BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF STATISTICS AND MATHEMATICS

TOPIC

Statistical Analysis Of Reported Tuberculosis Cases In Zimbabwe: A Case Study Ofmakoni District , Zimbabwe

BY

ALEX NYAKWIMA

(B1953934)

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE BACHELOR OF SCIENCE HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS.

(DECEMBER 2023)

SUPERVISOR: MRS. HLUPO

APPROVAL FORM

1. TO BE COMPLETED BY THE STUDENT

I certify that this dissertation meets the preparation guidelines as presented in the Faculty guide

and instruction for the presentation of the dissertation.

A:Nuak…………………………… **31/11/2023**

(Signature of Student) (Date)

2. TO BE COMPLETED BY THE SUPERVISOR

This dissertation is suitable for submission to the Faculty. It has been checked for conformity with the Faculty guidelines.

31/11/2023

(Signature of Supervisor) (Date)

3. T0 BE COMPLETED BY THE CHAIRMAN OF THE DEPARTMENT

I certify, to the best of my knowledge that the required procedures have been followed and the preparation criteria have been met for this dissertation

Dr Magodora 31/11/2023

(Date)

(Chair person's signature)

RELEASE FORM

REGISTRATION NUMBER: B1953934

DISSERTATION TITLE: STATISTICAL ANALYSIS OF REPORTED TUBERCULOSIS CASES IN ZIMBABWE; (CASE STUDY MAKONI DISTRICT ZIMBABWE.)

YEAR GRANTED: 2023

Permission is granted to Bindura University of Science Education Library and the Department of Statistics and Financial Mathematics to produce copies of this dissertation whenever it deems necessary for academic use only.

Signature of Student A.……………………………

Date Signed **31/11/2023**

Permanent Address:

House Number 2743

Mabvazuva Infills

Rusape,

Zimbabwe

DEDICATION

This research is dedicated to my mother, Mrs E. Nyakwima and my siblings for the unwavering support they gave throughout the academic time.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my supervisor, Mrs. Hlupo, for her valuable academic guidance and unwavering support throughout the completion of this project. Her commitment and patience have significantly contributed to its success. I would also like to extend my thanks to the lecturers Mr. Kanjodo, Mr. Kusotera, Mr. Mukonoweshuro Mrs. Pagan'a and Dr Magodora in the Statistics and Mathematics department at Bindura University, as well as my fellow students, for their support and assistance during my studies. Lastly, I am deeply indebted to my mother for her unwavering support, both financially and otherwise. Without her, this endeavor would not have been possible.

First and foremost, I express my deepest gratitude to the Almighty God, the Supreme Being, the one who holds ultimate power and authority. It is through the strength, guidance, and abilities bestowed upon me by God that I was able to successfully carry out my research report and navigate through the entire journey.

ABSTRACT

This research was carried out to statistically analyze and model Tuberculosis prevalence with relation to the determinants of Tuberculosis. STATA version 13.2 was used to fit a binary logistic regression model in which five independent variables were confirmed with 85.13% accuracy of predicting Tuberculosis prevalence. The researcher used secondary data to answer the research topic with 4,007 observations. The dependent variable is Tuberculosis result, dichotomous in nature, that is either TB negative or TB positive and the explanatory variables included location where client lives, age group, HIV status, mode of test used, and TB risk group. The research results discovered that all the explanatory variables have different significant to tuberculosis prevalence. The research also reviewed that there is a strong relationship between HIV status, type of test used and TB prevalence. It was recommended that the Ministry of Health must ensure that individuals in rural areas of Makoni have adequate access to healthcare facilities and services can help in early detection, diagnosis, and treatment of Tuberculosis cases government put in a place some measures, policies and strategies that would help in mitigating Tuberculosis prevalence. These include anti-stigmatization campaigns, contact tracing, and to improving data sharing between ministry of health and NGOs that are into the TB program.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

1.0 INTRODUCTION

Tuberculosis (TB) is a leading cause of death in Zimbabwe, resulting in 4,600 reported fatalities in 2019. The country has faced multiple economic crises over the past ten years, leading to a weakened economy, job loss, and increased poverty. The healthcare system has also been significantly affected by more than a decade of economic decline and rising expenses. Despite these challenges, Zimbabwe is striving to decrease the TB mortality rate to below 5% of all reported cases. The term "tuberculosis" is derived from the Latin word for "nodule" or any protruding structure. TB is recognized as a global pandemic, primarily affecting countries with high incidence rates, including Zimbabwe.

The dissertation consist of five chapters. The researcher firstly provides a global overview of Tuberculosis in chapter one including, study's background, a problem description, goals, objectives, the applicability of the study, and any restrictions. The next chapter, Chapter Two, reviews related literature. Chapter three deals with the research methodology, this chapter also gives an outline of the research methodology and design including the research design, the population, and the sample, sampling procedures, research instruments, data collection procedures, presentation of the data, and the analysis procedures, as well as a summary. Chapter Four presents analyses, interprets, and discusses data. Chapter five gives the summary of the whole project, conclusion, and recommendations.

1.1 BACKGROUND OF THE STUDY

Tuberculosis is a significant public health issue, causing around 10 million new cases and 1.5 million deaths yearly. Zimbabwe is a priority country for TB, as it experiences approximately 90,000 new cases annually and has an 11% TB/HIV coinfection rate. In 2014, the World Health Organization established the End Tuberculosis Strategy, aiming to decrease TB cases and deaths by 90% and 95% respectively by 2035 and minimize the economic impact on affected families. The United Nations also set a goal in 2015 to reduce TB-related deaths by 90% by 2030.

Zimbabwe's National Tuberculosis Control Program has implemented several strategies to manage the disease, leading to a decrease in incidence and mortality. However, there are still many challenges to address, including TB in prisons, TB/HIV coinfection, drug-resistant TB, comorbidities, high treatment abandonment rates, low adherence to treatment, low contact evaluation, and low coverage of rapid molecular diagnosis.

The success of global and national TB control and elimination strategies relies on professionals who generate, evaluate, and apply scientific knowledge correctly. In Zimbabwe, doctors and masters are trained within the National Postgraduate System, with Postgraduate Programs accredited and evaluated by the Coordination of Higher-level Personnel Improvement (CAPES). Understanding the quantitative evolution of theses and dissertations on human TB, as well as their spatiotemporal, thematic, and institutional distribution, can aid in creating strategic plans for professionals in this field. This study emphasizes the critical role of theses and dissertations in national scientific disclosures.

.

1.2 STATEMENT OF THE PROBLEM

Tuberculosis (TB) is a major public health problem in Zimbabwe, with high rates of incidence and mortality. Makoni district, located in the eastern part of Zimbabwe, has been identified as one of the high-burden areas for TB. Despite efforts to control TB in the district, the prevalence and incidence of the disease remain high, and treatment outcomes are suboptimal. There is a lack of comprehensive statistical data on TB in the Makoni district, which hinders the development of evidence-based interventions to address the problem. Therefore, there is a need for a statistical analysis of TB in Makoni district to better understand the burden, distribution, and determinants of the disease, as well as to identify gaps in the current control program. This study aims to fill this gap by conducting a comprehensive statistical analysis of TB in the Makoni district, which informs the development of targeted interventions to improve TB control and reduce the burden of the disease in the district.

1.3 OBJECTIVES OF THE STUDY

The overall objective of this research is *to analyze the prevalence of tuberculosis in Zimbabwe.*

The specific objectives of this research are as follows:

1. To identify the demographic factors associated with tuberculosis incidence, (thus age, gender, occupation, and socio-economic status).
2. To determine the impact of HIV co-infection on tuberculosis incidence and treatment outcomes.

1.4 RESEARCH QUESTIONS.

The following sub-problems had to be answered to answer the main problem:

1. How does the incidence of tuberculosis vary by demographic factors such as age, gender, occupation, and socio-economic status?
2. What is the prevalence of HIV co-infection among tuberculosis patients, how does it affect treatment outcomes?

1.5 ASSUMPTION OF THE STUDY

1 The economic, political, and cultural environment would remain consistent until the end of the research.

1.6 SIGNIFICANCE OF THE STUDY

The research conducted at the University (BUSE) will contribute to the existing knowledge base of the institution and serve as a foundation for future exploration in the field. It may also attract attention from the industry being studied, raising the institution's profile as a centre of academic excellence.

The research findings may also be utilized by statistical agencies to align their processes and procedures with the recommendations proposed in the study, if necessary.

1.7 LIMITATIONS OF THE STUDY

The research aims to examine whether the Tuberculosis prevalence rates are affected by demographic factors and HIV status with a primary focus on Makoni district of Zimbabwe. The investigation will utilize data obtained from Development Aid from People to People Total Control of Tuberculosis (DAPP TC TB). The research will be conducted within a specific time frame of five years, spanning from 2018 to 2022

1.8 DELIMITATION OF THE STUDY

During the research, the researcher encountered several hurdles that prevented them from attaining the desired level of conclusiveness. Readers are advised to exercise caution while reviewing this document, considering the limitations outlined below:

1. Time constraint: The researcher faced limited time availability to conduct the study, necessitating working extensively during both day and night hours.
2. Imperfect access to required information: The researcher experienced challenges in accessing certain data due to its sensitive nature, posing difficulties in obtaining the necessary information for the study.

Despite these limitations, the researcher aims to make the best use of the available resources and information.

1.9 DEFINITION OF TERMS

The following terms used in this project should be understood within the context of this study:

TB – Tuberculosis is a contagious bacterial illness that arises from Mycobacterium tuberculosis (MTB). This bacterium is transmitted among individuals through the respiratory pathway and primarily targets the lungs, although it has the potential to harm other tissues as well. (Murray, 2020).

Statistical- is the practice or science of collecting and analyzing numerical data in large quantities, especially for inferring proportions in a whole from those in a representative sample. (Fisher. 1915)

Health- the state of being free from illness or injury concerning a person’s mental or physical condition. (Expenditure, 2021)

1.10 CHAPTER SUMMARY AND ORGANIZATION OF THE PROJECT.

This chapter discusses the background to the study, the statement of the problem, the purpose, and objectives of the study, research questions/ sub-problems, the significance of the study, research assumptions, and definition of terms, delimitation, and limitations of the study, before chapter summary. This gave direction to the rest of the project.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

Emphasized in this chapter are the theoretical reviews and previous research studies that are related to this study. This chapter showed several contributions by many writers as well as analyses' of Tuberculosis infection rates and Tuberculosis death rates. The chapter brought to light and reviewed their views, and recommendations and fundamentally assessed and interpreted contributions towards the subject under review. Previous and current information was used concerning this chapter to reach and achieve reliable research.

2.2 TUBERCULOSIS

According to the World Health Organisation (WHO), Tuberculosis (TB) is caused by bacteria (Mycobacterium tuberculosis) that most often affect the lungs. Tuberculosis is a preventable communicable disease and ranks amongst the ***top ten*** causes of death globally. According to the World Health Organization (WHO), the world had an estimated 1,418,000 deaths in 2019, with 95% of these deaths being reported by low to middle-income nation-states. About 85% of people who develop TB can be effectively treated with a six-month drug regimen, and since the year 2000, the use of anti-tuberculosis medicines has managed to deter more than 60 million deaths. Death of patients during TB treatment is not merely a function of infection with Mycobacterium tuberculosis but there are parasites, diseases, and health system factors that cause a negative outcome. The majority of TB deaths are recognized as non-tuberculosis-related reasons. Therefore, treating TB and the underlying co-morbidities is vital to ensure positive results for the registered cases in 2015 and 2020. Globally an estimated 66 million lives were saved through TB diagnosis and treatment between 2000 and 2020.

Globally, close to one in two TB-affected households face costs higher than 20% of their household income, according to the latest national TB patient cost survey data. The world did not reach the milestone of 0% TB patients and their households facing catastrophic costs as a result of TB disease by 2020. By 2022, US$ 13 billion is needed annually for TB prevention, diagnosis, treatment and care to achieve the global target agreed upon at the UN high-level level meeting on TB in 2018.

Funding in low- and middle-income countries (**LMICs**) that account for 98% of reported TB cases falls far short of what is needed. According to the International Monetary Fund (IMF), spending in 2020 amounted to US$ 5.3 billion less than half (41%) of the global target. There was an 8.7% decline in spending between 2019 and 2020 (from US$ 5.8 billion to US$ 5.3 billion), with TB funding in 2020 back to the level of 2016. Ending up the TB prevalent by 2030 is among the health goals of the United Nations Sustainable Development Goals (**SDGs**).

People infected with TB bacteria have a lifetime danger of falling ill with TB of 10%. However, persons with compromised immune systems, such as people living with HIV, diabetes or malnutrition, or people who use or smoke tobacco, have a much higher risk of falling ill.

For people who progress active TB (disease), the symptoms (cough, fever, night sweats, weight loss etc.) may be minor for many months. This can lead to delays in the search for care and results in the transmission of the bacteria to others. People ill from TB can infect up to 10-15 other people through close interaction over the course of a year. Without proper treatment up to two-thirds of people ill from TB dies.

Tuberculosis is the third leading killer disease in Zimbabwe, and the country reported 4600 deaths in 2019.

2.2.1 STATISTICAL ANALYSIS

Mathematics has long been intertwined with the biological sciences. The importance of mathematical approaches in many areas of biology is obvious but it is less appreciated that biological questions have stimulated the emergence of a variety of new directions in mathematics. Among many others, the areas of mathematics fully or partially developed in response to demands of biology include branching processes, travelling wave solutions of reaction-diffusion systems, Turing bifurcation and diffusive instability, analysis of replicator equations, stochastic coalescent process, evolutionary game theory and analysis of variance.

2.2.2 TUBERCULOSIS INCIDENCE

Mycobacterium tuberculosis is the bacteria that causes tuberculosis. Mycobacterium tuberculosis is contracted from one person to another when someone with an active form of TB releases tiny, bacteria-containing droplets into the mid-air - either by sneezing, coughing, talking, laughing, singing, or similar. The bacteria can go on suspended in the air for hours, possibly infecting anyone who inhales them. When a patient who has never been exposed to TB inhales the bacteria, it results in an initial TB infection or primary infection. At this stage, some people have no symptoms, while others may experience fever or pulmonary symptoms. Even though TB is spread similarly to the flu or common cold, it is not as contagious.

In most people who have breathed in the bacteria, the immune system instantly kicks in and one recovers without further signs of the illness. The bacteria may then go on in a latent, or dormant, stage - it's in one's system, but not making them sick. But in some cases, the bacteria eventually reactivate and reproduce, leading to the active form of TB - when it makes the person indicative and infectious.

Latent TB shouldn't be ignored, though, because the disease can become active at any time if your immune system gets weakened. According to the Centers for Disease Control and Prevention (CDC), about 5% to 10% of people infected with **Latent TB** develop active TB at some point in their lives if they don't receive treatment. According to Asim A. Jani, MD, the Medical Director for communicable diseases at the Florida Department of Health in Orange County, each stage doesn't necessarily have to lead to the next.

In 2021, globally an estimated 10.6 million people suffered from TB, with 1.2 million being children and adolescents. Child and adolescent TB is often unnoticed by health providers and can be difficult to diagnose and treat. Also in 2021, the **30 high TB burden countries** accounted for 87% of new TB cases (*India, China, Russian Federation, Indonesia, Philippines, Pakistan, Bangladesh, Nigeria, and Democratic Republic of the Congo*). Multidrug-resistant TB **(MDR-TB**) remains a public health crisis and a health security threat. In 2021, 474,000 people fell in from MDR-TB, and most of them lived in just 3 nations: India (24%), China (13%) and the Russian Federation (10%) according to the Global stats and targets - TB Alert. Only about one in three people with drug-resistant TB accessed treatment in 2021. Here in 2021, Zimbabwe recorded 245 cases of TB a slight increase of 7.92% from 227 registered cases in 2020. Most TB patients are cured by a strictly followed six-month drug regimen. Mismanagement and negligence of the treatment and person-to-person transmission have driven up multidrug resistance. In Makoni district, there are 24 cases recorded for the year 2021 a drop from 29 registered cases from the previous year 2020.

According to the WHO 2020 Factsheets, globally, TB incidence is falling at about 2% per year and between 2015 and 2020 the cumulative reduction was 11%. Though this target failed the outcome was over halfway to the End Tuberculosis Strategy milestone of 20% reduction.

2.2.3 RISK FACTORS OF TUBERCULOSIS

Siigh, 2017, says risk factors for TB take into account anything that deteriorates a person's immune system or puts someone in frequent, close contact with a person who has active tuberculosis.

1. **Poverty;** People living in poverty often lack admission to quality healthcare. It's also likely that in the United States, people with little means could be living close to those who have recently emigrated from a country where tuberculosis is common.
2. **HIV Infection;** since HIV attacks the immune system, it positions people at greater risk of getting ill from other bacteria and viruses. The combination of HIV and TB can be deadly because these two diseases feed off each other. In 2021, about 187,000 people with HIV died of TB globally.
3. **Homelessness;** People who are displaced often live in crowded environments with little or no access towards healthcare.
4. **Being in Prison or Jail;** Incarcerated people are often in enclosed areas with a crowd, breathing the same air.
5. **Substance Abuse;** Intravenous (IV) drug use and alcoholism weaken the immune system.
6. **Taking Medication That Weakens the Immune System Autoimmune disorders;** medicine such as rheumatoid arthritis, psoriasis, and Crohn's disease, cause the body's immune system to attack itself. Treatments for these syndromes often involve medication that suppresses the immune system. But that means one's immune system may not be able to fight off tuberculosis after contact.
7. **Kidney Disease and Diabetes Chronic Conditions;** these include diseases such as kidney disease and diabetes, which weaken your immune system, making it difficult for the body to fight off tuberculosis.
8. **Organ Transplants;** the drugs people take to prevent the rejection of an organ transplant can weaken the immune system.

9. **Working in Healthcare Facilities;** Doctors, nurses, and other healthcare workforces get exposed to many patients regularly, which means they're also more likely to be near someone with TB.
10. **Cancer Chemotherapy** weakens the immune system.
11. **Smoking Tobacco Smoking** can increase your risk of getting TB and dying from it. The WHO estimates that 8% of TB cases globally can be connected to smoking.
12. **Babies, Young Children, and Elderly People;** The immune system can be more defenseless when someone is very young or very old.

2.2.4 TUBERCULOSIS MORTALITY

In 2021 about 1.6 million people around the world died from TB, and most of those deaths occurred in low- and middle-income countries, according to the World Health Organization (WHO). Areas of the world with higher rates of tuberculosis include Africa, Asia, Eastern Europe, Russia, Latin America and the Caribbean Islands. Over 25 per cent of tuberculosis occurs in The African Region.

These eight countries made up two-thirds of new TB cases in 2020; India, Indonesia, China, Philippines, Pakistan, Bangladesh, Nigeria and the Democratic Republic of the Congo. According to the Zimbabwe National TB Program in 2021, the tuberculosis mortality rate was 13 cases per 100,000 people. This death rate has fluctuated substantially in recent years, it tends to decrease from the 2002 – 2021 period ending at 13 cases per 100,000 people in 2021. According to reports from Makoni District Hospital, a death rate of 3% was recorded in the year 2021 following a decrease from 4.7 per cent the previous year 2020.

2.3 EMPIRICAL REVIEW

Chipinduro, 2020, studied "The National TB Prevalence Survey". The study aims to provide accurate estimates of bacteriologically confirmed pulmonary TB disease among adults aged ≥ 15 years in 2014. This study evaluated the proportion of Tuberculosis deaths, medical conditions at admission, and determinants of Tuberculosis of patients in the TB intensive care unit (TBICU) of a tertiary hospital in Western Uganda. The authors used the following materials and methods to carry out the study. A prospective cohort study of 351 consecutively enrolled TB prevalence was conducted from March to June 2019.

Moyo (2019) conducted a secondary data analysis focusing on tuberculosis (TB) deaths in Bulawayo province, Zimbabwe. The study revealed a significant increase in the TB death rate within the province, rising from 15.3% in 2016 to 14.2% in 2019, surpassing the threshold of 5%. The research followed a descriptive cross-sectional design. For data analysis, Moyo utilized Microsoft Excel 2007 to generate graphs and employed STATA 17 to perform a Chi-square test for trends. The results indicated that males accounted for 278 out of 469 TB-related deaths, representing 59.3% of the total. Based on these findings, the author concluded that the high mortality rates, particularly during the intensive phase of treatment, could be attributed to suboptimal clinical care.

In Wang's study titled "Data analysis and forecasting of TB prevalence rates for smart healthcare based on a novel-combination model" (2018), the author addressed the challenging and significant task of selecting an appropriate model for mining and analyzing relevant medical information. The research focused on analyzing time series data of tuberculosis (TB) prevalence rates for four income groups identified by the World Bank. To determine the statistical significance of the differences in TB prevalence rates among the income groups, Wang employed Kruskal-Wallis analysis of variance and conducted multiple comparison tests. Additionally, the study proposed a novel-combined forecasting model. The weights of this model were optimized using an artificial intelligent algorithm called cuckoo search. The purpose of this model was to forecast the hierarchical TB prevalence rate from 2013 to 2016. The results demonstrated that the developed combination model was not only simple but also capable of reasonably approximating the actual TB prevalence rate. Wang's research highlighted the potential of this novel approach for analyzing and forecasting TB prevalence rates in the context of smart healthcare.

Azeez (2016) conducted a study titled "Seasonality and trend forecasting of TB prevalence data in the Eastern Cape, South Africa using a hybrid model." The objective of the research was to compare the predictive capabilities of two models, SARIMA and SARIMA-NNAR, for forecasting tuberculosis (TB) incidence and analyze its seasonality in South Africa. To achieve this, Azeez obtained TB incidence data from January 2010 to December 2015 from the Eastern Cape health facility report. The study compared the performance of both models in predicting TB incidence. The combined model (SARIMA-NNAR) demonstrated superior performance compared to the single SARIMA model. Furthermore, the study examined the seasonality trend of TB incidence.

The forecasting results indicated a slight increase in seasonal TB incidence trend from the SARIMA-NNAR model compared to the single model. These findings shed light on the seasonality patterns of TB incidence in the Eastern Cape region of South Africa and highlighted the improved predictive power of the hybrid model.

In Wademan's study titled "TB is a disease which hides in the body: Qualitative data on conceptualization of TB recurrence among patients in Zambia and South Africa" (2021), the author explored how participants perceive the risk of TB recurrence, using a discursive analytics approach. The study examined the experiences of individuals who had multiple episodes of TB and the associated social, economic, and physiological vulnerabilities. These experiences challenged the participants' biomedical understanding of TB curability. The findings emphasized the importance of healthcare providers engaging in discussions with patients about the risk of TB recurrence. It was recommended that healthcare providers promote prevention strategies, as well as early detection and diagnosis of TB, to address the challenges posed by recurrent TB episodes. By addressing these issues, healthcare providers can better support patients in managing and preventing TB recurrence.

2.4 RESEARCH GAP AND CONCEPTUAL FRAMEWORK

Tuberculosis incidence and Tuberculosis mortality have attracted so much attention in the global health arena. Several researchers have felt the need to carry out research on the matter and studies have been carried out on the causes and treatments of TB, tuberculosis infection rates, TB mortality rates, risk factors of TB infections, and determinants of tuberculosis, just to mention a few. It is a disappointment that few to almost no study has been carried out in Zimbabwe more specifically at the district level, where the economy is unpredictable and the health sector is said to be weakening slowly. In the research that has been carried out, some were conducted in countries that have better health systems than Zimbabwe whilst others have much worse health systems than Zimbabwe and Makoni District. Those that were carried in Zimbabwe were carried when the health systems were better and some have actually been unsuccessful in being published and hence unnoticed. Therefore, this study is of significance since Makoni district is in Zimbabwe where the studies are taking place. This gives a clear picture of where we stand as a nation in the fight to decrease TB incidence and TB mortality as part of the Sustainable Development Goals and Vision 2030 at the district level. It also helps the policymakers and other related parties in decision-making processes about improving the health system and health services provision.

2.5 CONCLUSION

The epicenter of this chapter was on the literature review of the study. It elucidated the conceptual framework as well as the theoretical and empirical framework of the research. It directed its attention to Tuberculosis Incidence, the risk factors of tuberculosis, Tuberculosis Mortality and the research gap. The subsequent chapter three focuses on the research methodology.

CHAPTER 3

3.1 INTRODUCTION

This chapter looks at the research technique for this study. The primary goal of this chapter is to defend the writer's strategy for data collection to satisfy the study objectives. The chapter looks at the research design and data collecting techniques, sample population, data analysis procedure, data display, and the chapter summary.

3.2 RESEARCH DESIGN

Research design is an essential component of any quantitative analysis. It refers to the overall plan for collecting and analyzing data to answer a specific research question or to test a hypothesis. Research design encompasses all aspects of research including the formulation of the research question or hypothesis, the sampling of subjects, data collection methods, the design of experiments, data analysis procedures, and the writing up of results. In other words, research design provides a road map for a researcher to identify the information needed and method of collecting and analyzing to answer the research question or to test the hypothesis. Lee and Lings (2008) further defined the research design as a general plan of how the research is going to be carried out and mainly concentrates on giving solutions or objectives for testing the research hypothesis. Harwell (2007) states, “It is possible to characterize a research study’s methodology as either qualitative or quantitative or both”. The study uses quantitative and qualitative research methods, and a secondary data analysis method called logistic regression analysis.

3.3 STUDY POPULATION AND SAMPLE

3.3.1 TARGETED POPULATION

A population is the entire target group of a particular type being studied (Mudimu and Machegetwa, 2002). The population is explained by Wenger (2000) as the set of all observations of a random variable being studied and based on which one tries to conclude in fact. Kotler & Armstrong (2011) also defines a population as a group of people, things or events of interest to the researcher and what to study. In that context, this study focuses on Makoni district. This is where the majority of patients from different clinics of the district are tested for TB, and therefore significant enough data is available for this study. Diagnosis tests that take place at local and district clinics are then matched against provincial data for registration and statistical purposes.

3.3.2 RESEARCH SAMPLE

Sounders and Cornet (2007) define a sample as a representative part of the population. Kumar (2011) also defines the sample as a selected subset of the population of interest to the researcher, that is, the sample is a subset of the study population. 78.57% of the population was used as a sample for this study. This means that of the 5,100 patients tested, a sample size of 4,007 was used.

3.4 SOURCES OF DATA

This section presents the data sources used in the study. Data collection is the process of collecting, collecting, extracting, and storing huge amounts of data that may be in a structured or unstructured form such as text, video, audio, XML files, records, or Other image files used in later data phases. Analysis. In big data analysis, "collecting data" is the first step before starting to analyze patterns or useful information in the data. The data to be analyzed must be collected from various valid sources. The collected data is known as raw data which is not useful at present, but cleaning up the impurity and using that data for further analysis forms insights. We used data from TB patients who were tested in Makoni District between January 1, 2018, and December 31, 2022. The monitoring and evaluation officer of Development Aid from People to People Total Control of Tuberculosis (DAPP TC TB), the District Medical Officer (DMO), a laboratory technician, a clinic nurse in charge, and a tuberculosis nurse were the five key informants carefully selected. Actual data is mainly divided into two categories: primary data and secondary data.

3.4.1 SECONDARY DATA

Saunders, Phillip & Adrian, 2009 define secondary data analysis as the practice of identifying and collecting secondary data and analyzing and evaluating it objectively. According to Bhattacharyya & Kuma, 2006, secondary data is defined as data collected by individuals or institutions for purposes other than solving the problem at hand and answering research questions.

3.4.2 TYPES OF DATA COLLECTED

To better understand the factors affecting TB prevalence and devise effective intervention strategies, it is crucial to conduct a comprehensive statistical analysis using various types of data. This essay discusses different data types that can be employed in the research topic and how they contribute to a well-rounded understanding of the TB situation in the Makoni District.

1. **Demographic Data:**
   Demographic data, such as age, sex, and ethnicity, can help identify population groups that are more vulnerable to TB. This information is found on local health records. Analyzing demographic factors concerning TB cases can provide insights into specific risk factors and help target interventions more effectively.

2. **Geographical Data:**
   Geographical data can be used to map the spatial distribution of TB cases across the Makoni District. This includes information on rural versus urban areas and prisons, population density, and proximity to health facilities. By analyzing the geographical distribution of TB cases, researchers can identify hotspots and prioritize resource allocation for disease control and prevention.

3.5 DATA PRESENTATION AND ANALYSIS PROCEDURES

According to Alexopoulos (2010), data analysis tries to extract reliable information from raw data. (Kojo, 2011), unique data analysis is a method of gathering and combining data in a meaningful way for easy interpretation. Data presentation is the display of analyzed statistical information to aid in understanding trends. Tables, bar charts, graphs, and narrative elements were used to show the deductive and inferential statistics. The data analysis process cleared the way for the interpretation of results in the statistical study of TB prevalence.

STATA was used to examine the data. Tables, graphs, pie charts, bar charts, and line charts were used to present and analyze quantitative results. Data were first pooled at. The data in this study is analyzed using Logistic Regression Analysis.

A statistical method for examining data sets with one or more independent variables that affect the result is logistic regression analysis. Dichotomous variables—used when there are only two possible outcomes—are used to measure outcomes. The dependent variable in logistic regression might be either binary or binary. Various stages of the analysis process are followed by researchers:

3.5.1 STEP 1: DESCRIPTIVE STATISTICS

The analyst is attempting to calculate clear insights to come up with a list of the features of the components. Graphic statistics are about what we say about the data you have gathered from your test, (Garth, 2008). The least, most extreme, cruel, and standard deviation are among the unmistakable insights that must be calculated.

3.5.2 STEP 2: AIMS OF LOGISTIC REGRESSION

The researcher is interested in creating a model that accounts for the demographic, geographic, socioeconomic, clinical, and epidemiological independent factors as well as the TB incidence in Zimbabwe's Makoni district. As a result, apply the logistic regression model. The researcher is also interested in figuring out how the independent factors and prevalence interact. It is suggested that a logistic regression analysis be used to evaluate the relationships.

To apply the regression procedure, the researcher selected

Y: The response variable is the TB test result, which is binary (dichotomous) in nature

X: The explanatory variables, all variables are nominal

X_1 – location cluster of the clinic

X_2 – age of patient

X_3 – HIV status

X_4 – TB risk group

X_5 – type of TB test

3.5.3 STEP 3: ANALYSIS OF LOGISTIC REGRESSION

Selection of Variables

According to Bangley (2001), Concato (1993), and Peduzzi (1995), the model must contain all pertinent variables, but it's also crucial that it doesn't begin with more than is warranted for the number of observations. More variables generally result in a better model fit for the set of data. However, an overabundance of variables affects the model's coefficient and make it fit too well. The le predictive power and frequently challenging data interpretation may be the outcomes of a complex model with numerous unimportant variables. There are two approaches to choosing variables: filter and statistical (Austin, 2004; Genuer, 2010). The variables are condensed according to how significant the independent 31 variables are when using the filter approach. The risk factors are diminished by determining the independent variables included in the model. Several steps must be taken. Correlation analysis is used in statistical procedures. Any regression analysis has challenges when two predictor variables are significantly associated with one another (Bangley, 2001; Feinstein, 1996). The logistic regression analysis may not be accurate if the variables are heavily associated with one another. The statistical approach of variable selection has two procedures.

1. **The Interaction Test**

   In the regression model, the interactions are expressed as product terms, which are terms that are the result of two predictors rather than one predictor alone (Bangley, 2001; Hosmer & Lemeshow, 2000; Kleinbaum, 1994). To determine the significant values for each variable, interactive tests were run. Engagement's significance is quantified and reported. The values used in the cross-tabulation test were from the Pearson Chi-Square.
2. **The Co-Linearity Analysis**

   The variance associated with these coefficients increases leading to a loss of statistical significance (Bangley, 2001). Collinearity analysis is based on the critical values of the interaction test. The significant value for each variable should be less than 0.20 (Hosmer & Lemeshow, 2000). Variables with a significance of less than 0.20 were selected in the logistic regression analysis
3. **The Stationarity Test**

   Data from a time series must be stationary. It is not permitted to model, predict, or forecast using non-stationary data. To check for stationarity, the researcher applies the Augmented Dickey-Fuller Test (ADF). The ADF test is a statistical test that determines whether a time series is stationary or not based on its autocorrelation structure. The test is based on the null hypothesis that the time series contains a unit root, meaning that it is non-stationary. The alternative hypothesis is that the time series is stationary. To perform the ADF test, the first step is to estimate the regression equation of the time series on its lagged values.

Validation

According to Bangley (2001) and Feinstein (1996), the validation analysis was carried out to determine whether the logistic regression analysis was appropriate or not. The main samples' prediction rate for correct cases 32 must be higher than or equal to that of the verified samples. The percentage of accurate cases is calculated during validation using additional sample data that share the same coefficient values as the primary data.

3.5.4 STEP 4: LOGISTIC REGRESSION ASSUMPTIONS

Both the connection as a whole and the specific variables (dependent and independent) are subject to the assumptions of **logistic regression analysis**. To satisfy the requirements of logistic regression, this stage focuses on assessing the relationship between the dependent and independent variables. Both before and after estimating the model, the assumptions are evaluated. The following assumptions need to be assessed:

1. **Linearity**
   The logit function is the natural logarithm of the odds ratio, which is the ratio of the probability of success to the probability of failure. This assumption implies that the logit of the dependent variable is a linear function of the independent variables. If this assumption is violated, the model may produce incorrect estimates, and the interpretation of the results may be misleading. Before estimating the regression equation, the researcher evaluates linearity using scatterplots and the correlation matrix. After estimating the equation, the researcher plots the standardised residuals as a function of the standardised expected values to look at the residual plots.
2. **Independence**
   According to this presumption, the observations ought to be independent. The value of the dependent variable for one observation should not be influenced by the value of the dependent variable for a different observation, according to this. The model parameters

may be estimated with bias if this premise is broken. The researcher applied the Durbin-Watson test to determine whether the error term is independent.

3. **Homoscedasticity**

   The researcher plots the standardised residuals against the anticipated dependent values.

4. **Multicollinearity**

   Multicollinearity occurs when two or more independent variables are highly correlated with each other. This makes it difficult to determine the independent effect of each of the variables on the dependent variable, leading to unstable and unreliable estimates of the model parameters. Multicollinearity is a metric of how dependent the model's explanatory variables are, and if it is "high" it violates the BLUE principle. As a result, the estimates' variance and covariance are high, which makes precise estimation challenging. The variables' multicollinearity are examined using the correlation matrix. Although multicollinearity in economic data is predictable, if the multicollinearity measures R2 ≤ 0.8 is acceptable, otherwise the issue is grave and requires attention. multiple correlations between explanatory factors renders the model unsuitable for estimation, resulting in inflated standard errors, type 1 errors, and an unacceptably high goodness-of-fit 33 even in the absence of strong independent variable explanatory power. The researcher evaluates the degree of multicollinearity.

3.6 MODEL FOR ANALYSIS

3.6.1 DEPENDENT VARIABLE

The dependent variable is the tuberculosis incidence rate, with the following link function representing a general multivariate binary logit model with five independent variables.

$$\textbf{Logit (P)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + e_i \qquad \textbf{equation 3a}$$

Where **(Y = 1/ ∀)**, then **Y = 0** given **1 – P**

The likelihood that response variable Y in a database classifies as a new TB case given **k** predictors is denoted by the symbol **P**, whereas the probability that a patient would be classified as surviving is denoted by the symbol **(1 - p)**.

3.6.2 SELECTING FITTING MODELS

Variable selection is a step in the model fitting process that helps identify which variables are important to the model. This procedure is carried out to make sure that the final model has a small variance and correctly matches the data. A model with low variance has a higher capacity for prediction than a model with high variance.

The likelihood ratio (LR) test is to choose the model variables for the probit model. The theoretical probit model serves as the foundation for the model are used in the variable selection procedure. The poverty model, which is the proposed empirical model, is mathematically denoted as **Logit (Y) =** $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e_i}$ **equation 3b**

Where X_i, for $\boldsymbol{i}$, equals 1, 2, 3, 4, or 5 depending on the location where the patient lives, the age group of the client, the HIV status of the client, the TB risk group of the tested person, and the type of test used for diagnosis. Finding the best-fit model in binary logistic regression is extremely improbable. The researcher eliminates unimportant variables from each level of the model during the modelling process. If there are still non-significant factors, the researcher should repeat the process until they are all eliminated before creating a second model that includes the significant variables from the previous model. Stay with the "best" fit model if all variables are not significant (i.e., $p>0.05$). Where the regression coefficients for the variable X_n are, respectively, 0 and n. In logistic regression, we forecast the likelihood that Y occurs given known values of X_i (or X_s) rather than the value of one variable Y from one predictor X_i or more predictors (X_s). The bracketed portion of the equation, which has a constant b_0, a predictor X_1, and a coefficient (or weight) associated with this predictor of 1, intuitively resembles a multiple linear regression. The equation becomes: when there are numerous predictor variables.

$$\mathbf{P(Y)} = \frac{\mathbf{1}}{\boldsymbol{1+e\ (\beta 0+\ \beta 1\ i+\beta 2 i+\cdots+\beta ¿\ )}}$$

equation 3c

Is the same as the equation used when there is only one predictor**,** except that the linear combination has been expanded to include any number of predictors? Therefore, the multiple prediction version of the logistic regression equation also contains the multiple regression equation, but the single-predicted form only contains the single linear regression equation.

Although logistic regression and linear regression have many similarities, there is a valid reason why linear regression cannot be used in situations when the outcome variable is absolute. The rationale is that one of the fundamental premises behind linear regression is that there exists a linear relationship between the variables. For linear regression to be a reliable model, there must be a linear relationship in the observed data. This presumption is broken when the result variable is absolute (Berry, 1993). Given that it covers the range between positive and negative infinity, the logit function is limitless. The association function between probabilities is called logit. The explanatory variables indicated in the research model of the first chapter are represented by the variable n of the logit 35 algorithm, symbolized by the symbol X. The study model's explanatory variables include both categorical and continuous data. The independent variables, often known as model parameters since they represent the weights of each predictor variable in the logit, have comparable beta coefficients.

3.7 RESEARCH VALIDITY AND RELIABILITY

3.7.1 SIGNIFICANCE OF TESTING OF THE LOGISTIC REGRESSION COEFFICIENTS

After estimating the model parameters, the researcher examines the significance of the regression coefficients. The following theories are put to the test:

The formula yields the Wald statistic for any coefficient ***b***. A Wald statistic is a measure of the significance of the regression coefficients in a logistic regression model. It is calculated by dividing the square of the coefficient estimate by the square of its standard error.

$$\text{Wald} = \frac{b}{se\ c}$$

equation 3d

Where the explanatory variable and's estimated coefficient ***β*** of determination is represented by ***b*.** Its standard error is E (***b***). The theories include:

$H_0 : \beta_j = \beta_1 = 0$

$H_1 : \beta_j = \beta_1 \neq 0$.... Where $j = 1, 2 \ldots., k$

3.7.2 OVERALL TEST FOR THE SIGNIFICANCE: ANOVA TEST THE COEFFICIENTS OF THE DETERMINATION

The F ratio is used to test the hypothesis that more variation is explained by the regression model than by the average (i.e., that R2 is greater than zero).

Calculating the test statistic F ratio is as follows:

$$F_{ratio} = \frac{\dfrac{\sum of\ errors\ of\ Regression}{Degrees\ of\ freedom\ of\ Regression}}{\dfrac{\sum of\ errors\ of\ Residual}{Degrees\ of\ freedom\ of\ Residual}}$$

The hypothesis to be tested

H0: β j=β1=0,   H1: β j=β1≠0.... Where *j* = 1, 2…., k

3.8 CHAPTER SUMMARY

This chapter describes the methodology used to conduct the research. The study design and data sources used in the study were also highlighted. This chapter proceeded to consider the data analysis steps in the next chapter. The next chapter covers data presentation and analysis.

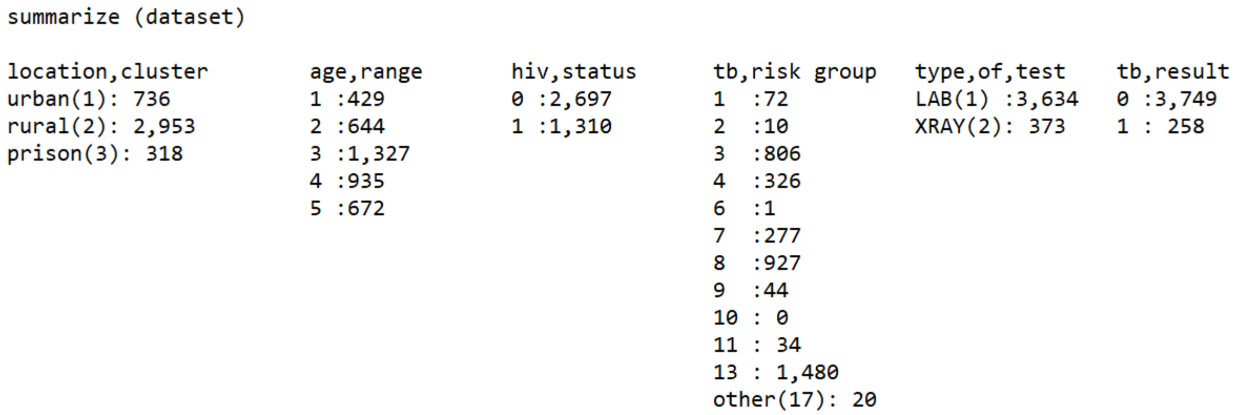
CHAPTER 4: DATA ANALYSIS PRESENTATIONS AND DISCUSSIONS

4.0 INTRODUCTION

This chapter focuses on presenting the analysis findings and their interpretation. To explore the factors related to tuberculosis contraction, through utilization of multivariable binary logistic regression, employing different pre-test and validation techniques: the Interaction Test, Co-Linearity Analysis, Stationarity Test, Validation Test, Linearity Test, Independent Durbin Watson Test, Homoscedasticity Test, Multicollinearity Test, and Wald Test. Additionally, descriptive statistics and model estimation outputs based on STATA 13.2 Statistical Computing Software are provided. The results are interpreted in relation to the study's objectives.

4.1.1 DATA EXPLORATION

TABLE 4.1



```
summarize (dataset)

location,cluster        age,range      hiv,status     tb,risk group   type,of,test     tb,result
urban(1): 736           1 :429         0 :2,697       1   :72         LAB(1) :3,634    0 :3,749
rural(2): 2,953         2 :644         1 :1,310       2   :10         XRAY(2): 373     1 : 258
prison(3): 318          3 :1,327                      3   :806
                        4 :935                        4   :326
                        5 :672                        6   :1
                                                      7   :277
                                                      8   :927
                                                      9   :44
                                                      10 : 0
                                                      11 : 34
                                                      13 : 1,480
                                                      other(17): 20
```

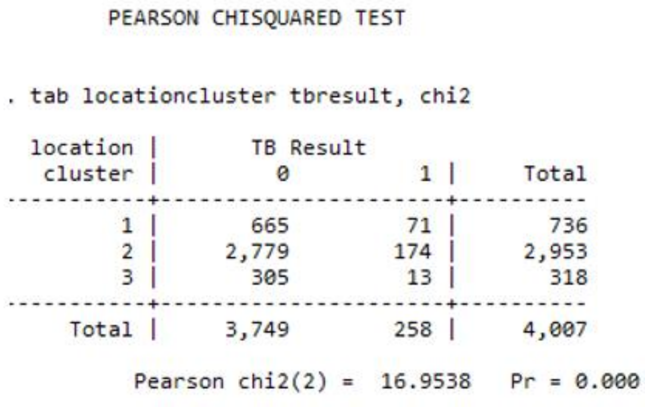
In table 4.1 above, (1) represents the location of Makoni District urban with a total population of 736. (2) Represents the location of Makoni District rural with a total population of 2953. (3) Represents the section of Makoni District Prisons with a total population of 318 and the total number of people tested for TB was 4007. The age group 1 is of people between 1 to 18 years and it has 429 people tested for TB. The age group 2 is of people between 19 to 30 years and it has 644 people tested for TB. The age group 3 is of people between 31 to 45 years and it has 1327 people tested for TB. The age group 4 is of people between 46 to 60 years and it has 935 people tested for TB. The age group 5 is of people between 61 to 120 years and it has 672 people tested for TB. HIV 0 means negative and HIV 1 means positive. TB 1 to 13 are the TB standards risk group and their population is given in the table above. There are two types of test used which are laboratory test represented by LAB (1) and X-ray test represented by XRAY (2). LAB (1) has a total population of 3634 and XRAY (2) has a total population of 373. On TB results 0 represent negative and 1 represent positive. There are 258 people confirmed TB positive and 3749 TB negative.

4.2.0 INTERACTION TEST

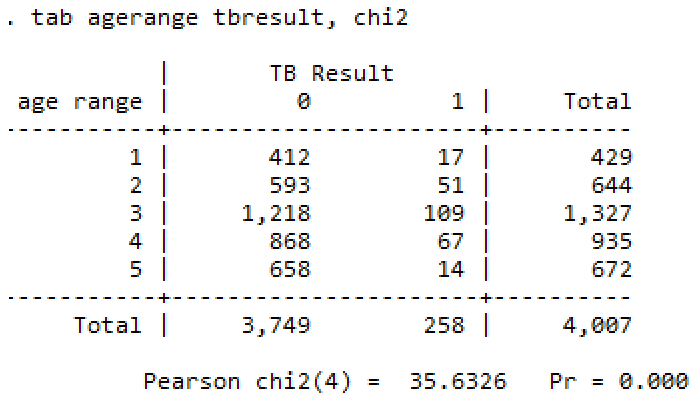
The interaction test refers to a statistical analysis that examines whether the relationship between two variables varies depending on the levels or values of other variables. It helps determine if the effect of one variable on the outcome is different across different levels of another variable. The values used in the cross-tabulation test were from the **Pearson Chi-Square.**

4.2.1 PEARSON CHI-SQUARE

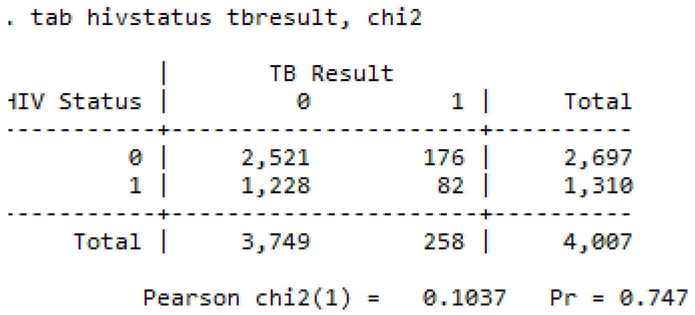
The Pearson chi-squared test was employed to analyze various variables in relation to the TB result. The outcomes are summarized as follows:



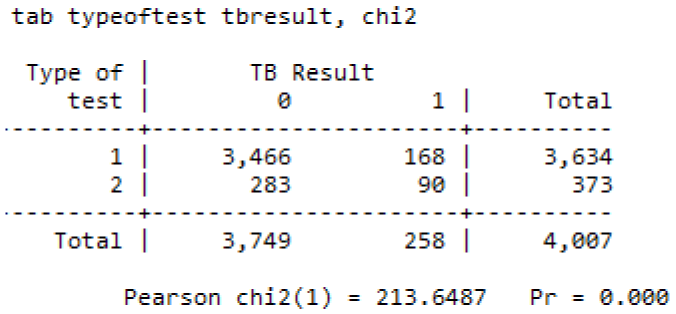
```
        PEARSON CHISQUARED TEST

. tab locationcluster tbresult, chi2

  location |        TB Result
   cluster |         0          1 |     Total
-----------+----------------------+----------
         1 |       665         71 |       736 
         2 |     2,779        174 |     2,953 
         3 |       305         13 |       318 
-----------+----------------------+----------
     Total |     3,749        258 |     4,007 

          Pearson chi2(2) =  16.9538   Pr = 0.000
```



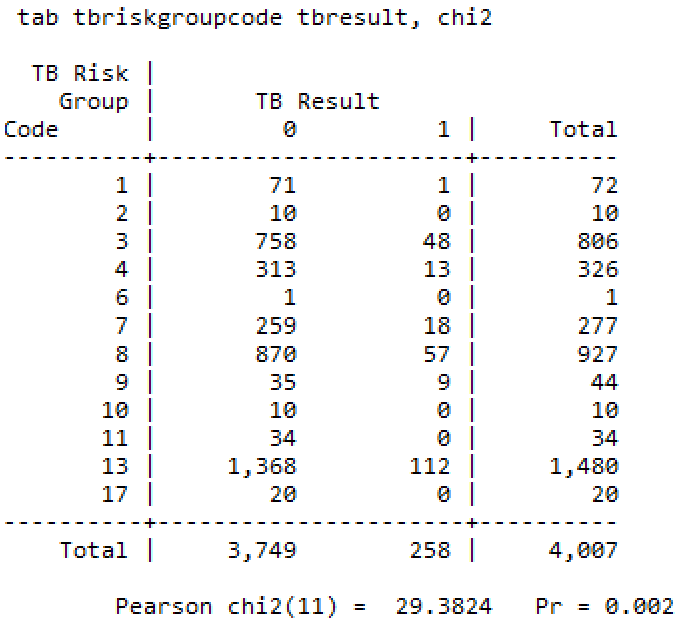
```
. tab agerange tbresult, chi2

           |       TB Result
 age range |         0          1 |     Total
-----------+----------------------+----------
         1 |       412         17 |       429 
         2 |       593         51 |       644 
         3 |     1,218        109 |     1,327 
         4 |       868         67 |       935 
         5 |       658         14 |       672 
-----------+----------------------+----------
     Total |     3,749        258 |     4,007 

          Pearson chi2(4) =  35.6326   Pr = 0.000
```



```
. tab hivstatus tbresult, chi2

           |       TB Result
HIV Status |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,521        176 |     2,697 
         1 |     1,228         82 |     1,310 
-----------+----------------------+----------
     Total |     3,749        258 |     4,007 

          Pearson chi2(1) =   0.1037   Pr = 0.747
```



```
 tab typeoftest tbresult, chi2

  Type of |       TB Result
     test |         0          1 |     Total
----------+----------------------+----------
        1 |     3,466        168 |     3,634 
        2 |       283         90 |       373 
----------+----------------------+----------
    Total |     3,749        258 |     4,007 

        Pearson chi2(1) = 213.6487   Pr = 0.000
```



```
tab tbriskgroupcode tbresult, chi2

  TB Risk |
    Group |       TB Result
Code      |         0          1 |     Total
----------+----------------------+----------
        1 |        71          1 |        72
        2 |        10          0 |        10
        3 |       758         48 |       806
        4 |       313         13 |       326
        6 |         1          0 |         1
        7 |       259         18 |       277
        8 |       870         57 |       927
        9 |        35          9 |        44
       10 |        10          0 |        10
       11 |        34          0 |        34
       13 |     1,368        112 |     1,480
       17 |        20          0 |        20
----------+----------------------+----------
    Total |     3,749        258 |     4,007

        Pearson chi2(11) =  29.3824   Pr = 0.002
```

4.2.3 TEST SUMMARY

The provided table presents the outcomes of a pairwise correlation analysis (specifically, Pearson correlation) conducted on several variables: ***locationcluster, age-range, hivstatus, tbriskgroupcode, typeoftest, and tbresult.***

The table displays the correlation coefficients, with asterisks denoting the significance levels. Here is a summary of the findings:

1. ***locationcluster*** has a correlation coefficient of 1.0000 with itself, which is expected.
2. ***Agerange*** exhibits a weak negative correlation (-0.0671) with *locationcluster*, reaching statistical significance at the 5% level (*).
3. ***Hivstatus*** demonstrates a weak positive correlation (0.0558) with *locationcluster*, which is statistically significant at the 5% level. It also displays a weak positive correlation (0.1662) with *agerange*, also statistically significant at the 5% level.
4. ***Tbriskgroupcode*** showcases a moderate negative correlation (-0.2112) with locationcluster, significant at the 5% level. It displays a weak positive correlation (0.1177) with *agerange*, also significant at the 5% level. Additionally, it exhibits a weak negative correlation (-0.1704) with *hivstatus*, also statistically significant at the 5% level (*).

5. ***Typeoftest*** reveals a moderate negative correlation (-0.1849) with *locationcluster*, reaching statistical significance at the 5% level. It shows a weak positive correlation (0.0581) with *agerange*, which is statistically significant at the 5% level. It also demonstrates a weak negative correlation (-0.1244) with *hivstatus*, significant at the 5% level. Furthermore, it displays a moderate positive correlation (0.2872) with *tbriskgroupcode*, also statistically significant at the 5% level.

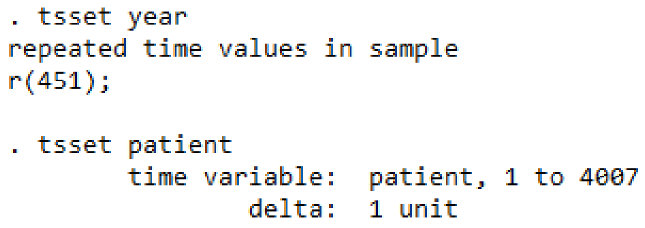
6. ***Tbresult*** indicates a weak negative correlation (-0.0629) with *locationcluster*, reaching statistical significance at the 5% level. It exhibits a weak negative correlation (-0.0337) with *agerange*, which is statistically significant at the 5% level. It reveals a very weak positive correlation (0.0051) with *hivstatus*, which is not statistically significant. It showcases a weak positive correlation (0.0373) with *tbriskgroupcode*, significant at the 5% level. Additionally, it demonstrates a moderate positive correlation (0.2309) with *typeoftest*, reaching statistical significance at the 5% level.

These results provide insights into the strength and direction of the linear relationships between the variables. Nonetheless, it is crucial to note that correlation does not imply causation, and further analysis may be necessary to comprehend the underlying relationships among these variables.

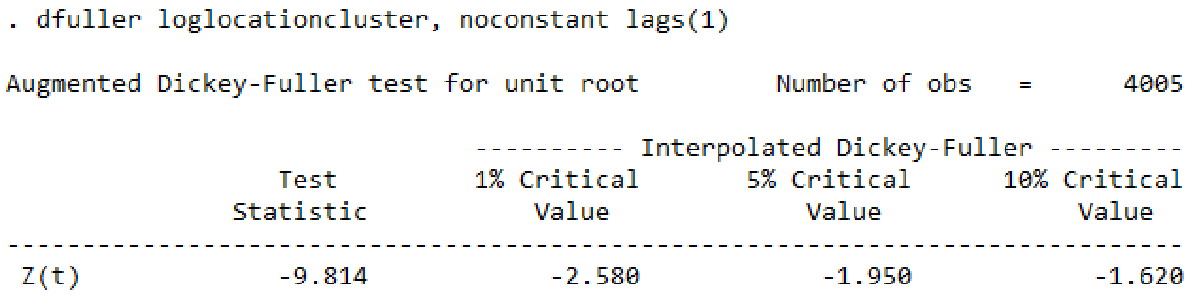
4.3 COLLINEARITY TEST

Due to advancements in STATA 14.2, the researcher did not conduct the co-linearity test since the software internally runs the test on the variables before predicting the outcome models

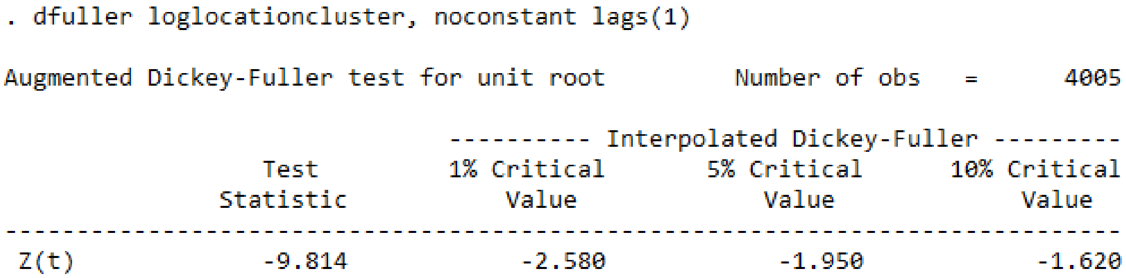
4.4.0 STATIONARITY TEST (ADF TEST)



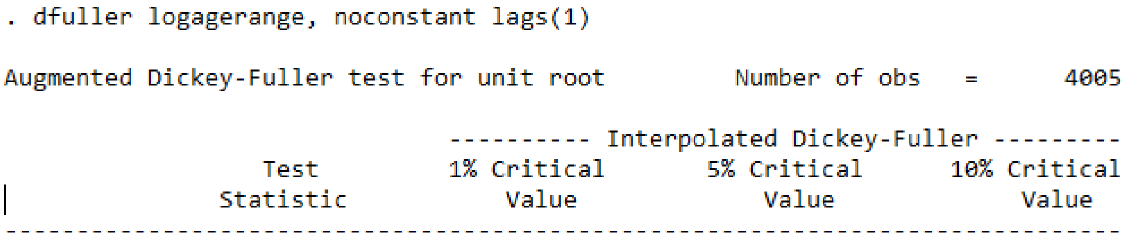
```
. tsset year
repeated time values in sample
r(451);

. tsset patient
        time variable:  patient, 1 to 4007
                delta:  1 unit
```



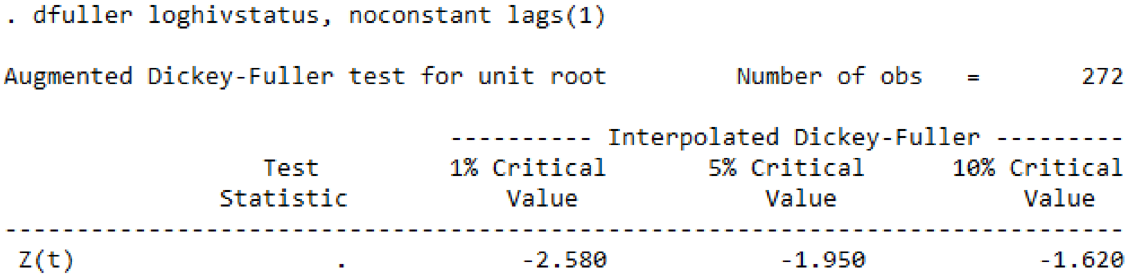
```
. dfuller loglocationcluster, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -9.814            -2.580            -1.950            -1.620
```



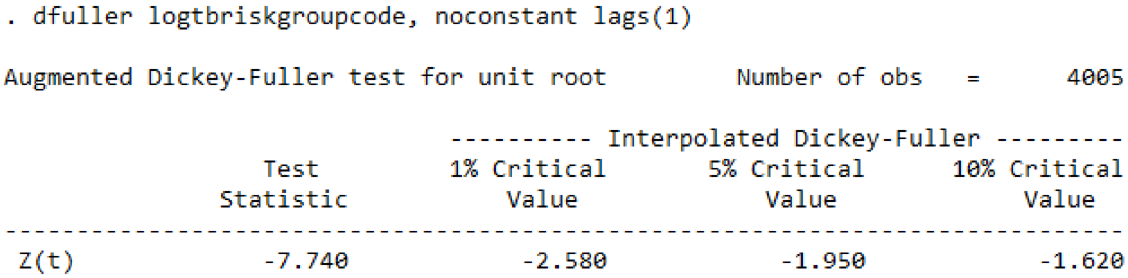
```
. dfuller loglocationcluster, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -9.814            -2.580            -1.950            -1.620
```



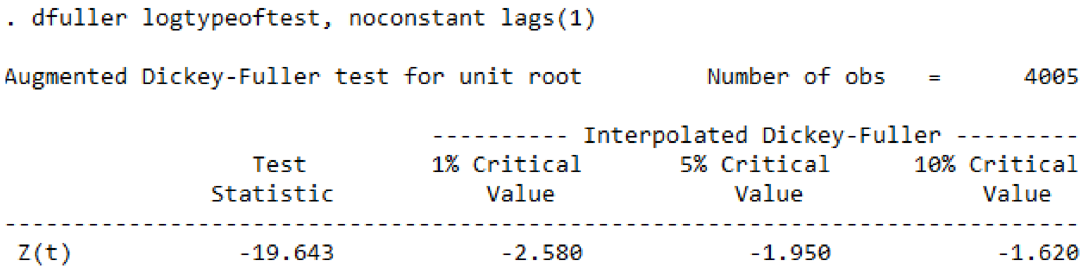
```
. dfuller logagerange, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
```



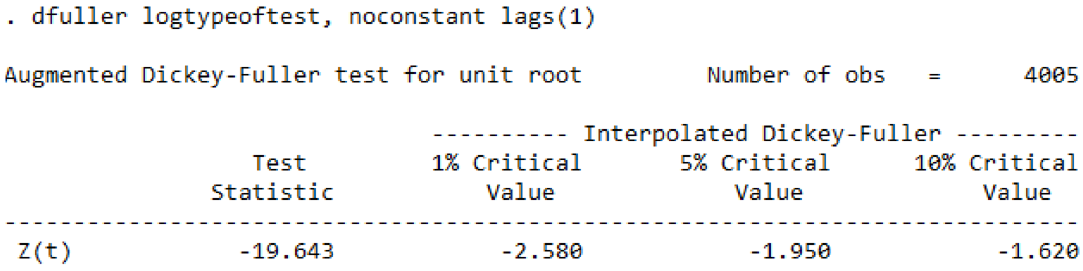
```
. dfuller loghivstatus, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =       272

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)                  .            -2.580            -1.950            -1.620
```



```
. dfuller logtbriskgroupcode, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -7.740            -2.580            -1.950            -1.620
```



```
. dfuller logtypeoftest, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -19.643            -2.580            -1.950            -1.620
```



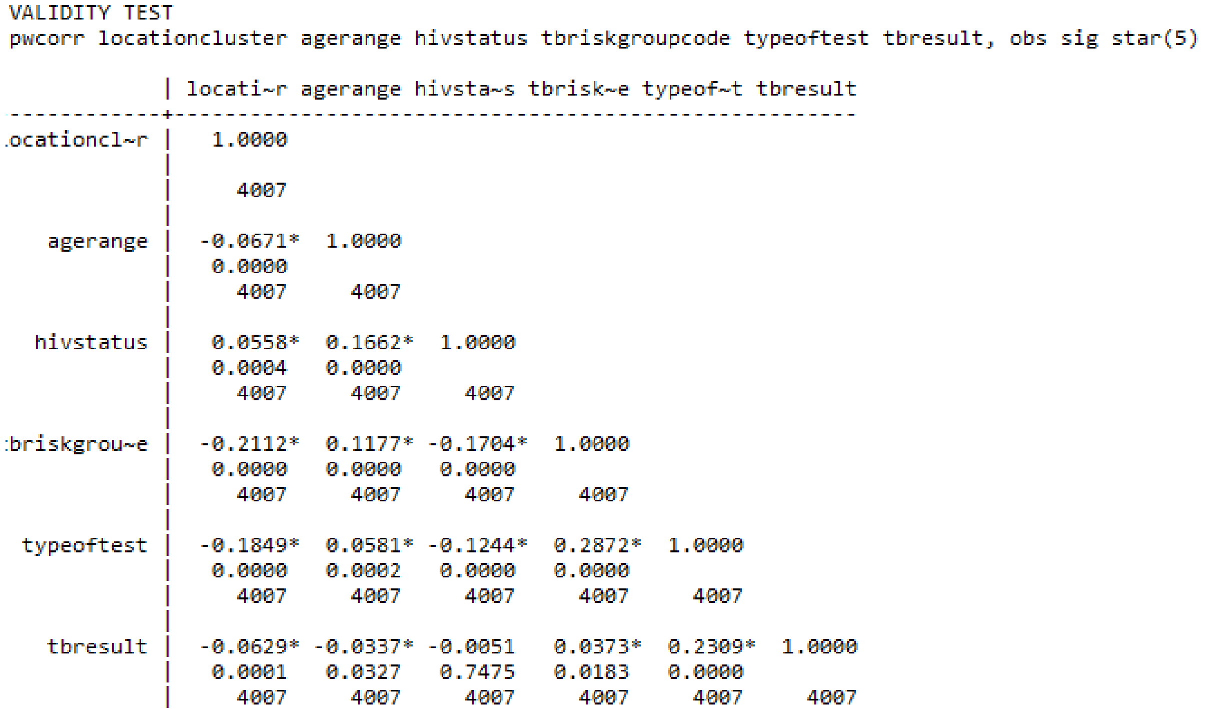
```
. dfuller logtypeoftest, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =      4005

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -19.643            -2.580            -1.950            -1.620
```

4.4.1 TEST SUMMARY

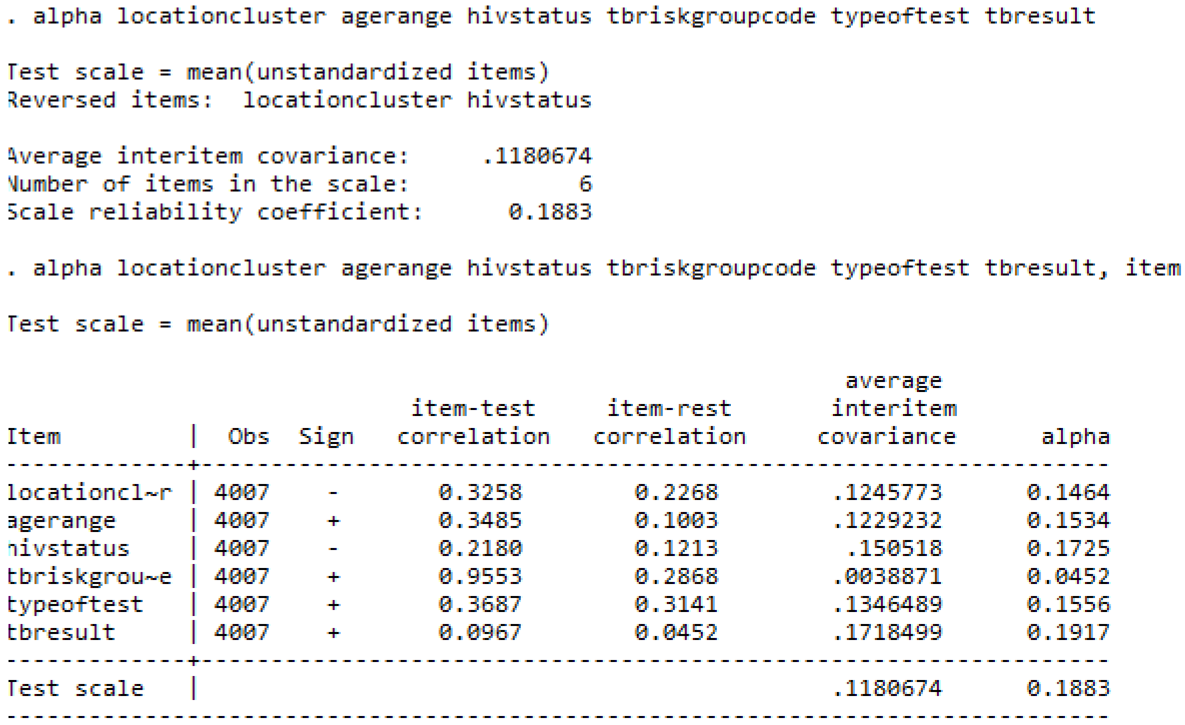
The variable ***location cluster*** is stationary since the Test statistic of 5.063 is greater than 5% critical value of 1.950. Furthermore the variable ***agerange*** is stationary since its test statistic of 9.648 is greater than its 5% critical value of 1.950. In addition the variable ***HIV status*** is stationary because its test statistic is 26.319 which is greater than 5% critical value which is 1.950. The variable ***tb risk group code*** is stationary because test statistic is 11.324 and the 5% critical value is 1.950 and less than the test statistic. In addition the variable ***type of test*** is also stationary since its test statistic of 4.816 is greater than the 5% critical value of 1.950. Lastly the variable ***tb result*** is also stationary since its test statistic of 30.258 is greater than the 5% critical value of 1.950. In conclusion to the model results, it is shown that all variables are stationary

.

4.5.0 VALIDITY TEST



```
VALIDITY TEST
pwcorr locationcluster agerange hivstatus tbriskgroupcode typeoftest tbresult, obs sig star(5)

             | locati~r agerange hivsta~s tbrisk~e typeof~t tbresult
-------------+------------------------------------------------------
.ocationcl~r |   1.0000
             |
             |     4007
             |
    agerange |  -0.0671*  1.0000
             |   0.0000
             |     4007     4007
             |
   hivstatus |   0.0558*  0.1662*  1.0000
             |   0.0004   0.0000
             |     4007     4007     4007
             |
:briskgrou~e |  -0.2112*  0.1177* -0.1704*  1.0000
             |   0.0000   0.0000   0.0000
             |     4007     4007     4007     4007
             |
  typeoftest |  -0.1849*  0.0581* -0.1244*  0.2872*  1.0000
             |   0.0000   0.0002   0.0000   0.0000
             |     4007     4007     4007     4007     4007
             |
    tbresult |  -0.0629* -0.0337* -0.0051   0.0373*  0.2309*  1.0000
             |   0.0001   0.0327   0.7475   0.0183   0.0000
             |     4007     4007     4007     4007     4007     4007
```



```
. alpha locationcluster agerange hivstatus tbriskgroupcode typeoftest tbresult

Test scale = mean(unstandardized items)
Reversed items:  locationcluster hivstatus

Average interitem covariance:     .1180674
Number of items in the scale:            6
Scale reliability coefficient:      0.1883

. alpha locationcluster agerange hivstatus tbriskgroupcode typeoftest tbresult, item

Test scale = mean(unstandardized items)

                                                            average
                             item-test     item-rest       interitem
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
locationcl~r | 4007    -       0.3258        0.2268        .1245773      0.1464
agerange     | 4007    +       0.3485        0.1003        .1229232      0.1534
hivstatus    | 4007    -       0.2180        0.1213         .150518      0.1725
tbriskgrou~e | 4007    +       0.9553        0.2868        .0038871      0.0452
typeoftest   | 4007    +       0.3687        0.3141        .1346489      0.1556
tbresult     | 4007    +       0.0967        0.0452        .1718499      0.1917
-------------+-----------------------------------------------------------------
Test scale   |                                            .1180674      0.1883
-------------------------------------------------------------------------------
```

4.5.1 TEST SUMMARY

In order to evaluate the validity of the correlation analysis, a validity test was performed. The findings are outlined below:

Assessment of Scale Reliability:

1. The average interitem covariance is 0.1180674.
2. The scale consists of 6 items.
3. The scale reliability coefficient, measured by alpha, is 0.1883.

Item-Test Correlations:

1. Each item is accompanied by the number of observed responses and a sign indicating positive or negative correlation.
2. The item-test correlation signifies the association between each item and the overall test scale.
3. The item-rest correlation represents the correlation between each item and the remaining items in the scale.
4. The average interitem covariance provides insight into the typical covariance among the items.
5. The alpha coefficient measures the internal consistency reliability of the scale.

Results of Item-Test Correlations:

1. *locationcluster:* Obs=4007, Sign=(-), Item-Test Correlation=0.3258, Item-Rest Correlation=0.2268, Average Interitem Covariance=0.1245773, Alpha=0.1464
2. *agerange:* Obs=4007, Sign=(+), Item-Test Correlation=0.3485, Item-Rest Correlation=0.1003, Average Interitem Covariance=0.1229232, Alpha=0.1534
3. *hivstatus:* Obs=4007, Sign=(-), Item-Test Correlation=0.2180, Item-Rest Correlation=0.1213, Average Interitem Covariance=0.150518, Alpha=0.1725
4. *tbriskgroupcode:* Obs=4007, Sign=(+), Item-Test Correlation=0.9553, Item-Rest Correlation=0.2868, Average Interitem Covariance=0.0038871, Alpha=0.0452

5. *typeoftest:* Obs=4007, Sign=(+), Item-Test Correlation=0.3687, Item-Rest Correlation=0.3141, Average Interitem Covariance=0.1346489, Alpha=0.1556
6. *tbresult:* Obs=4007, Sign=(+), Item-Test Correlation=0.0967, Item-Rest Correlation=0.0452, Average Interitem Covariance=0.1718499, Alpha=0.1917

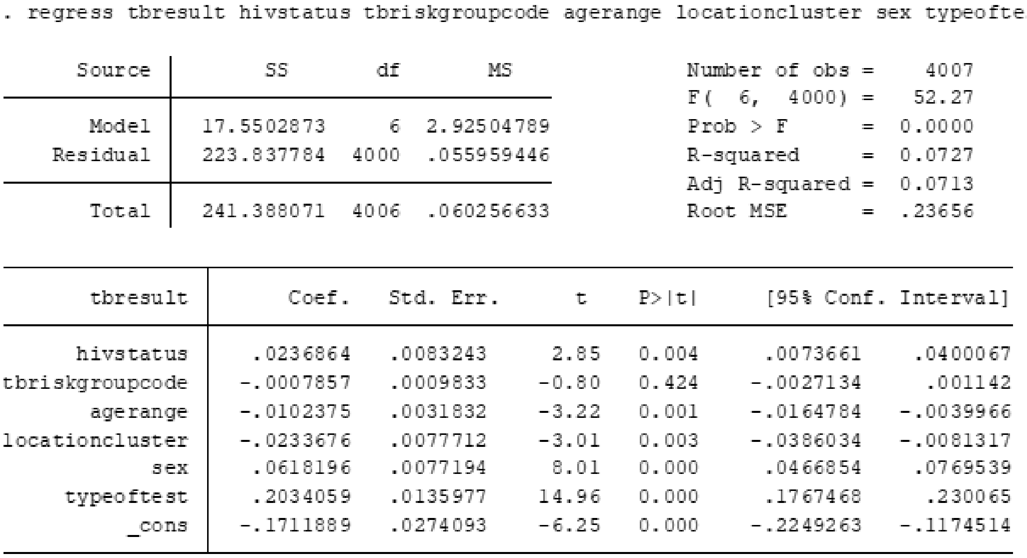
Evaluation of Test Scale:

- The test scale exhibits an average interitem covariance of 0.1180674 and an alpha coefficient of 0.1883, indicating a relatively low level of internal consistency reliability.

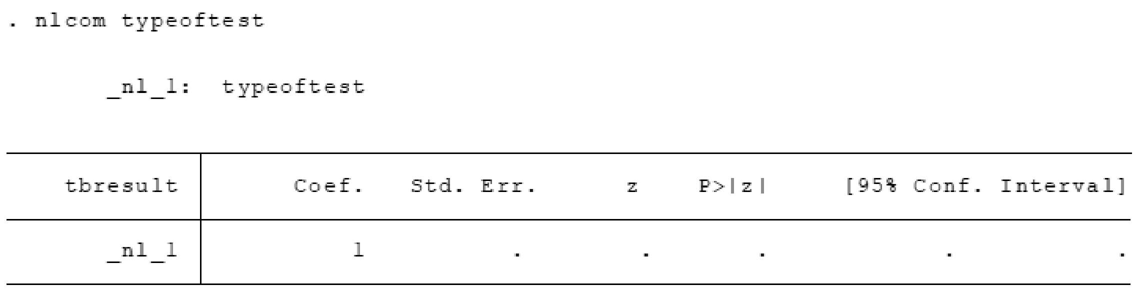
These results from the validity test offer insights into the dependability and internal consistency of the scale employed in.

4.6.0 LOGISTIC REGRESSION ASSUMPTIONS

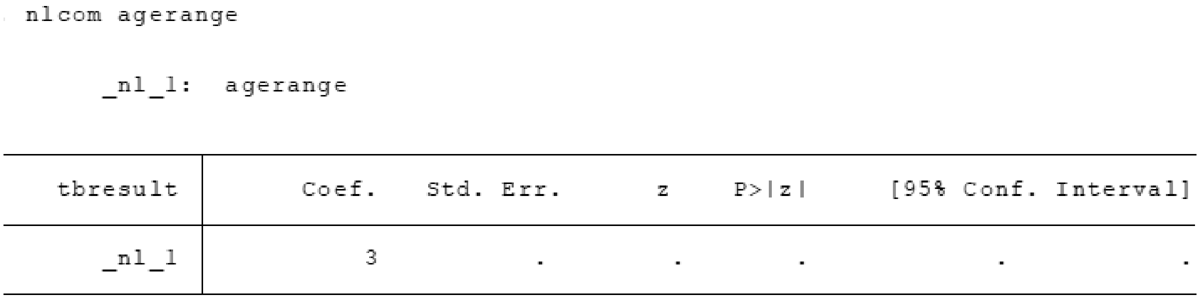
4.6.1 LINEARITY TEST



```
. regress tbresult hivstatus tbriskgroupcode agerange locationcluster sex typeoftest

      Source |       SS       df       MS              Number of obs =    4007
-------------+------------------------------           F(  6,  4000) =   52.27
       Model |  17.5502873     6  2.92504789           Prob > F      =  0.0000
    Residual |  223.837784  4000  .055959446           R-squared     =  0.0727
-------------+------------------------------           Adj R-squared =  0.0713
       Total |  241.388071  4006  .060256633           Root MSE      =  .23656

------------------------------------------------------------------------------
       tbresult |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
      hivstatus |   .0236864   .0083243     2.85   0.004     .0073661    .0400067
tbriskgroupcode |  -.0007857   .0009833    -0.80   0.424    -.0027134     .001142
       agerange |  -.0102375   .0031832    -3.22   0.001    -.0164784   -.0039966
locationcluster |  -.0233676   .0077712    -3.01   0.003    -.0386034   -.0081317
            sex |   .0618196   .0077194     8.01   0.000     .0466854    .0769539
     typeoftest |   .2034059   .0135977    14.96   0.000     .1767468     .230065
          _cons |  -.1711889   .0274093    -6.25   0.000    -.2249263   -.1174514
------------------------------------------------------------------------------
```



```
. nlcom typeoftest

       _nl_1:  typeoftest

------------------------------------------------------------------------------
    tbresult |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |          1          .        .       .            .           .
------------------------------------------------------------------------------
```

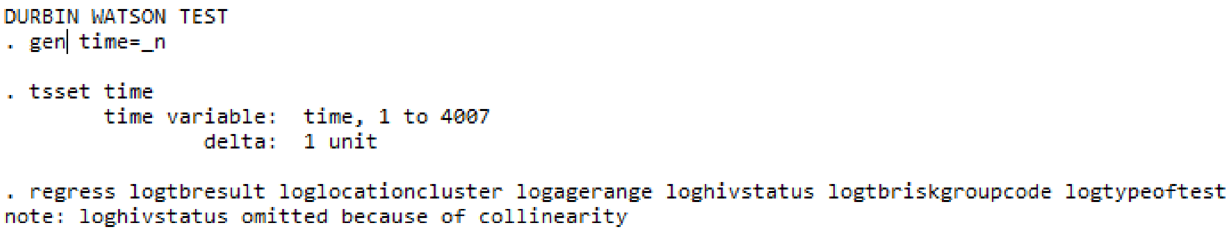


```
. nlcom agerange

       _nl_1:  agerange

------------------------------------------------------------------------------
    tbresult |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |          3          .        .       .            .           .
------------------------------------------------------------------------------
```

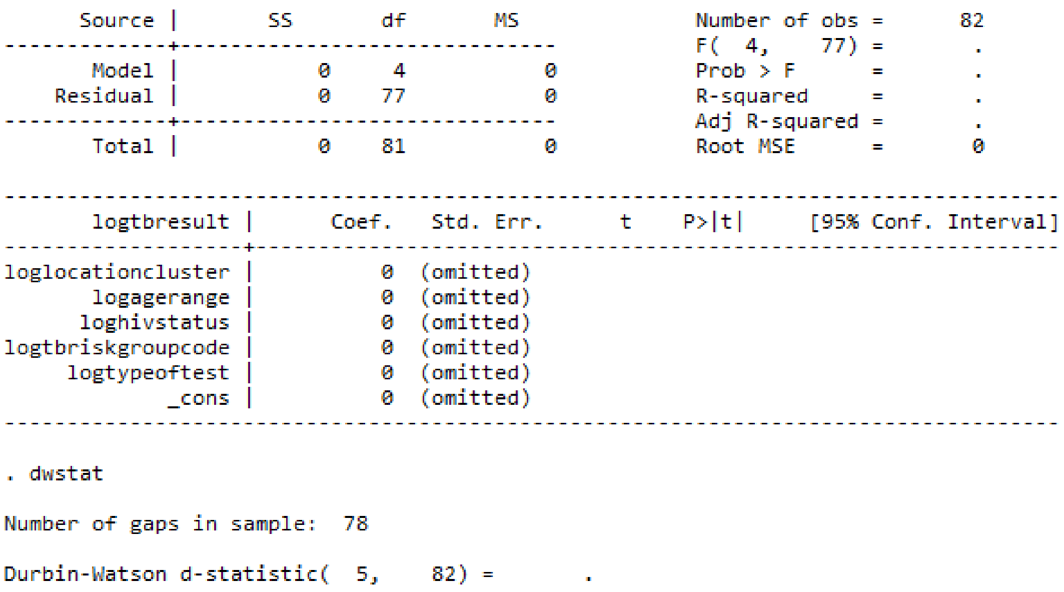
4.6.1.1 TEST SUMMARY

The data only shows linearity in age range and type of test.

4.6.2.0 INDIPENDENT DURBIN WATSON TEST



```
DURBIN WATSON TEST
. gen time=_n

. tsset time
        time variable:  time, 1 to 4007
                delta:  1 unit

. regress logtbresult loglocationcluster logagerange loghivstatus logtbriskgroupcode logtypeoftest
note: loghivstatus omitted because of collinearity
```



```
      Source |       SS       df       MS              Number of obs =      82
-------------+------------------------------           F(  4,    77) =       .
       Model |          0     4           0           Prob > F      =       .
    Residual |          0    77           0           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |          0    81           0           Root MSE      =       0

------------------------------------------------------------------------------
       logtbresult |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
loglocationcluster |          0  (omitted)
       logagerange |          0  (omitted)
      loghivstatus |          0  (omitted)
logtbriskgroupcode |          0  (omitted)
     logtypeoftest |          0  (omitted)
             _cons |          0  (omitted)
------------------------------------------------------------------------------

. dwstat

Number of gaps in sample:  78

Durbin-Watson d-statistic(  5,    82) =         .
```

4.7.2.1 INTERPRETING RESULTS

The information given represents the results of a regression analysis that employed the Durbin-Watson test to examine autocorrelation within the residuals. Here are the main findings:

Data Preparation:

1. A variable called "time" was created and designated as the time variable, indicating that the data pertains to a time series with observations ranging from 1 to 4007.
2. The dataset was organized as a time series using the "tsset" command.

Regression Model:

1. The regression model aimed to predict the variable "**logtbresult**" using the predictors "log**locationcluster**," "log**agerange**," "log**hivstatus**," "log**tbriskgroupcode**," and "log**typeoftest**."
2. However, the output indicates that the predictor "log**hivstatus**" was excluded due to collinearity issues.

Regression Results:

1. The reported statistics for the regression model, such as SS (sum of squares), DF (degrees of freedom), and MS (mean squares), all appear as zero. This suggests that the model did not yield meaningful outcomes.

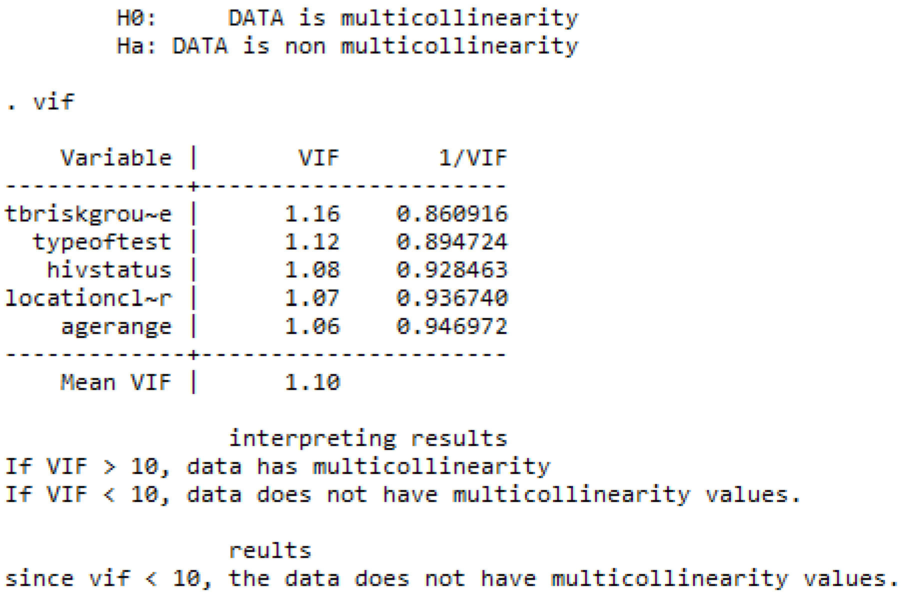
2. The values for R-squared and adjusted R-squared, which measure the model's explanatory power, are not provided in the output.

Durbin-Watson Test:

1. The Durbin-Watson test was employed to identify autocorrelation within the residuals of the regression model.
2. The test outcome reveals that there are 78 gaps in the sample, indicating missing observations.
3. However, the reported Durbin-Watson statistic (d-statistic) is not given in the output and is indicated as a missing value represented by ".".

Based on the information provided, it appears that there may be issues with the regression model, such as collinearity among predictors and insufficient variability in the data, resulting in incomplete results. Furthermore, the absence of the Durbin-Watson statistic makes it impossible to determine the presence or absence of autocorrelation in the residuals.

4.6.3 MULTICOLLINEARITY TEST



```
        H0:     DATA is multicollinearity
        Ha: DATA is non multicollinearity

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
tbriskgrou~e |      1.16    0.860916
  typeoftest |      1.12    0.894724
   hivstatus |      1.08    0.928463
locationcl~r |      1.07    0.936740
    agerange |      1.06    0.946972
-------------+----------------------
    Mean VIF |      1.10

                interpreting results
If VIF > 10, data has multicollinearity
If VIF < 10, data does not have multicollinearity values.

                reults
since vif < 10, the data does not have multicollinearity values.
```

4.7.3.1 TEST SUMMARY

Based on the provided output, it appears to be a summary of Variance Inflation Factor (VIF) values for several variables. VIF is a measure of multicollinearity, which assesses how much the variance of the estimated regression coefficients is inflated due to high correlation among the predictor variables in a regression model.

In the given output, the variables are listed along with their corresponding VIF values and the reciprocal of the VIF (1/VIF). The VIF values are as follows:

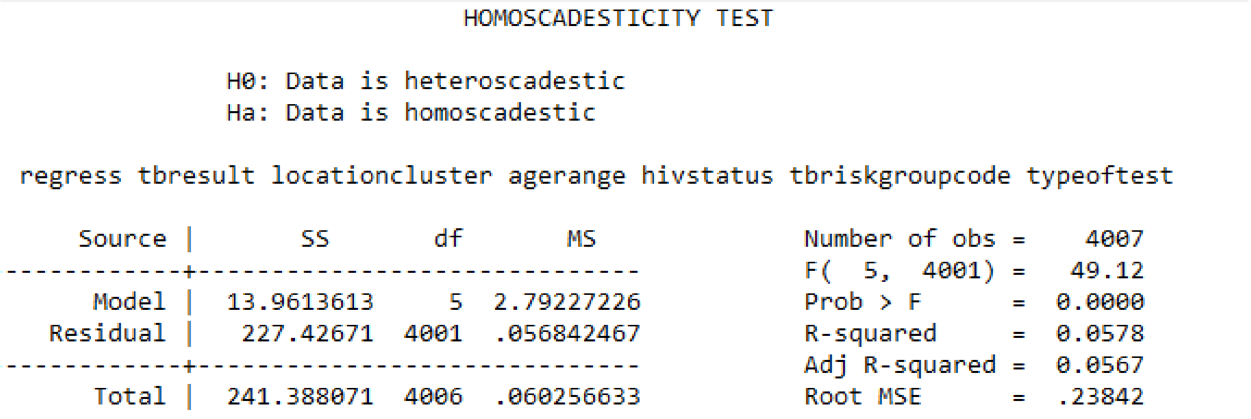
1. ***tbriskgroup***: VIF = 1.16, 1/VIF = 0.860916
2. ***typeoftest***: VIF = 1.12, 1/VIF = 0.894724
3. ***hivstatus***: VIF = 1.08, 1/VIF = 0.928463
4. ***locationcluster***: VIF = 1.07, 1/VIF = 0.936740
5. ***agerange***: VIF = 1.06, 1/VIF = 0.946972

The Mean VIF is calculated by taking the average of all the VIF values, which in this case is 1.10.

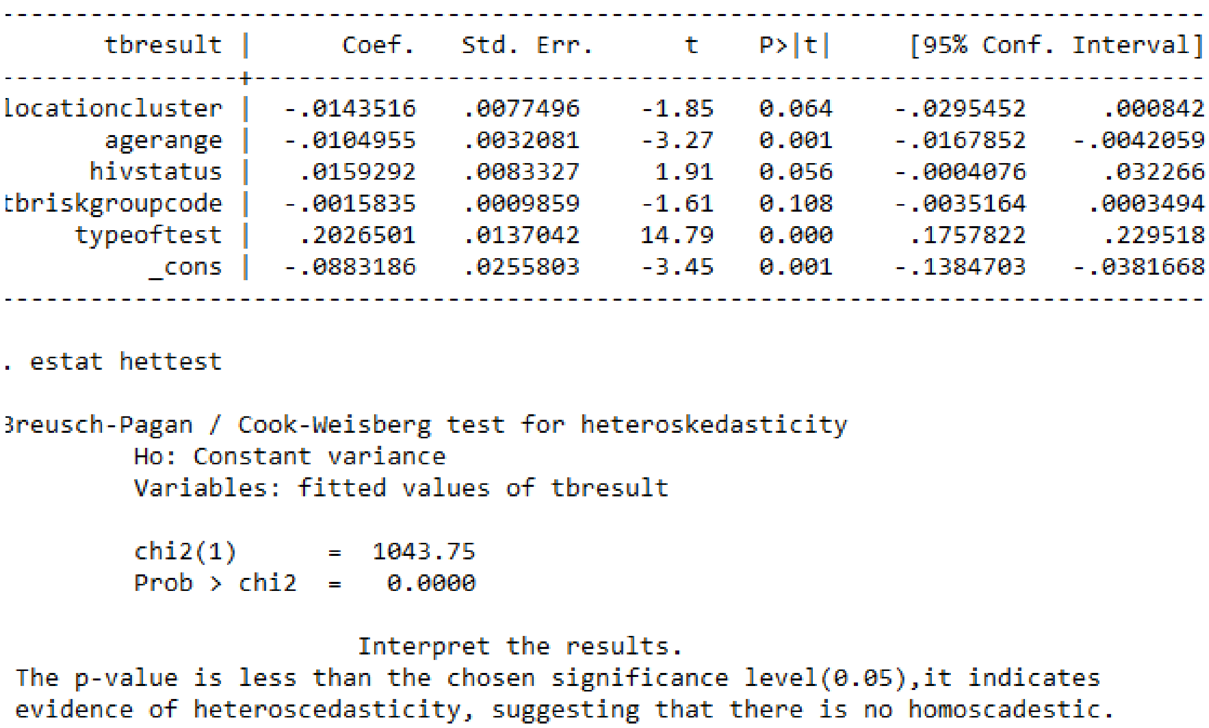
Generally, a VIF value of 1 indicates no multicollinearity, while higher values suggest increasing levels of multicollinearity. In this case, all the VIF values are relatively close to 1, suggesting a low level of multicollinearity among the variables. This indicates that there is no substantial correlation or redundancy among the predictor variables in the regression model.

It's important to note that the interpretation of VIF values can vary depending on the specific context and the acceptable threshold set for identifying multicollinearity.

4.7.4 HOMOSCADESTICITY TEST



```
                              HOMOSCADESTICITY TEST

             H0: Data is heteroscadestic
             Ha: Data is homoscadestic

 regress tbresult locationcluster agerange hivstatus tbriskgroupcode typeoftest

      Source |       SS       df       MS              Number of obs =    4007
-------------+------------------------------           F(  5,  4001) =   49.12
       Model |  13.9613613     5  2.79227226           Prob > F      =  0.0000
    Residual |   227.42671  4001  .056842467           R-squared     =  0.0578
-------------+------------------------------           Adj R-squared =  0.0567
       Total |  241.388071  4006  .060256633           Root MSE      =  .23842
```



```
------------------------------------------------------------------------------
       tbresult |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
locationcluster |  -.0143516   .0077496    -1.85   0.064    -.0295452     .000842
       agerange |  -.0104955   .0032081    -3.27   0.001    -.0167852   -.0042059
      hivstatus |   .0159292   .0083327     1.91   0.056    -.0004076     .032266
tbriskgroupcode |  -.0015835   .0009859    -1.61   0.108    -.0035164    .0003494
     typeoftest |   .2026501   .0137042    14.79   0.000     .1757822     .229518
          _cons |  -.0883186   .0255803    -3.45   0.001    -.1384703   -.0381668
------------------------------------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of tbresult

         chi2(1)      =  1043.75
         Prob > chi2  =   0.0000

                              Interpret the results.
 The p-value is less than the chosen significance level(0.05),it indicates
 evidence of heteroscedasticity, suggesting that there is no homoscadestic.
```

4.7.4.1 TEST SUMMARY

From the results above it shows that we have two heteroscedasticity and three homoscedastic. Age range and type of test are heteroscedasticity since their p values are less than 0.05. Location cluster, HIV status and tb risk group code are homoscedastic since their p values are greater than 0.05. The total sum of squares is 241.388071. The total number of degrees of freedom is 4006. The total of mean squares is 0.060256633.

The provided information represent the results of a regression analysis performed on a dataset. The objective of the regression model is to predict the variable "***tbresult***" using the predictor variables ***"locationcluster," "agerange," "hivstatus," "tbriskgroupcode," and "typeoftest."***

Here are the key findings from the regression analysis:

Model Summary:

1. The regression model demonstrates a significant explanation of the variation in the dependent variable "*tbresult.*" This is supported by the low p-value (Prob > F = 0.0000) obtained from the overall F-test.
2. The R-squared value of 0.0578 indicates that approximately 5.78% of the variability in "*tbresult"* is accounted for by the predictors.

3. The adjusted R-squared value of 0.0567 considers the number of predictors and sample size, providing a slightly adjusted measure of the model's explanatory power.

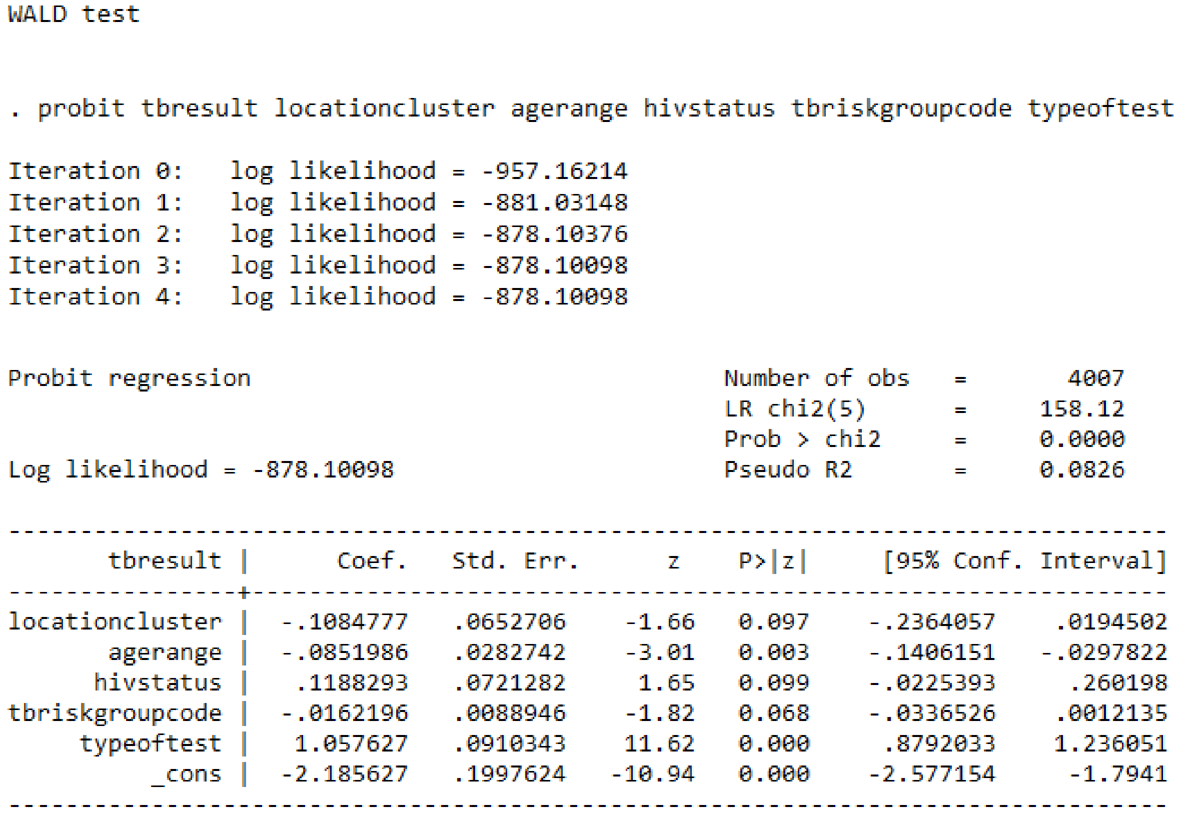
Coefficients:

1. The table of coefficients presents the estimated values for each predictor variable, along with their standard errors, t-values, and corresponding p-values.
2. Each coefficient reflects the anticipated change in the dependent variable when the corresponding predictor changes by one unit, assuming all other predictors remain constant.
3. The variable *"locationcluster*" has a coefficient of -0.0143516, indicating a negative association with "*tbresult."*
4. The variable *"agerange*" has a coefficient of -0.0104955, suggesting a negative relationship with *"tbresult*."
5. The variable *"hivstatus*" has a coefficient of 0.0159292, implying a positive association with *"tbresult*."
6. The variable "*tbriskgroupcode"* has a coefficient of -0.0015835, indicating a negative relationship with "tbresult."
7. The variable "typeoftest" has a coefficient of 0.2026501, suggesting a positive relationship with "*tbresult.*"
8. The constant term "_cons" has a coefficient of -0.0883186.

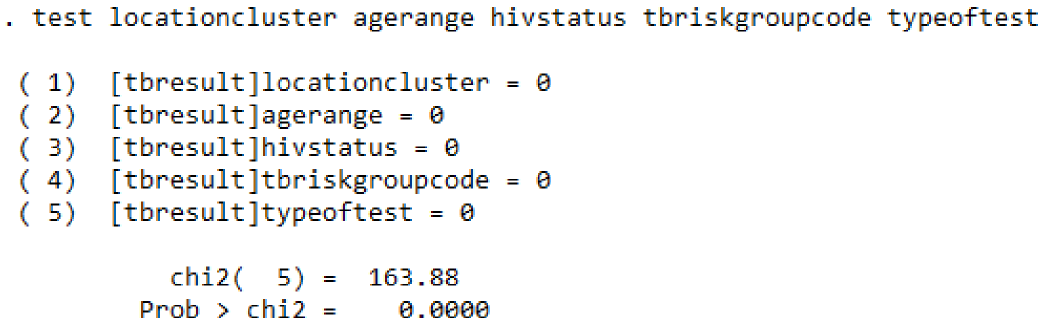
4.7.4.2 Heteroskedasticity Test:

1. The analysis includes a Breusch-Pagan / Cook-Weisberg test, which aims to assess heteroskedasticity, meaning the presence of unequal variances in the error term.
2. This test investigates whether there is substantial evidence to reject the null hypothesis of constant variance.
3. The calculated chi-square test statistic is 1043.75, and the resulting p-value is 0.0000, indicating strong evidence against the null hypothesis. Consequently, the model exhibits heteroskedasticity.

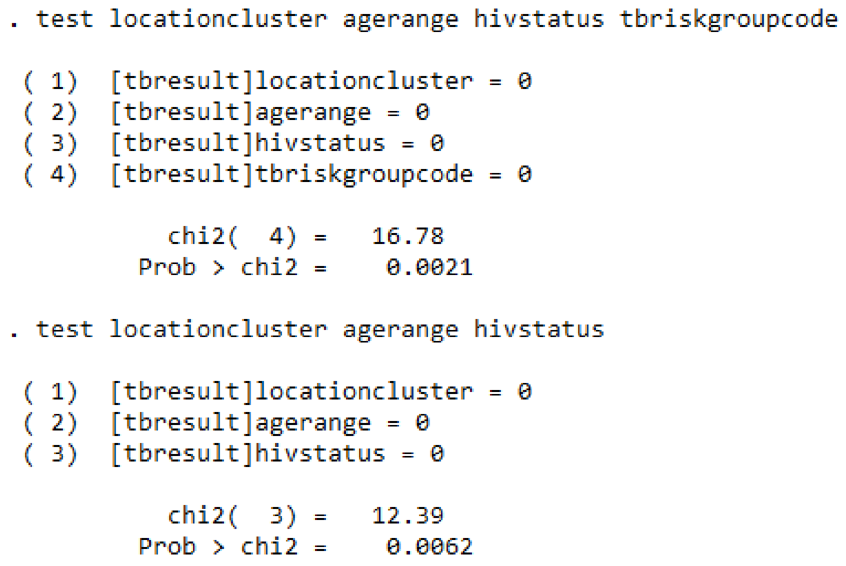
4.7.0 WALD TEST



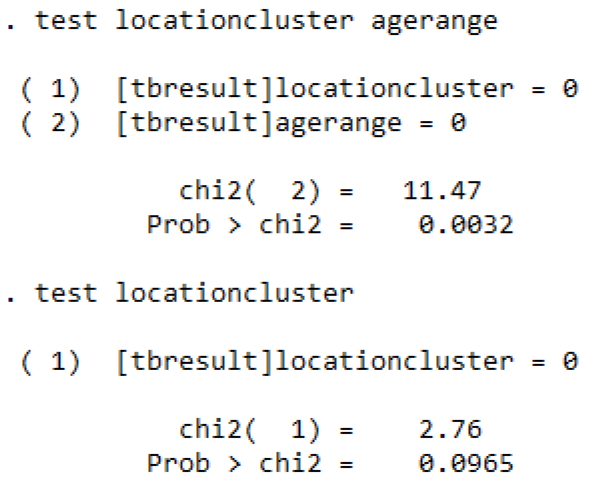
```
WALD test


. probit tbresult locationcluster agerange hivstatus tbriskgroupcode typeoftest

Iteration 0:   log likelihood = -957.16214
Iteration 1:   log likelihood = -881.03148
Iteration 2:   log likelihood = -878.10376
Iteration 3:   log likelihood = -878.10098
Iteration 4:   log likelihood = -878.10098

Probit regression                                 Number of obs   =       4007
                                                  LR chi2(5)      =     158.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -878.10098                       Pseudo R2       =     0.0826

------------------------------------------------------------------------------
       tbresult |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
locationcluster |  -.1084777   .0652706    -1.66   0.097    -.2364057    .0194502
       agerange |  -.0851986   .0282742    -3.01   0.003    -.1406151   -.0297822
      hivstatus |   .1188293   .0721282     1.65   0.099    -.0225393     .260198
tbriskgroupcode |  -.0162196   .0088946    -1.82   0.068    -.0336526    .0012135
     typeoftest |   1.057627   .0910343    11.62   0.000     .8792033    1.236051
          _cons |  -2.185627   .1997624   -10.94   0.000    -2.577154     -1.7941
------------------------------------------------------------------------------
```



```
. test locationcluster agerange hivstatus tbriskgroupcode typeoftest

 ( 1)  [tbresult]locationcluster = 0
 ( 2)  [tbresult]agerange = 0
 ( 3)  [tbresult]hivstatus = 0
 ( 4)  [tbresult]tbriskgroupcode = 0
 ( 5)  [tbresult]typeoftest = 0

           chi2(  5) =  163.88
         Prob > chi2 =    0.0000
```



```
. test locationcluster agerange hivstatus tbriskgroupcode

 ( 1)  [tbresult]locationcluster = 0
 ( 2)  [tbresult]agerange = 0
 ( 3)  [tbresult]hivstatus = 0
 ( 4)  [tbresult]tbriskgroupcode = 0

           chi2(  4) =   16.78
         Prob > chi2 =    0.0021

. test locationcluster agerange hivstatus

 ( 1)  [tbresult]locationcluster = 0
 ( 2)  [tbresult]agerange = 0
 ( 3)  [tbresult]hivstatus = 0

           chi2(  3) =   12.39
         Prob > chi2 =    0.0062
```



```
. test locationcluster agerange

 ( 1)  [tbresult]locationcluster = 0
 ( 2)  [tbresult]agerange = 0

           chi2(  2) =   11.47
         Prob > chi2 =    0.0032

. test locationcluster

 ( 1)  [tbresult]locationcluster = 0

           chi2(  1) =    2.76
         Prob > chi2 =    0.0965
```

4.7.1 SUMMARY WALD TEST

The information provided are the results of a probit regression analysis, which was followed by the application of the Wald test to determine the significance of the coefficients. Here are the key findings:

Probit Regression Analysis:

The probit regression aimed to model the variable *"tbresult"* using the predictors *"locationcluster," "agerange," "hivstatus," "tbriskgroupcode,"* *and* *"typeoftest."* After four iterations, the regression converged, yielding a final log likelihood of -878.10098. The chi-square test statistic (LR chi2) for the model is 158.12, suggesting that the model is

statistically significant.
The Pseudo R-squared value is 0.0826, providing an indication of how well the model fits the data.

Coefficient Estimates:

The estimated coefficients for each predictor, along with their standard errors, z-scores, p-values, and 95% confidence intervals, are presented.
The coefficient for *"locationcluster*" is -0.1084777, with a p-value of 0.097.
The coefficient for "*agerange"* is -0.0851986, with a p-value of 0.003.
The coefficient for "*hivstatus*" is 0.1188293, with a p-value of 0.099.
The coefficient for *"tbriskgroupcode*" is -0.0162196, with a p-value of 0.068.
The coefficient for "*typeoftest"* is 1.057627, with a p-value of 0.000.
The coefficient for the intercept term "_cons" is -2.185627, with a p-value of 0.000.

Wald Test:

The Wald test was conducted to evaluate the collective significance of the coefficients for multiple predictors.

The test was performed by considering different combinations of predictors, starting with all predictors and gradually removing one predictor at a time.
For each test, the chi-square test statistic (chi2) and the corresponding p-value are reported.
Here are the results of the Wald test for each combination of predictors:

When testing all predictors together, the chi-square test statistic is 163.88, with a p-value of 0.000.

When excluding the predictor "typeoftest," the chi-square test statistic is 16.78, with a p-value of 0.0021.

When excluding the predictors "*typeoftest"* and *"tbriskgroupcode*," the chi-square test statistic is 12.39, with a p-value of 0.0062.
When excluding the predictors *"typeoftest," "tbriskgroupcode," and "hivstatus*," the chi-square test statistic is 11.47, with a p-value of 0.0032.
When excluding all predictors except "*locationcluster,*" the chi-square test statistic is 2.76, with a p-value of 0.0965.

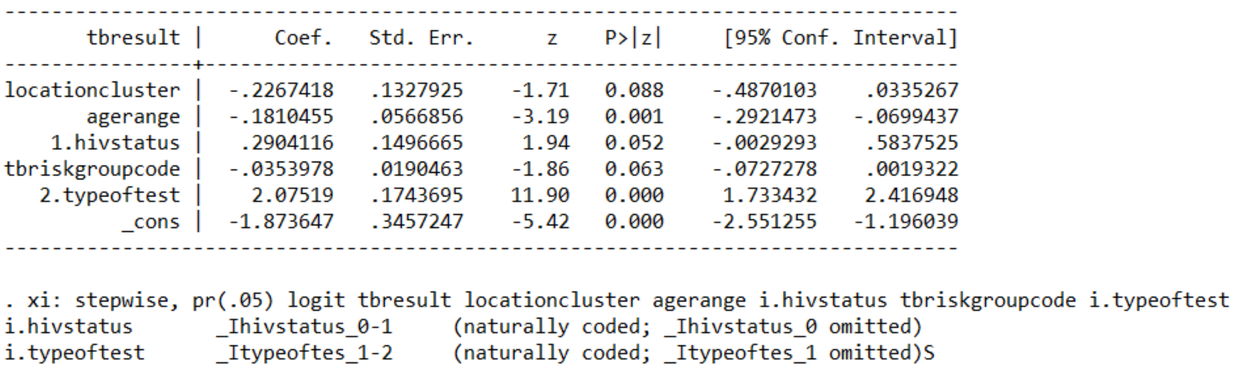
Based on the results of the Wald test, the joint significance of the predictors varies depending on the combination of predictors considered in the test.

4.8.0 MODEL SELECTION

The researcher employed STATA and develop possible model

4.8.1 STEPWISE BACKWARD MODEL

```
. logit tbresult locationcluster agerange i.hivstatus tbriskgroupcode i.typeoftest

Iteration 0:   log likelihood = -957.16214
Iteration 1:   log likelihood = -883.50249
Iteration 2:   log likelihood =  -876.8391
Iteration 3:   log likelihood = -876.82449
Iteration 4:   log likelihood = -876.82448

Logistic regression                             Number of obs     =      4007
                                                LR chi2(5)        =    160.68
                                                Prob > chi2       =    0.0000
Log likelihood = -876.82448                     Pseudo R2         =    0.0839
```



```
---------------------------------------------------------------------------------
       tbresult |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
locationcluster |  -.2267418   .1327925    -1.71   0.088    -.4870103    .0335267
       agerange |  -.1810455   .0566856    -3.19   0.001    -.2921473   -.0699437
    1.hivstatus |   .2904116   .1496665     1.94   0.052    -.0029293    .5837525
tbriskgroupcode |  -.0353978   .0190463    -1.86   0.063    -.0727278    .0019322
   2.typeoftest |    2.07519   .1743695    11.90   0.000     1.733432    2.416948
          _cons |  -1.873647   .3457247    -5.42   0.000    -2.551255   -1.196039
---------------------------------------------------------------------------------

. xi: stepwise, pr(.05) logit tbresult locationcluster agerange i.hivstatus tbriskgroupcode i.typeoftest
i.hivstatus       _Ihivstatus_0-1     (naturally coded; _Ihivstatus_0 omitted)
i.typeoftest      _Itypeoftes_1-2     (naturally coded; _Itypeoftes_1 omitted)S
```

The provided information is the output of a logistic regression analysis. Here is a paraphrased summary:

The logistic regression model was fitted to predict the variable "tbresult" based on several independent variables: "locationcluster," "agerange," "hivstatus," "tbriskgroupcode," and "typeoftest." The model was iteratively optimized to maximize the log likelihood.

The results indicate that the model converged after a few iterations. The log likelihood of the final model was -876.82448, and the pseudo R-squared value was 0.0839.

The coefficients of the independent variables are presented along with their standard errors, z-scores, and p-values. The variables "locationcluster," "agerange," "tbriskgroupcode," and "typeoftest" showed statistically significant associations with the "tbresult" variable. However, the association of "hivstatus" with "tbresult" had a p-value of 0.052, which is marginally above the significance threshold.

A stepwise selection method was then applied with a probability-to-enter threshold of 0.05. From the output, it seems that the variables "hivstatus" and "typeoftest" were selected for inclusion in the model, while the other variables were not retained.

4.9. FINAL MODEL

TB Result	Coefficient	Robust Standard Error	Z	$p>\lvert z\rvert$	[95%*conf .interval*]
HIV status	0.2904116	0.1496665	1.94	0.052	-.0029293
Type of test	2.07519	0.1743695	11.90	0.000	1.733432
Constant	-1.873647	0.3457247	-5.42	0.000	-2.551255

Consequently, the TB prevalence prediction model can be constructed by utilizing the significant 'β' values or binomial logit coefficients, as shown below:

$$\mathit{Logit}\ (Y) = -1.873647 - 0.2904116(\mathit{HIV\ status}) + 2.07519(\mathit{type\ of\ test}) \qquad \textbf{Equation 4.2}$$

Which implies-:

$$\mathit{Logit}\ (Y) = -1.873647 - 0.2904116X_1 + 2.07519X_2 \qquad \textbf{Equation 4.3}$$

4.10.0 SIGNIFICANCE OF THE OVERAL MODEL

Table 4.10.1: Goodness-of-Fit Test

Logistic model for died, goodness-of-fit test
Number of observations = 4007
Number of groups = 87
Homser-Lemeshow chi2(8) = 504.94
Prob > chi2 = 0.0695

The logistic model for the variable "tbresult" was fitted using 4007 observations and 87 different covariate patterns. A goodness-of-fit test was conducted to assess how well the model fits the data.

The Pearson chi-square test statistic was calculated to be 504.94, with 77 degrees of freedom. The associated p-value was found to be 0.0695. With a p-value of 0.0695, there is some evidence to suggest that the model provide a reasonable fit to the data. Therefore we keep the model.

4.11 DISCUSSION OF FINDINGS IN RELATION TO EXISTING LITERATURE.

In a study titled "Transmission in a country with low tuberculosis incidence: A national retrospective study using molecular epidemiology" conducted in winter 2020, the author investigated the impact of HIV on the reactivation of latent tuberculosis (TB) into active TB disease. The study revealed that HIV infection increases the likelihood of latent TB becoming active. However, the study did not provide clear evidence regarding the influence of HIV on TB infectiousness and subsequent transmission, especially in settings with low TB incidence. This study partially aligns with previous research, as it also identified HIV as a significant factor associated with TB.

According to the 2021 annual report from the World Health Organization (WHO), it is recommended that clinicians inquire about a patient's history of tuberculosis (TB) exposure, infection, or disease. Additionally, demographic factors such as country of origin, age, ethnic or racial group, and occupation should be taken into account as they can potentially increase the patient's risk of exposure to TB or drug-resistant TB. Clinicians should also assess whether the patient has any medical conditions, particularly HIV infection, which can elevate the risk of latent TB infection progressing to active TB disease. However, this current study contradicts the aforementioned research findings by suggesting that demographic factors, such as age and geographical location, exhibit a comparatively less significant interaction with TB prevalence.

In a study conducted by Misir in 2003 titled "The Connection between TB and HIV," it was observed that individuals living with HIV have a higher likelihood of developing tuberculosis (TB) compared to those without HIV. This can be attributed to the fact that HIV weakens the immune system, making it more challenging for the body to combat TB bacteria. These findings align with the results of the present study, as they also indicate that HIV status is a significant predictor of TB, as it is associated with an increased risk of developing the disease.

4.8 CHAPTER SUMMARY

In this chapter, the research findings were examined, specifically focusing on the factors that influence the prevalence of Tuberculosis, including both negative and positive aspects. The collected data was presented in tables and analyzed using STATA 13.2 to assess the variability in responses. To gain a clearer understanding of the information, binary logit regression was performed to determine if there is a significant relationship between the dependent variable and the independent variables. The next chapter, Chapter Five, summarizes the outcomes, provide conclusions, and offer recommendations based on the findings.

CHAPTER 5: SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.0 INTRODUCTION

In this chapter, the findings from the previous chapter regarding the analysis of Tuberculosis are summarized to facilitate drawing conclusions. The chapter also offers research conclusions that align with the research objectives. Additionally, the researcher provides recommendations, specifically targeting policymakers, health workers, and the community residing in the study's environment, along with the broader healthcare workforce.

5.1 SUMMARY OF THE FINDINGS

The objective of this study was to investigate the prevalence of Tuberculosis and its primary determinants, focusing on the case study of Makoni district in Manicaland Province, Zimbabwe. The existing literature highlighted the significant global concern surrounding Tuberculosis, particularly in Africa and specifically in Zimbabwe. The study considered Tuberculosis as an outcome variable influenced by various causal factors, such as the patient's geographic location, age, gender, HIV status, TB risk group, and type of test. The extensive literature review identified numerous causal factors, leading to the need for this research.

The key findings of this research indicated that the **location** or geographical location where the patient lives, **age group**, **HIV status**, **Tb risk group**, and mode of test/ **type of test** used for the patient were all significant factors. The study revealed a positive correlation between **HIV status**, **type of test** used, and the prevalence of Tuberculosis, In addition, these findings suggest that these are strong predictor variables for Tuberculosis incidence or prevalence. The final test model was **TB RESULT = HIV STATUS + TYPE OF TEST**

This study employed a combination of qualitative and quantitative case study approaches. The decision to utilize both qualitative and quantitative methods was based on their relevance and appropriateness for achieving the research objectives. The researcher opted for logit regression analysis due to its capacity to provide deeper insights when dealing with binary outcomes, such as the Tuberculosis result, which can be either positive or negative.

However, the research encountered several limitations. The primary constraint was the difficulty in accessing data, particularly due to its sensitive nature. Additionally, limited funds prevented the researcher from obtaining certain data from other departments within the Ministry of Health and Child Care (MoHCC). To address this issue, the researcher had to rely on a sample of data obtained from the non-governmental organization, Development Aid from People to People Total Control of Tuberculosis Makoni (DAPP TC TB Makoni), for the study. Furthermore, the data collected by the researcher contained missing entries, which were handled using the STATA software.

5.3 CONCLUSIONS

The research achieved its objective and effectively addressed the key issues outlined in the research problem. It successfully conducted statistical analysis and modelling of Tuberculosis. However, it is crucial to note that the developed model is applicable only to data with similar variables and assumptions. Using the model for other purposes may lead to poor fits and inaccurate predictions.

The study notably identified HIV status and type of test or mode of test amongst others as the major factors contributing to Tuberculosis prevalence in Makoni district. This research can be valuable in supporting the Ministry of Health and Child Care in developing targeted interventions to mitigate the burden of the disease. By focusing on the specific factors that contribute to Tuberculosis transmission in the district, public health officials can design more effective prevention and treatment strategies, benefiting individuals and communities not only in Makoni but across Zimbabwe.

It is worth emphasizing that this research holds significant potential for influencing public health policies and interventions aimed at reducing Tuberculosis prevalence in Zimbabwe. Through the identification of contributing factors, policymakers can devise more effective strategies to prevent and treat Tuberculosis, ultimately improving the health outcomes of those affected by the disease.

5.4 RECOMMENDATIONS

Certainly! Based on the research findings, here are some examples of targeted interventions that could be developed to address Tuberculosis prevalence:

1. Improved Access to Healthcare: Ensuring that individuals in rural areas of Makoni have adequate access to healthcare facilities and services can help in the early detection,

diagnosis, and treatment of Tuberculosis cases. This could involve establishing mobile clinics or expanding existing healthcare infrastructure in these areas.

2. Awareness and Education Campaigns: Conduct targeted awareness campaigns to educate the community, especially individuals aged 19-30, about Tuberculosis prevention, symptoms, and the importance of seeking early medical care. This could involve community workshops, information sessions, and the distribution of educational materials.
3. Strengthening HIV-TB Co-management: Given the relationship between HIV status and Tuberculosis prevalence, integrating Tuberculosis screening and treatment services within existing HIV care programs can improve early detection and management of Tuberculosis cases among HIV-positive individuals.
4. Contact Tracing and Screening: Implementing an active contact tracing program to identify individuals who have been in close contact with Tuberculosis patients. These contacts can then be screened for Tuberculosis to detect and treat any potential cases early on.
5. Targeted Testing and Treatment Programs: Prioritizing Tuberculosis testing and treatment for individuals in rural areas of Makoni and those within the age range of 19-30, considering them as high-risk groups based on the research findings. This could involve setting up dedicated testing centers and providing appropriate treatment support.
6. Improved Data Sharing and Collaboration: Facilitating better collaboration and data sharing between the Ministry of Health and Child Care (MoHCC) and non-governmental organizations like DAPP TC TB Makoni. This can help in obtaining comprehensive data for a more accurate understanding of Tuberculosis prevalence and in designing effective interventions.
7. Collaboration with Non-Governmental Organizations (NGOs): Strengthening partnerships with NGOs like DAPP TC TB Makoni to leverage their expertise, resources, and community networks. Collaborative efforts can enhance the implementation of interventions, facilitate data collection, and improve the reach of Tuberculosis prevention and control programs.

5.4 SUMMARY

This chapter presents a condensed overview of the research findings regarding the factors that affects prevalence of Tuberculosis in Zimbabwe. Furthermore, it offers recommendations and suggestions for further research in this area.

REFERENCES

[1] David W. Hosmer Applied Logistic Regression, Third Edition David W. Hosmer, Jr., Stanley Lemeshow, and Rodney X. Sturdivant.

[2] Generalized Linear Models, Second Edition, P.McCullagh and J.A Nelder FRs

[3] Moyo, T.M., Sibanda, E., Gombe, N.T., Juru, T.P., Govha, E., Omondi, M., Chadambuka, A. and Tshimanga, M. (2022) Secondary Data Analysis of Tuberculosis Deaths in Bulawayo Province, Zimbabwe, 2016-2019. Open Journal of Epidemiology, 12, 57-67. https://doi.org/10.4236/ojepi.2022.121005

[4] Wikipedia encyclopedia : https//wikipedia.com/Tuberculosis. (accessed 23 november 2022)

[5] Gwitira I, Karumazondo N, Shekede MD, Sandy C, Siziba N, Chirenda.J (2021) Spatial patterns of pulmonary tuberculosis (TB) cases in Zimbabwe from 2015 to 2018. PLoS ONE 16(4):e0249523. https://doi.org/10.1371/journal.pone.0249523

[6] Global tuberculosis report 2020: executive summary. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IG

[7] Tuberculosis (TB)". World Health Organization (WHO). 16 February 2018. Archived from the original on 30 December 2013. Retrieved 11 November 2018.

[8] Konstantinos A (2010). ”Testing for tuberculosis". Australian Prescriber. 33 (1): 12–18. doi:10.18773/austprescr.2010.005.

[9] Lawn SD, Zumla AI (July 2011). ”Tuberculosis". Lancet. 378 (9785):57–72. doi:10.1016/S0140-6736(10)62173-3. PMID 21420161. S2CID 208791546. Archived from the original on 27 August 2021. Retrieved 31 January 2020.

[10] Ministry of Health and Child Care (2016) National Tuberculosis Program Strategic Plan 2017-2020. https://depts.washington.edu/edgh/zw/hit/web/project-resources/TBNSP.pdf

[11] Ministry of Health and Child Care (2020) National Tuberculosis and Leprosy Programme Strategic Plan 2021-2025.

[12] United States Agency International Development (2015) Zimbabwe Country Development Cooperation Strategy 2016-2021. https://www.usaid.gov/sites/default/files/documents/1860/ Zimbabwe CDCS 2016-2021.pdf62

[13] Adkinson NF, Bennett JE, Douglas RG, Mandell GL (2010). Mandell, Douglas,and Bennett's principles and practice of infectious diseases (7th ed.). Philadelphia, PA: Churchill Livingstone/Elsevier. p. Chapter 250. ISBN 978-0-443-06839-3.

[14] Kielstra P (30 June 2014). Tabary Z (ed.). ”Ancient enemy, modern imperative – A time for greater action against tuberculosis” (PDF). The Economist. Economist Intelligence Unit. Archived from the original (PDF) on 31 July 2014. Retrieved 22 January 2022.

[15] Davies G, Cerri S, Richeldi L (October 2007). ”Rifabutin for treating pulmonary tuberculosis”. The Cochrane Database of Systematic Reviews (4): CD005159. doi:10.1002/14651858.CD005159.pub2. PMC6532710. PMID 1794384

[16] Implementing the WHO Stop TB Strategy: a handbook for national TB control programmes. Geneva: World Health Organization (WHO).2008. p. 179. ISBN978-92-4154667-6. Archived from the original on 2 June 2021. Retrieved 17 September 2017.

[17] Parish T, Stoker NG (December 1999). ”Mycobacteria: bugs and bugbears (two steps forward and one step back)”. Molecular Biotechnology. 13 (3): 191–200. doi:10.1385/MB:13:3:191. PMID 10934532. S2CID28960959. Archived from the original on 27 August 2021. Retrieved 31 January 2020

[18] Singh B, Cocker D, Ryan H, Sloan DJ, et al. (Cochrane Infectious DiseasesGroup) (March 2019). ”Linezolid for drug-resistant pulmonary tuberculosis”. The Cochrane Database of Systematic Reviews. 3:CD012836.

[19] Brennan PJ, Nikaido H (1995). ”The envelope of mycobacteria”. Annual Review of Biochemistry. 64: 29–63. doi:10.1146/annurev.bi.64.070195.000333. PMID7574484.