high to the HIV status of the participant but the direction of their relationship is opposite that of the logistic model. Hence the two models diverge in the direction of their relationship and the constants for the two models which have opposite signs. The logistic and discriminant models, converge in that the following symptoms (fever, TB, severe lower abdominal pain, age group, sores/ulcers on genitals, genital warts and difficulty swallowing solids) contribute significantly in determining the HIV status of a patient especially if he/she has been on treatment for TB in the last 3 months or currently on TB treatment.

Chapter 6: Conclusion

The logistic model has proved to be an efficient tool for classification of participants into different categories. It can correctly classify 89.1% of the participants as HIV positive or negative whilst the discriminant model can correctly classify 87.4%. the discriminant model had a higher diagnostic power (93.4%) compared to logistic regression's 85.0% meaning that it could correctly classify 93.4% of the respondents as HIV positive or HIV negative whilst the logistic model could only do the same to 85% of the respondents. However, the logistic model has an error rate of 0.13 implying that the logistic model is a better tool to use when classifying participants into two categories although the difference is not statistically significant. Also, the logistic model has a higher positive predictive power of 79.6% compared to 30.5% of the discriminant model meaning that it has more power to correctly classify the HIV positive participants. The discriminant model has higher negative predictive power (97.3%) of correctly classifying the HIV negative participants compared to 89.8% of the logistic model.

The two models converged in correctly classifying patients as HIV positive or HIV negative (82.4% for logistic regression and 79.5% for discriminant analysis have insignificant difference, i.e. in explaining the variation in predicting the HIV status of a patient using the Receiver Operating Curve (ROC) where the area under the curve was 0.7784 for logistic regression and 0.7554 for discriminant analysis and the difference is not significant. They also converged in coming up with the same symptoms that can be used to predict the HIV status of a patient. However, we observed that the two models diverged in the direction of the relationships. The two models showed some divergence in the magnitude of their constants and coefficients, which were extremely different in that those for discriminant analysis were all negative (except for sector, pneumonia, burning urine, sores/ulcers on genitals and constant), whilst those for logistic regression were all positive except for sector, pneumonia, burning urine, sores/ulcers on genitals and constant.

The wide use of discriminant analysis and logistic regression for diagnosis and identification of people at risk emphasises the importance of employing the method that minimises errors in classification. Recommendations as to which method to use, are often guided by the apparent model assumptions. The most commonly employed discriminant analysis procedure assumes equal population covariance structures for the groups and, in certain situations, multivariate normality of predictor variables. Logistic regression, on the other hand, does not make these assumptions and is recommended for predicting group memberships because it does not make stringent assumptions and has minimum errors, higher positive predictive power and correct classification rate.

Reference

1. Tabachnink BG and Fidell LS, *Using Multivariate Statistics*, HarperCollins, New York, NY, USA, 1996.

2. Hair JF, Anderson RE, Tatham RL, and Black WC, *Multivariate Data Analysis with Readings*, Prentice-Hall, Englewood Cliffs, NJ, USA, 5th edition, 1998.

3. Stevens JP, *Applied Multivariate Statistics for the Social Sciences*, Lawrence Erlbaum, Hillsdale, NJ, USA, 4th edition, 2002.

4. Pampel FC, *Logistic Regression: A Primer*, Sage, Thousand Oaks, Calif, USA, 2000.

5. Wright RE, "Logistic regression," in *Reading and Understanding Multivariate Statistics*, Grimm LG and Yarnold PR, Eds., pp. 217–244, American Psychological Association, Washington, DC, USA, 1995.

6. Pohar M, Blas M, and Turk S, "Comparison of logistic regression and linear discriminant analysis: a simulation study," *Metodološki Zvezki*, vol. 1, no. 1, pp. 143–161, 2004.

7. Brenn T and Arnesen E, "Selecting risk factors: a comparison of discriminant analysis, logistic regression and Cox's regression model using data from the Tromsø heart study," *Statistics in Medicine*, vol. 4, no. 4, pp. 413–423, 1985.

8. Montgomery ME, White ME, and Martin SW, "A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows," *Canadian Journal of Veterinary Research*, vol. 51, no. 4, pp. 495–498, 1987.

9. Batter V, Matela B, Nsuami M. et al. High HIV-1 incidence in young women masked by stable overall seroprevalence among childbearing women in Kinshasa, Zaire: estimating incidence from serial seroprevalence data. AIDS 1994, 8: 811 -817.

10. Gregson S, Donnelly CA, Parker CG, Anderson RM. Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions. AIDS 1996, 10: 1689 -1697.

11. Cleghorn FR, Jack N, Murphy JR. et al. Direct and indirect estimates of HIV-1 incidence in a high-prevalence population. Am J Epidemiol 1998, 147: 834 -839.

12. Saidel T, Sokal D, Rice J, Buzingo T, Hassig S. Validation of a method to estimate age-specific human immunodeficiency virus (HIV) incidence in developing countries using population-based seroprevalence data. Am J Epidemiol 1996, 144: 214 -223.

13. US Bureau of Census. *HIV/ AIDS Surveillance Data Base.* Washington, DC: US Bureau of Census, 1999

14. Zaba B, Gregson S. Measuring the impact of HIV on fertility in Africa. AIDS 1998, 12: (supple. 1): S41 -S50.

15. Fontanet AL, Messele T, Dejene A. et al. Age- and sex-specific HIV-1 prevalence in the urban community setting of Addis Ababa, Ethiopia. AIDS 1998, 12: 315 -322.

16. K. N. Priftis, D. B. Panagiotakos, G. Antonogeorgos, et al., “Factors associated with asthma symptoms in schoolchildren from Greece: the Physical Activity, Nutrition and Allergies in Children Examined in Athens (PANACEA) study,” Journal of Asthma, vol. 44, no. 7, pp. 521–527, 2007.

17. AIDS Education & Training Centers National Resource Center, 2009 HIV Classification: CDC and WHO Staging Systems. AETC 2009 April, Pennsylvania/MidAtlantic, US, Available from: http://www.aidsetc.org/aidsetc?page=cm-105_disease

18. Centers for Disease Control and Prevention. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults . MMWR Recomm Rep. 1992 Dec 18;41(RR-17):1-19. Available online at www.cdc.gov/mmwr/preview/mmwrhtml/00018871.htm.

19. Centers for Disease Control and Prevention. Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome . MMWR Recomm Rep. 1999 Dec 10;48(RR-13):1-28, 29-31. Available online at www.cdc.gov/mmwr/preview/mmwrhtml/rr4813a1.htm

20. Hare CB. WHO Staging System for HIV Infection and Disease in Adolescents and Adults, Table 3; Clinical Overview of HIV Disease . In: Peiperl L, Coffey S, Volberding PA, eds. HIV InSite Knowledge Base [textbook online]. San Francisco: UCSF Center for HIV Information; 2006. Accessed March 23, 2006. Available online at http://hivinsite.ucsf.edu/InSite?page=kb-03-01-01.

21. World Health Organization. WHO Case Definitions of HIV for Surveillance and Revised Clinical Staging and Immunological Classification of HIV-Related Disease In Adults and Children . 2007. Accessed March 30, 2009. Available online at www.who.int/hiv/pub/guidelines/HIVstaging150307.pdf.

22. NLM Gateway. Utility of the WHO clinical staging of HIV infection and disease for the identification of HIV-related illness and mortality, In: Sewankambo NK, Wawer MJ,

Gray RH, Serwadda D, Konde-Lule J, Li CJ, Nalugoda F, Lutalo T; International Conference on AIDS. 1996. Jul 7-12; 11: 300 (abstract no. Th.B.4283) Availabe online at http://gateway.nlm.nih.gov/MeetingAbstracts/ma?f=102220572.html

23. NLM Gateway. Age-specific patterns of HIV infection and STI symptom history among young Zimbabwean males, In: Loo V, McNaghten AD, Dube H, Hader S; International Conference on AIDS (15th : 2004 : Bangkok, Thailand). *Int Conf AIDS.* 2004 Jul 11-16; 15: (abstract no. MoPeC3492)

24. AVERT. HIV and AIDS in Zimbabwe. Available online at http://Avert.org Last updated March 10, 2010

Appendix 1: Sample Size

The 23 wards were numbered 1 to 23 and 13 wards were randomly selected for the BSS study using EPI TABLE calculator in EPI INFO version 6. Households within each ward were systematically selected using the calculated sampling interval. The sample size was calculated proportional to the size of the ward (weighted by population size of each ward based on the BRTI 2003 OVC census data) Assuming a 50% response rate and a target of 1000 participants in each age group, the target group sample size was increased to 1 500 and the ward sample size was calculated as follows:

Sample size = $\frac{No.of\ households\ in\ the\ ward}{Total\ number\ of\ households\ in\ the\ 13\ wards}$ X 1500 (targeted sample size)

Sampling Interval (SI) = $\frac{No.of\ households\ in\ the\ ward}{Ward\ sample\ size}$

The sampling of households was based on systematic interval of 9 households. On entering the selected ward, the research team selected an outstanding feature (mountain, school, river, business centre, borehole, etc) to the furthest north-west point and randomly select going in an easterly direction. Appointments for the return visit were made if respondent was not there and no call-backs for those who refused to be interviewed and not at home on the second visit.

After identifying the household, the head or representative of the household would describe the household composition. The interviewer then stratified household members into their age groups. If there were two or more persons in an age group and/or sex in the household, their names were written on pieces of paper, put in a hat, shuffled and one name drawn by the supervisor. Interviewers also alternated in choosing a male or female respondent from one household to the other in order to have equal numbers of respondents per each age group. The areas of focus of the study were demographic, orphanhood status, rites of passage, marital status, sexual experience, HIV/AIDS-related behaviour and condom use, knowledge, attitude and perceived risk of HIV/AIDS, OVC care issues, PLWHA and Health status (including history of hospitalisation)

Appendix 2: Discriminant Analysis

Discriminant analysis focuses on the association between multiple independent variables and a categorical dependent variable by forming a composite of the independent variables. This type of multivariate analysis can determine the extent of any of the composite variables discriminates between two or more pre-existing groups of subjects and also can derive a classification model for predicting the group membership of new observations [1]. The simplest type of discriminant analysis is when the dependent variable has two groups. In this case, a linear discriminant function that passes through the means of the two groups (centroids) can be used to discriminate subjects between the two groups. When there are more groups, the number of groups minus one function is needed to classify an observation among them. For each of the groups, linear discriminant analysis assumes the explanatory variables to be normally distributed with equal covariance matrices. For each case, the estimated coefficient for an independent variable is multiplied by the case's score on that variable. These products are summed and added to the constant, and the result is a composite score, that is, the discriminant score for that case.

The linear discriminant function (LDF) is represented by

$$LDF = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \dots \dots + b_k x_{ik}$$

where b_j is the value of the jth coefficient, j = 1,…., k and x_{ik} is the value of the ith case of the jth predictor. The LDF can also be written in standardized form which allows comparing variables measured on different scales. In the standardized LDF, each variable is adjusted by subtraction of its mean value and division by its standard deviation. Coefficients with large absolute values reflect greater discriminating ability to their corresponding variables. From the LDF, scores can estimate predicted probabilities and predicted group membership for every case on the dependent variable. This approach is based on the rationale that it is more likely that the independent and dependent variables are related as the between-groups sum of square is larger relative to within-group sum of squares. Also the ratio of between-group divided by total sum of squares (beta-squared statistic or explained variability) or of within-group divided by total sum of squares (Wilks' lambda statistic or unexplained variability) is used to assess the relationship. As we can see, the ratio of between-group divided by within-group sum of squares is an analogue to the ratio of variances, which is the *F* statistic, a test that controls the possibility that the observed relationship is due to chance.

The principle by which the discriminant coefficients (or weights) are selected is that they maximize the distance between the two group means (centroids) $|\bar{y}_1 - \bar{y}_2|$. Fisher [2] was the first who suggested to transform the multivariate observation *x* to univariate observations *y* in such way that the *y*'s derived from groups 1 and 2 have the maximum distance between them. Thus, the linear combination $y = a'x$ is the one that maximizes the ratio (squared distance between sample means)/(sample variance *y*). The vector of coefficients is given by the eigenvectors of the matrix $B * S^{-1}$, where $B = (\bar{x}_1 - \bar{x}_2)'$is the between-group matrix and S is an estimate of Σ. A very important characteristic of these composite sums of squares is that they enclose the variability and the covariability of each variable. The discriminant coefficients can be calculated in unstandardized or standardized form but they are irrelatively of the form, less informative than those in regression. Assuming that there are 2 groups, $\bar{x}_1$, $\bar{x}_2$ are the means of each group, and S is the pooled covariance matrix, the allocation rule based on Fisher's discriminant functions is the following: (2)

$$X_i \in \begin{cases} \text{group 1, if } y = (\overline{x}_1 - \overline{x}_2)' S^{-1} X_i \\ \qquad\qquad \geq \frac{1}{2}(\overline{x}_1 - \overline{x}_2)' S^{-1} (\overline{x}_1 + \overline{x}_2), \\ \text{group 2, if } y = (\overline{x}_1 - \overline{x}_2)' S^{-1} X_i \\ \qquad\qquad < \frac{1}{2}(\overline{x}_1 - \overline{x}_2)' S^{-1} (\overline{x}_1 + \overline{x}_2). \end{cases}$$

Appendix 3: Logistic Regression

Logistic regression is a form of regression which is used when we want to predict probabilities of the presence or absence of a particular disease, characteristic, or an outcome in general based on a set of independent of explanatory variables of any kind (continuous, discrete, or categorical)[2]. Since the predicted probability must lie between 0 and 1, simple linear regression techniques are insufficient to achieve that, because they allow the dependent variable to pass these limits and to produce inconsistent results. Defined as P_1, the probability of an object is belonging to group 1, and as P_0, the probability of an object is belonging to group 0. The logistic regression model has the form of

$$Zi = log\left(\frac{P1}{P0}\right)$$

$$= b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \dots \dots + b_k x_{ik}$$

where P_{i1} / P_{i0} is called the odds ratio, $b_{j.}$ is the value of the *j*th coefficient, j = 1,.....,k, and x_{ij} is the value of the *i*th case of the *j*th predictor. The parameters (b_0 to b_k) of the logistic model are estimated with the use of maximum likelihood method. The probability of an event to occur can be calculated using the logistic regression model

$$P(Y_i = 1 \setminus X_i) = \frac{e^{b^T X_i}}{1 + (e^{b^T X_i})} = \frac{1}{1 + e^{-b^T X_i}},$$

where $e^{b^T X_i}$ is the linear predictor of the logistic regression function, and Y_i is the event under study (dependent variable).

If we use a probability cut off of 0.5, then we can classify an object to group 1 if the estimated $p_1 > .5$and to group 0 if $p_1 < .5$. In order to estimate the parameters of the logistic regression model, the method of maximum likelihood maximizes the coefficients of the log-likelihood function, a statistic which summarizes the information of the predictor variables.

Appendix 4: Principal Components Analysis

Principal Components Analysis is one of the best known and most used Multivariate Exploratory Analysis technique.

Goal of Principal Components Analysis

Given a data set described by a set of **numerical** variables $\{x_1, x_2, ..., x_p\}$, the goal of Principal Components Analysis is to describe this data set with a smaller set of new, synthetic variables. These variables will be **linear combinations** of the original variables, and are called **Principal Components**.

Quite generally, reducing the number of variables used to describe data will lead to some loss of information. PCA operates in a way that makes this loss **minimal**, in a sense that will be given a precise meaning.

Therefore, PCA may be regarded as a dimensionality reduction technique.

Properties of the Principal Components

Although the ultimate goal is to use only a small number of Principal Components, PCA first identifies p such components, that is, the same number as the number of original variables. Only later will the analyst decide on the number of Components to be retained. "Retaining k Principal Components" means "Replacing the observations by their orthogonal projections in the k-dimensional subspace spanned by the first k Principal Components".

Orthogonality of the Principal Components

The Principal Components define **orthogonal** directions in the space of observations. In other words, PCA just makes a change of orthogonal reference frame, the new variables being replaced by the Principal Components.

Uncorrelatedness of the Principal Components

It will turn out that the Principal Components are pariwise uncorrelated.

Ordering the Principal Components, optimal projection subspaces

The fundamental property of the Principal Components is that they can be ordered by decreasing order of "importance" in the following sense :

1. If the analyst decides to describe the data with only *k* (*k* < *p*) linear combinations of the original variables and yet loose as little information as possible in the process, then these *k* linear combinations have to be the first *k* Principal Components.

So the fundamental property of the PCs is that the best *k*-dimensional projection subspace is spanned by the first *k* Principal Components. In other words, the optimal subspaces are **nested**, a strong, useful and not at all obvious property.

Applications of Principal Components Analysis

Exploratory data analysis

PCA is mostly used for making 2-dimensional plots of the data for visual examination and interpretation. For this purpose, data is projected on **factorial planes** that are spanned by pairs of Principal Components chosen among the first ones (that is, the most significant ones).

From these plots, one will try to extract informations about the data structure, such as:

1. The detection of outliers (observations that are very different from the bulk of the data).
2. The identification of clusters that suggest that several subpopulations might coexist within the data set.

3. Interpretation of the Principal Components. Whereas the original variables have a "native" interpretations, Principal Components derive from a mathematical definition. A succesful PCA will allow interpreting the Principal Components in terms of realistic, if not measured, properties of the observations. When this is possible, it is sometimes said that PCA has revealed the existence of "latent variables".

Data pre-processing, dimensionality reduction

All multivariate modelisation techniques are prone to the bias-variance tradeoff, which states that the number of variables entering a model should be severely restricted. Data is often decribed by many more variables than necessary for building the best model. Sometimes, specific techniques exist for selecting a "good" subset of variables (see for instance Mutiple Linear Regression), but dimensionality reduction techniques such as PCA may also be considered for feeding the model with a reduced number of variables. For example, Multiple Linear Regression may be replaced by a model using only a reduced number of Principal Components as regressors (Principal Components Regression).

Data compression and data reconstruction

The table describing the data with coordinates on the first *k* Principal Components is smaller than the original data table. Therefore, Principal Components Analysis may be used as a (lossy) data compression technique.

PCA incorporated in other techniques

Although created from a downright applicative perspective (data visualization), the mathematical machinery of PCA is quite general and is at the heart of other important modelization techniques. Let's mention :

1. Discriminant Factor Analysis, that can be regarded as a PCA on the ponderated class barycenters using the Mahalanobis distance.
2. Ridge Regression, that receives a very illuminating interpretation in terms of PCA.

Generalizations of Principal Components Analysis

PCA is just a change of orthogonal reference frame. Therefore, it relies on simple Linear Algebra as its main mathematical engine, and is quite easy to interpret geometrically. But this strength is also a weakness, for it might very well be that other synthetic variables, more complex than just linear combinations of the original variables, would lead to a more economical data description.

In fact, PCA can be generalized in many ways, mostly based on **non linear** transforms of the original variables. This issue is not addressed in this Glossary, but the interested reader may seek information on :

1. Independent Components Analysis (ICA), that generates new variables that not just uncorrelated (as are the Principal Components), but genuinely independent.
2. Curvilinear Components Analysis, a non linear projection technique whose priority is to respect the distances between observations.
3. PCA on latent variables, that describes data by linear combinations of a small latent (unobserved) variables.
4. Kernel-based PCA, that sends data into a high-dimensional space by an appropriate non-linear projection, and then performs an ordinary PCA in this high-dimensional space.

Besides, Kohonen Maps may be regarded as a non linear dimensionality reduction technique.

Appendix 5: The different stages of HIV infection

HIV infects cells in the immune system and the central nervous system. The main type of cell that HIV infects is the T helper lymphocyte. These cells play a crucial role in the immune system, by coordinating the actions of other immune system cells. A large reduction in the number of T helper cells seriously weakens the immune system.

HIV infects the T helper cell because it has the protein CD4 on its surface, which HIV uses to attach itself to the cell before gaining entry. This is why the T helper cell is sometimes referred to as a CD4+ lymphocyte. Once it has found its way into a cell, HIV produces new copies of itself, which can then go on to infect other cells.

Over time, HIV infection leads to a severe reduction in the number of T helper cells available to help fight disease. The process usually takes several years.

HIV infection can generally be broken down into four distinct stages: primary infection, clinically asymptomatic stage, symptomatic HIV infection, and progression from HIV to AIDS.

STAGE 1 : Primary HIV infection

This stage of infection lasts for a few weeks and is often accompanied by a short flu-like illness. In up to about 20% of people the HIV symptoms are serious enough to consult a doctor, but the diagnosis of HIV infection is frequently missed.

During this stage there is a large amount of HIV in the peripheral blood and the immune system begins to respond to the virus by producing HIV antibodies and cytotoxic lymphocytes. This process is known as seroconversion. If an HIV antibody test is done before seroconversion is complete then it may not be positive.

STAGE 2 : Clinically asymptomatic stage

This stage lasts for an average of ten years and, as its name suggests, is free from major symptoms, although there may be swollen glands. The level of HIV in the peripheral blood drops to very low levels but people remain infectious and HIV antibodies are detectable in the blood, so antibody tests will show a positive result.

Research has shown that HIV is not dormant during this stage, but is very active in the lymph nodes. A test is available to measure the small amount of HIV that escapes the lymph nodes. This test which measures HIV RNA (HIV genetic material) is referred to as the viral load test, and it has an important role in the treatment of HIV infection.

STAGE 3 : Symptomatic HIV infection

Over time the immune system becomes severely damaged by HIV. This is thought to happen for three main reasons:

1. The lymph nodes and tissues become damaged or 'burnt out' because of the years of activity;
2. HIV mutates and becomes more pathogenic, in other words stronger and more varied, leading to more T helper cell destruction;
3. The body fails to keep up with replacing the T helper cells that are lost.

As the immune system fails, symptoms develop. Initially many of the symptoms are mild, but as the immune system deteriorates the symptoms worsen.

Symptomatic HIV infection is mainly caused by the emergence of opportunistic infections and cancers that the immune system would normally prevent. This stage of HIV infection is often characterised by multi-system disease and infections can occur in almost all body systems.

Treatment for the specific infection or cancer is often carried out, but the underlying cause is the action of HIV as it erodes the immune system. Unless HIV itself can be slowed down the symptoms of immune suppression will continue to worsen.

STAGE 4 : Progression from HIV to AIDS

As the immune system becomes more and more damaged the illnesses that occur become more and more severe leading eventually to an AIDS diagnosis.

At present in the UK an AIDS diagnosis is confirmed if a person with HIV develops one or more of a specific number of severe opportunistic infections or cancers. In the US, someone may also be diagnosed with AIDS if they have a very low count of T helper cells in their blood. It is possible for someone to be very ill with HIV but not have an AIDS diagnosis.

Examples of opportunistic infections and cancers

The table below shows examples of common opportunistic infections and cancers and the body systems that they occur in.

System	Examples of Infection/Cancer
Respiratory system	1. Pneumocystis jirovecii Pneumonia (PCP) 2. Tuberculosis (TB) 3. Kaposi's Sarcoma (KS)
Gastro-intestinal system	1. Cryptosporidiosis 2. Candida

	3. Cytomegolavirus (CMV) 4. Isosporiasis 5. Kaposi's Sarcoma
Central/peripheral Nervous system	1. Cytomegolavirus 2. Toxoplasmosis 3. Cryptococcosis 4. Non Hodgkin's lymphoma 5. Varicella Zoster 6. Herpes simplex
Skin	1. Herpes simplex 2. Kaposi's sarcoma 3. Varicella Zoster

WHO clinical staging of HIV disease in adults and adolescents (2006 revision)

In resource-poor communities, medical facilities are sometimes poorly equipped, and it is not possible to use CD4 and viral load test results to determine the right time to begin antiretroviral treatment. The World Health Organisation (WHO) has therefore developed a staging system for HIV disease based on clinical symptoms, which may be used to guide medical decision making.

Clinical Stage I:

1. Asymptomatic
2. Persistent generalized lymphadenopathy

Clinical Stage II:

1. Moderate unexplained* weight loss (under 10% of presumed or measured body weight)**
2. Recurrent respiratory tract infections (sinusitis, tonsillitis, otitis media, pharyngitis)
3. Herpes zoster
4. Angular chelitis
5. Recurrent oral ulceration
6. Papular pruritic eruptions
7. Seborrhoeic dermatitis
8. Fungal nail infections

Clinical Stage III:

1. Unexplained* severe weight loss (over 10% of presumed or measured body weight)**
2. Unexplained* chronic diarrhoea for longer than one month
3. Unexplained* persistent fever (intermittent or constant for longer than one month)
4. Persistent oral candidiasis
5. Oral hairy leukoplakia
6. Pulmonary tuberculosis
7. Severe bacterial infections (e.g. pneumonia, empyema, pyomyositis, bone or joint infection, meningitis, bacteraemia)
8. Acute necrotizing ulcerative stomatitis, gingivitis or periodontitis
9. Unexplained* anaemia (below 8 g/dl), neutropenia (below 0.5 billion/l) and/or chronic thrombocytopenia (below 50 billion/l)

Clinical Stage IV:***

1. HIV wasting syndrome
2. Pneumocystis pneumonia
3. Recurrent severe bacterial pneumonia
4. Chronic herpes simplex infection (orolabial, genital or anorectal of more than one month's duration or visceral at any site)
5. Oesophageal candidiasis (or candidiasis of trachea, bronchi or lungs)
6. Extrapulmonary tuberculosis
7. Kaposi sarcoma
8. Cytomegalovirus infection (retinitis or infection of other organs)
9. Central nervous system toxoplasmosis
10. HIV encephalopathy
11. Extrapulmonary cryptococcosis including meningitis
12. Disseminated non-tuberculous mycobacteria infection
13. Progressive multifocal leukoencephalopathy
14. Chronic cryptosporidiosis
15. Chronic isosporiasis
16. Disseminated mycosis (extrapulmonary histoplasmosis, coccidiomycosis)
17. Recurrent septicaemia (including non-typhoidal Salmonella)
18. Lymphoma (cerebral or B cell non-Hodgkin)
19. Invasive cervical carcinoma
20. Atypical disseminated leishmaniasis
21. Symptomatic HIV-associated nephropathy or HIV-associated cardiomyopathy

Revised World Health Organization (WHO) Clinical Staging of HIV/AIDS For Adults and Adolescents (2005)

(This is the interim African Region version for persons aged 15 years or more who have had a positive HIV antibody test or other laboratory evidence of HIV infection) (It must be noted that the UN defines adolescents as persons aged 10−19 years but for surveillance purposes, the category of adults and adolescents comprises people aged 15 years and over)

Primary HIV infection

1. Asymptomatic
2. Acute retroviral syndrome

Clinical stage 1

1. Asymptomatic
2. Persistent generalized lymphadenopathy

Clinical stage 2

1. Moderate and unexplained weight loss (<10% of presumed or measured body weight)
2. Recurrent respiratory tract infections (such as sinusitis, bronchitis, otitis media, pharyngitis)
3. Herpes zoster
4. Recurrent oral ulcerations
5. Papular pruritic eruptions
6. Angular cheilitis
7. Seborrhoeic dermatitis
8. Fungal finger nail infections

Clinical stage 3

Conditions where a presumptive diagnosis can be made on the basis of clinical signs or simple investigations

1. Unexplained chronic diarrhoea for longer than one month
2. Unexplained persistent fever (intermittent or constant for longer than one month)
3. Severe weight loss (>10% of presumed or measured body weight)
4. Oral candidiasis
5. Oral hairy leukoplakia
6. Pulmonary tuberculosis (TB) diagnosed in last two years
7. Severe presumed bacterial infections (e.g. pneumonia, empyema, meningitis, bacteraemia, pyomyositis, bone or joint infection)
8. Acute necrotizing ulcerative stomatitis, gingivitis or periodontitis

Conditions where confirmatory diagnostic testing is necessary

1. Unexplained anaemia (< 80 g/l), and or neutropenia (<500/µl) and or thrombocytopenia (<50 000/ µl) for more than one month

Clinical stage 4

Conditions where a presumptive diagnosis can be made on the basis of clinical signs or simple investigations

HIV wasting syndrome

Pneumocystis pneumonia

Recurrent severe or radiological bacterial pneumonia

Chronic herpes simplex infection (orolabial, genital or anorectal of more than one month's duration)

Oesophageal candidiasis

Extrapulmonary Tuberculosis

Kaposi's sarcoma

Central nervous system toxoplasmosis

HIV encephalopathy

Conditions where confirmatory diagnostic testing is necessary

Extrapulmonary cryptococcosis including meningitis

Disseminated non-tuberculous mycobacteria infection

Progressive multifocal leukoencephalopathy

Candida of trachea, bronchi or lungs

Cryptosporidiosis

Isosporiasis

Visceral herpes simplex infection

Cytomegalovirus (CMV) infection (retinitis or of an organ other than liver, spleen or lymph nodes)

Any disseminated mycosis (e.g. histoplasmosis, coccidiomycosis, penicilliosis)

Recurrent non-typhoidal salmonella septicaemia

Lymphoma (cerebral or B cell non-Hodgkin)

Invasive cervical carcinoma

Visceral leishmaniasis

Original proposal in 1990

Clinical Stage I

Asymptomatic

Generalised lymphadenopathy

Performance scale: 1: asymptomatic, normal activity.

Clinical Stage II

Weight loss, < 10% of body weight

Minor mucocutaneous manifestations (seborrheic dermatitis, prurigo, fungal nail infections, recurrent oral ulcerations, angular cheilitis)

Herpes zoster within the last five years

Recurrent upper respiratory tract infections (i.e. bacterial sinusitis)

And/or performance scale 2: symptomatic, normal activity.

Clinical Stage III

Weight loss, > 10% of body weight

Unexplained chronic diarrhoea > 1 month

Unexplained prolonged fever (intermittent or constant), > 1 month

Oral [candidiasis] ([thrush])

Oral hairy leucoplakia

Pulmonary tuberculosis

Severe bacterial infections (i.e. pneumonia, pyomyositis)

And/or performance scale 3: bedridden < 50% of the day during last month.

Clinical Stage IV

The declaration of AIDS

HIV wasting syndrome *

Pneumocystis carinii pneumonia

Toxoplasmosis of the brain

Cryptosporidiosis with diarrhoea > 1 month

Cryptococcosis, extrapulmonary

Cytomegalovirus disease of an organ other than liver, spleen or lymph node (ex: retinitis)

Herpes simplex virus infection, mucocutaneous (>1 month) or visceral

Progressive multifocal leucoencephalopathy

Any disseminated endemic mycosis

Candidiasis of esophagus, trachea, bronchi

Atypical mycobacteriosis, disseminated or lungs

Non-typhoid Salmonella septicemia

Extrapulmonary tuberculosis

Lymphoma

Kaposi's sarcoma

HIV encephalopathy **

And/or performance scale 4: bedridden > 50% of the day during last month.

(*) HIV wasting syndrome: weight loss of > 10% of body weight, plus either unexplained chronic diarrhoea (> 1 month) or chronic weakness and unexplained prolonged fever (> 1 month).

(**) HIV encephalopathy: clinical findings of disabling cognitive and/or motor dysfunction interfering with activities of daily living, progressing over weeks to months, in the absence of a concurrent illness or condition other than HIV infection which could explain the findings.

Appendix 6: The Kappa Agreement Coefficient

Cohen's Kappa coefficient is a statististical measure of inter-rater agreement or *inter-annotator agreement* for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Some researchers have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can have the effect of understanding agreement for a category that is also commonly used; for this reason, κ is considered an overly conservative measurement of agreement.

Others contest the assertion that kappa "takes into account" chance agreement. To do this effectively would require an explicit model of how chance affects rater decisions. The so-called chance adjustment of kappa statistics supposes that, when not completely certain, raters simply guess – a very unrealistic scenario.

Nevertheless, and despite potentially better alternatives, Cohen's kappa enjoys continued popularity. A possible reason for this is that kappa is, under certain conditions, equivalent to the intra-class correlation coefficient.

Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892). The equation for κ is:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then

$\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

Note that Cohen's kappa measures agreement between **two** raters only. For a similar measure of agreement (Fleiss' kappa) used when there are more than two raters. The Fleiss kappa, however, ia a multi-rater generalization of Scottt's pi statistic, not Cohen's kappa.

Example: **Suppose that you were analysing data related to people applying for a grant. Each grant proposal was read by two people and each reader either said "Yes" or "No" to the proposal. Suppose the data were as follows, where rows are reader A and columns are reader B:**

	Yes	**No**
Yes	20	5
No	10	15

Note that there were 20 proposals that were granted by both reader A and reader B, and 15 proposals that were rejected by both readers. Thus, the observed percentage agreement is Pr(a)=(20+15)/50=0.70. To calculate Pr(e) (the probability of random agreement) we note that:

Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.

Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

Therefore the probability that both of them would say "Yes" randomly is 0.50*0.60=0.30 and the probability that both of them would say "No" is 0.50*0.40=0.20. Thus the overall probability of random agreement is Pr("e") = 0.3+0.2=0.5.

So now applying our formula for Cohen's Kappa we get:

K = $\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} = \frac{0.70-0.50}{1-0.50}$ = 0.40

One of the problems with Cohen's Kappa is that it does not always produce the expected answer. For instance, in the following two cases there is equal agreement between A and B (60 out of 100 in both cases) so we would expect the relative values of Cohen's Kappa to reflect this. However, calculating Cohen's Kappa for each:

	Yes	**No**
Yes	45	15
No	25	15

K = $\frac{0.60-0.54}{1-0.54}$ = 0.1304

	Yes	**No**
Yes	25	35
No	5	35

K = $\frac{0.60-0.46}{1-0.46}$ = 0.2593

We find that it shows greater similarity between A and B in the second case, compared to the first.

Significance

Landis and Koch gave the following table for interpreting κ values. This table however by *no means* universally accepted; Landis and Koch supplied no evidence to support it, basing it instead on personal opinion. It has been noted that these guidelines may be more harmful than helpful, as the number of categories and subjects will affect the magnitude. The kappa will be higher when there are fewer categories.

K	**Intepretation**
< 0	No agreement
0.0 – 0.20	Slight agreement
0.21 - 0.40	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement