UNIVERSITY OF THE WITWATERSRAND

FACULTY OF HEALTH SCIENCES

SCHOOL OF PUBLIC HEALTH

Title: Investigating the effect of different weighting methods on HIV prevalence estimation in the SABSSM V survey of 2017 in South Africa

Student: **Moraka Ephraim Maake**

Student number: **2332653**

MSc Epidemiology and Biostatistics

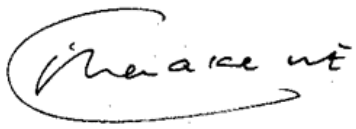
Field of Study: MSc (Epidemiology & Biostatistics)

Supervisor: Prof Jonathan Levin & Dr Edmore Marinda

Johannesburg, November 2024

DECLARATION

I Moraka Ephraim Maake (Student number: 2332653) declare that this Research report is my own, unaided work. It is being submitted for the Degree of Master of Science (Epidemiology & Biostatistics) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.



_____

Signature

_03_day of __November_2024_ in __Johannesburg, South Africa

Dedication

I dedicate this research report to my lovely wife Bella for being a pillar of strength and continually being my cheerleader. Kwenadi!!

A special thanks to my parents Vida and Maleema Maake for raising and supporting me through and through.

To my children ‘Hlalo, Mali’ & ‘Kgadi, I thank you for your patience and understanding.

In Memory of my aunt, Rakgadi Grace Mamabolo, nee Maake.

To my sister Makgotlo ‘la Mokhabane! Nogane! Ke tla ba eng?

To my brothers Saku le Mosima. Bakgaga‘a Makubela, kea le leboga.

My former teachers *Meneer* Gafane and *Meneer* Motloutsi, thank for the inspiration.

To my cousin and someone I look up to, Joseph Mphekgoane. I am grateful to grow surrounded by great brothers like you

I can’t forget my colleagues who have accommodated my study demands. Thank you Mam’Grace; Aus Joyce & Dr Kenyeres.

To many more that were not mentioned, *Kea le leboga*!

ABSTRACT

Introduction: In the health systems, prevalence surveys play a key role in identifying key priority areas. They provide valuable information which is important for monitoring and evaluation of health programs and interventions. The South African Behavioural, Sero-surveillance and Media communication (SABSSM) survey is one of the biggest health surveys in South Africa. However, the SABSSM continues to experience a high number of participants refusing to take HIV tests (non-response). This high non-response can introduce bias in HIV prevalence estimates. Sampling weights are introduced to address bias taking into account non-response. This study was aimed to investigate the effect of sampling weights and explore the use of weights derived from predictors of non-response as an alternative method to reduce bias.

Methods: A cross-sectional population-based household survey data analysis was conducted. Stratified multistage cluster random sampling was used. Two sets of HIV prevalence estimates derived from questionnaire sampling weights and HIV data sampling weights were compared. Factors associated with non-response were investigated and used to determine a set of weights. The third set of weights were used to calculate new HIV prevalence estimates.

Results: A change in HIV prevalence estimates when using different weights were observed. Using HIV sampling data resulted in a lower HIV prevalence estimate, slightly high SEs and wide 95%CI compared to using questionnaire sampling weights. Factors associated with non-response include mental depression and being sexually active. Third weights resulted in slightly lower HIV prevalence estimates compared to estimates obtained from HIV sampling weights with a lower SE and narrower 95%CI.

Conclusion: Using different sampling weights changed prevalence estimates. The study also demonstrate that non-response is not entirely at random.

Key words: non-response; sampling weights, HIV prevalence estimate

Acknowledgement

I acknowledge my supervisor, Prof Levin for the foundation he laid through his teachings in Biostatistics. I am also grateful for his guidance and patience throughout this project.
I would also like to mention and acknowledge my co-supervisor, Dr Marinda.

I express my gratitude to the Human Sciences Research Council for granting access to the dataset, this work would not have been possible without you.

Table of contents

List of Figures

List of Tables

List of Appendices

Nomenclature (List of Abbreviations/Acronyms)

Abbreviation/ Acronym	Definition
HIV	Human Immune-deficiency Virus
SADC	Southern African Development Communities
ART	Antiretroviral therapy
SABSSM V survey	The fifth South African National HIV Prevalence, Incidence, Behaviour and Communication Survey
HSRC	Human Sciences Research Council
UNAIDS	Joint United Nations Programme on HIV and AIDS
UTI	Urinary tract infection
MCAR	missing completely at random
MAR	missing at random
MNAR	missing not at random
SE	Standard error
MRP	Multilevel regression and poststratification
CI	Confidence Interval
SALs	small area layers
VPs	visiting points
DEFF	design effect
ROC	Receiver Operating Characteristics
AUC	area under curve
VIF	Variance Inflation Factor
REC	Research Ethics Committee
HTN	Hypertension

Chapter 1

1. Introduction

This research report is presented in five chapters. The first chapter (**Chapter 1**) introduces the problem to be examined. It will give an overview of HIV and AIDS burden both in the region and in South Africa. It also touches on challenges encountered in the fight against HIV and AIDS. Additionally, this chapter also narrows its focus on the role played by population-based surveys in the fight against the disease and problems encountered in such studies. It continues to give a brief critical review of the surveys and problems identified when conducting surveys. The chapter finishes by introducing the study aims and objectives as well as giving the justification for this research project.

The second chapter (**Chapter 2**) presents methodology by giving a description of the study design, sampling, data management, analysis, and ethical matters. This is followed by **Chapter 3** which presents the study results. Discussion of key findings is presented in **Chapter 4** and **Chapter 5** presents conclusions and recommendations.

1.1. Background

By the end of 2022, the World Health Organization (WHO) estimated that 39 million people were infected with HIV globally.[1] The WHO further reported that two thirds of all people living with HIV reside in the Africa region. In South Africa, 7.8 million people were estimated to be living with HIV according to the recent prevalence survey results released in 2023 in South Africa.[2] This number of people constitute about fifth of all people living with HIV worldwide.

Prevalence surveys are an integral part of the health systems. They play an important role of identifying key priority areas essential for health planning purposes such as areas in need of prioritisation or intervention. During assessment, monitoring or evaluation of health practices or health related interventions, surveys can be vital as they provide valuable information. It is therefore crucial that results from surveys are credible and valid. Failure to obtain credible results when conducting surveys, may lead to a waste of resources and may ultimately affect health delivery negatively. Methodologies used in surveys must be continuously scrutinised to attain the highest level of credibility and aid resource allocation.[3, 4]

1.2. Literature review

1.2.1. Burden of HIV disease

The number of people living with Human Immune-deficiency Virus (HIV) in South Africa is among the highest globally, even though the prevalence is lower than that of some neighbouring countries. For instance, the SA national HIV prevalence of 14.0%[5] is lower than the HIV prevalence of neighbouring Eswatini and Lesotho whom individually have a reported HIV prevalence of 30% (95%CI: 29 to 32)[6] and 25.6% (95%CI:24.7 to 26.4),[7] respectively. Despite not having the highest HIV prevalence in the SADC region, South Africa has the biggest ART program globally owing to the high number people who are HIV positive.

The burden of HIV requires a targeted plan to address its impact on the health system and on society. The health system continues to face challenges of emerging and re-emerging diseases, in particular the emergence of Covid-19 challenged many health systems worldwide and took centre stage globally.[8] Part of the reason new health challenges such as Covid 19 receive so much attention is availability of credible health data which is crucial for intervention strategies to succeed. For instance, available data shows that the new drivers of HIV in some parts of Africa include urbanisation, poverty, and substance abuse.[9] Such information can guide strategies to tackle the disease burden and inform priority areas.

1.2.2. Role of population-based surveys

In many countries, population-based surveys are used to investigate burden of diseases such as the prevalence of HIV. When compared to other survey types such as sentinel surveys, population based surveys are considered to be the gold standard when estimating HIV prevalence.[10] One of the reasons why these surveys are considered to be stronger is because large samples of data are collected and the samples are representative of the entire study population, including minorities. The questions in surveys are standardised, thus improving reliability. Results from surveys are therefore reliable and can be generalised to the entire population of interest.

Like most study designs, population-based surveys have their own shortcomings. Some of the challenges encountered in these surveys include non-response. Non-response occurs when an individual refuses to participate in a study or where an individual participates but chooses not to answer certain questions. For example, non-response was encountered in the South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017 conducted by the Human Sciences Research Council (HSRC) in a survey also known as the SABSSM V survey.[5] In that survey, some individuals refused to participate when selected, whereas others participated but refused to give blood for HIV testing or refused to answer certain questions. One of the objectives of the SABSSM V survey was to determine the prevalence of HIV. Surveys like the SABSSM V survey are desirable as they do not only rely on self-report of HIV status, but they also do HIV antibody blood testing. This is done to avoid bias such as desirability bias and recall bias. In desirability bias, participants have the tendency to give responses that will be considered in a favourable way by the people conducting the interview. There might not be honest reporting of private details pertaining to participant's lives. Furthermore, current HIV status may have changed and not known if the last HIV test was negative. Whereas in recall bias, participants have difficulty remembering important facts pertinent to the research question. In the SABSSM V survey, to circumvent challenges relating to HIV, blood tests were conducted to measure the outcome directly as opposed to relying on self-reported status.

1.2.3. Non-response as missingness of data in surveys

Despite the advantages of surveys, challenges remain. The SABSSM V survey is confronted by a relatively high non-response rate, especially for an HIV test. For example, in this survey, only 61.1% of eligible participants provided blood specimens for an HIV test. The remainder include those refusing to provide blood sample for an HIV test, those who refused to be interviewed and those who were absent from the household. The proportion of those who refused to provide a blood sample for an HIV test was 32.4% of the eligible study population. To address this challenge in the SABSSM V survey, there was an adjustment of individual weights for HIV non-response during analysis. HIV sampling weights were incorporated during analysis when computing the final prevalence estimates.

The high non-response rate encountered in the SABSSM V survey has the potential to introduce bias. Bias occurs when part of research or the entire research is skewed towards a specific study result due to systematic error in the research process. The results from such a process will be different from the true values. Non-response can introduce bias if factors influencing non-response are associated with the HIV status. Bias can be influenced by many other factors. For instance, information bias in the form of errors when self-reporting. A study found that HIV positive participants were more likely to refuse participation in surveys than HIV-uninfected participants.[11] In such situations, bias is introduced and can be a big issue if the proportion of HIV positive participants refusing participation is high. The uncertainty that is introduced by the 32.4% who refused to give a blood sample for HIV testing in SABSSM V survey can bias the prevalence estimates. For instance, if most of the people that refused to test for HIV were HIV infected, then the actual HIV prevalence would be much higher than that estimated by the survey. It is therefore important to investigate factors associated with refusal to provide a blood sample for HIV-testing in SABSSM V survey and explore ways of using this information when estimating HIV prevalence to reduce bias.

In cases where refusal to participate is associated with HIV status, biased estimates may be obtained if factors influencing the non-response are disregarded during analysis. For some population-based surveys, the effect of non-response pertaining to HIV testing is minimal due to the small proportion of the study

population refusing to test for HIV.[12] However, in the SABSSM V survey, the proportion of non-responders is large, as almost a third of participants didn't agree to have their blood samples drawn for HIV testing.[5] It is important that bias is minimized in surveys because results elicited by these surveys are crucial for planning. For programs focused on HIV, surveys are important in determining progress towards global goals such as the 95-95-95 targets set by UNAIDS. These targets refer to 95% of all people living with HIV to be tested for HIV and know their status, 95% of all people who tested positive for HIV to be enrolled for antiretroviral treatment and 95% of all people on antiretroviral treatment to be virally suppressed, by the year 2025. It is important to have an unbiased estimate of the HIV prevalence to measure whether these goals are being achieved. In addition, factors impeding HIV testing should be explored. Some of the factors that have previously been associated with testing for HIV are perception of elevated risk of HIV infection such as increased number of sexual partners and engaging in sex after alcohol intake, additionally, having recently visited a health facility was also associated with testing for HIV.[13, 14] A study conducted in Taiwan reported that having urinary tract infections (UTI) is associated with HIV testing.[15] It is possible that people being infected in the urinary tract are more likely to present to a health facility and due to association between UTI and HIV risk will be more likely to test for HIV.

There are many other factors that can influence non-response. Interviewers can influence non-response among participants[16]; if interviewers are not trained properly or are not randomly assigned, it can exaggerate the influence of refusal bias. Besides interviewers, some of the drivers of non-response for participants are urban dwelling, high education attainment and wealth.[10] This could be due to the association between education, wealth and dwelling. Education attainment can lead to the ability to afford a highly secured dwelling which may not be accessible to field workers. Also, the upper middle class may have the perception that they do not benefit from public health services as they have greater access to private health care. They would thus refuse to participate in such surveys. Furthermore, in the SABSSM V survey non-response was associated with race, with observation of low response rate among whites and Indians. It is possible that HIV is not perceived to be a challenge in these groups.

Non-response should be viewed in the context of missing data. The mechanism of missingness of data is key during analysis and it is important to determine whether one is faced with data missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). The MCAR mechanism depends on random omission of data irrespective of observed covariates while MAR depends on observed outcomes. For example, assuming after data was collected, there is ***X (age)*** as one variable that does not have missing values and then there's ***Y (HIV status)*** which for some observations is missing. In such a case, MCAR occurs when the missing ***Y (HIV status)*** is independent of ***X (age)*** and ***Y (HIV status).*** In such a case, participants with HIV status missing are not different in terms of ***age*** or ***HIV status*** (Whether they are HIV infected or not). Regardless of ***age***, proportion of missing ***HIV status*** observation is similar across age groups. This could have resulted from the loss of specimens or malfunction of some test. If it is found that there is an association between ***age*** and ***HIV status***, then it can be concluded that the mechanism of missing data is not MCAR.

If the missingness of ***Y (HIV status)*** differs across ***X (age)*** groups, then the mechanism of missing data is MAR. In MAR, a larger proportion of one age group experiences missing data compared to the other groups. For example, finding that older age groups have more missing values than younger. In such a case, it would mean age as an observed variable has an influence on the missingness of data. When there is an association between certain demographic characteristics between ***Y (HIV status)*** with ***X (age)*** or even race or sex, it can also lead to MAR. In such cases, post-stratification can be applied by assessing and adjusting the characteristics through weights for non-responders.

There are times when the mechanism of missing data is neither MAR nor MCAR, in such cases the mechanism is Missing not at random or MNAR. This mechanism occurs when the missingness of ***Y (HIV status)*** is dependent on ***Y (HIV status)*** even when conditioning for X(age) or other variables.

The MCAR is ignorable because the missingness of HIV status is at random, as it is not likely to lead to biased estimates.[17] It is often that refusal is non-random, such that it depends on observed or non-observable covariates. If present, MNAR can lead to bias because it can be dependent on outcome variable(s). MAR would

correspond to ***Y (HIV status)*** being more likely to be missing depending on certain observable characteristics of participants. For example, socio-demographic or socio-economic characteristics such as education level or employment status. It is therefore important to establish the missing data mechanism. If the missing data mechanism is classified as MAR, then observed covariates can be used to adjust the sampling weights.

1.2.4. Sample weights

When conducting complex population-based surveys, sample weights are used because they reduce bias when inferring survey findings to the general population.[18] Sample weights are calculated by computing the inverse of the probability of selection and are used in complex surveys because not everyone has an equal chance of having been sampled. Surveys are conducted with the aim of being representative of the population of interest, thus ensuring the inclusion of minorities through oversampling. Sampling weights are required to get the correct SEs and correct 95%CIs by adjusting for the unequal selection probability in different strata or clusters. In the SABSSM V survey, the sample weights were derived from demographic information and were calculated separately for the questionnaire and for the HIV-testing results, where the sample was restricted to those who provided a blood sample.[5]

Several statistical approaches can be employed to tackle missingness of data. Among the methods, the three methods are firstly to simply omit the non-responders, secondly to impute the missing values and thirdly to modify the sampling weights using the probability of responding. Omission should not affect estimates if non-response results from MCAR. However, if non-response is MAR or MNAR, it can lead to biased estimates. Application of imputation involves using observable characteristics in the estimation of missing values. When the missing data mechanism is MAR, assigning statistical weights to non-responders may be used. For example, in the SABSSM V survey, for the demographic questionnaire, information used to calculate the questionnaire sampling weights included province, age, sex and race. For the analysis of the HIV testing results, the weights were recalibrated using these and other sociodemographic characteristics. In addition to the socio-demographic characteristics that were used to recalibrate the weights, there could be other factors, particularly behavioural characteristics, that could contribute to the HIV results being missing.

In the SABSSM V survey, weights were calculated separately for the questionnaire data and for the HIV test results. For the questionnaire, the weights were adjusted so that for all the subgroups, weights are added up to the total population size. For instance, for sex, the weights for all males added up to the total

population size for males. This was done for all participants. For HIV sampling weights, the same procedure was used, but the sample was restricted to those who agreed to have their blood tested for HIV. The use of sample weights are recommended particularly if variability between groups on certain elements is large.[3, 19] There are different strategies for calculating sample weights in surveys, for example raking and cell weighting. [18]

When performing raking, sample weights are adjusted to represent the population by relying on population parameters, so that results can be inferred to the population. The raking technique was used in adjusting for demographic characteristics of the population of Armenia in a study aimed at improving poverty measurements.[20, 21] When analysing complex surveys in situations where there is non-response among study population, cell weighting can be used. Cell weighting involves identifying similarities between those who responded and those who did not.[21] A study by Pike found that weighting can lead to high precision for some parameters while for other parameters it does not.[22] Ideally, one would want to know which parameters affect estimates when conducting a study. When comparing different estimation approaches for population estimates, Downes et al found that Multilevel regression and poststratification (MRP) were accurate.[23] In their study, MRP estimates were compared with those found using simple weighted, unweighted, and raking models.

There is no clear consensus on how sample weights should be adjusted for factors such as non-response. Some argue that variation between sources of nonresponse or refusal to participate fully in a study is crucial to decide on the use of sampling weights.[22] When carrying out an analysis, it is important to determine the type of missing data that one is dealing with, and to explore different approaches with the aim of reducing bias.

This study aims to investigate the potential impact of different sampling weights on the estimation of HIV prevalence in the SABSSM V survey. By using the HIV sampling weights and the questionnaire survey weights, the effect of different weights on the overall prevalence estimate was explored. Then behavioural factors that influence the decision of whether an individual provided blood sample for HIV testing were investigated. These factors were used to adjust the original sampling

weights. The effect of using these new weights on the estimated HIV prevalence was investigated and compared with the HIV prevalence estimates determined using HIV sampling weights and questionnaire survey weights.

1.3. Conceptual framework

Some factors which could influence prevalence estimates in population-based surveys are shown in the conceptual framework in **Figure 1**. Interviewer effects could lead to differential refusal rates. Participant's characteristics such as sexual behaviour, socio-economic status and general health status could also influence the non-response rate. At analysis phase, high non-response rates on the outcome of interest (***HIV status***) could lead to refusal bias. Other factors such as stigma which are not routinely measured in surveys can also introduce bias. Refusal bias can be addressed at the analysis stage through the adjustment of the sample weights.

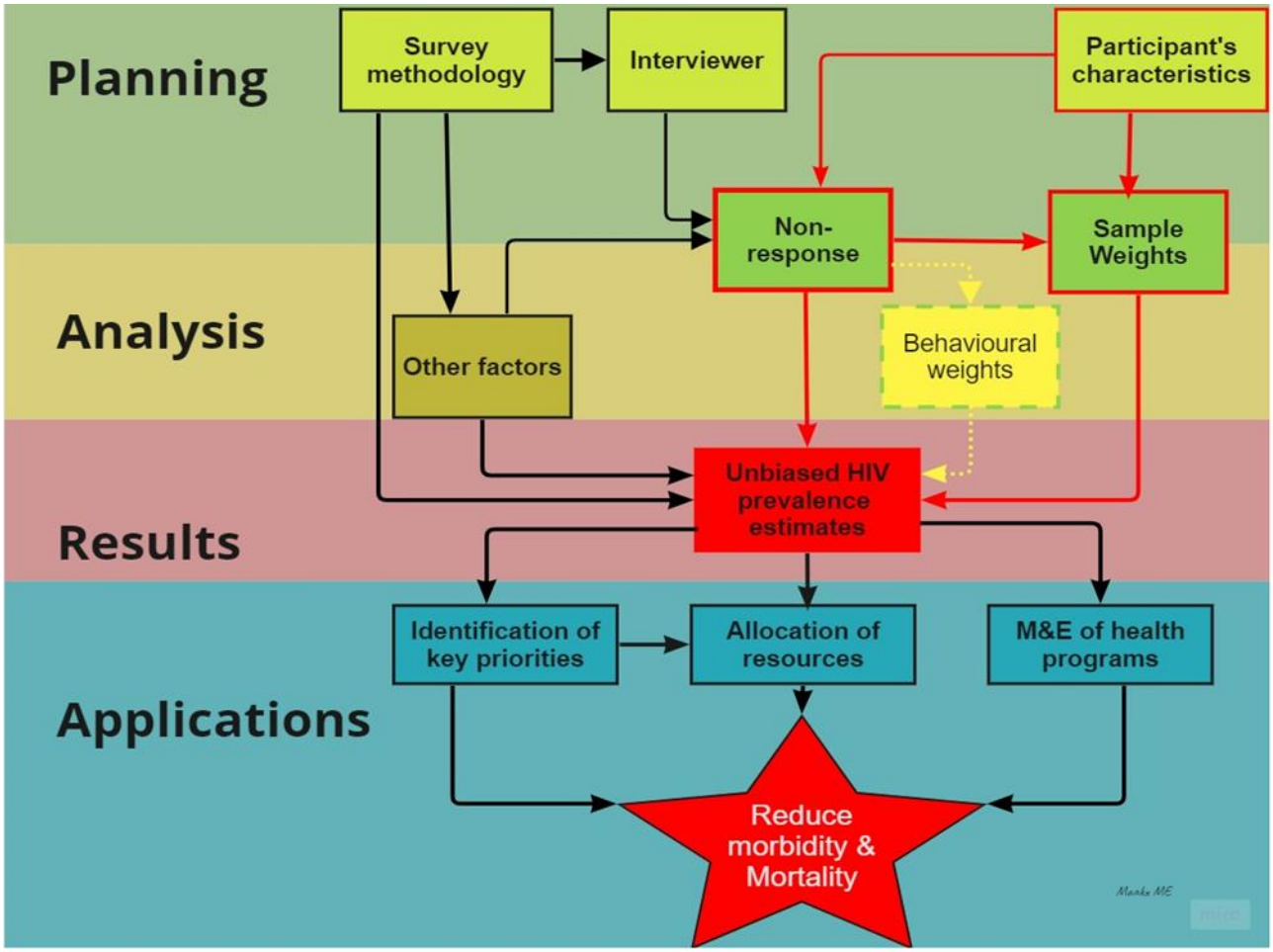


Figure 1: Conceptual framework of factors and management of non-response bias

1.4. Problem statement

During analysis of the SABSSM V survey, the sample weights were taken into account. Although the sample weights were recalibrated to account for HIV non-response, the adjustments did not consider behavioural factors which could impact on both non-response and the actual HIV status. This could lead to biased HIV prevalence estimates, particularly if non-response is associated with variables which are associated with the outcome (***HIV status***).

1.5. Justification

Underestimation of HIV prevalence estimates due to non-response may conceal the true burden of HIV and result in missed opportunities to intervene, leading to high morbidity and mortality. Conversely, overestimation of the prevalence estimates may lead to a waste of resources as resources will be committed to an exaggerated health threat. HIV prevalence surveys play an important role in health planning, so it is important to minimize bias in the prevalence estimates.

1.6. Research question

The analysis sought to answer the three questions:

- Do HIV prevalence estimates differ when using different sampling weights?
- Which behavioural factors influence non-response (Refusal to take HIV test)?
- Does the HIV prevalence estimate change when using the new weights?

1.7. Aim and Objectives

1.7.1. Aim:

To investigate the potential impact of the use of different sampling weights in the estimation of HIV prevalence in the SABSSM V survey.

1.7.2. Objectives:

i. To estimate overall HIV prevalence as well as HIV prevalence broken down by socio-demographic factors such as *sex, race, age, province* and *locality type* using:
   1. The HIV sampling weights from survey
   2. The questionnaire sampling weights from the survey.

ii. To investigate the impact of socio-demographic and behavioural variables on non-response for HIV testing by fitting a robust multiple logistic regression model.

iii. To investigate the effect of adjusting the sample weights using inverse probability weighting as a way of dealing with missing HIV results. The goal is to provide a third estimate of HIV prevalence and compare this with the prevalence estimates from **Objective i**.

Chapter 2

2. Methodology

2.1. Study Setting

The primary study, SABSSM V survey, was conducted in the year 2017 in South Africa which is a country situated in the southern part of the African continent as shown in **Figure 2**. South Africa is divided into nine provinces. When the study was conducted, the population size of South Africa was 56.43 million.[5] The racial profile of the country is comprised of Black Africans who are in the majority, followed by White, Coloured and Indian groups. The country has the highest number of HIV infected people in the world and the Department of Health in the country runs the biggest HIV and AIDS program globally. This study consists of secondary data analysis using the dataset from the SABSSM V survey.

Figure 2: Map of South Africa

2.2. Study design

2.2.1. Description of primary study design

The primary study was a cross-sectional, population-based household survey and stratified multistage cluster random sampling was used. The data was collected between December 2016 and February 2018. The HSRC Ethics committee approval details were as follows: REC: 4/18/11/15.

2.2.2. Secondary study design

This project involved secondary data analysis of the SABSSM V survey dataset. The HIV prevalence estimates calculated using the HIV sampling weights were compared to the HIV prevalence estimates calculated using the questionnaire sampling weights. Factors associated with non-response regarding HIV testing were estimated by fitting robust multiple logistic regression models that take the sample design of the survey into account. The chosen regression model was used to predict the probability of non-response which in this context means refusal to providing a blood sample for HIV-testing. These predicted probabilities were then used to adjust the original sampling weights and HIV prevalence was re-estimated.

2.3. Sample

2.3.1. **Study population**

All people who were eligible to participate in the SABSSM V survey.

2.3.2. **Sampling**

Primary study

SABSSM V survey used a multistage sampling design, with disproportionate sampling as explained below. The primary sampling units were small area layers (SALs). The SALs replaced enumeration areas (EAs) as the spatial area used for the collection of census information in South Africa and they exist at a higher level than the previously used EAs. EAs consisted of between approximately 80 households in rural areas and 180 households in urban areas. In forming the SALs, EAs that had few visiting points were merged to form a single SAL. The SALs were sampled with probability proportional to size, with the measure of size being the number of visiting points (VPs) within each SAL. The selection of SALs was stratified by province and locality type. Oversampling was performed among SALs that were predominantly inhabited by minority racial groups and in the sparsely populated Northern Cape Province to get sufficient sample sizes in these groups.

Secondary study

The whole dataset from the SABSSM V survey was analysed to answer the research question.

2.3.3. **Sample size**

In this study population, the total sample size for those who completed the questionnaire was 36 609 and the total sample size for those who tested for HIV was 23 910. The whole dataset formed part of this study. In calculating power and sample size estimations, different domains were considered. In the primary study, when determining the sample size, the power was set at 80%, level of significance was 5%. Taking account of the survey design, a moderate design effect (DEFF) of 2 was assumed. The sample size in the primary study was approximately 36 609.

Therefore, the effective sample size = $36609/_{deff}$ ≈ 18000. Assuming roughly equal numbers of males and females, a sample size of 9000 males and

9000 females would have been sufficient to detect an absolute difference of 3.5% in the proportion refusing to test between males and females with over 90% power when one assumes that overall, roughly 35% refused to test. The sample size was sufficient to detect factors associated with refusal to test for HIV. In investigating the effect of different weighting schemes on the estimate of HIV prevalence, a total sample size of 23 910 who tested for HIV was considered sufficient.

2.3.4. Sample weights

Primary study

The individuals included in the SABSSM V survey dataset were sampled with different probabilities. To take these different sampling probabilities into account, sampling weights were calculated. The sampling weights are proportional to the inverse of the sampling probability. The weights reflect the disproportionate allocation of SALs and differ by race, province and locality type (geotype).

Secondary study

To investigate the potential impact of different sampling weights, the overall HIV prevalence was estimated and broken down by socio-demographic and behavioural factors: (i) using the HIV sampling weights from the survey and (ii) using the survey weights from the questionnaire. For the questionnaire, the weights were adjusted so that for any sub-group (e.g. African males) the weights added up to the total population of African males. Initially for the questionnaire this was done for all respondents. For the HIV results a similar procedure was carried out but restricted to those who agreed to provide a blood sample for an HIV test. In summary, for those who agreed to test for HIV, the weights for each subgroup added up to the population total for that subgroup.

A robust multiple logistic regression model which took the survey design into account was fitted to find factors associated with refusal to take HIV test. The model was then used to predict the probability of refusal for each individual and hence the probability of providing a blood sample to test for HIV for each individual, as the complement of the probability of refusal. These predicted probabilities were then used to adjust the original sampling weights for those who provided a blood sample. This is known as “Inverse probability weighting”, which is one way of dealing with

missing outcome data. The effect of using the new third weights on the HIV prevalence was also investigated.

2.4. Data management

2.4.1. Data collection and management

In the SABSSM V survey, the dataset was collected between December 2016 and February 2018. The dataset was requested from gatekeepers at the HSRC (Appendix 4) and upon receipt, it was stored on a computer and was password protected. The dataset could only be accessed by the relevant researchers. All personal identifiers were removed in the original study to ensure confidentiality. Some of the explanatory variables were recoded as part of the model building process when finding factors associated with non-response for HIV test.

Some categories were combined for ease of analysis, for instance, stratum was created from geotype and province. Some codes were combined into one, for example when generating hivtest from hivstat. From the dataset, hivstat variable referred to HIV status and was coded such that “1” was for HIV positive and “2” was for HIV negative. When creating the new variable hivtest which referred to agreeing to test for HIV, both code 1 and 2 were coded 1 (Both negative and positive for HIV indicate that they were tested) for HIV test and the rest (missing) was coded 0 (not tested for HIV). From hivtest, an inverse variable refusehivtest was created with 1 indicating refusal to test for HIV, while code 0 indicated agreeing to test for HIV.

Confounding was addressed by adjusting for potential confounding variables when running the multiple logistic regression model. Potential confounding variables included age, race and sex.

2.4.2. Data sources

Dataset from the SABSSM V survey was collected directly from participants through a population-based survey and was used in the current study.

Data processing involved deletion of some observations that were not going to be utilized in the study. For example, observations were deleted if the value of province was missing. If two variables measured essentially the same quantity only one of the variables was considered.

2.5. Analysis

2.5.1. Objectives

Data was analysed using STATA inter-cooled version 16.0 software package. Based on the SBSSM V report, a substantial amount of data is categorical.

2.5.1.1. Objective 1

The influence of sampling weights on HIV prevalence estimates was investigated by comparing the estimates obtained utilising the two weights provided (overall questionnaire weights and the HIV sample weights). The HIV prevalence was reported across different levels of sex, geotype, race, age and province.

2.5.1.2. Objective 2

Multiple logistic regression models were fitted to find factors associated with participants refusal to take HIV test. The outcome of interest was refusehivtest which referred to refusal to provide a blood sample for HIV testing. Screening of variables to include in the regression model was performed through bivariate analysis. This was performed to identify variables that predicted refusehivtest using a liberal cutoff (p-value < 0.20). The liberal p-value was used with the aim of detecting all possible confounders.[24] When carrying out the screening for variables, the design variables, namely province and geotype, as well as sex, race and age, were included in the model for each variable screened.

The design variables were fixed when performing both the forward and backward stepwise regression. Only variables preselected when screening for variables were considered for stepwise regression. When performing stepwise regression, a stringent p-value (P<0.05) was used to select potential variables to include in the model. The model selected through backward stepwise regression was chosen. Backward stepwise was selected because of its high R^2 statistic compared to the forward stepwise regression. This indicated that backward stepwise regression explained better the proportion of the variation in the refusetesthiv that is predicted by the behavioural factors. The **Wald** test was used to check if addition of the covariates improved the fit of the model using a stringent p-value (p<0.05). This model still included the five fixed socio demographic variables comprising of the

design variables such as province and geotype as well as variables that were needed for tabulations, namely age, race and sex.

Logistic regression diagnostics were performed to check if assumptions were met. Omission of important variables was investigated using **linktest**. The **Hosmer-Lemeshow** goodness of fit test was used to check the goodness of fit of the final model.[24] Multicollinearity which results from multiple explanatory variables being linearly determined by combination of other explanatory variables was investigated. Multicollinearity was investigated using **collin** command which is a user-written command downloadable from **Stata**. Influential observations were investigated using **Pearson** residuals and **Pregibon** delta beta statistics (**Pregibon's dbeta**). The final model was tested to check how it performed in predicting non-response. Receiver Operating Characteristics (***ROC***) curve was used to check how well the model predicted the outcome of interest.

2.5.1.3. Objective 3

The logistic regression model of the outcome of interest (refusetesthiv) was reversed such that the model predicts testing for HIV (hivtest) or the probability of agreeing to test for HIV. The logistic regression model predicting probability of testing for HIV was then used to determine the probability of providing blood sample for an HIV test (pr_tested). The final behavioural weights were determined by first dividing questionnaire weights (*ibreal12_combined*) by the probability (*prtest*) of agreeing to an HIV test to get the behavioural weights. When dividing the mean of new weights by prtest, the new weights were inflated and was scaled by multiplying by constant derived by dividing the mean of HIV weights by the mean of behavioural weights. To determine the new behavioural weights (third weights) the formula as shown be was used:

$$final\_behav\_wt = \frac{ibreal12\_combined}{prtest} * \frac{mean\ of\ ibreal1\_combined}{mean\left[\frac{ibreal12\_combined}{prtest}\right]}$$

Final weights were used to estimate the overall prevalence of HIV and stratified according to sex, geotype, age, race and province. Brief description of analysis for each objective is shown in **Table 1**.

Table 1: Summary of the analysis for the objectives.

Objective	Analysis
1. To investigate the influence of sampling weights on HIV prevalence estimates by comparing utilising the two weights provided (**overall questionnaire weights** and the **HIV sample weights**)	The two sets of HIV prevalence estimates, together with the SEs and 95% confidence intervals, were compared
2. To investigate behavioural variables associated with non-response (Refusal to give blood for HIV test).	i. Robust multiple logistic regression (Bivariate & Multivariate) ii. For initial model selection (Bivariate regression analysis), the cut-off p-value was set at 0.20 to identify candidate variables and a more stringent ($p<0.05$) was used when selecting variables to include in the multivariate regression model using the Wald test. Model was checked for logistic regression assumptions. iii. Final model was used to derive new weights, and these were scaled to determine final weights.
3. By using weights resulting from the behavioural factors in addition to questionnaire weights and sample weights.	The model from **Objective 2** predicted the probability of refusing to give blood test for HIV testing (non-response), from which the probability of having agreed to test for HIV was calculated. The questionnaire sampling weights (i.e. the survey weights) were divided by the probability of agreeing to test (response) for each participant. These weights were used after scaling to provide a third estimate of HIV prevalence.

2.5.2. **Outcome variables**

HIV status (Self-reported or blood tested)-hivnew, HIV testing response (Y/N)-hivtest or refusehivtest, with the first outcome hivtest used to investigate the probability of doing the HIV test and while refusehivtest being used to investigate the probability of refusing to do the HIV test.

2.5.3. Predictor variables (Some)

The predictor variables that were explored included:

- Whether the respondent is sexually active
- Condom use at last sexual encounter
- Consumption of alcohol

2.6. Confidentiality

In this study, confidentiality was dealt with in the following manner. All identifiers were removed from the dataset, and this was done before the dataset was released to the researcher for analysis.

2.7. Ethical clearance

Permission to access the dataset which was obtained from the HSRC and the Ethics certificate for SABSSM V survey was attached **(Appendix 3).** The study commenced after approval of ethical clearance (Clearance certificate No: M230231 MED22-10-134) issued by the Human Research Ethics Committee from the University of the Witwatersrand (**Appendix 3**). The dataset was anonymised, so that no individual was identifiable.

2.8. Dissemination

The final research report will be submitted to the University of the Witwatersrand for storage.

CHAPTER 3

3. Results

In the SABSSM V survey, the non-responders comprised of those who were not at home (2.4%) and those who refused to be interviewed (4.1%). After accessing the dataset from the HSRC, a total of 1014 observations were excluded due to province being missing. Province and geotype were combined to generate the composite variable stratum. The two HIV prevalence estimates calculated using two different weights from the SABSSM V survey, namely questionnaire sampling weights are (ibreal12_combine) and HIV specimen sampling weights (ibreal1_combine) were compared. The HIV prevalence estimate calculated using HIV specimen sampling weights were compared to the HIV prevalence estimate calculated using the new behavioural sampling weights derived in this study.

3.1. HIV prevalence estimates from current weights

A comparison of HIV prevalence estimates using weights derived from HIV data and weights derived from questionnaire data was carried out. A higher estimate of the overall HIV prevalence was obtained when using the weights for the questionnaire data compared to the estimated prevalence using the weights for the HIV test results. When using the questionnaire weights, the estimated prevalence of HIV in South Africa was 14.97% with SE = 0.38 (95%CI: 14.26 to 15.71) while with the HIV data derived weights the estimated HIV prevalence was 14.01% (SE=0.40; 95%CI:13.22 to 14.80). The 95% confidence intervals were also wider in the HIV data weighted estimates compared to the questionnaire data weighted estimation. Similarly, when HIV prevalence was estimated across provinces using questionnaire data, HIV prevalence in Mpumalanga province remained the highest (21.15%; SE=0.93 [95%CI: 19.32 to 22.97]) followed by KwaZulu-Natal province at 20.20% SE=0.84 (95%CI:18.54 to 21.85). When using HIV data weights, HIV prevalence reduced to 19.37% (SE=1.28; 95%CI: 16.85 to 21.89) in Mpumalanga Province and KwaZulu-Natal province also reduced (18.04%).

Table 2: Comparison of HIV prevalence estimates derived using questionnaire weights and those derived using HIV sample weights

HIV prevalence comparing estimates derived from questionnaire data weights and HIV sample weights							
		Questionnaire weights			**HIV sample weights**		
	Sub-category	Prevalence %	SE	95%CI	%Prevalence	SE	95%CI
Overall	Overall	14.97	0.38	14.23-15.71	14.01	0.40	13.22-14.80
Sex	Male	11.10	0.42	10.28-11.92	10.79	0.44	9.93-11.65
	Female	18.38	0.48	17.45-19.31	17.09	0.59	115.94-18.23
Race	African	16.91	0.42	16.09-17.73	16.15	0.45	15.26-17.04
	Coloured	2.46	0.86	4.68-7.46	5.90	0.76	4.41-7.39
	Indian	2.46	0.86	0.78-4.14	2.43	0.95	5.70-4.29
	White	2.60	0.45	1.71-3.48	4.59	1.13	2.37-6.80
Geotype	Urban	14.43	0.54	13.37-15.50	13.03	5.16	12.02-14.05
	Rural (tribal)	15.65	0.50	14.66-16.63	15.28	0.68	13.95-16.62
	Rural (farms)	16.81	1.26	14.35-19.28	17.77	1.31	15.20-20.34
Province	WC	9.22	1.38	6.53-11.92	8.61	1.16	6.33-10.88
	EC	15.83	0.97	13.93-17.74	15.91	1.56	12.84-18.98
	NC	8.61	0.80	7.03-10.18	8.21	0.75	6.75-9.67
	FS	17.68	1.19	15.34-20.01	17.01	1.25	14.55-19.47
	KZN	20.20	0.84	18.54-21.85	18.04	0.80	16.47-19.62
	NW	16.26	0.85	14.60-17.91	14.82	0.86	13.13-16.50
	GP	13.79	0.90	12.02-15.56	12.05	0.85	10.39-13.72
	MP	21.15	0.93	19.32-22.97	19.37	1.28	16.85-21.89
	LP	10.15	0.70	8.77-11.54	10.13	0.85	8.46-11.80
Age groups	15-19 yrs	5.05	0.47	4.12-5.98	4.87	0.48	3.93-5.81
	20-24 yrs	10.77	0.68	9.44-12.11	10.54	0.82	8.94-12.14
	25-29 yrs	19.80	0.99	17.87-21.73	18.95	1.13	16.73-21.17
	30-34 yrs	28.10	1.26	25.62-30.57	26.45	1.54	23.44-29.47
	35-39 yrs	33.96	1.55	30.91-37.00	32.99	1.65	29.75-36.22
	40-44 yrs	30.85	1.49	27.93-33.76	29.31	1.58	26.22-32.40
	45-49 yrs	29.09	1.67	25.81-32.38	26.95	1.95	23.12-30.78
	50-54 yrs	21.71	1.50	18.77-24.66	21.31	1.77	17.84-24.78
	55-59 yrs	15.26	1.41	12.49-18.02	15.01	1.64	11.79-18.22
	60 yrs +	7.21	0.69	5.86-8.56	6.60	0.68	5.27-7.94

The lowest HIV prevalence was estimated for the Northern Cape province at 8.61% (SE=0.80; 95%CI: 7.04 to 10.18) when using questionnaire data derived weights and this was also higher than the prevalence (8.21%; SE=0.75; 95%CI: 6.75 to 9.67) estimated using weights derived from HIV data.

Of all the geotypes, rural farms experienced the highest HIV prevalence (16.81%; SE=1.26 [95%CI: 14.35 to 19.28]) when estimated using the questionnaire data derived weights. When using the weights from the HIV test results, rural farms still had the highest prevalence, although it was lower than that found using the weights from the questionnaire data. Using the questionnaire derived weights, HIV prevalence was at its peak (33.96%; SE=1.55 [95%CI: 30.91 to 37.00]) from the age of 35 years to 39 years, and gradually decreased at both ends to reach lowest HIV prevalence of 5.05% (SE= 0.47; 95%CI: 4.12 to 5.98) in the age group 15 to 19 years. A similar pattern was observed using HIV sample derived weights, although the estimates were slightly lower. Using the questionnaire weights, Africans (16.91%; SE=0.42; 95%CI 16.09 to 17.73) had a higher HIV prevalence than other races. HIV is also more prevalent among females (18.32%; SE=0.48; 95%CI: 17.45 to 19.31) than males (11.10; SE=0.42; 95%CI: 10.28 to 11.92) when using both HIV data weights and questionnaire weights. In summary, HIV prevalence estimates from HIV sampling weights had slightly elevated SEs and wider confidence intervals than those derived from questionnaire weights.

3.2. Factors associated with Refusal to take HIV test

Predictor variables associated with refusal to take HIV test were investigated. Screening was performed to investigate predictor variables that predicted non-response using liberal selection criterion ($p<0.20$). During screening of variables, 81 variables were considered. After screening, 60 predictor variables were considered for inclusion when performing stepwise regression. The model chosen by stepwise regression used a more stringent selection criterion ($P<0.05$) and included seventeen (17) variables including the five variables which were fixed in the model, namely province, geotype, sex, race and age. The final model was selected through Wald test using a stringent selection criterion ($p<0.05$) and comprised of eleven (11) variables including the five fixed variables. Logistic regression diagnostics were performed to assess if the logistic regression assumptions were met.

3.2.1. Explanatory variables associated with refusal to give blood sample for HIV testing

The final multivariable logistic regression model resulted in the inclusion of eleven (11) predictor variables comprising of five (5) fixed variables and six (6) selected behavioural predictor variables.

Table 3: Variables predicting ***refusetesthiv*** including the five fixed variables.

	Variable/code	**Description**	**Variable**
1	province	The name your province	Province
2	geotype	Geographical location (rural [farm]; rural[tribal] or urban)	Geotype
3	sex_q	Sex of respondent	Sex
4	race_q	Race of respondent	Race
5	agecat_13	Age of respondent	Age
6	q13_1	Health status (Poor/Fair/Good/Excellent)	Health status
7	q13_4	Been hospitalised for illness in the past 12 months	Hospitalisation
8	q9_6	Received counselling before recent HIV test	HIV test pre-counselling
9	q13_7a_i	Diagnosed with Hypertension	Hypertension diagnosed
10	q14_7	Feeling depressed	Feel depressed
11	q16_2	Being away from place of residence for at least a month in one year	Away from usual residence

All these variables were fitted into the logistic regression predicting refusal to give blood for HIV testing (***refusetesthiv***). Running the multivariate regression analysis resulted in a model which predicted the outcome of interest, which was

refusetesthiv. Covariates that formed part of the **final model** predicting the probability of ***refusetesthiv*** are as shown in **Table 3.**

3.2.2. Model performance

The logistic regression model was tested using the Receiver Operating Characteristics (ROC) curve to check how it performs in predicting non-response (***refusetesthiv)*** as shown in **Figure 3**.

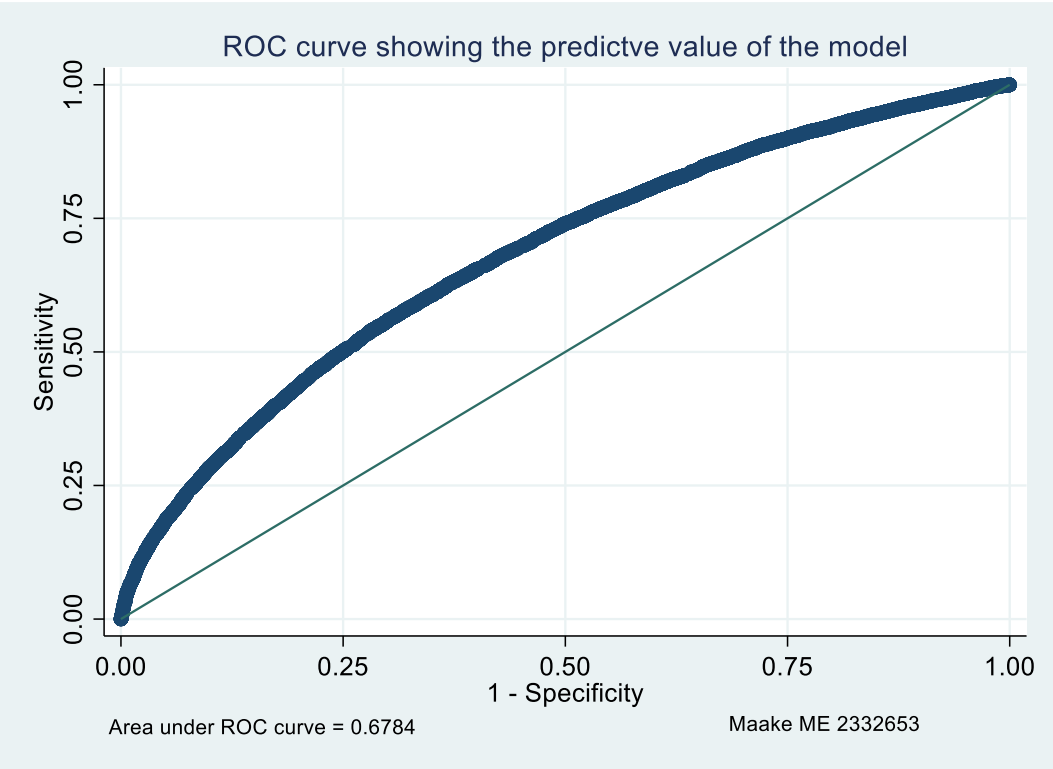


Figure 3: ROC curve showing performance of the model.

The ROC curve analysis in **Figure 3** indicated a 67.84% chance that the model will correctly predict those who refuse to test for HIV. The Area Under Curve (AUC) of 0.6784 indicates a satisfactory performance of the model.

3.2.3. Logistic regression diagnostics

Goodness of fit

To check the goodness of fit of the model, the ***Hosmer and Lemeshow*** test was performed. There was evidence of lack of fit (P=0.002).

Omission of important terms from model

Omission of important variables from the model was checked using ***linktest*** to test whether any variables were omitted or if there was the need for an interaction term that was excluded from the model. When running the ***linktest***, the ***_hat*** was statistically significant (P<0.001) and therefore the model was correctly specified. The ***_hatsq*** was also overwhelmingly significant (P<0.001), therefore the ***linktest*** was significant and this suggested that there are relevant variable(s) that were omitted. An attempt to create composite variable between medical aid cover and hospitalisation did not yield a change. In summarily, the model was properly specified but the relevant variable(s) may have been omitted.

Multicollinearity

The model was checked for multicollinearity to investigate if any two or more predictor variables were determined by linear combination of other predictor variable in the model. There was slight difference in VIF statistics for some covariates. For example, for the three covariates, namely hypertension, feel depressed and medical aid cover. For these three variables, the Variance Inflation Factor (VIF) was slightly elevated as it ranged from 3.84 for hypertension to 5.27 for feel depression. Furthermore, tolerance for the three variables was closer to zero compared to the rest, feel depressed had the lowest tolerance of 0.19 followed by medical aid cover at 0.20. The mean VIF= 2.27. All these results did not indicate conclusive evidence of correlation for the three variables as none of the variables had a VIF of 10 and all had tolerance of greater than 0.1.

Influential observations

Relying on residual measures, Pearson residuals were checked to investigate influential observations that might have significantly impacted the model as shown in **Figure 4**. Some observations appear to be slightly detached from most of the other observations slightly detached from the rest. These observations in **Figure 4** appearing above the yellow line and those below the orange line appear slightly detached from the rest. Those that appear above the yellow line (stdres greater than five) were twelve (12) in total and those that appear below the orange line (stdres less than negative 5) were 73. For Pearson residual analysis, a cutoff point of values greater than 3 is often used for outliers.[25] This is commonly the case when the end goal is just to get factors associated with outcome of interest. There are many residuals in the scatter plot that appear to be greater than three (3) but are not detached from the rest. Exclusion of those that are of those that appear detached was explored. The exclusion of these observations did not change the model. Also, in this study the end goal is to use the regression model to determine weights which will be used to get HIV prevalence estimates. Hence, the researcher paid attention only to those that appeared to be slightly detached from the rest of the observations.

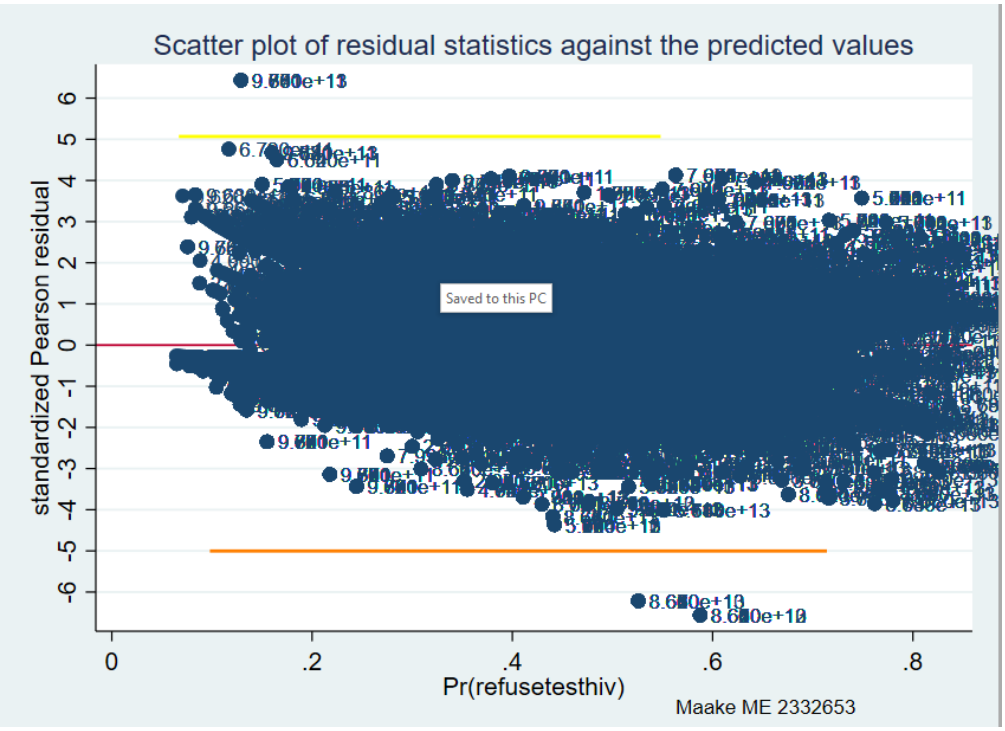


Figure 4: Pearson residuals scatter plot.

The second plot for residual as shown in **Figure 5** was an index plot which plotted statistics against index ID (**idnum**). This plot also showed a similar pattern where some observations appeared to be detached from other observations when visually inspecting the scatter plot. These were the same observations that were previously deleted without a change in the regression model. All observations were therefore kept in the model and not removed to allow further investigation of their influence.

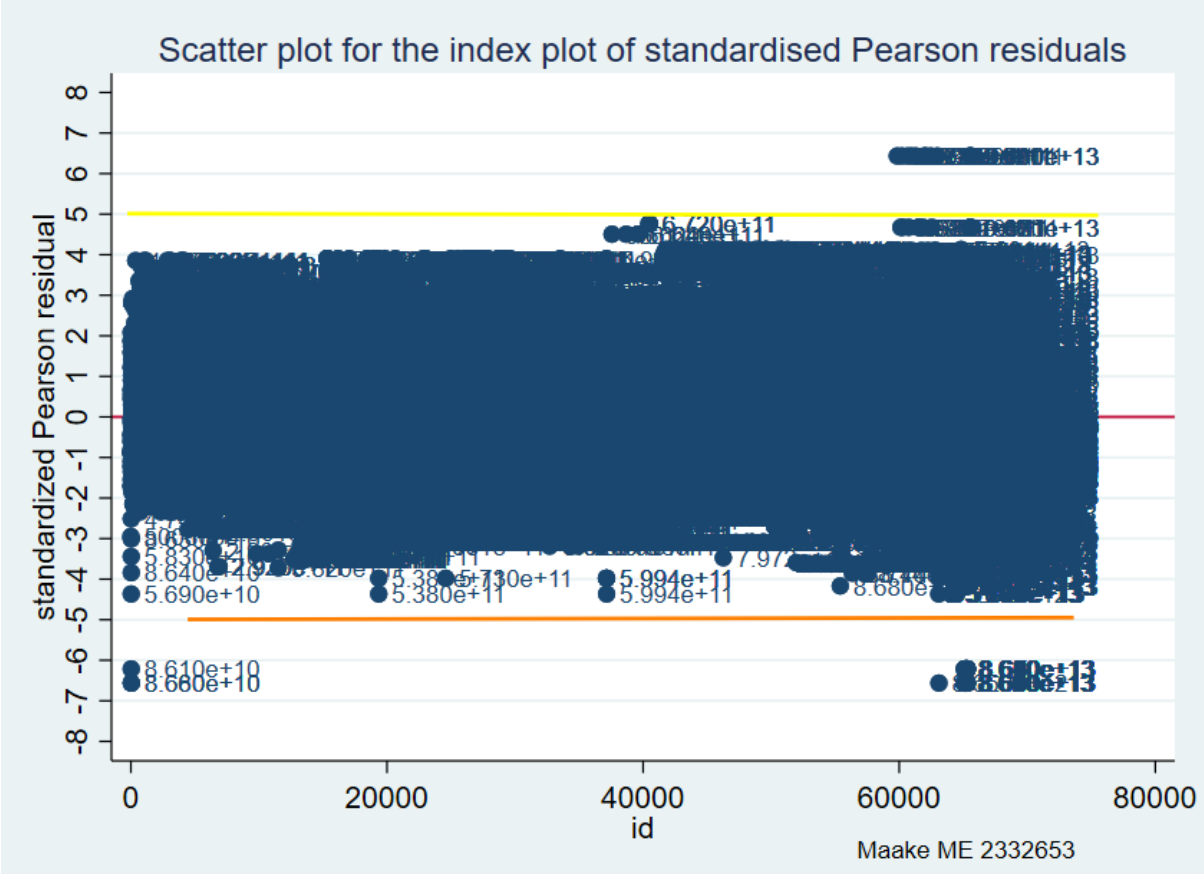


Figure 5: Index plot standardized Pearson residuals

The next diagnostic test performed to explore how individual observations influence parameter estimates was *Pregibon delta beta statistics* (***Pregibon's dbeta)***. This was done to investigate observations with high leverage which also when omitted may lead to a change on one or more coefficients. Visual inspection showed that there are observations that are far from most observations. These are the observations with Pregibon's dbeta greater than one (1) that appear above the yellow line in **Figure 6**. Further probing revealed that these observations were 39 in total. When all these 39 observations were removed from the dataset, no change

was noted on the regression model. All covariates were still significant. These observations did not have an impact on the logistic model. It was decided that the observations should be kept in the dataset.

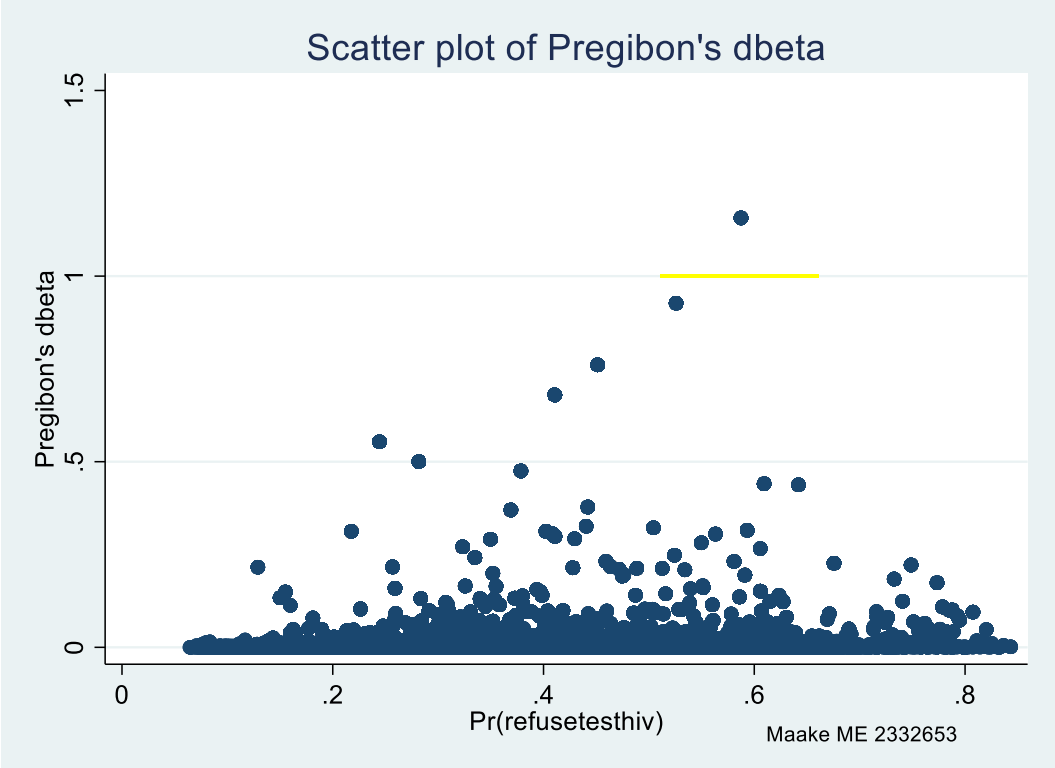


Figure 6: *Pregibon's delta beta* statistics.

In addition, although these observations appear detached from the rest of the observations and are greater than the **Pregibon's** cut-off value of one(1),[26] they did not change the regression model when they were removed. For these reasons and reasons mentioned previously, it was decided that the observations should be kept in the model and not removed to allow further investigation of their influence. The **final model** was kept as it was and was used to determine behavioural weights which would be used to calculate HIV prevalence estimates. After doing the above diagnostic tests, the final logistic regression model seemed to be overall satisfactory.

3.2.4. Description of the final logistic regression model

The multivariate logistic regression model above was overwhelmingly significant (P<0.001). The outcome of interest was refusing to provide blood sample for HIV test.

3.2.4.1. Fixed variables

Comparison within the fixed variables as shown in **Table 4**, revealed that for provinces, residents of some provinces had greater adjusted Odds Ratio (aOR) of refusing to test than others. For example, using the Western Cape province as reference showed overwhelmingly (P<0.001) that the odds (aOR=2.32; 95%CI:1.54 to 3.48) of refusing to provide blood sample for an HIV test for residents of KwaZulu-Natal province were much greater than for those who resided in the Western Cape. Those who resided in Mpumalanga and North-West followed in terms of the odds of refusing to provide blood sample for HIV test compared to the Western Cape province. The former had the second greater odds (aOR=2.05; 95%CI: 1.47 to 2.85) while the latter followed in third place (aOR=1.54; 95%CI: 1.04 to 2.29) and for both provinces, evidence was overwhelming (p<0.001) and strong (p=0.031) respectively.

When looking at the three geotypes, the results showed greater odds (aOR= 1.50; 95%CI: 1.14 to 1.96) of refusing to provide blood sample for HIV test for urban dwellers when compared to rural dwellers in tribal areas (informal) and the evidence was very strong (p=0.003). The difference was also noticed between the sexes whereby, males had greater odds (aOR=1.21; 95%CI: 1.16 to 1.28) of refusing to provide blood sample for an HIV test compared to females (p<0.001). Comparison of the racial groups showed overwhelming evidence (p<0.001) of greater odds (aOR=2.26; 95%CI: 1.80 to 2.83) of refusing to take an HIV test among the White racial group compared to the African racial group.

Changes were also noted when age groups were compared. Using the elderly (60 years +) participants as reference, there was overwhelming evidence (P<0.001) of greater odds (aOR=1.48; 95%CI: 1.22 to 1.81) of refusing to provide blood sample for an HIV test for age groups 35 to 39 years when compared to those 60 years old and older. The odds of refusing to provide blood sample for an HIV test were lower for younger age groups particularly those aged from zero to twenty-four years compared to those aged 25 years and older.

Table 4: The final regression model to predict refusal to give blood for an HIV test with fixed variables

Factor	Level	aOR	95% C.I.	P-value
Province	Western cape	1	Reference	-
	Eastern Cape	1.31	0.95 – 1.79	0.096
	Northern cape	1.14	0.83 – 1.55	0.442
	Free state	1.03	0.72 – 1.49	0.863
	KwaZulu-Natal	2.32	1.54 – 3.48	<0.001
	North-west	1.54	1.04 – 2.29	0.031
	Gauteng	1.45	1.07 – 1.96	0.017
	Mpumalanga	2.05	1.47 – 2.85	<0.001
	Limpopo	0.85	0.58 – 1.25	0.413
Geotype	Rural (Tribal)	1	Reference	-
	Urban	1.50	1.14 – 1.96	0.003
	Rural (Farm)	1.21	0.87 – 1.71	0.260
Sex	Female	1	Reference	1
	Male	1.21	1.16 – 1.28	<0.001
Race	African	1	Reference	-
	Coloured	1.29	0.99 – 1.67	0.055
	Indian	1.34	0.99 – 1.79	0.055
	White	2.26	1.80 – 2.83	<0.001
Age	60 + years	1	Reference	-
	0 -4 years	0.61	0.36 – 1.02	0.059
	5 - 9 years	0.30	0.17 – 0.51	<0.001
	10 - 14 years	0.19	0.11 – 0.34	<0.001
	15 - 19 years	0.72	0.60 – 0.86	<0.001
	20 – 24 years	0.89	0.74 – 1.07	0.227
	25 - 29 years	1.12	0.94 -1.34	0.213
	30 – 34 years	1.17	0.97 – 1.42	0.096
	35 – 39 years	1.48	1.22 – 1.81	<0.001
	40 – 44 years	1.27	1.06 – 1.52	0.008
	45 – 49 years	1.21	1.01 – 1.44	0.034
	50 – 54 years	1.31	1.05 – 1.64	0.016
	55-59 years	1.06	0.86 – 1.30	0.589

3.2.4.2. Behavioural variables

The association between refusing to provide blood sample for an HIV test and behavioural variables was investigated. A total of six behavioural variables were found to be associated with refusal to take an HIV test. All selected variables were significant at cut-off point ($p<0.05$). The behavioural factors associated with refusing to provide blood sample for HIV test are presented with sublevel categories in **Table 5** to allow comparison within each covariate.

Table 5: The final regression model to predict refusal to give blood for an HIV test with behavioural variables

Factor	Level	aOR	95% C.I.	P-value
Health status	Poor	1	Reference	-
	Excellent	1.85	1.46 – 2.35	<0.001
	Good	1.44	1.14 – 1.80	0.002
	Fair	1.19	0.95 – 1.49	0.125
	No response	0.11	0.03 – 0.40	0.001
Hospital admission (for illness)	Yes	1	Reference	-
	No	1.04	0.91 – 1.19	0.535
	No response	0.90	0.30 – 2.69	0.852
HIV test pre-counselling	Yes	1	Reference	-
	No	0.80	0.70 – 0.90	<0.001
	No response	1.39	1.26 – 1.55	<0.001
Hypertension diagnosed	Yes	1	Reference	-
	No	1.12	0.99 – 1.28	0.082
	Refused to answer	1.24	0.65 – 2.36	0.517
Feeling depressed	None of the time	1	Reference	-
	A little of the time	0.80	0.71 – 0.90	<0.001
	Some of the time	0.75	0.65 – 0.86	<0.001
	Most of the time	0.60	0.49 – 0.75	<0.001
	All the time	0.58	0.41 – 0.82	0.002
	No response	1.61	0.80 3.24	0.186
Away from place of residence	Yes	1	Reference	1
	No	1.08	0.97 – 1.21	0.167
	No response	1.50	0.98 – 2.30	0.061

There was a strong association between health status and refusing to provide blood sample for an HIV test. It is shown in **Table 5**, that those who rated their general health status to be “excellent” had greater odds (aOR=1.85; 95%CI: 1.46 to 2.35) of refusing to provide blood sample for an HIV test compared to those who

rated their general health to be poor and the evidence was overwhelming (p<0.001). Those who rated their health status to be “good” had greater odds (aOR=1.44; 95%CI: 1.14 to 1.80) of refusing to provide blood sample for an HIV test compared to those who rated theirs to be “poor”.

There was an association between receiving counselling before testing for HIV and refusing to provide blood sample for HIV test. Those who did **not** receive counselling before being tested for HIV previously had reduced odds (aOR=0.80; 95%CI: 0.70 to 0.90) of refusing to test for HIV compared to those who received counselling before taking an HIV test previously (p<0.001). Essentially, receiving counselling before testing for HIV is associated with greater odds of refusal to take an HIV test.

When adjusted for age, those who previously received HIV counselling had greater odds of refusing to test for HIV. Furthermore, the odds ratio of refusing to test for HIV for those who previously received counselling increased when adjusted for age. The odds ratio of refusing to test for HIV when one previously received counselling increased. Also, the odds of non-response increased with age until age groups 35 to 39 years. For example, from age groups 15 to 19 years the odds increased and peaked at age groups 35 to 39 years. The peak was followed by a gradual decline, thus the association between refusal to test for HIV and age assumed an inverted “U” shape. For age, only those aged 15 years and above were included in this analysis they only these age groups had the autonomy to decide about taking an HIV test or not.

There was some evidence (p=0.082) showing that those who were not diagnosed with hypertension had increased (aOR=1.12; 95%CI: 0.99 to 1.28) odds of refusing to take an HIV test. Conversely, those who were diagnosed with hypertension had less odds of refusing to provide blood sample for HIV test.

There was evidence of association between feeling depressed and refusing to provide blood sample for an HIV test. Those who were “depressed all the time” had almost half the odds (aOR=0.58; 95%CI: 0.41 to 0.82) of refusing to provide blood sample for an HIV test compared to those who did not feel depressed or those reported that they felt depressed “none of the time” (P<0.002). The odds of refusing to test for HIV decreased with the intensity of the depression as demonstrated by the

overwhelming evidence (P<0.001) in **Table 5** that those who felt depressed "A little of the time" had slightly greater odds (aOR=0.80; 95%CI: 0.71 to 0.90) of refusing to take an HIV test compared to those did not feel depressed. Supported by overwhelming evidence, the odds (aOR=0.60; 95%CI: 0.49 to 0.75) of refusing to provide blood sample for an HIV test reduced further for those who felt depressed "most of the time" compared to those who did not feel depressed (p<0.001).

Although without being supported by strong evidence, with regards to place of residence, those who have not been away from their current place of residence for at least a month, had greater odds (aOR=1.08; 95%CI: 0.97 to 1.21) of refusing to take an HIV test compared to those who have been away for at least a month in the past year (P=0.167).

3.3. HIV prevalence estimate

3.3.1. Determination of new behavioural weights

The above final model shown in both **Table 4 and Table 5** was used to derive **third weights** for estimating the prevalence of HIV in South Africa. The weights (***final_behav_wt)*** were used to re-calculate HIV prevalence estimates.

3.3.2. HIV prevalence estimates

When estimating the overall prevalence of HIV using the weights, a slightly lower HIV prevalence of 14.64% (SE=0.39, 95%CI=13.89 to 15.40) was obtained as shown in **Table 6**. This was lower compared to questionnaire sampling weights HIV prevalence estimates but greater than HIV sampling weights HIV prevalence estimates. The overall HIV prevalence estimate derived using the third weights are greater than the 14.01% estimated when using the HIV sampling weights are but lower than the 14.97% estimated using questionnaire weights. Using the new weights led to a slightly decreased standard errors and hence the narrower 95%CIs compared to when HIV sampling weights were used.

This observation of reduced SEs persisted when estimating HIV prevalence for other sub-categories except for age group greater than 60 years, urban geotype and most provinces, including Western Cape, KwaZulu-Natal and Gauteng. In a pattern like when HIV sample weights were used, across age groups, HIV prevalence

peaked in age group 35-39 years followed by age group 40-44 years. Overall, for all age groups, the prevalence pattern was similar when using the third weights compared to when using HIV sampling weights as age group 0-4 years had the lowest prevalence.

Table 6: HIV prevalence estimated from the third weights compared to estimates from HIV sampling weights.

		HIV sample data			Behavioural factors data		
Factors	**Sub-category**	**%Prev**	**SE**	**95%CI**	**% Prev.**	**SE**	**95%CI**
HIV prevalence	National	14.01	0.40	13.22-14.80	14.64	0.39	13.89 – 15.40
Sex	Male	10.79	0.44	9.93-11.65	10.65	0.48	9.72 – 11.59
	Female	17.09	0.59	15.94-18.23	17.56	0.48	16.62 – 18.49
Race	African	16.15	0.45	15.26-17.04	16.04	0.42	15.22 – 16.87
	Coloured	5.90	0.76	4.41-7.39	5.88	0.66	4.57 – 7.19
	Indian	2.43	0.95	0.57-4.29	2.59	1.03	0.58 – 4.61
	White	4.59	1.13	2.37-6.80	2.28	0.40	1.49 – 3.07
Geotype	Urban	13.03	0.52	12.02-14.05	13.32	0.58	13.18 – 15.45
	Rural informal	15.28	0.68	13.95-16.62	14.83	0.54	13.78 – 15.89
	Rural (farms)	17.77	1.31	15.20-20.34	16.58	1.21	14.21 – 18.96
Province	WC	8.61	1.16	6.33-10.88	10.70	1.64	7.48 – 13.91
	EC	15.91	1.56	12.84-18.98	15.86	1.00	13.91 – 17.82
	NC	8.21	0.75	6.75-9.67	8.04	0.78	6.51 – 9.56
	FS	17.01	1.25	14.55-19.47	15.82	1.19	13.47 – 18.16
	KZN	18.04	0.80	16.47-19.62	20.37	0.88	18.65 – 22.10
	NW	14.82	0.86	13.13-16.50	17.46	0.96	15.57 – 19.35
	GP	12.05	0.85	10.39-13.72	14.11	0.99	12.17 – 16.05
	MP	19.37	1.28	16.85-21.89	18.73	0.88	17.01 – 20.44
	LP	10.13	0.85	8.46-11.80	11.24	0.81	9.65 – 12.83
Age groups	0 – 4 years	3.09	0.55	2.01 – 4.16	3.50	0.80	1.92 – 5.08
	5 - 9 years	2.66	0.32	2.03 – 3.29	2.58	0.43	1.75 – 3.42
	10-14 years	2.69	0.32	2.07 – 3.31	2.34	0.43	1.75 – 3.42
	15-19 years	4.87	0.48	3.93-5.81	5.60	0.60	4.43 – 6.78
	20-24 years	10.54	0.82	8.94-12.14	11.52	0.82	9.91 – 13.13
	25-29 years	18.95	1.13	16.73-21.17	20.07	1.14	17.83 – 22.30
	30-34 years	26.45	1.54	23.44-29.47	29.14	1.65	25.91 – 32.37
	35-39 years	32.99	1.65	29.75-36.22	36.08	1.70	32.73 – 39.42
	40-44 years	29.31	1.58	26.22-32.40	34.15	1.72	30.77 – 37.53
	45-49 years	26.95	1.95	23.12-30.78	30.61	1.88	26.93 – 34.29
	50-54 years	21.31	1.77	17.84-24.78	22.99	1.71	19.64 – 26.35
	55-59 years	15.01	1.64	11.79-18.22	16.59	1.66	13.32 – 19.85
	60 years +	6.60	0.68	5.27-7.94	7.79	0.71	6.38 – 9.19

Using the **third weights** also showed that among all provinces in South Africa, KwaZulu-Natal Province recorded the highest HIV prevalence (20.37%, SE=0.88; 95%CI:18.65 to 22.10) of HIV in South Africa. This observation was different to the HIV prevalence estimates recorded when using the HIV sampling weights which recorded Mpumalanga province to have the highest HIV prevalence estimates (19.37%, SE=1.28; 95%CI: 16.85-21.89) among all provinces in South Africa. The **third weights** also agreed with HIV sampling weights regarding the province with the least HIV prevalence and showed that the Northern Cape province had the lowest HIV prevalence with 8.04% (SE=0.78; 95%CI 6.51 to 9.56).

Rural farming areas continued to have the highest HIV prevalence (16.58%, SE=1.21; 95%CI:14.21 to 18.96). On the other hand, the lowest HIV prevalence (14.83%; SE=0.54 95%CI: 13.78 to 15.89) was estimated for urban areas.

A similar pattern was observed for both HIV sampling weights and the **third weights** regarding sexes. Using estimates derived from the third weights showed that more females compared to males were infected with HIV as 17.56% (SE=0.48; 95%CI: 16.62 to 18.49) of women were infected with HIV compared to the 10.65% (SE=0.48 ;95%CI: 9.72 to 11.59) HIV prevalence in males. The HIV prevalence estimates for the sex category also showed slightly decreased SEs for females on HIV prevalence estimates derived from the **third weights** compared to prevalence estimates derived from HIV sample data derived weights. In general, HIV prevalence estimates decreased, standard error became less, and confidence intervals narrowed after incorporation of the **third weights**. It should be noted that the difference between the two estimates were modest.

CHAPTER 4

4. DISCUSSION

This study sought to first compare HIV prevalence estimates determined using questionnaire sampling weights to HIV prevalence estimates determined using HIV sampling weights from the SABSSM V survey dataset. Secondly, it investigated factors associated with non-response, which here meant failure to provide a blood sample for HIV testing. The third objective was to determine prevalence estimates using third weights derived from behavioural factors associated with refusal to take an HIV test. For the first objective, the effect of weights that were being investigated were the questionnaire sampling weights (***ibreal12_combined***) and the HIV data sampling weights (***ibreal1_combine***). The questionnaire sampling weights were based on those who answered the questionnaires which included questions on demographic information. The HIV specimen weights were based on the HIV test results deduced from HIV testing after agreeing to give a blood sample.

Comparison of HIV prevalence estimates derived through HIV sampling weights and questionnaire sampling weights revealed a modest difference between the two estimates. The overall HIV prevalence estimates were lower when using the HIV sampling weights compared to the questionnaire sampling weights. Using the questionnaire weights, the results showed that the overall HIV prevalence was estimated to be 14.97%. However, when using the HIV specimen weights, the prevalence reduced to 14.01%. This shows that the choice of weights does have an influence on HIV prevalence estimates and estimates do change when using different weights. This difference was notable across all strata, particularly those with relatively small population size such as age group ***60 years +*** and the White race group.

Investigating factors associated with refusing to provide blood sample for HIV test or non-response revealed eleven (11) explanatory variables (including the five fixed variables). The final logistic regression model was assessed to determine if assumptions for logistic regression are met. Overall, the model satisfied assumptions for logistic regression except for the **Hosmer-Lemeshow** goodness of fit test and omission of variables. When checking for omission of relevant variables it was found that relevant variables were omitted. This poses a challenge considering the

substantial number of variables in the dataset. However, composite variables were created as an attempt to look for relevant variables omitted, however, this could not be resolved. The number of variables to select from also posed a challenge. There was a non-response for some variables which could possibly affect the model. This could be due to some questions being not applicable to certain age groups. Multicollinearity was possible and this could be due to a similar proportion of missing observations for some variables. Some were recoded as non-responses, but the challenge remained. However, none of variables had conclusive evidence of multicollinearity. When the regression model was checked with ROC, the model performance was satisfactory.

When focusing on provinces, KwaZulu-Natal had the greatest odds of refusing to take an HIV test compared to the Western Cape and it was followed by Mpumalanga. It should be noted that these two provinces (KwaZulu-Natal and Mpumalanga) had the highest HIV prevalence in the country. Furthermore, the Western Cape province had the lowest HIV prevalence compared to all the other provinces. The odds of refusal to take an HIV test were greater in all the eight provinces when compared to the Western Cape. However, this observation of people infected with HIV refusing to take an HIV test was observed elsewhere. It was previously found that HIV positive participants were more likely to refuse participation in surveys than HIV-uninfected participants.[11] Perhaps people who are infected in provinces with high HIV prevalence do not see the purpose of testing as already they know the HIV status.

Despite being more impacted by HIV infections than their male counterparts, females seem to be more willing to take an HIV test than males. The results showed that the odds of refusing to take HIV test were lower in females compared to males. This is despite constituting the biggest proportion of infections. However, the health seeking behaviour of females is different from that of men, partly due to greater concern with reproductive health. They would visit health facilities more frequently and this could also influence their decision to test. Visiting health facilities could give them the opportunity to gain more knowledge about HIV. Knowledge about HIV is associated with increased chances of taking an HIV test.[14]

For geotype, the findings are similar to observations on females. Despite having the highest proportion of people infected with HIV among all geotypes, rural informal dwellers living in tribal land had reduced odds of refusing to test for HIV. Living in a rural area where illiteracy rates are commonly high in among rural communities may increase likelihood being illiterate. Being less literate, rural dwellers might find it more difficult to refuse a request to take an HIV test. The element of high poverty prevalence among rural dwellers would be expected to decrease the odds of refusing to take an HIV test.[14] In rural areas of SA, the odds of refusing to take an HIV test are low. Resources are limited in rural places and unemployment is high due to lack of industries which are a common feature in urban areas. Also, for many rural communities, there are no alternatives to public health facilities. Therefore, rural dwellers would likely depend solely on public health facilities for their health care needs. Mishra et.al. found that factors associated with non-response include socio-economic factors such as education and wealth.[10] Majority of rural dwellers commonly have a low socio-economic status and will likely be keen to agree to an HIV test. Perhaps they would also feel compelled to agree to take HIV test as they may feel this would be to their benefit especially seeing those who are infected getting the treatment and care.

Compared to Africans, White racial group had greater odds of refusing to take an HIV test. This could be due to the history of SA whereby policies of the previous government resulted in Whites enjoying a higher socio-economic status. Studies have reported that those with a higher socio-economic status are more likely to refuse to take an HIV test.[10] This could explain the greater odds of refusal among Whites as compared to Africans.

The results showed an association between health status and refusal to take an HIV test. Those who rated their health status as "poor" had lower odds of refusing to agree to test for HIV compared to those who rated their health status as excellent. This could be due to the perceived risk of HIV infection. Those who feel less healthy may be keen to know more about their health. It has been reported that those who are at an elevated risk of infection are reported to be keen to test for HIV.[13] It is expected that due to the apparent risk of HIV infection associated with being unwell could drive people to seek an HIV test so that they know what could be cause of their illness.

The same could be said about feeling of depression. The results revealed an association between depression and refusing to provide blood sample for an HIV test. Those with less depression had greater odds of refusing to provide blood sample for an HIV test and those who were more depressed were keener to agree to an HIV test. It is possible that some of the depression could be linked to anxiety about possible HIV infection. The solution for such individuals could be to test for HIV. Other possible reason why those with depression are willing to take an HIV test may be that they are less assertive and are unable to refuse.

It seems for people who are depressed, fear is not likely to influence their decision as it appears that the intensity of stress is indirectly related to refusal to take an HIV test. HIV has always been associated with stigma and discrimination.[27] Without counselling, it may be difficult to overcome fear of being discriminated against and one may not have the courage to take the HIV test. Therefore, another possible explanation could be that, perhaps those who feel depressed get more counselling which could provide them with much needed information to take better decision such as taking an HIV test. Counselling may have the benefits of overcoming stigma and empowering to face discrimination.

A confusing association was for those who received counselling before undergoing their previous HIV test and refusal to take HIV test. For this group, those who did not receive counselling before their previous HIV test seem to be keen to take an HIV test. One would expect those who previously received counselling to be keen to take HIV test because, knowledge about HIV is strongly associated with HIV testing.[28] Those have received counselling would be expected to have a better understanding of the disease and the benefits of taking the HIV test. It is possible that those who have already been counselled before testing, know their HIV results and may not see a need of doing another HIV test. They perceive testing again as a waste of resources or an unnecessary inconvenience.

It was expected that those who previously received counselling would take an HIV test, but it doesn't seem to have the expected effect. Therefore, the negative association could be due to some participants not seeing the need to test for HIV again as they already know their HIV status. The previous counselling received could have been good, but it seems to fail to influence the decision to test in the current

study. In addition, those who previously tested positive for HIV, could be reluctant to retests fearing stigmatization. It is possible that those who had been counselled would recently have had an HIV test. They would have been counselled before their recent HIV test and this would be the reason why they refused another test.

The odds of refusal seem to peak at middle ages and this corelates with another study conducted in Ethiopia.[29] The odds of refusal to test for HIV seem to decrease on both ends of the age groups, for example, among the young and the eldest. Some of the reasons could be that these age groups may not be assertive enough compared to the middle age group. The younger groups who may not yet know much about their rights may be less likely to refuse. Also, these age group(young) may not consider themselves to be at risk of being HIV infected and may be keen to test at younger ages. Those older than 39 years could also be less assertive due to being more dependent on the health system. Additionally, the older groups could be already familiar with health systems as the older commonly visit health facilities more frequently due to the onset of chronic conditions such as diabetes and hypertension. These may lead to them being keener to accept other health services available to them. Considering the above observation where counselling fails to influence testing for HIV, it is imperative that innovative ways of encouraging participants to test should be investigated

Those who have not been away from their place of residence for at least one month had increased odds of refusing to provide blood sample for an HIV test. How ever, this was not statistically significant. If this association is true, it might have been because they do not spend time away from their communities and may be more fearful of the discrimination, they would receive from communities should they test positive. Fear of stigma could influence their decision to refuse testing. Fear could result from stigma because, HIV has always been associated with stigma and discrimination.[27] It may be difficult to overcome fear of being discriminated against and one may not have the courage to take the HIV test. Therefore, people at risk may be fearful that knowing one's status if infected could lead to stigmatisation from their communities. It is possible that stigma could also play a role in the association between being away from place of residence for at least a month and refusal to take HIV test.

The results reported that the odds of refusing to test for HIV were greater for those who were **not** diagnosed with hypertension. Conversely, those diagnosed with HTN had greater odds of agreeing to take an HIV test. People diagnosed with HTN are often put on chronic medication and usually visit health facilities regularly. Their familiarity with health services and their contact with health practitioners may influence their decision to agreeing to test for HIV. Agreeing to test for HIV could also be due to the knowledge and information that they would receive when they go to a hospital which may help them to understand the importance of taking tests such as HIV tests. This observation would be in line with a study conducted in East Africa which reported that women who visited a health facility within a year had greater odds of consenting to an HIV test than those who have not visited a clinic.[14] Being diagnosed with HTN may also lead to self-perception of elevated health risk and such people may seek to know where they stand on health matters, including HIV.

Based on the results of the logistic regression model, it does appear that risk of infection and familiarity with HIV may influence the decision to take an HIV test or refusal to test for HIV. Stigma could also affect some of the factors.

The HIV prevalence was re-estimated, after using the **third weights**. Estimations revealed a slightly greater prevalence estimate of 14.64% compared to the previous prevalence estimates derived using HIV data sample weights (14.01%) as reported by the SABSSM V survey. An HIV prevalence of 14.64% is equivalent to 0.63% more people infected with HIV than reported in the SABSSM V survey.

An increase of 0.63% translates into 355 509 more people being HIV positive. This number is substantial in a country that at the time of publication of the results from SABSSM V survey, its population size was estimated to be 56.43 million.[5] This number can be important in terms of its impact on planning with regards to human resources and therapeutic agents' availability. It may be argued that in a country with a population of more than 55 million, it might not significantly impact planning and resource mobilization. However, it can influence attaining health targets such as the 95-95-95 target. A prevalence of 14.01% translates into 7.91 million people infected. The 355 509 more people infected is equivalent to 4.49 % of the total 7.91 million people infected. This may theoretically negatively impact achieving the 95-95-95 targets. Adjusting for this third estimates should be considered if it leads to the

reduction of bias when estimating HIV prevalence. Particularly, because using the third weights led to smaller SEs. This would be crucial if the proportion of non-response persists as a greater level of non-response may introduce biased estimates, so we need to consider measures to reduce bias.

Using the **third weights**, the 95%CI of the overall HIV prevalence estimate slightly narrowed compared to HIV prevalence estimates attained using HIV data sample weights. The narrow confidence interval found may reduce uncertainty as the number of people infected with HIV may vary less. Such small differences in confidence intervals may be expected when using different weights.

In summary, application of additional behavioural weights has slightly impacted HIV prevalence estimates by moderately reducing the HIV prevalence estimates. The change in HIV prevalence estimates was marginal and would arguably not affect programs that are aimed at reducing HIV prevalence considering HIV is a disease that is targeted for eradication. The change in this HIV prevalence estimates may not massively impact the targets set for HIV but can potentially extent affect strategy aimed at eradicating HIV. This study has demonstrated that factoring behavioural predictors of non-response could affect HIV prevalence estimates. Non-response remains an issue that needs to be looked in light of the high response rate attained in some countries across the continent.[12] As long as the non-response remains high, more strategies such as the one explored in this study need to be investigated to minimize bias.

CHAPTER 5

5. CONCLUSION

The study demonstrated the effect of inverse probability weights on HIV prevalence estimate. Sampling weights do have an effect and can reduce bias. The effect of sampling weights also noticeable on SEs and 95%CI. This study also explored factors associated with non-response with respect to refusing to provide blood sample for an HIV test. The study managed to explore another method of potentially reducing bias by factoring behavioural factors that may influence non-response. It is important to explore other ways of reducing bias in future studies by exploring measures to also reduce non-response. Additionally, to encourage participation and reduce non-response, innovative ways of encouraging participants to test should be investigated. However, more exploration of MAR and addition of other weights can also be considered as another way in reducing bias in research. This study was able to also demonstrate that non-response is not entirely at random. Future studies may probe whether non-response is also affected by HIV status.

5.1. Limitations

Some variables may have not been included in the dataset, and these variables may influence results. Other limitations of this study are shared with the primary study such as recall bias and social desirability bias. Some variables had high proportion of missing observation, and this can affect selection of predictors and impact on the weights to be used. Another limitation is that participants who refused to answer the questionnaire could not be included in the study. Since this involved a small proportion of sampled individuals. The analysis showed that adjusting the weights proportional to the probability of missing values had a small impact. Therefore, the 4.1% non-responders are unlikely to have an impact.

REFERENCES

1. HIV and AIDS [Internet]. WHO. 2023 [cited 2024 January 15]. Available from: https://www.who.int/news-room/fact-sheets/detail/hiv-aids.
2. The sixth South African National HIV prevalence, incidence and behavior survey (SABSSM VI) [press release]. 2023.
3. Greenberg M, Weiner M. Keeping surveys valid, reliable, and useful: a tutorial. Risk Anal. 2014;34(8):1362-75.
4. Boerma J, Ghys P, Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. Lancet. 2003;362(9399):1929-31.
5. Simbayi L, Zuma K, Zungu N, Moyo S, Marinda E, Jooste S, et al. South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017. South Africa: Human Sciences Research Council; 2019.
6. Nkambule R, Philip N, Reid G, Mnisi Z, Nuwagaba-Biribonwoha H, Ao T, et al. HIV incidence, viremia, and the national response in Eswatini: Two sequential population-based surveys. PLoS One. 2021;16(12).
7. Thina K, Frederixb K, McCrackenc S, Letsiea M, Lowd A, Patelc H, et al. Progress toward HIV epidemic control in Lesotho. AIDS. 2019;33(15).
8. Litewka SG, Heitman E. Latin American healthcare systems in times of pandemic. Dev World Bioeth. 2020;20(2):69-73.
9. Bajunirwe F, Akakimpa D, Tumwebaze FP, Abongomera G, Mugyenyi PN, Kityo CM. Persistence of traditional and emergence of new structural drivers and factors for the HIV epidemic in rural Uganda; A qualitative study. PLoS One. 2019;14(11):e0211084.
10. Mishra V, Barrere B, Hong R, Khan S. Evaluation of bias in HIV seroprevalence estimates from national household surveys. Sex Transm Infect. 2008;84 Suppl 1:i63-i70.
11. Floyd S, Molesworth A, Dube A, Crampin AC, Houben R, Chihana M, et al. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. AIDS. 2013;27(2):233-42.
12. Mishra V, Rathavuth H, Khan S, Gu Y, Liu L, Rathavuth H, et al. Evaluating HIV Estimates from National Population-Based Surveys for Bias Resulting from Non-Response. Calverton, Maryland, USA: Macro International Inc: DHS Analytical studies; 2008.
13. Thomas P, Voetsch A, Song B, Calloway D, Goode C, Mundey L, et al. HIV Risk rehaviors and testing history in historically black college and university settings. Public Health Reports. 2008;123:115-25.
14. Worku M, Tesema G, Teshale A. Prevalence and associated factors of HIV testing among reproductive-age women in eastern Africa: multilevel analysis of demographic and health surveys. BMC Public Health. 2021;21(1262).
15. Shih H, Ko N, Hsu H, Wu C, Huang C, Lee H, et al. Rapid Human Immunodeficiency Virus screening in an emergency department in a region with low HIV seroprevalence. Jpn J Infect Dis. 2015;68:305-11.
16. McGovern ME, Marra G, Radice R, Canning D, Newell ML, Barnighausen T. Adjusting HIV prevalence estimates for non-participation: an application to demographic surveillance. J Int AIDS Soc. 2015;18:19954.
17. Fielding S, Fayers P, Ramsay C. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. Health Qual. 2009;7(57).
18. Korn E, Graubard B. Examples of differing weighted and unweighted estimates from a sample survey. Am Stat. 1995;49(3):291-5.
19. Van der Heyden J, De Bacquer D, Gisle L, Demarest S, Charafeddine R, Drieskens S, et al. Additional weighting for education affects estimates from a National Health Interview Survey. Eur J Public Health. 2017;27(5):892-7.

20. Arushanyan A, Kouyumjyan K, Martirosova D, Mushtaq A, Scheuren F. Sustaining improved estimates of poverty in Armenia: An application of multivariate raking ratio adjustment. Statistical journal of the IAOS. 2014;30(4).
21. Graham G, Flores-Cervantes I. Weighting methods. JOS. 2003;19(2):81-7.
22. Pike G. Using weighting adjustments to compensate for survey nonresponse. Res High Educ. 2008(49):153-71.
23. Downes M, Carlin JB. Multilevel Regression and Poststratification Versus Survey Sample Weighting for Estimating Population Quantities in Large Population Health Studies. Am J Epidemiol. 2020;189(7):717-25.
24. Hosmer DW LS. Applied Logistic Regression. 3rd ed ed. Chicester: Wiley; 2013 2013.
25. Habshah M, Syaiba B. Modified standardized Pearson residual for the identification of outliers in logistic regression model. Journal of Applied Sciences. 2013;13(6):828-36.
26. Pregibon D. Logistic regression statistics. The Annals of Statistics.9(4):705-24.
27. Genberg B, Kawichai S, Chingongo A, Sendah M, Chariyalertsak S, Konda K, et al. Assessing HIV/AIDS stigma and discrimination in developing countries. AIDS Behav. 2008;12:772-80.
28. Alem A, Liyew A, Guadie H. Spatial pattern and associated factors of HIV testing and counselling among youths (15–24 years) in Ethiopia. BMC Public Health. 2021;21(1).
29. Reniers G AT, Berhane Y, Davey G, Sanders EJ. Disclosure of HIV test results and non-response bias in seroprevalence surveys. Population program working papers (POP2007-05). 2007.

Appendix 1: Plagiarism declaration report

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I MAAKE MORAKA EPHRAIM (Student number: 2332653) am a student registered for the degree of MASTER OF SCIENCE (EPIDEMIOLOGY) in the academic year 2024.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:  Date: 28-03-2024

1

Appendix 2: Ethical certificate-HREC(Wits)

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

DECLARATION:

Adherence to HREC (Medical) Ethics Application Terms and Conditions

I, the undersigned, hereby declare that I have not collected data/ done secondary data analysis or any other form of research, prior to obtaining clearance certificate from the HREC (Medical) for study no: M22-10-134

I have read and understood the terms and conditions on page 8-9 of the HREC (Medical) application form. I confirm that it is my responsibility to ensure that I have received final HREC (Medical) Clearance before commencing any research.

MORAKA EPHRAEM MSAKE

Name, Surname and Signature

Student/Staff no if applicable: 2332653

Date: 12 APRIL 2023

Name, Surname and Signature

Supervisor (if applicable)

Date:

Appendix 3: Ethical certificate (HSRC)

HSRC
Human Sciences Research Council

Research Coordination, Ethics and Integrity (ReCEI)

RESEARCH ETHICS COMMITTEE ADMINISTRATION

Room 1345 - HSRC Building
134 Pretorius Street, Pretoria
Gauteng, South Africa
Tel: 27 12 3022012 - Fax: 27 12 3022005
Email: ksithole@hsrc.ac.za - Website: research.ethics@hsrc.ac.za
REC toll free no 0800 212 123

30 June 2016

To: Prof Leickness Simbayi
Deputy Chief Executive Officer for Research
Human Sciences Research Council
Private Bag X41
Pretoria,
0001
South Africa

Dear Prof Simbayi

Ethics Clearance of HSRC Research Ethics Committee Protocol No REC 4/18/11/15: THE FIFTH SOUTH AFRICAN NATIONAL HIV PREVALENCE, INCIDENCE AND BEHAVIOUR SURVEY, 2016

Thank you for your application for ethics approval of the above study. This was considered by the Research Ethics Committee at its meeting on 15 November 2015.

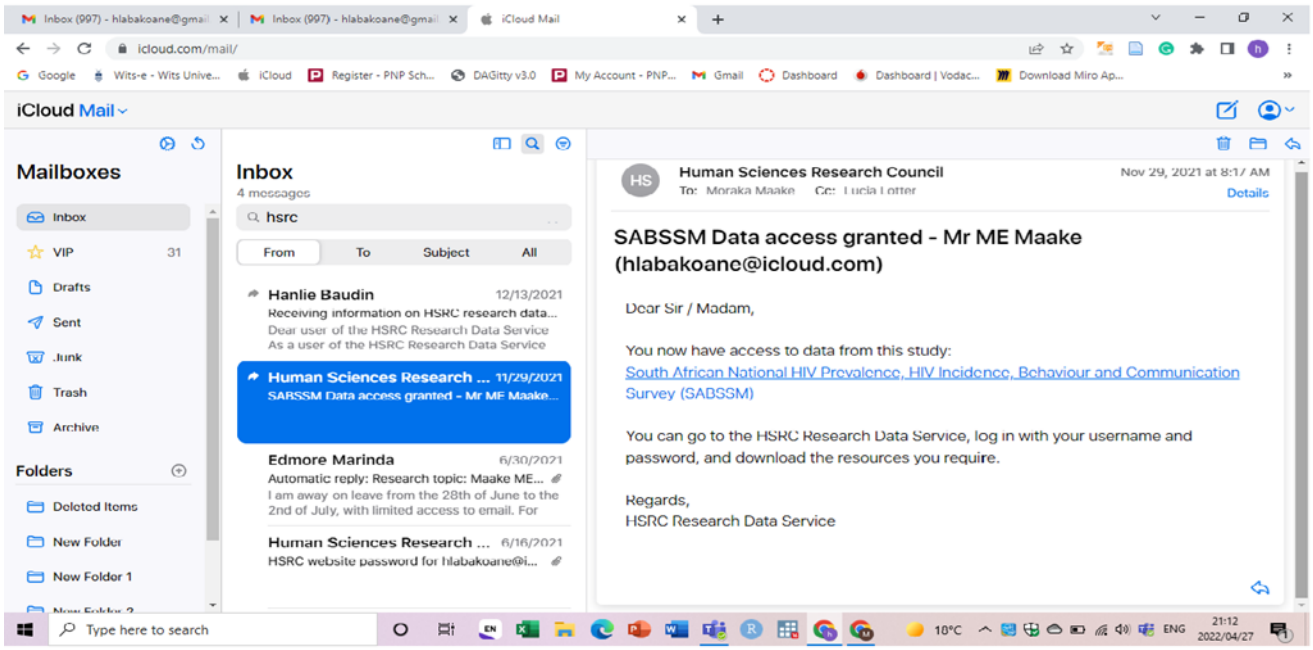
The study was provisionally approved pending appropriate responses to queries raised. Your responses dated 28 June 2016 to the queries raised on 10 December 2015 have been noted by a sub-committee of the Research Ethics Committee.

The conditions have now been met and the study is given full ethics Approval and may begin as from 30 June 2016.

Principal Investigator: Prof Leickness Simbayi

Organisation: Human Sciences Research Council

Appendix 4: Permission access the dataset



FWA number: 0000 6347

IRB number: 0000 3962

Cooperative Agreement name: Support the strengthening and utilization of strategic information activities in the Republic of South Africa under the President's Emergency Plan for AIDS Relief (PEPFAR)

Cooperative Agreement number: 1U2GGH001629-02

Protocol title: The Fifth South African National HIV Prevalence, Incidence and Behaviour Survey, 2016 (SABSSM V)

Protocol version: 3.1

Protocol date: 06 November 2015

Protocol approval date: 30 June 2016

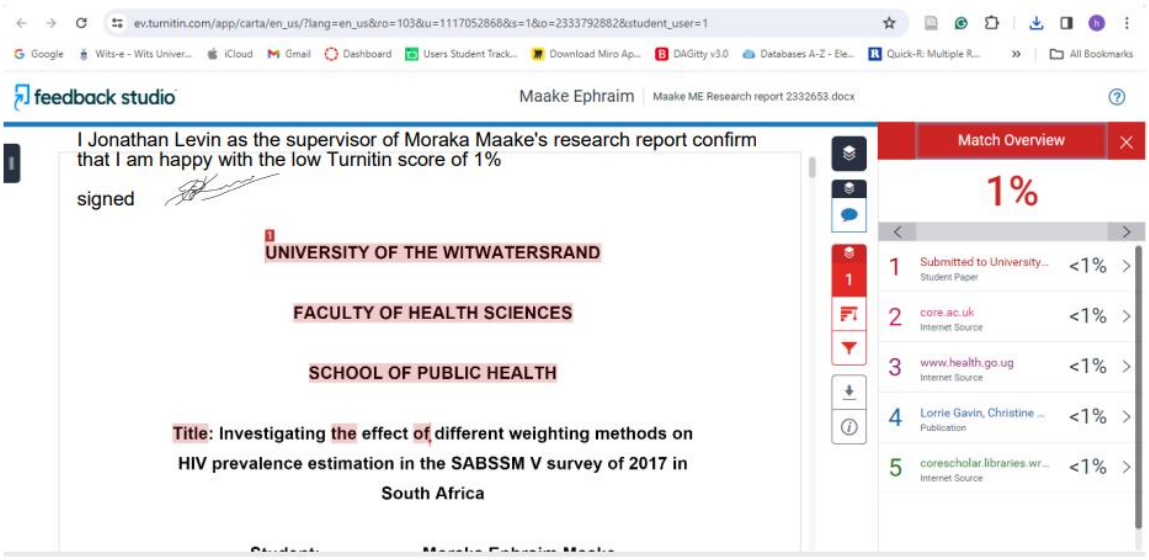
Date of expiry of protocol approval: 30 June 2017

The Committee wishes you success in your research.

Yours sincerely,

Prof. D R Wassenaar PhD
Chairperson: HSRC REC

Appendix 5: Turnitin report



Turnitin - Originality Report -