Predicting District Level HIV Prevalence in South Africa Using Medicine Ordering Data

Author: Juandre Liebenberg (LBNJUA001)

Minor Dissertation presented in partial fulfilment of the requirements for the degree of

Master of Science (Data Science)

in the Faculty of Science at the University of Cape Town

Supervisor: A/Prof Sheetal Silal

Co-supervisor: Dr Şebnem Er

September 5, 2024

Plagiarism Declaration

I, Juandre Liebenberg, know the meaning of plagiarism and declare that all of the work in the minor dissertation, save for that which is properly acknowledged, is my own.

Signed by candidate

Signed

2024/09/05

Date

Abstract

The Human Immunodeficiency Virus has been at the forefront of South Africa's public health challenges, placing the healthcare system under immense pressure. As a result of HIV planning by policymakers, more than 5.5 million People Living with HIV have access to antiretroviral treatment at present day. Dynamic, mechanistic models such as the Thembisa and Naomi Bayesian models have been used to generate provincial and district-level estimates such as HIV prevalence, People Living with HIV, and the number of residents on antiretroviral treatment. An alternative methodology for estimating drug utilisation and predicting HIV estimates was explored by using medicine ordering data as the primary input for analysis from 2020 to 2022. Two objectives were set out, the first being a drug utilisation analysis aimed at approximating the number of individuals per 1000 inhabitants per day taking antiretroviral drugs to determine if the adequate stock was ordered at district and provincial levels. The second was to predict HIV prevalence by fitting panel data and spatial linear models to predict district prevalence and People Living with HIV; the estimations for People Living with HIV were converted to prevalence to compare the direct estimation of prevalence to the calculated. Results from the drug utilisation analysis suggested that district municipalities hold insufficient stock to meet the demands of those inflicted with the disease. In contrast, larger metropolitan municipalities hold excess medication, implying that people travel across district boundaries to receive treatment. The fitted spatial models generated better prevalence estimates than fixed-effect panel data models for the predicted and calculated prevalence with root mean square error metrics of 0.009 (0.87%) and 0.012(1.24%) compared to that of 0.012(1.21%) and 0.015(1.53%) from the fixed-effect panel data models. The impact of high quantities of antiretroviral drugs ordered by metropolitan municipalities resulted in an underestimation of prevalence in those regions due to the negative relationship between the dependent variable Prevalence and the independent Quantity variable. From the spatial models fitted, the best performing spatial model accurately estimated the prevalence rates for 51 out of 52 districts, which fell within the acceptable range defined by the Naomi Model. The results of the study have shown that the use of ordering data to predict disease prevalence has the potential to serve as an alternative methodology in the absence of established models.

Contents

List of Figures

List of Tables

List of Abbreviations

ARV: **A**ntiretroviral
ART: **A**ntiretroviral **T**herapy
ATC: **A**natomical **T**herapeutic **C**hemical
HIV: **H**uman **I**mmunodeficienc **V**irus
PLHIV: **P**eople **L**iving **HIV**
TEE: **T**enofovir **E**mtricitabine and **E**favirenz
TLD: **T**enofovir **L**amivudine and **D**olutegravir
WHO: **W**orld **H**ealth **O**rganisation

Chapter 1

Introduction

1.1 Background

Estimates predict that 7.5 million people living with Human Immunodeficiency Virus (HIV) reside in South Africa, making it the country with the highest number of people infected with the disease. The South African National HIV Prevalence, Incidence, Behaviour and Communication Survey of 2017 reported the country's HIV prevalence at 14%, amounting to 1 in 7 people on average affected by the disease(Simbayi et al., 2017).

In 2004 the country's antiretroviral therapy (ART) programme was launched to assist in managing infection rates and improving the lives of People Living with HIV (PLHIV). The drive to combat HIV has resulted in approximately 5.5 million patients having received treatment as of 2021 (Kitenge et al., 2023).

The Thembisa and Naomi models have assisted policymakers in making informed decisions regarding HIV planning and resource allocation. Published in 2014, the Thembisa model was developed to assess diagnosis and treatment targets on national and provincial levels, factors accounting for the variations in HIV prevalence between provinces, and the impact of ART on mortality (Johnson and Dorrington, 2021). The Naomi model provides district-level estimates such as HIV prevalence, PLHIV, ART coverage, and new HIV infections outputs for current and one-year projections used for HIV planning (Eaton et al., 2021). The Naomi model is calibrated to the Thembisa model by aggregating district estimates to match the Thembisa provincial estimates to provide a holistic view of the country's HIV landscape.

South African district-level estimates currently produced by the Naomi model require various input data such as household surveys, ART service delivery, and Antenatal clinic (ANC) HIV testing indicators by district and year data to generate HIV estimates in addition to the model outputs from the Thembisa model. These data sources may not readily be available which has opened the possibility of using alternative data sources to predict HIV estimates.

In collaboration with the Clinton Health Access Initiative (CHAI), this study set out to explore whether pharmaceutical sales ordering data may be used to develop a framework for assessing antiretroviral (ARV) drug availability and HIV prevalence on a district level. Using pharmaceutical ordering data, the application of drug utilisation methodologies and statistical modelling techniques were implemented to estimate ARV allocation and HIV prevalence based on the quantities of ARV drugs ordered by health facilities such as district hospitals, clinics, and medical depots over three years from 2020 to 2022.

1.2 Research Aims and Objectives

This study aimed to determine if the ordering of ARV drugs within each district could be used as an alternative data input to analyse ARV availability, as well as estimate HIV prevalence. To achieve this aim, the following objectives were met:

1. Using the Defined Daily Dosage (DDD); the average maintenance dose required per day by an adult. The number of individuals using ARV drugs was approximated using the DDD per 1000 inhabitants per day drug utilisation methodology to calculate the proportion of the population using ARV drugs. This was directly compared to the Naomi model estimates for the number of residents on ART in each district to determine whether a sufficient amount of medication was available to meet the needs of PLHIV.
2. Estimating district-level HIV prevalence using panel data and spatial modelling techniques. By transforming the ordering data to a panel dataset for the years 2020 to 2022 it was possible to study the relationship between HIV prevalence and the quantity of ordered ARV drugs over time. For each modelling technique, two measures of prevalence were obtained; an estimation obtained from predicting prevalence from the quantity of ARV drugs ordered per district, and a measure obtained from calculating prevalence from the estimation of PLHIV from the quantity of ARV drugs ordered per district.

1.3 Research Significance

The significance of the development of a framework to estimate ARV drug availability and HIV prevalence by using medicine ordering data may lead to a simplified and time-effective method for estimating HIV related figures by using ordering data as a viable predictor of the same output as the Naomi model, so that it could be used as an alternative to the data-intensive Naomi model.

In addition, model-based prevalence estimates are not readily available for all health conditions in South Africa. The ordering data contains the ordering information of different medications used to treat various diseases such as diabetes and tuberculosis. This would therefore open the possibility of applying this framework and providing policymakers with a method of gaining insights on diseases where little or no mathematical models exist.

1.4 Thesis Preview

The outline of the thesis and contents of each chapter are as follows:

Chapter 2 - Literature Review

This chapter contains the literature study on various topics relevant to the problem statement. The chapter focuses on providing an overview of HIV and antiretroviral viral therapy in the South African context. The literature review also provides a summary of previous research studies and findings relating to drug utilisation, panel data and spatial modelling.

Chapter 3 - Exploratory Data Analysis

The exploratory data analysis chapter is dedicated to describing the datasets utilised in this study. This includes the characteristics of the data and the pre-processing that was applied to transform the data into a

usable format. The latter part of the chapter focuses on presenting key insights that were used to guide the direction of the study methodology.

Chapter 4 - Methodology

In Chapter 4, the methodology used to achieve the objectives is presented. The methodology contains a detailed description of the DDD per 1000 inhabitants drug utilisation methodology used to calculate ARV availability. An overview of panel data modelling techniques and statistical tests are provided, in addition to the spatial analysis and modelling techniques implemented.

Chapter 5 - Results

This chapter is devoted to presenting and discussing the results of the implemented methodology from Chapter 4. In this chapter, the results of the drug utilisation analysis, and fitted panel data and spatial linear models are interpreted.

Chapter 7 - Discussion Conclusion

This chapter summarises the findings, and how they relate to the study objectives. Lastly, the limitations of the study as well as recommendations are discussed.

Chapter 2

Literature Review

2.1 Introduction

South Africa's first recorded death related to HIV/AIDS was in December of 1981. Even after being classified as an epidemic in 1992, action was only taken in 1994 when the government accepted the national AIDS plan lobbied by the Networking HIV/AIDS Community of South Africa (NACOSA). At this point HIV prevalence had already risen to 4.3% with no sign of slowing down; to where the estimated HIV prevalence for South Africa has risen to 18.3% for ages 15 to 49 according to Naomi model estimation outputs for 2022 (Simelela and Venter, 2014).

Figure 2.1 illustrates the estimated district-level prevalence for ages 15 to 49 taken from the Naomi Model estimates for 2022. It can be observed that the Western Cape which consists of six districts, had the lowest recorded prevalence in the Namakwa District Municipality of 6.5% while the uMkhanyakude Municipal District in KwaZulu-Natal had the highest estimated HIV prevalence of 30.4% which is not surprising when considering its proximity to Eswatini which is the country with the highest average HIV prevalence in the world at 25.9% for adults aged 15 to 49 (UNAIDS, 2023). Other districts such as the Zululand, Ehlanzeni and Gert Sibande district municipalities which border Eswatini have some of the highest prevalence rates in South Africa making this region a HIV hotspot in Southern Africa (Mukuna et al., 2024).

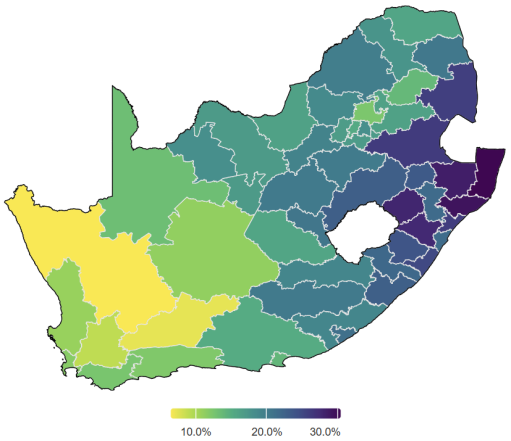


Figure 2.1: Map of South Africa Displaying District-Level HIV Prevalence Produced from the Naomi Model Estimates for 2022 for ages 15-49 (UNAIDS, 2023)

Based on HIV prevalence estimates for 2022, there is variation in prevalence among provinces and districts. Geographical variation among districts is attributed to factors such as population density, socioeconomic factors, and high migration to urban areas by young adults (Van Schalkwyk et al., 2021).

2.2 Antiretroviral Therapy

Antiretroviral therapy treatment has resulted in a significant reduction in HIV-related mortality and morbidity. The development of ART guidelines and treatment has gone through different iterations since the first clinical trials in the mid-1980s. The development of nucleoside reverse transcriptase inhibitors (NRTI), a class of drug that incorporates into viral DNA thus terminating DNA chain growth and viral replication in the human body was the first breakthrough in ART treatment (Vella et al., 2012).

Initial ART treatments however had the drawback of high toxicity levels and adverse side effects; it was not until the development of more tolerable drugs such as lamivudine (3TC) that this improved. Lamivudine was found to be a drug that was synergistic with other nucleosides which allowed for the formulation of safer and easier to adhere to regimens (Quercia et al., 2018).

In the mid-1990s new classes of ARV drugs, namely non-nucleoside reverse transcriptase inhibitors (NNRTI) and Protease Inhibitors (PI) were developed and approved. NNRTIs worked by blocking the reverse transcriptase (RI) enzyme to prevent viral replication, whereas PI drugs inhibited the protease enzyme responsible for cutting viral proteins into smaller pieces required for viral replication (Maggiolo et al., 2017).The introduction of the new drug classes led to combination therapies involving the administration of one or more drugs from each class to create a synergistic effect that would effectively reduce RT activity but also hinder the virus's ability to develop resistance against drugs.

The success of triple drug therapy (a combination of three ARV drugs in specific doses) especially in mother-to-child transmission was reported in 1996; this combination could for example include two NRTIs combined with one drug from the other classes (Vella et al., 2012). Since then, triple drug therapy has become the golden standard and backbone of ART with newer and safer drugs being used in different combinations.

2.2.1 World Health Organisation Guidelines

The World Health Organisation (WHO) reports on standardised ARV drug regimens across all populations by suggesting well-studied treatment regimens; the 2019 WHO guidelines gave a new recommendation for ART treatment (WHO, 2021). The new first-line ART recommendation for adults and adolescents was the use of a combination of dolutegravir (DTG), a class of drugs known as an integrase strand transfer inhibitor (INSTI) that blocks the viral enzyme that allows the viral DNA to insert itself in the host cell DNA, with a NRTI backbone (Two NRTIs).

The alternative, second-line regimen for adults and adolescents was the use of a 400mg efavirenz (EFV) dose in combination with a NRTI backbone. The safety and efficacy of DTG in pregnant women and those coinfected with tuberculosis (TB) have been driving factors in recommending DTG as a first-line ART. Compared to EFV, which is naturally resistant to HIV-2, DTG has a higher genetic barrier to developing drug resistance (Dooley et al., 2020). High to moderate certainty evidence was found with a regimen of two NRTIs and DTG, with higher rates of viral suppression and better adherence than EFV-based regimens.

The adoption of DTG-based regimens has seen a steady incline since its introduction. In 2017 DTG was in the process of being licensed to multiple generic manufacturers to replace EFV over the next few years even though at this point more than 300 000 people initiated ART annually on the efavirenz/tenofovir disoproxil/emtricitabine (EFV/TDF/FTC) fixed-dose, abbreviated as TEE (Venter et al., 2017).

A 2021 study by Patel et al. (2021) had shown that since its introduction, the DTG fixed-dose combination of dolutegravir/lamivudine/ tenofovir alafenamide (DTG/3TC/TAF) commonly referred to as TLD has become pivotal in several ART programmes in sub-Saharan Africa.

The TLD dose comprises a fixed dose of 50mg DTG, 300mg of TDF, and 300mg of 3TC, prescribed to patients older than 10 years of age and weighing more than 35kg (WHO, 2021). Due to the benefits of TLD, South Africa has adopted it as the first line of treatment for HIV as per the WHO recommendation as of 2019.

2.3 Defined Daily Dosage for Consumption Studies

Approximating district-level ARV drug consumption can be performed using drug utilisation methodologies. These methodologies generally focus on volume, cost, and defined daily dosage (DDD) to calculate region-specific consumption (Truter, 2008).

2.3.1 Defined Daily Dosage

The defined daily dosage (DDD) is a set unit of measurement used in healthcare and pharmacology. The measurement provides a quantity for the average maintenance dose of a drug consumed by an average adult. The DDD can also define a quantifiable value for drug consumption independent of factors such as price, currencies, and package size (WHO, 2003). This measurement provides a method for analysing drug utilisation trends across different regions, populations, and health systems which provides researchers and policymakers with data to assess the effectiveness of treatment regimens.

The DDD is a generalised guideline dosage meant to address drug consumption between different geographical regions, it is however not ideal when estimating utilisation for children as there is too much variability in the dosages given (Zhang et al., 2012). Using the DDD of a drug being studied, the DDD per 1000 inhabitants per day methodology originally developed in Scandinavia has become a strategy used to estimate drug utilisation in a region by using the sales of a drug in a geographical area and dividing that by a per-person denominator; this result would serve as a proxy for drug consumption (Truter, 2008).

2.3.2 DDD Methodology Applications and Studies

From studies reviewed, the use of the DDD per 10000 inhabitants per day methodology has shown to be one of the more popular methods of estimating drug consumption. The first study by Criado-Alvarez et al. (2017) aimed to estimate the prevalence of attention deficit hyperactivity disorder (ADHD) in Castile-La Mancha, Spain for patients aged 5 to 8, while a study by Laugesen et al. (2017) focused on the consumption of glucocorticoids for the entire Danish population.

These studies utilised data recorded over more than 20 years. In the study by Criado-Alvarez et al. (2017) the consumption of ADHD drugs dispensed by pharmacies in the region was calculated for the years 1992 to 2015; methylphenidate, atomoxetine, and lisdexamfetamine were studied since they are usually the first choice of treatment for ADHD.

The results of the study were able to show the pattern of ADHD medication consumption based on the DDD methodology calculations; these showed that an increase of 98.92% was estimated from 1992 to 2015 with 13.22 DDDs calculated for the year 2015. Building on this; an autoregressive integrated moving average (ARIMA) model was built to forecast DDD per 10000 inhabitants per day till 2020 using the years 2000,2009 and 2012 as join points. For 2020, a prevalence of 14.11 DDDs per 1000 inhabitants was estimated, with a 95% confidence interval ranging from 10.18 to 18.06 DDDs per 1000 inhabitants.

Laugesen et al. (2017) was interested in studying consumption patterns to determine the annual prevalence of prescription users of glucocorticoids within the Danish population from Jan 1991 to 31 December 2015. The results of the study showed that the use of glucocorticoid drugs remained stable over the period studied with approximately 3% of all Danish citizens using a glucocorticoid drug. It was observed that when analysing different age groups the prevalence increased among the elderly from 7% to 8.2% for those aged 65 to 79, and 8.4% to 10% for those over 80 years of age.

From these two studies, there is evidence that the DDD per 10000 inhabitants per day methodology could be used to reveal patterns in drug consumption. The DDD per 10000 inhabitants per day methodology has also been shown to work beyond drug consumption; Olveira et al. (2009) however adapted this method to estimate the trend in consumption, cost, and prevalence of home enteral nutrition from 2000 to 2007 in Andalusia, Spain.

The DDD per 10000 inhabitants per day calculation was modified to account for kilocalories per package rather than drug dosage. The results of the study showed a sharp increase in patients using home enteral nutrition from 66.4 to 1315.7 cases per inhabitant per day. The estimates produced from this study indicated that DDD per 10000 inhabitants per day method could be used as an approximation, as the results were found to be in line with other European countries.

2.4 Panel Data Modelling

Panel data is a term used in econometrics and statistics to describe datasets with repeated observations over time for the same individuals. In the case of this study, the individuals refer to the 52 districts in South Africa. The benefits of transforming data to a panel format lie in the fact that, unlike time series data that contains no distinct individual differences and cross-sectional data that contain no time-period specific differences, panel data can account for both individual and temporal distinctions (Biørn, 2017).

Panel data analysis has also been extended by combining it with spatial analysis to study the individual and time effects while taking spatial dependence into account. In spatial analysis, observations may differ due to space-specific time-invariant variables that impact the dependent variable. A rural area could have a completely different socioeconomic profile relating to poverty, education, and household income than an urban area; by accounting for this a spatial analysis assists in reducing biased estimation results by incorporating spatial correlation (Elhorst, 2017).

2.4.1 Modelling Applications and Studies

The section looks into the application of panel data and spatial linear analysis performed in health sector-related studies.

2.4.1.1 Panel Data Analysis

Panel data analysis is used for many different applications and is implemented in various fields of study. Studies on panel data modelling place importance on selecting models that take the appropriate model effects into account, whether they be random or fixed. A study by Astawesegn et al. (2022) aimed to analyse the trends and effects of ART coverage during pregnancy on mother-to-child transmission in Sub-Saharan Africa. For this study, data collected from 2010 to 2019 was used to establish this trend on a country level.

Using publicly available data from the United Nations Programme on HIV/AIDS, The World Health Organisation, and the World Bank; a total of 41 countries could be included in the study. A panel dataset of 410 observations was created from the transformed data. Astawesegn et al. (2022) made use of variables such as

HIV prevalence, HIV incidence-to-prevalence ratio, population size, Mother-to-child transmission (MTCT) rate, ART coverage for prevention MTCT of HIV, and married women's satisfaction towards modern family planning services to assess the independent variable being MTCT HIV transmission rate.

The Breusch-Pagan Lagrange multiplier (LM) test was used to confirm that a panel regression model was appropriate in this instance; with a p-value < 0.001 the null hypothesis of the LM test could be rejected thus indicating significant differences across countries in terms of the variables studied. A series of linear fixed and random effect models were then built. From the Hausman test, it was confirmed that a fixed effects model approach would be more consistent than a random effect model since the null hypothesis of a random effect model being more consistent could be rejected. The model output revealed that ART coverage for HIV-positive pregnant women and the HIV incidence-to-prevalence ratio were significantly associated with MTCT HIV transmission rate.

Another panel data analysis by Rahman et al. (2023) aligns with this study as it aims to determine the effects of antibiotics usage on the prevalence of resistance in humans to these drugs. An eleven year panel of data on usage and resistance for 26 different antibiotic-bacteria combinations in 26 European countries was analysed.

The antibiotic usage for these countries for the years 2008 to 2018 was retrieved from IQVIA's MIDAS database which reports the volume of sales for antibiotics. These antibiotic sales were converted using the ATC/DDD index to DDDs per 1000 inhabitants per day to account for population consumption in each region using the World Banks population estimates. The resistance data was obtained from the European Antimicrobial Resistance Surveillance System which collects the antimicrobial resistance in bacterial pathogens for various antibiotic classes.

The fixed effects model included the bacteria class and country as entity fixed effects and year as a time-fixed effect. Following this methodology with the DDD per 1000 inhabitants per day representing usage over the respective years it was determined that on average prevalence of resistance bacteria increases immediately after usage and continues to increase for four years, while a decrease in usage has no immediate impact.

Rahman et al. (2023) made use of three empirical models to estimate the effect of antibiotic usage on the prevalence of resistance. A distributed-lag fixed effect, distributed first differences, and an event-study model were built for this purpose; although the results for all models are presented in supplementary material, the paper reviewed focused on the fixed-effect estimation methodology.

Considering the two studies presented it is apparent that a panel data analysis modelling approach can be successful in predicting specific disease indicators from sales ordering data. In the case of this study, using ARV drug ordering as a predictor for HIV indicators such as prevalence may foreshadow a surge or decrease in HIV prevalence, but also act as a measure of transmission rates therefore making ARV drug procurement a potentially useful indicator of HIV trends.

2.4.1.2 Spatial Data Analysis

Each district has different social and economic characteristics that must be accounted for. Linear modelling ignores spatial autocorrelation which may lead to incorrect reasoning. A spatial linear modelling study by Muleia et al. (2020) looking at the spatial distribution of HIV prevalence among young people aged 15 to 24 in Mozambique using 2009 Aids survey data used HIV sero-sites as the dependent binomial variable, with 1 indicating of a positive HIV status and 0 negative.

The feature variables used were a combination of socio-demographic factors such as age, sex, and education level; biological factors such as sexually transmitted infection in the last 12 months and behavioral factors

such as alcohol and condom use. By incorporating a geographical component in the traditional logistic regression function Muleia et al. (2020) developed a generalised geoadditive model to study the geographical variation of HIV risk among young people. The geographical component made use of covariance functions to explain the variation of variables depending on spatial separation.

The modelling approach used was to first identify the most appropriate feature variables using a backward selection technique and then make use of the Akaike information criterion (AIC) as the metric to select the best-fit model. Multiple model scenarios were built to obtain the best covariance function; the first model was a normal logistic regression model, the next three were logistic regression models using exponential, gaussian, and spherical covariance functions for the geographical components while the last was a logistic regression model with a gaussian covariance function where the spatial random effect was kept constant.

The models were all compared to one another using the corrected AIC_c as the comparison metric as the regular AIC could ignore spatial dependence in data (Muleia et al., 2020). It was found that the models including the geographical component produced the lowest AIC_c but that there was no distinct advantage in using one covariance function over another. Ultimately, the Gaussian covariance function was selected and the receiver operating characteristic (ROC) curve with accuracy measures was used to test model performance.

The results of the study concluded that southern, central, and north-west regions of Mozambique displayed the highest prevalence which could be attributed to the number of child marriages in these areas leading to early exposure to unprotected sex which contributed to females being six times more likely to be infected than their male counterparts.

2.5 Implications of Study

This study provides an alternative framework to estimate ARV drug utilisation and HIV prevalence at a district-level in South Africa. From the literature reviewed, no study focussing on using medicine ordering data to achieve this has been found. Although the use of the DDD per 1000 inhabitants per day methodology has been used for the analysis of drugs used to treat other diseases, no studies on ARV drugs have been found. Similarly, no method used to estimate HIV prevalence by using medicine ordering data exists.

This study contributes to the literature by providing an alternative method of estimating ARV drug utilisation and HIV prevalence at a district-level. The adoption of such a framework can be used as a supplementary tool to established methods such as the Naomi model; furthermore, it provides a methodology for identifying the shortcomings and successes of South Africa's ART programme that could be considered when aiming to improve ART accessibility.

From the literature reviewed it is apparent that the DDD per 1000 inhabitants per day remains the best method when aiming to approximate drug usage or consumption, and that panel data and spatial analysis have both yielded positive results in other medical-related studies aiming to predict prevalence.

Chapter 3

Data

3.1 Introduction

This chapter introduces the datasets used throughout the study and how they can be used to provide a meaningful analysis. Furthermore, the insights from the exploratory data analysis (EDA) are used to formulate the study methodology and a way for interpreting the results in the chapters to follow.

3.2 Data Description

The primary datasets used were the RSA Pharma, DHIS (District Health Information System), MPHL(Master Health Product List), and Naomi Model Indicator datasets.The RSA Pharma and DHIS datasets are not publicly available, but the MPHL and Naomi Model datasets can be found on the National Department of Health and SA District HIV Estimates (Estimates, 2024) websites. The RSA pharma dataset provided by the Clinton Health Access Initiative (CHAI) is a large dataset containing the sales data of different medications purchased by different health facilities. The DHIS and MHPL datasets were used to enrich the RSA Pharma dataset with additional health facility and medication information. Merging the DHIS and RSA Pharma datasets provided additional information on each health facility, such as the provincial, district, and sub-district locations. The MPHL dataset contains information on all the products procured in the public sector such as drug category information and the Anatomical Therapeutic Chemical (ATC) codes. The HIV estimates from the Naomi Model Indicator datasets contain the estimated HIV estimate outputs for the years 2020 to 2022. These yearly estimates are included in the panel dataset to provide year-on-year estimates for each district that are used as the baseline for comparison for the models built. The Naomi model estimates are provided with a mean estimation with lower and upper limit bounds.

3.2.1 RSA Pharma Data

The RSA Pharma dataset contains the sales orders for different medications and medical supplies. Each row in the dataset represents a single order containing information such as the Supplier Name, NSN (National Serial Number), Description, Brand Name, and Customer; using these details it was possible to determine the total quantity of a medication ordered by a customer in a given period. Section 3.3 describes the pre-processing performed on the RSA Pharma dataset to get it into a usable format, Tables 3.1 to 3.3 provide a view of a subset of the original dataset.

Supplier Name	NSN	Description
Pharma Dynamics (pty) Ltd	181817608	Rifampicin, Pyrazinamide, Ethambutol, Isoniazid; 150mg, 400mg, 275mg, 75mg; Tablet; 112 Tablets
Pharma Dynamics (pty) Ltd	181817608	Rifampicin, Pyrazinamide, Ethambutol, Isoniazid; 150mg, 400mg, 275mg, 75mg; Tablet; 112 Tablets
Pharma Dynamics (pty) Ltd	181817608	Rifampicin, Pyrazinamide, Ethambutol, Isoniazid; 150mg, 400mg, 275mg, 75mg; Tablet; 112 Tablets
Pharma Dynamics (pty) Ltd	181817608	Rifampicin, Pyrazinamide, Ethambutol, Isoniazid; 150mg, 400mg, 275mg, 75mg; Tablet; 112 Tablets
Pharma Dynamics (pty) Ltd	181817608	Rifampicin, Pyrazinamide, Ethambutol, Isoniazid; 150mg, 400mg, 275mg, 75mg; Tablet; 112 Tablets

Table 3.1: First 5 Rows of the RSA Pharma Dataset. Each Observation Indicates the Supplier Name, National Serial Number, and Product Description of the Medication Ordered

Brand Name	Order Qty (per Item)	Province	Customer name
Risopet 112'S	14	KwaZulu-Natal	KZN - PORT SHEPSTONE HOSPITAL
Risopet 112'S	36	KwaZulu-Natal	KZN - PORT SHEPSTONE HOSPITAL
Risopet 112'S	125	KwaZulu-Natal	KZN - EAST BOOM CHC
Risopet 112'S	108	KwaZulu-Natal	HLENGISIZWE COMM HEALTH CENTRE
Risopet 112'S	18	KwaZulu-Natal	KZN-GENERAL JUSTICE GIZENZA MPANZA HOSP

Table 3.2: RSA Pharma Dataset Continued. Each Observation Indicates the Brand Name, Quantity Ordered, Province where the Health facility is Located, and the Customer/Health Facility Name

Price Per Pack (ZAR)	Total Price of Transaction (ZAR)	Quanity Delivered (per Item)	Date Order Delivered
251.34	3518.76	14	2022/10/05
251.34	9048.24	36	2022/10/05
251.34	31417.50	125	2022/10/05
251.34	27144.72	108	2022/10/14
251.34	4524.12	18	2022/10/05

Table 3.3: RSA Pharma Dataset Continued. Each Observation Indicates the Price Per Pack, the Total Price of the Transaction Derived from the Quantity Delivered and the Date of Delivery

3.2.2 DHIS Data

The DHIS dataset contains detailed information about the health facilities in South Africa with each observation providing specific information on each health facility as seen in Table 3.4, which presents a subset of the dataset. The dataset contained 14033 observations and 17 features of which 3 were used for this study; the Province, District, and Short Facility Name features were required to match the customers from the RSA Pharma dataset to the correct province and district.

Province	District	Short Facility Name
kz KwaZulu-Natal Province	Amajuba District Municipality	Niemeyer Mem Hosp
kz KwaZulu-Natal Province	Amajuba District Municipality	Madadeni Hosp
kz KwaZulu-Natal Province	Amajuba District Municipality	Mediclinic Newcastle Hosp
kz KwaZulu-Natal Province	Amajuba District Municipality	Newcastle Hosp
kz KwaZulu-Natal Province	eThekwini Metropolitan Municipality	Addington Hosp

Table 3.4: First 5 Rows of the District Health Information System Dataset. Each Observation Provides Geographical Information for Each Health Facility

The features in the table above are the Province, District, and Short Facility Name. The Province and District features provide the provincial and district locations of each health facility given in the Short Facility Name feature.

3.2.3 MHPL Data

The MHPL data contains information relating to product descriptions. The dataset contained 1166 observations and 20 features. From this dataset, the product NSN, Description, ATC, and Category were the features of interest as they could be used to identify which drug classification each observation in the RSA Pharma dataset belongs to. Table 3.5 provides a view of the first 5 rows of the dataset.

NSN	Description As Per Contract	ATC	Category
181781208	Abacavir; 20mg/ml; Solution; 240 ml	J05AF06	Anti-Retroviral Medicines
181781208	Abacavir; 20mg/ml; Solution; 240 ml	J05AF06	Anti-Retroviral Medicines
181896191	Abacavir; 300mg; Tablet; 56 Tablets	J05AF06	Anti-Retroviral Medicines
181901076	Abacavir; 60mg; Tablet, dispersible; 56 Tablets	J05AF06	Anti-Retroviral Medicines
222001257	Atazanavir, Ritonavir; 300mg, 100mg; Tablet; 28 Tablets	J05AE08	Anti-Retroviral Medicines

Table 3.5: First 5 Rows of the Master Health Product List Dataset Providing Medical Product Information

The features in the table above describe characteristics of each product. The National Serial Number (NSN) feature is a unique identifier for each medication. The Description as Per Product feature states the medication name and information regarding dosage size, form and quantity pack. The Anatomical Therapeutic Chemical (ATC) code is used to classify medications based on the organ or system they treat while the Category feature indicates the name of the organ or system the medication treats.

3.2.4 Naomi Model Indicators Data

The Naomi model, which is a Bayesian small area model used to estimate and make inferences on administrative areas where direct observations may be limited, estimates HIV prevalence indicators stratified by factors such as geographical area, sex, and age groups to provide current and one-year projections for HIV

planning (Eaton et al., 2021). The Naomi model estimates for 2020,2021 and 2022 were used in this study by incorporating the district-level estimates into the panel dataset.

Obtained from Estimates (2024); district-level HIV estimates for HIV prevalence, PLHIV, and ART number (residents) are some of the HIV indicators generated by the Naomi Model. For each district, these specific indicators were added to the dataset to have each of the indicators for the years 2020 to 2022. Table 3.6 is a view of a subset of the Naomi Indicators dataset for 2020 showing relevant features; each estimate indicator is provided with a mean, lower, and upper bound estimation.

Area Name	Indicator Label	Mean	Lower	Upper
Buffalo City MM	Population	407263.82	407263.82	407263.82
Buffalo City MM	HIV prevalence	0.20	0.18	0.23
Buffalo City MM	PLHIV	82967.46	71958.19	94994.45
Buffalo City MM	ART coverage	0.63	0.53	0.72
Buffalo City MM	ART number (residents)	51621.50	48150.36	55223.29

Table 3.6: First 5 Rows of the Naomi Indicators Dataset for 2020 Providing HIV Indicator Estimates for Each District

The Naomi model requires six different inputs to produce the district estimates. These estimates are stratified for geographical area, sex, age, and time. For this study both sexes were included for those aged 15+; a detailed description of the model inputs and outputs is provided here:

Model Inputs

1. Area hierarchy: List of administrative areas used for health planning with geographic boundaries.
2. Population: Population estimates stratified by district, sex, and five-year age groups.
3. Household survey: Data on HIV prevalence, ART coverage, and recent infections from surveys tabulated by district, sex, and age group.
4. ART service delivery data: Number of people receiving ART at health facilities in each district at the end of each yearly quarter.
5. Antenatal clinic (ANC) HIV testing indicators by district and year.
6. Spectrum Estimates: Outputs from national or district level Spectrum files exported from Shiny90.

Model Outputs

1. Population
2. HIV prevalence/PLHIV
3. ART coverage, Number of people on ART and number of untreated PLHIV
4. Number and percent of those aware and unaware of status
5. HIV incidence rate of new annual infections
6. Annual ANC clients by HIV status and ART status at first ANC visit

Stratification

1. Geographical Areas: Province and District
2. Sex: Male and female
3. Age groups: (0 to 14, 15 to 24, 25 to 34, 35 to 49, 50 to 64, 65+) and (15 to 49, 15 to 64, 15+, 50+, all ages, <1, 1 to 4)
4. Time Points: T1: most recent household survey, T2: current quarter, T3: 9-month projection to 1 year

Figure 3.1 is an overview of the model's components, and the process flow that the Naomi model follows.

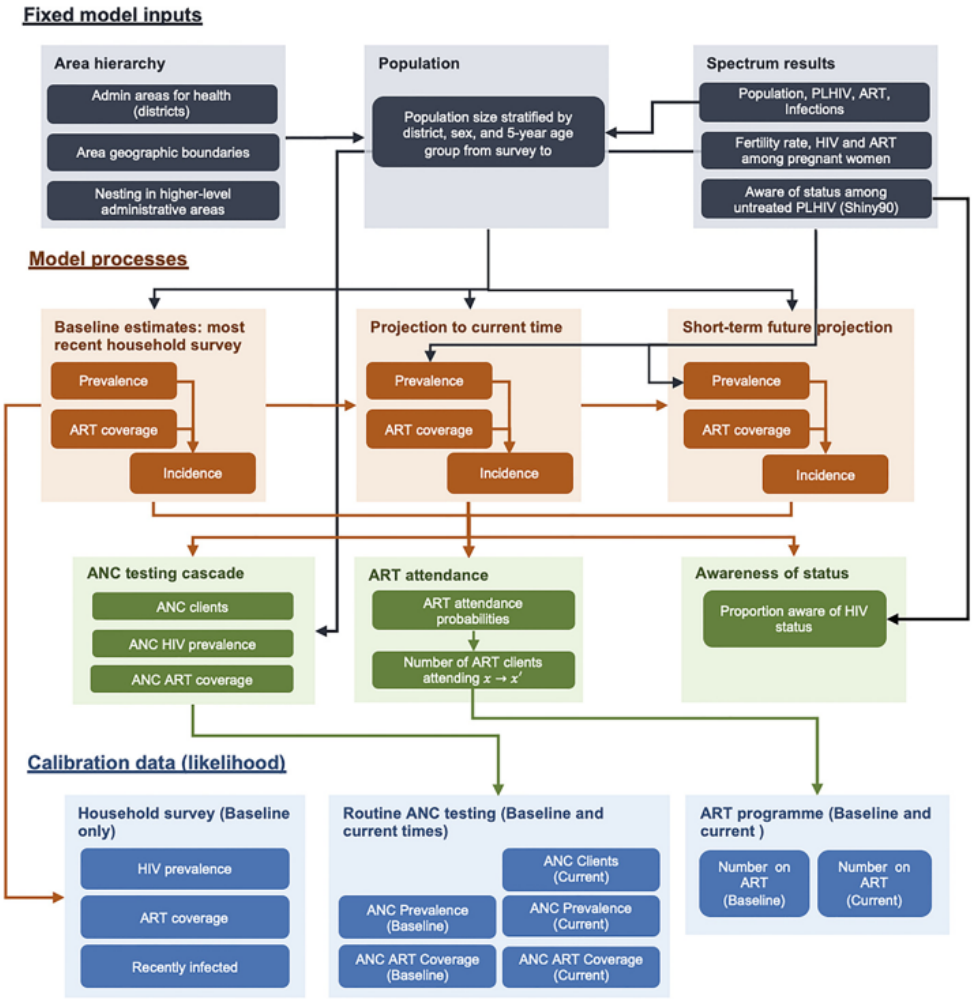


Figure 3.1: Naomi Model Components and Processes Overview (Eaton et al., 2021)

The figure shows how HIV metrics are estimated at three different points in time; the last household survey, present, and future projection. The model uses data from the household survey to estimate HIV prevalence, ART coverage, and incidence rates by district, sex, and age group. For the current period, it updates these estimates using recent data on new infections and survival rates. In the last period for the future projection, the model forecasts trends based on current estimates and short-term program plans. This assists in understanding the current and future state to better plan for future needs.

The outputs of the Naomi model are reviewed by the South Africa Estimates and Modelling Technical

Working Group, chaired by the National Department of Health and South Africa National AIDS Council (Estimates, 2024). By being a trusted source for HIV estimates, the Naomi model estimates were used to assess if medicine ordering data could be used to produce simple estimates that are acceptable.

3.3 Data Pre-processing

Pre-processing was required to get the datasets in a usable panel data format required for the analysis. The data transformation and handling steps followed are as follows:

1. The RSA Pharma dataset which is only limited to ordering information was first expanded by merging it with the DHIS dataset. The customer name in the RSA Pharma dataset was matched with the appropriate facility name from the DHIS dataset.

   Many inconsistencies in the customer names were present, as they were manually inputted. This human error was dealt with by joining the datasets on the Customer and Short Facility Name columns through fuzzy matching using the fuzzyjoin R-package (Robinson, 2020). Table 3.7 provides an example of this matching; a dist value of 0 indicates a perfect match with larger discrepancies occurring with larger dist values.

   To improve the accuracy of the fuzzy join, regular expressions via the string package were used (Wickham, 2023). String patterns were found in customer names so that the customer names could be changed to match the facility names in the DHIS dataset; this resulted in dist values of 0 for many customer names thus improving the accuracy of the fuzzyjoin.

RSA Pharma Name	DHIS Data Name	Dist
charlotte maxeke jhb ac hospit	charlotte maxeke hosp	0.0000000
charlotte maxeke jhb acc hospital	charlotte maxeke hosp	0.0000000
charlotte maxeke jhb ac hospit	charlotte maxeke hosp	0.0000000
charlotte maxexe jhb ac hospit	charlotte maxeke hosp	0.1269841
charlotte maxeke jhb academic hospital	charlotte maxeke hosp	0.1491228

Table 3.7: Fuzzyjoin Match Illustration Matching Facility Names from the RSA Pharma Dataset to the DHIS Dataset

2. The MHPL dataset was then joined to the newly generated RSA Pharma dataset with DHIS information. The datasets were joined on the NSN column which is a unique serial number given to each product. The result of this data join was a RSA Pharma dataset with ATC codes for each item sold as well as the drug classification.

3. All observations with medications not classified as antivirals for systemic use were filtered out so that only ARV drugs remained.

4. The final step was to then convert the dataset to a panel data format. This was done by grouping the dataset by year and district to obtain a panel format with three distinct years for the 52 districts resulting in a panel with 156 observations.

   The quantity ordered for each year was summed for each district. The Naomi Model Indicator data was then joined to the panel dataset with each district having the appropriate estimates linked to them

for the years 2020 to 2022. Table 3.8 provides a view of the panel dataset.

Year	District	QuantityOrdered	PLHIV	Prevalence	Population
2020	Amajuba District Municipality	143280,00	72829,58	0,24	497792,06
2021	Amajuba District Municipality	129293,00	75618,65	0,25	501794,20
2022	Amajuba District Municipality	482796,00	71392,04	0,23	506207,31
2020	Amathole District Municipality	4310,00	68018,46	0,19	711490,66
2021	Amathole District Municipality	15884,00	70990,89	0,20	703575,82
2022	Amathole District Municipality	12248,00	68904,16	0,19	696899,67

Table 3.8: SA District Ordering Panel Dataset, 2020 to 2022

3.4 Exploratory Data Analysis

After filtering out all medications not classified as antiviral for systemic use, it was found that the dataset contained ordering information for 18 unique ARV drugs. Although possible, it would not be feasible to perform an analysis on each ARV as there is complexity behind how different ARV drug combinations are prescribed. Chapters 4 and 5 focus on the Naomi model data from 2020 to 2022, which is the data currently available. However, the following sections' exploratory data analysis (EDA) covered a longer period, from 2017 to 2022. Analysing this extended period helped to show the trends in ARV ordering in South Africa over time, giving a better understanding of the situation and assisting in planning the next steps more clearly.

Figure 3.2 below shows the total percentage of each ARV drug ordered from 2017 to 2022. From this, it is apparent that two drugs J05AR27 and J05AR06 represented by their ATC codes were the two drugs with the highest order quantities accounting for 40.12% and 31.27% of all orders respectively.

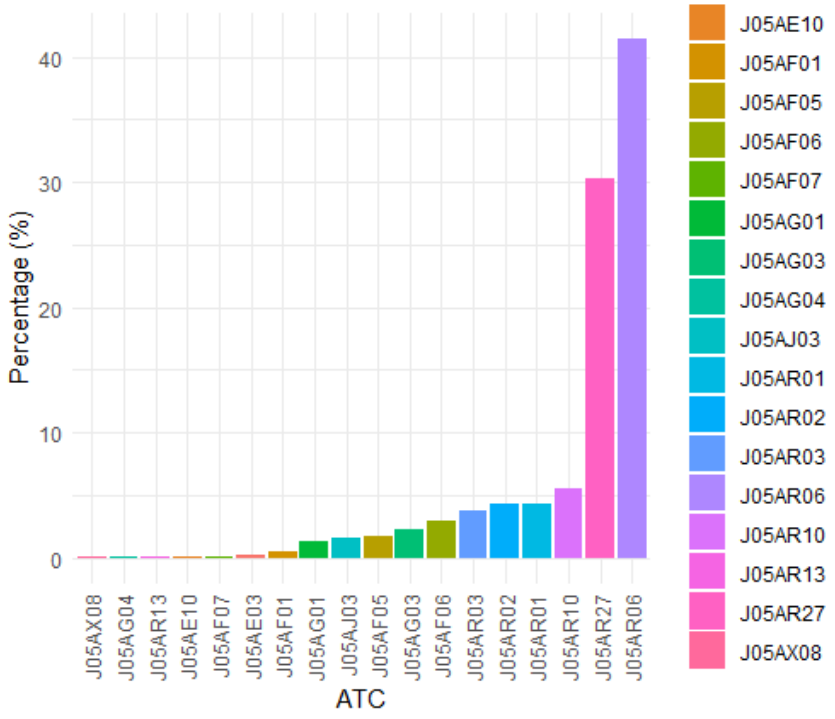


Figure 3.2: Percentage Breakdown of Individual ARV Drugs Ordered, 2017 to 2022

The J05AR27 and J05AR06 ATC codes correspond to dolutegravir/lamivudine/tenofovir disoproxil (TLD) and efavirenz/tenofovir disoproxil/emtricitabine (TEE). As highlighted in section 2.2.1, these two drugs are the first and second lines of treatment for HIV in adults as recommended by the World Health Organisation (WHO).

Based on the information gathered, these two drugs were ideal for this study for the following reasons:

1. They accounted for 71.39% of the orders.
2. By being fixed-dose drugs, they are typically not consumed in combination with other drugs, so no drug combinations had to be considered.
3. Patients are required to consume these ARV drugs daily for the rest of their lives which makes them ideal for long-term analysis
4. These drugs could be analysed simultaneously as there would be no overlap in consumption. A patient would be on one of the treatment regimens but not both.

TEE was the first line treatment before TLD was rolled out in 2019 as the new treatment regimen (Venter et al., 2017). Figure 3.3 is a stacked bar chart showing the quantity of these two drugs ordered from 2018 to 2022. In 2019 the TLD treatment regimen began picking up while TEE treatment remained the most popular, it wasn't until 2020 that the quantity of TLD ordered increased substantially indicating that South Africa had begun adopting the WHO recommendations. In 2021 TLD orders overtook TEE orders, and in 2022 it could be seen that in comparison to TLD, TEE had become negligible.

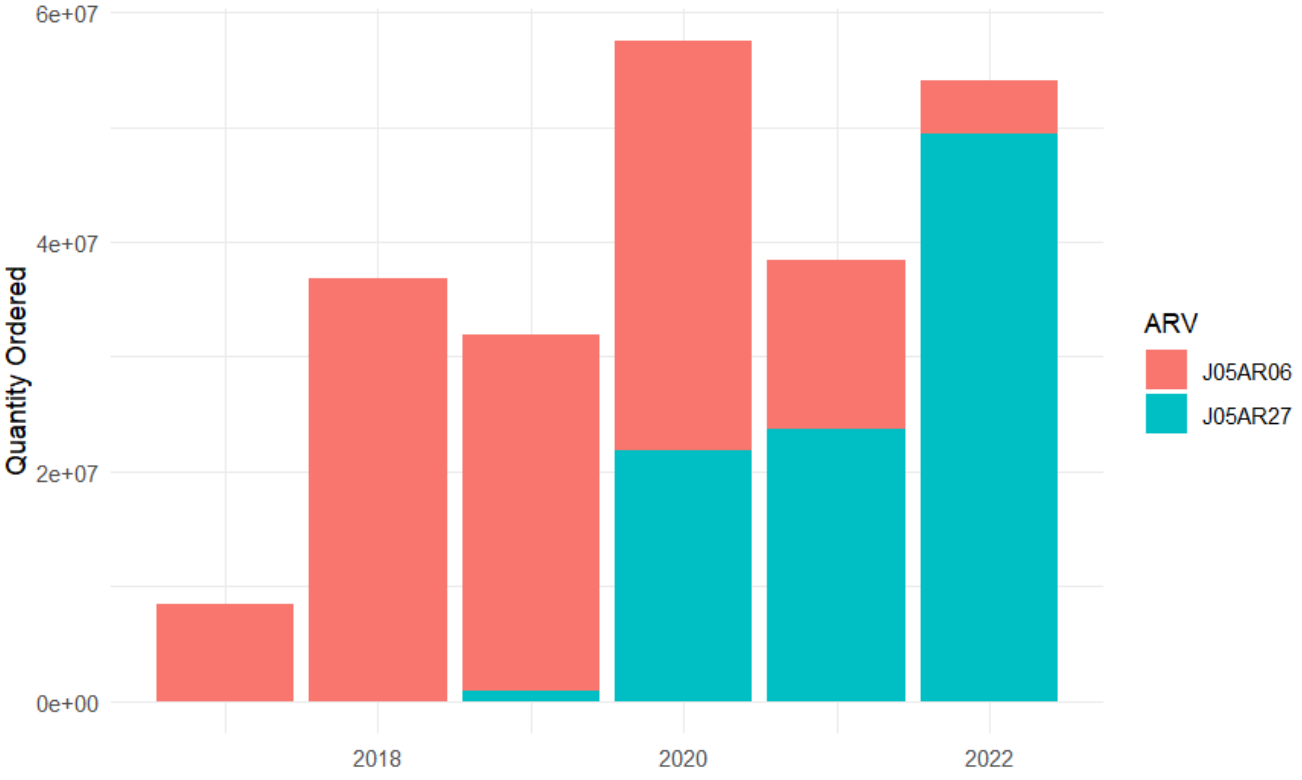


Figure 3.3: Stacked Bar Chart of TEE and TLD Units Ordered from 2018 to 2022

Probing into Figure 3.3 by observing ordering trends on a provincial level, it could be observed that the Western Cape, Eastern Cape, Gauteng, and KwaZulu-Natal ordered the most TEE and TLD units. Gauteng in particular ordered the most ARV drugs each year followed by KwaZulu-Natal and the Eastern Cape.

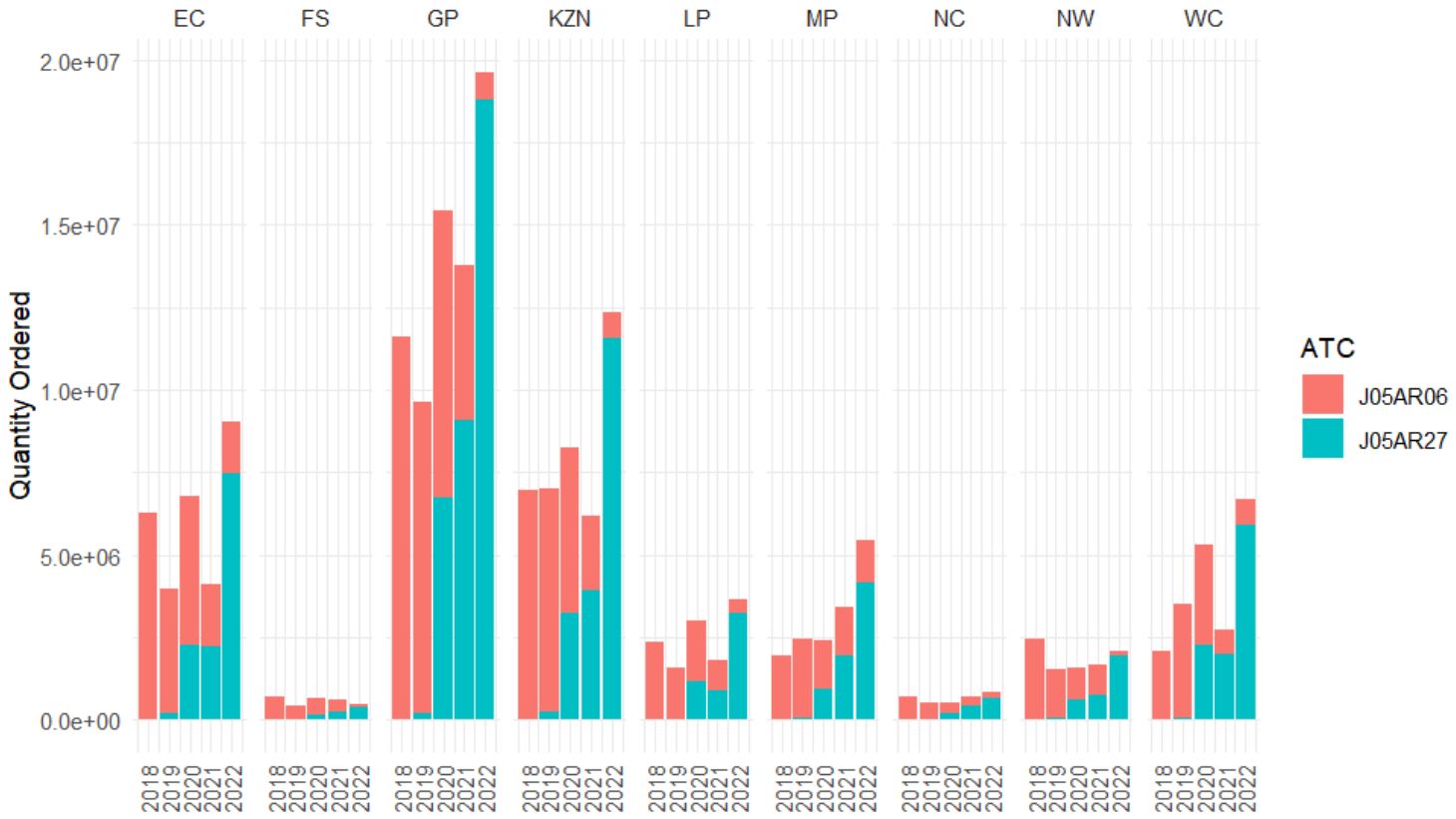


Figure 3.4: Stacked Bar Chart of TEE and TLD Units Ordered per Province, 2018 to 2022

From the figure above it was apparent that higher populated provinces ordered the largest quantities of TEE and TLD. The ordering trends followed the same trend as the population figures of South Africa with Gauteng having the highest population followed by Kwa-Zulu Natal, Eastern Cape, Western Cape, and Limpopo. This trend could be seen in the ordering data as the quantities ordered followed this pattern.

From a provincial level to a district and eventually health facility level, the largest purchasers of these two ARV drugs were identified. Figure 3.5 below provides a view of the top 10 districts by the quantity ordered; from this figure, it was found that six of the districts identified were metropolitan municipalities which typically have larger populations compared to the smaller district municipalities.

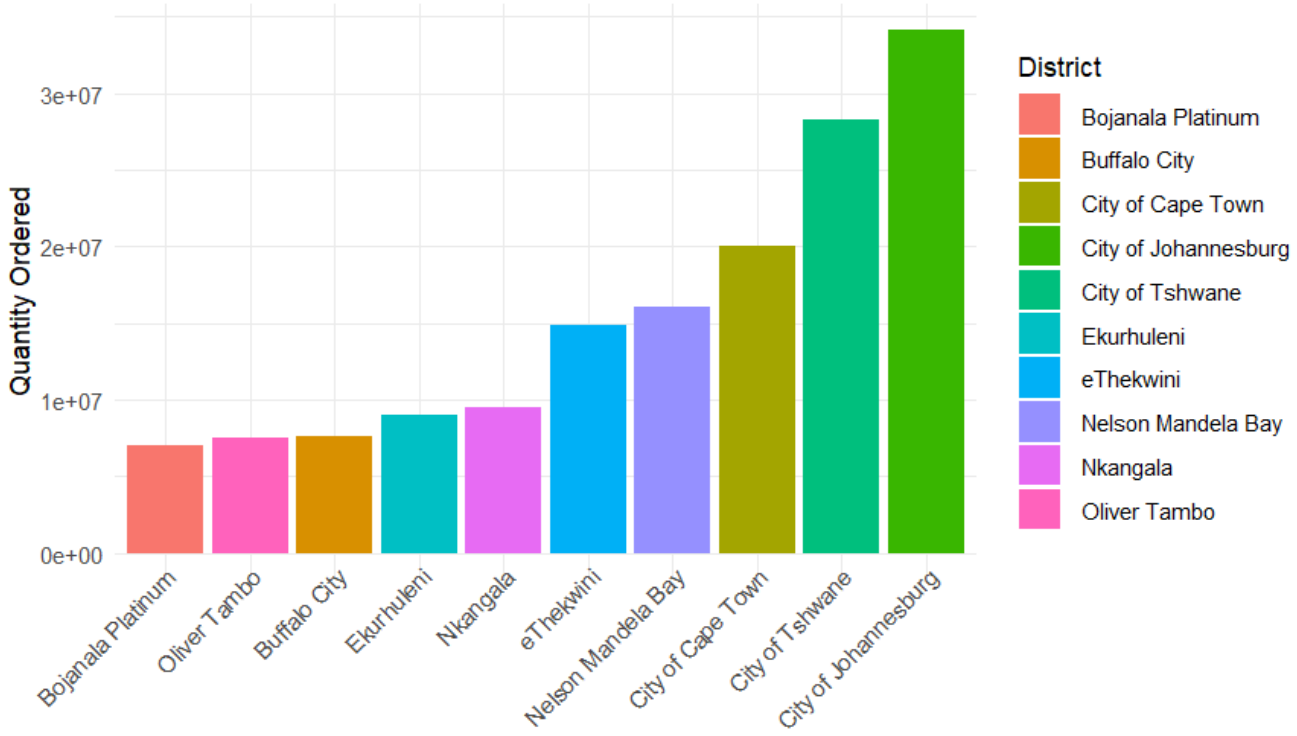


Figure 3.5: Top 10 Districts Ordering TEE and TLD by Quantity Ordered, 2018 to 2022

As defined by the Local Government: Municipal Structures Act of 1998 (Government, 1998) a municipal district is defined as a metropolitan if it meets the criteria of having a high population density, extensive movement of people, goods, and services as well as extensive development. Thus the analysis indicated that ARV drug ordering was not necessarily based on district HIV prevalence but on factors such as the number of People Living with HIV (PHLIV) in a district which would increase in higher populated districts. When considering that for 2022 the City of Johannesburg Metropolitan Municipality had a prevalence of 15.2%, which is low compared to the likes of the uMkhanyakude District Municipality at 30.2% for ages 15+. Although prevalence was higher in the Umkhanyakude District Municipality when translated to PLHIV the City of Johannesburg Metropolitan Municipality had approximately 5.52 times more infected residents.

Figure 3.6 shows the top 10 facilities ordering ARV medication throughout the country. It was seen that the only patient-facing facility was Frere Hospital located in the Buffalo City Metropolitan Municipality while the remaining facilities were medical depots or government facilities.

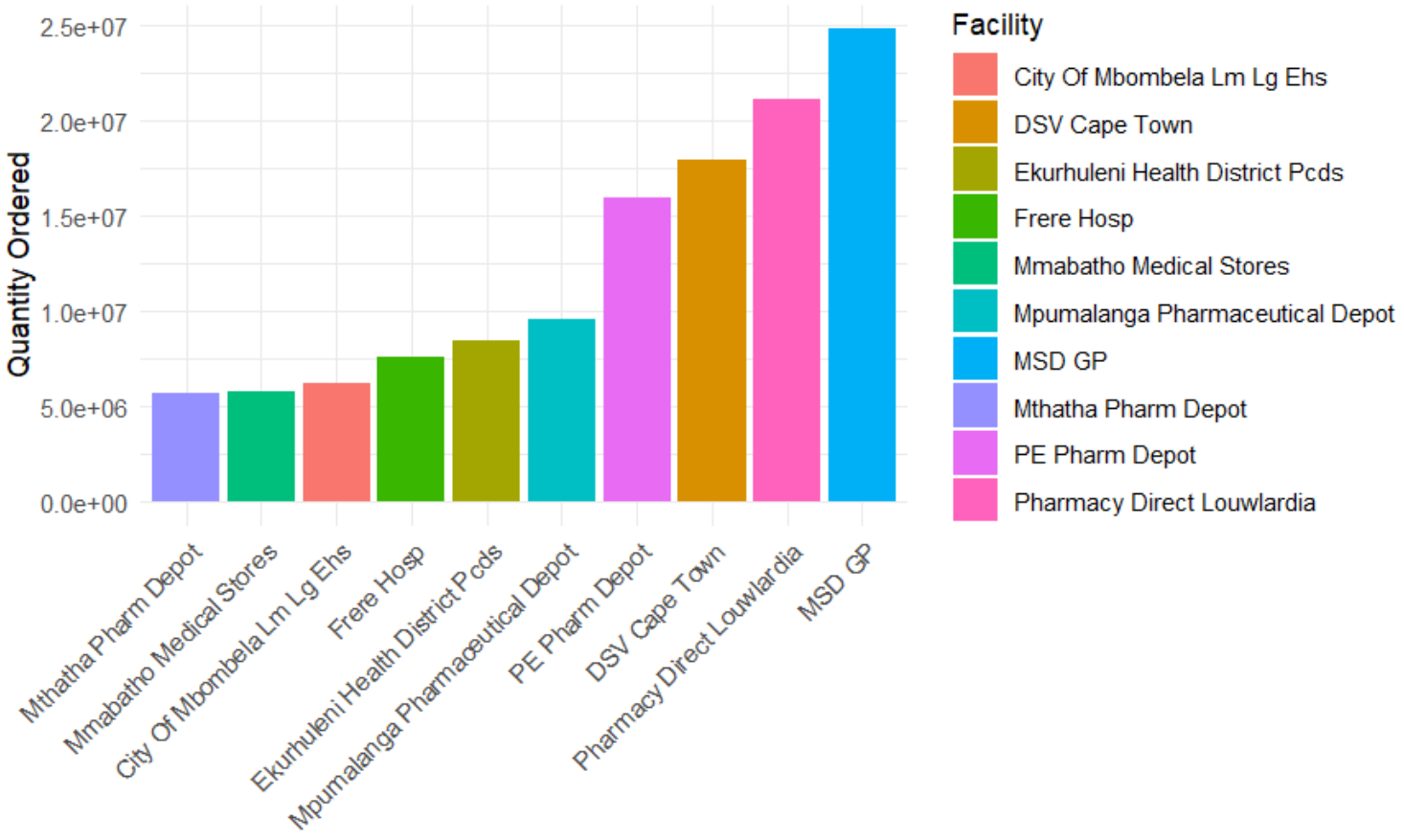


Figure 3.6: Top 10 Ordering Facilities of TEE and TLD, 2018 to 2022

Knowing that medical depots held a majority of the ARV drug stock supports the assumption that ARV drugs are ordered in large quantities by these facilities in highly populated areas. The distribution of supplies to surrounding facilities and district municipalities may originate from metropolitan municipalities since extensive movement of people and goods occurs in these areas.

Figure 3 shows the relationship between quantity and prevalence for 2020 with a fitted regression line. It could be seen that the impact of metropolitan municipalities results in a negative correlation due to the influence of the high quantities ordered by these municipalities. Figure 4 showed a stronger correlation between quantity and PLHIV with the metropolitan municipalities having higher numbers of PLHIV, even though no clear relationship could be observed between quantity and prevalence it was apparent that ARV drugs are ordered to meet the requirements of populations with a higher number of residents infected with HIV. The plots for years 2021 and 2022 are can be found in Appendix D.

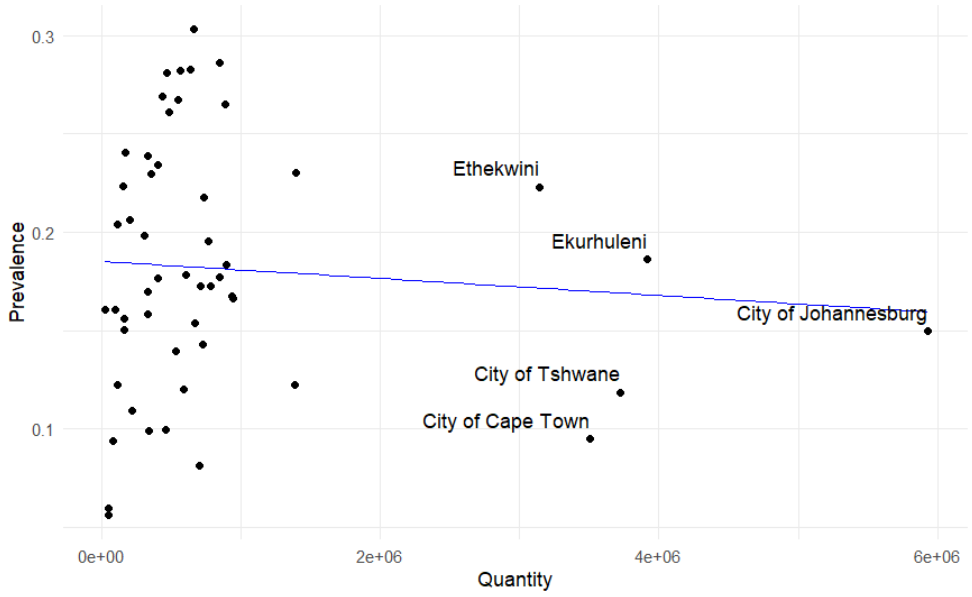


Figure 3.7: Scatterplot with a Fitted Regression Lines Showing the Relationship Between Quantity and Prevalence for 2020

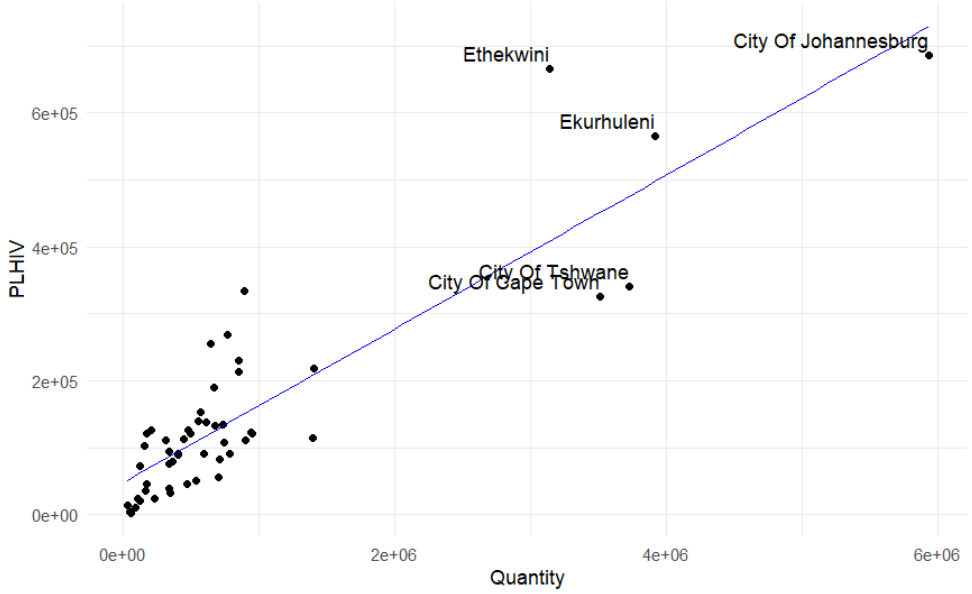


Figure 3.8: Scatterplot with a Fitted Regression Line Showing the Relationship Between Quantity and PLHIV for 2020

3.4.1 Summarised Findings

The exploratory data analysis resulted in findings that informed the methodology and analysis.

1. The results of the analysis and the literature indicated that TEE and TLD were the most relevant ARV drugs to study as they make up a majority of the orders, they are fixed-dose drugs, and analysis on both drugs could be performed as there is no overlap in the consumption of these drugs by patients.
2. The quantity of ARV drugs ordered is correlated to population. Apart from HIV prevalence, PLHIV may be a good HIV estimate to predict, as it can be used to calculate prevalence.
3. The largest orders were made by medical depots which are generally located in highly populated metropolitan municipalities. These facilities are not patient-facing facilities which implies that there is an extra step in the supply chain that makes the medication available to patients.
4. The large quantities of ARV drugs ordered by metropolitan municipalities result in a negative correlation between HIV prevalence and the quantity of drugs in a district. This goes against the expected trend as increasing ARV availability should increase the number of PLHIV thus increasing prevalence. This study however required that all districts be accounted for and none be excluded from the analysis.

Chapter 4

Methodology

4.1 Introduction

The following section serves as a road map outlining the steps taken to achieve the defined objectives. In this section the strategies, thought processes, and tools used are explained; this includes decisions made around modelling techniques and data considerations. The chapter focuses on the five areas, being:

1. Drug consumption using the DDD per 1000 inhabitants per day methodology.
2. Panel data modelling techniques; specifically pooled, random, and fixed effect models.
3. Panel data model statistical hypothesis testing.
4. Spatial autocorrelation testing and spatial modelling.
5. Modelling and data considerations

4.2 Implementation of the DDD Methodology

A Defined Daily Dosage (DDD) is allocated to a drug by the World Health Organisation Collaborating Centre in Oslo, which also assigns anatomical therapeutical chemical (ATC) codes to drugs. The ATC code is a unique identifier of a substance, and the classification system divides substances into different groups depending on the system or organ it is used to treat (WHOCC, 2024).

Table 4.1 below provides a breakdown of the levels for efavirenz/tenofovir disoproxil/emtricitabine (TEE); the first level J indicates that TEE is for anti-infective systemic use, the second level J05 classifies it as an antiviral drug for systemic use, the third level J05A indicates that it is direct-acting ARV drug, the fourth level classifies it as an antiviral for treatment of HIV infections, and the last level provides the name of the drug.

ATC Level	Code	Level Description
1st	J	Anti-infectives for systemic use
2nd	J05	Antivirals for systemic use
3rd	J05A	Direct-acting antivirals
4th	J05AR	Antivirals for treatment of HIV infections
5th	J05AR06	Efavirenz,Tenofovir Disoproxil,Emtricitabine

Table 4.1: ATC Level Breakdown for TEE

When using a DDD methodology there are important steps that must be completed. First, each pharmaceutical product must be linked to its correct ATC code and DDD followed by calculating the DDDs per package. For the effective implementation of the DDD methodology, it is recommended that the data used for the analysis contain specific product information such as the National Serial Number (NSN), product name, pack size, number of dosages per pack, active ingredients, DDD, pharmaceutical form and ATC code (WHOCC, 2024).

Through the data pre-processing described in Chapter 3, the data used in this study met these criteria, allowing for the ARV drug utilisation analysis to be performed. Drug utilisation can be presented differently depending on the health context being evaluated, and is expressed in DDDs as a unit of measurement. Three of these health contexts are given below:

1. DDD per 1000 inhabitants per day – Drug sales presented in DDDs can be used to approximate an estimation of the proportion of a population that uses a drug. The DDD in a group of 1000 inhabitants per day can be analysed as follows; if a value of 20 DDDs was calculated it could be interpreted that 20 out of 1000 people or 2% of the population being studied uses the drug.
2. DDDs per inhabitant per year – This metric is useful for analysing drugs usually used for a brief period; it estimates the number of days each patient is treated annually. If 9 DDDs per inhabitant per year were calculated, it is estimated that the consumption lasted for 9 days during a certain year.
3. DDD per 100 bed days – The DDDs per 100 bed days are applied when analysing drug use for inpatients. If a value of 50 DDDs per day was calculated, it estimates that 50% of inpatients receive the drug.

The DDD per 1000 inhabitants per day is the methodology followed in this study as successful implementation of this methodology has been identified in the literature reviewed. This methodology provides an approximation for consumption and must be implemented with the knowledge that not all drugs sold will be consumed in the period of study and that not all drugs are suitable for all age groups.

Using the ordering data available, the drug consumption of the two ARV drugs (TLD and TEE) was estimated to determine if there was an oversupply or undersupply of ARV drugs in each of the 52 districts, and on a provincial level. The equation for calculating DDD per 1000 inhabitants per day was defined in a study by Hollingworth and Kairuz (2021) as follows:

$$\texttt{DDD per 1000 inhabitants per day} = \frac{(Q * N * D * 1000\ \texttt{inhabitants})}{(\texttt{DDD} * 365\texttt{days} * \texttt{Pop})} \tag{4.1}$$

where $\begin{cases} Q & \text{: Quantity of drug sold in the study period} \\ N & \text{: Number of capsules/tablets per package} \\ D & \text{: Dosage of a single capsule/tablet} \\ DDD & \text{: Daily Defined Dose} \\ Pop & \text{: The population of the area studied} \end{cases}$

The result of equation 4.1 is a single DDDs value. Using this value the number of people using the ARV drugs in a specific district or province could be approximated using equation 4.2 below:

$$\texttt{Population Usage} = \frac{\text{Pop}}{1000} \times \texttt{DDD per 1000 inhabitants} \tag{4.2}$$

From results of equation 4.2, the population using TEE and TLD was interpreted as the number of people living with HIV (PLHIV) who are on antiretroviral therapy (ART). Using this approximation with the estimation figures for the number of residents on ART from the Naomi model, an Allocation Ratio (AR) could be calculated as the ratio of people consuming the drugs and residents on ART which provides a view of drug availability within a region, refer to Equation 4.3

$$\texttt{Allocation Ratio} = \frac{\text{Population Usage}}{\text{Residents on ART (Naomi Model Estimate)}} \tag{4.3}$$

where $\begin{cases} AR < 1 & \text{: Indicates that there are not enough ARV drugs to meet the demand} \\ AR{=}1 & \text{: Indicates that the supply perfectly meets demand} \\ AR > 1 & \text{: Indicates that the supply exceeds the demand} \end{cases}$

4.3 Panel Data Modelling

As mentioned, the RSA Pharma dataset contained individual orders for ARV drugs by health facilities across the country. These orders could be aggregated per district allowing for the total stock ordered to be calculated for each district over three years resulting in balanced panel data. Panel data is beneficial for many reasons. Firstly, it can measure effects that traditional cross-section and time-series data might miss due to its ability to capture both individual-specific and time-specific effects, making it ideal for improving estimations and analysing trends over time. Additionally, panel data analysis enables the study of the dynamics of change by examining a repeated cross-section of observations across multiple periods (Gujarati and Porter, 2009).

The models used for panel data regression are all derived from the ordinary least squares (OLS) regression model. Panel data regression is a combination of time series and cross-section data where the same unit is measured at different times, we would therefore have T time periods (t=1,2,...T) and N number of individuals (i=1,2,...N) resulting in a panel data of N x T observations (Croissant and Millo, 2008). The three model types used for panel data analysis are pooled OLS, fixed and random effect models.

4.3.1 Pooled OLS

The pooled OLS model is used to estimate panel data but has the inherent issue of not taking individual and time effects into account. It assumes that the relationship is the same for all variables across all individual entities removing heterogeneity amongst entities; this is a drawback to this model as there will most likely

be a difference among entities (Colonescu, 2017). The result of this is that the intercept α_{it}=α for all observations, the pooled OLS model is given by Equation 4.4.

$$Y_{it} = \alpha + \beta^{\top} X_{it} + \mu_i \tag{4.4}$$

where
$$\begin{cases} Y_{it} & : \text{Dependent variable for entity } i \text{ at time } t \\ X_{it} & : \text{Independent variable for entity } i \text{ at time } t \\ \alpha & : \text{Intercept} \\ \beta & : \text{Regression coefficient for } X_{it} \\ \mu_{it} & : \text{Entity-specific effect for entity } i \\ T & : \text{The time index} \end{cases}$$

The individual component μ_i may be independent or correlated to regressors. If correlated, the β estimation could be inconsistent due to endogeneity resulting in underestimated standard errors. therefore the use of a fixed or random effect model could be more appropriate.

4.3.2 Fixed Effect Models

Three types of fixed effect models are generally used in panel data analysis. These models include the individual, time, and two-way fixed effects models.

Individual Fixed Effects Model

Individual Fixed Effects Models differ from the pooled OLS by introducing a fixed effect term α_i for each entity i. The intercept may vary across entities but an entity's intercept does not vary over time. By keeping the intercept constant the differences between individuals are accounted for thus capturing time-invariant characteristics that differ across entities, thus any unobserved heterogeneity is accounted for (Gujarati and Porter, 2009). The individual fixed effects model is given by equation 4.5.

$$Y_{u_{it}} = \beta^{\top} X_{u_{it}} + \alpha_{u_i} + \mu_{u_{it}} \tag{4.5}$$

where
$$\begin{cases} Y_{u_{it}} & : \text{Dependent variable for entity } u_i \text{ at time } t \\ X_{u_{it}} & : \text{Independent variable for entity } u_i \text{ at time } t \\ \beta & : \text{Regression coefficient for } X_{u_{it}} \\ \alpha_{u_i} & : \text{Individual fixed effects for entity } u_i \\ \mu_{u_{it}} & : \text{Error term for entity } u_i \text{ at time } t \\ T & : \text{The time index} \end{cases}$$

Time Fixed Effects Model

The time fixed effect model aims to eliminate bias by controlling the unobserved time-specific factors that change over time but remain constant across entities and account for factors that differ across entities but are constant over time. The difference between a time and individual fixed effect model is the intercept term that becomes a time-specific fixed effect α_t. The time-fixed effects model is given by equation 4.6.

$$Y_{u_{it}} = \beta^{\top} X_{u_{it}} + \gamma_t + \mu_{u_{it}} \tag{4.6}$$

where

$$\begin{cases} Y_{u_{it}} & : \text{Dependent variable for entity } u_i \text{ at time } t \\ X_{u_{it}} & : \text{Independent variable for entity } u_i \text{ at time } t \\ \beta & : \text{Regression coefficient for } X_{u_{it}} \\ \gamma_t & : \text{Time fixed effects at time } t \\ \mu_{u_{it}} & : \text{Error term for entity } u_i \text{ at time } t \\ T & : \text{The time index} \end{cases}$$

Two-way Fixed Effects Model

Two-way fixed effect models can simultaneously include individual and time-specific effects. Two-way models work well when there is unobserved heterogeneity specific to each entity and unobserved time-specific factors that impact entities simultaneously. The two-way effects model is given by equation 4.7.

$$Y_{u_{it}} = \beta^\top X_{u_{it}} + \alpha_{u_i} + \gamma_t + \mu_{u_{it}} \tag{4.7}$$

where

$$\begin{cases} Y_{u_{it}} & : \text{Dependent variable for entity } u_i \text{ at time } t \\ X_{u_{it}} & : \text{Independent variable for entity } u_i \text{ at time } t \\ \beta & : \text{Regression coefficient for } X_{u_{it}} \\ \alpha_{u_i} & : \text{Fixed effects for entity } u_i \\ \gamma_t & : \text{Time fixed effects at time } t \\ \mu_{u_{it}} & : \text{Error term for entity } u_i \text{ at time } t \\ T & : \text{The time index} \end{cases}$$

The residual error assumptions are the same across the three fixed effect models defined. The assumptions are as follows:

1. **Mean of Zero:**
$$\mathrm{E}[\mu_{it}] = 0$$
The residual error term is assumed to have an expected value of zero in all three models.

2. **Constant Variance:**
$$\mathrm{Var}(\mu_{it}) = \sigma^2$$
The residual error term is assumed to have constant variance across all observations.

3. **Uncorrelated with Predictors:**
$$\mathrm{Cov}(\mu_{it}, X_{it}) = 0$$
The residual error term is assumed to be uncorrelated with the independent variables X_{it} in the three models.

4.3.3 Random Effect Models

Unlike pooled OLS and FE models which use the principle of OLS, this uses the principle of maximum likelihood or generalised least squares. Unobserved heterogeneity is accounted for by random effect models through the assumption that some parameters vary randomly across entities Croissant and Millo (2008). The random effects model is given by Equation 4.8.

$$Y_{u_{it}} = \alpha + \beta^{\top} X_{u_{it}} + \alpha_{u_i} + \mu_{u_{it}} \tag{4.8}$$

where
$$\begin{cases} Y_{u_{it}} & : \text{Dependent variable for entity } u_i \text{ at time } t \\ X_{u_{it}} & : \text{Independent variable for entity } u_i \text{ at time } t \\ \alpha & : \text{Intercept} \\ \beta & : \text{Regression coefficient for } X_{u_{it}} \\ \alpha_{u_i} & : \text{Random effects for entity } u_i \\ \mu_{u_{it}} & : \text{Error term for entity } u_i \text{ at time } t \end{cases}$$

4.3.4 Model Selection

The Chow, Hausman and Lagrange Multiplier (LM) statistical tests are performed to compare pooled, FE and RE models. Through a process of elimination, the models are compared until one modelling technique remains. Figure 4.1, modified from a paper by Zulfikar (2018) illustrates this process.

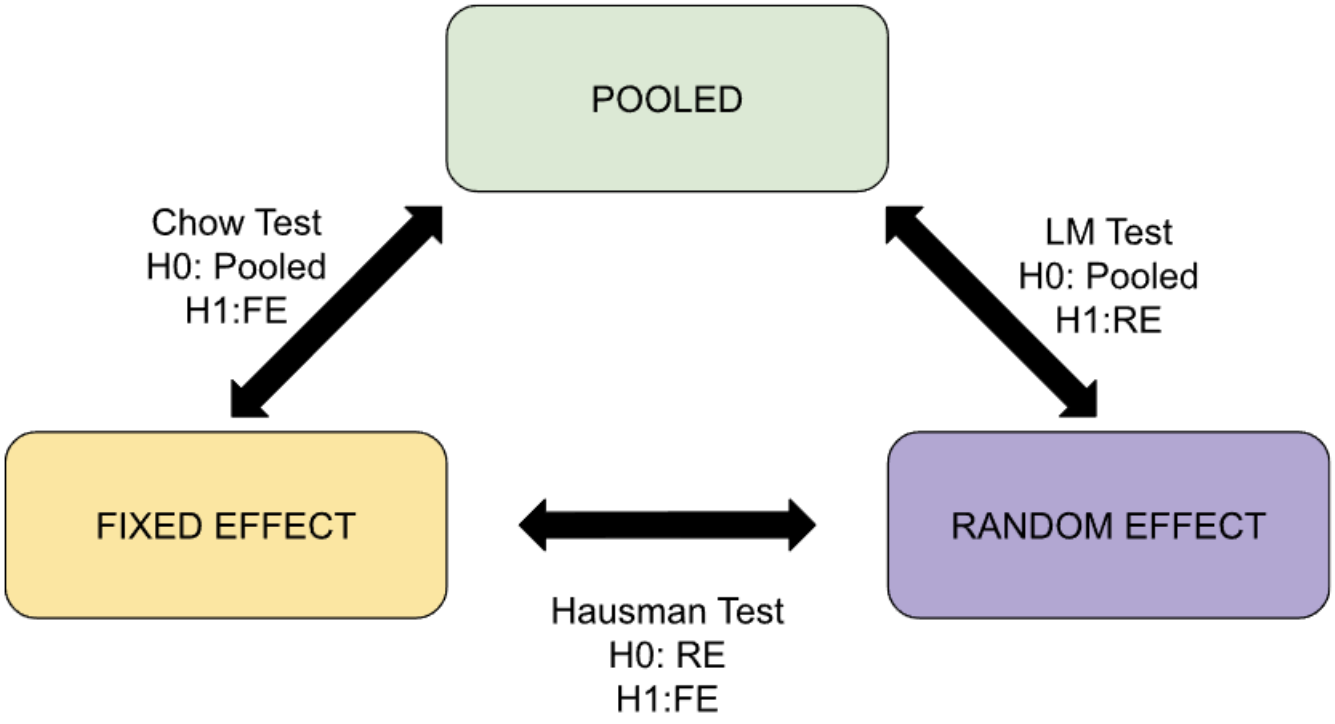


Figure 4.1: Panel Model Test Decision Flow for Selecting Panel Modelling Techniques (Zulfikar, 2018)

- **Chows Test**

  The Chow Test or pooltest is a test of poolability. This test determines if the same coefficients apply to the models, if so then the regression models compared would be the same. This test compares a pooled OLS model to an FE model, the hypothesis testing for the Chow test is as follows:

Let β_1 and β_2 be the coefficients for each model.

$$H_0 : \beta_1 = \beta_2$$

select pooled model

$$H_1 : \beta_1 \neq \beta_2$$

select fixed effect model

The Chow Test statistic is given by equation 4.9 (Lee, 2008).

$$CT = \frac{(RSS_1 - (RSS_2 + RSS_3))/k}{(RSS_2 + RSS_3)/(n_2 + n_3 - 2k)} \tag{4.9}$$

where
- RSS_1 is the Residual Sum of Squares from the pooled regression.
- RSS_2 is the Residual Sum of Squares from the regression on the first period.
- RSS_3 is the Residual Sum of Squares from the regression on the second period.
- k is the number of parameters estimated in each group
- n_2 is the number of observations in the first period.
- n_3 is the number of observations in the second period.

- **Hausman Test**

The Hausman Test is a statistical test that checks the consistency of the estimators in the models being compared. In this test, FE and RE models are compared. The hypothesis testing for the Hausman Test is as follows:

$$H_0 : \textit{Both models are consistent}$$

select random effect model

$$H_1 : \textit{One model is inconsistent}$$

select fixed effect model

The Hausman Test statistic is given by equation 4.10 (Greene, 2012).

$$HT = (b_{\mathrm{FE}} - b_{\mathrm{RE}})' \left[\mathrm{Var}(b_{\mathrm{FE}}) - \mathrm{Var}(b_{\mathrm{RE}})\right]^{-1} (b_{\mathrm{FE}} - b_{\mathrm{RE}}) \tag{4.10}$$

where
- b_{FE} is the coefficient vector from the fixed effects model,
- b_{RE} is the coefficient vector from the random effects model,
- $\mathrm{Var}(b_{\mathrm{FE}})$ is the variance-covariance matrix of the fixed effects estimator,
- $\mathrm{Var}(b_{\mathrm{RE}})$ is the variance-covariance matrix of the random effects estimator.

- **Lagrange Multiplier Test**

  The LM test, more specifically the Breusch-Pagan Lagrange Multiplier test for heteroscedasticity assesses if the variance of the error terms is constant or not. If heteroscedasticity is present then the validity of the pooled regression model can be questioned. The hypothesis testing for the LM test is as follows:

$$H_0 : Homoscedasticity\ is\ present$$

  select pooled model

$$H_1 : Heteroscedasticity\ is\ present$$

  select random effect model

The Lagrange Multiplier Test statistic is given by equation 4.11 (Baltagi et al., 2012).

$$LM = \frac{nT}{2(T-1)} \left(\frac{\sum_{i=1}^{n} \left(\sum_{t=1}^{T} e_{it} \right)^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right) \tag{4.11}$$

where $\begin{cases} n & = \text{Number of cross-sectional units,} \\ T & = \text{Number of time periods,} \\ e_{it} & = \text{Residuals from the OLS regression.} \end{cases}$

For each of the tests, it would remain true that the null hypothesis may be rejected if a p-value < 0.05 is obtained, else the null hypothesis is assumed to be true.

4.4 Spatial Analysis

The panel data modelling approaches are useful in predicting HIV prevalence but would not account for spatial dependence and interactions among districts. By transforming the data to the panel data format, the geographical locations for each district could be included; this would allow for spatial analysis to be performed if spatial autocorrelation is observed based on the quantities of ARV drugs ordered by each district. The following sections outline the approach taken to include spatial analysis and modelling.

4.5 Determining Spatial Autocorrelation

Before any spatial modelling could be performed, the presence of spatial dependence would need to be tested. The Moran's I was the method selected to determine if this phenomenon exists within the data. Moran's I determines neighbourhood relations as defined by the neighbourhood matrix. Neighbourhood matrices define the neighbourhood structure for an area of study, in this case, it would be the 52 districts in South Africa. The w_{ij} term denotes the (i, j) elements for the spatial weights matrix $\mathbf{W}$ that connects areas i and j, which is determined by the method in which the spatial matrix is structured.

Observation points can be deemed as neighbours through various methods based on contiguity, distance or k nearest neighbours. For this study, the contiguity method was selected as provides an intuitive method for defining neighbours.

Two regions are considered neighbours if they share a common edge; the rook and queen contiguity are often used to determine neighbours. As shown in Figure 4.2, according to the queen contiguity two observation points are considered neighbours if they share a common vertical, horizontal or diagonal edge. In the case of the rook contiguity, two observation points are considered neighbours only if they share horizontal or vertical edges or borders.

The queen contiguity was used for this study as it would not only be limited to considering vertical and horizontal edges when defining neighbours. With a spatial matrix constructed from the queen contiguity, the Moran's I test could be completed; note that the spatial matrix **W** is a **nxn** or **52x52** matrix in this case where a value of 1 at the intersection of area i and j would indicate that they are neighbours.

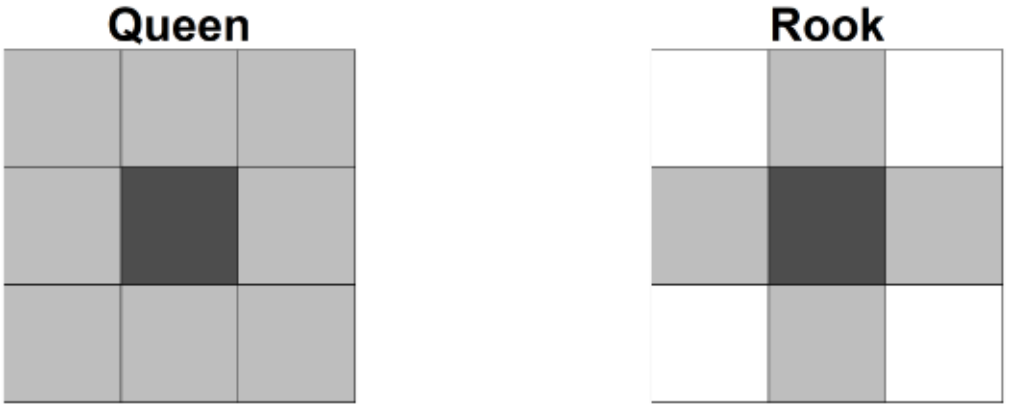


Figure 4.2: Queen and Rook Neighbours Based on Spatial Contiguity

The Moran's I measures the correlation between neighbouring observations which results in outputted values between -1 and 1. Negative values closer to -1 indicate significant negative autocorrelation, a value closer to 1 indicates significant positive autocorrelation and a value equal to or close to 0 indicates no spatial autocorrelation. Moran's I is given by Equation 4.12 (Mathur, 2015).

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4.12}$$

where
N : Number of observations
w_{ij} : The spatial weight between units i and j,
x_i : Values of the variable at locations i,
x_j : Values of the variable at locations j,
$\bar{x}$: mean of the variables

The value of the Moran's I itself cannot be used to determine if spatial autocorrelation is present, the p-value produced by the Moran's I must be used for this. In general, if a p-value < 0.05 is obtained then the null hypothesis may be rejected; the hypothesis testing for the Moran's I test is as follows:

$$H_0 : \textit{No spatial autocorrelation present}$$

$$H_1 : \textit{Spatial autocorrelation is present}$$

The Moran's I gives a global view of the spatial autocorrelation for the entire region being studied; the local Moran's I identifies specific locations with spatial autocorrelation. Using this local indicator of spatial association (LISA) the similarity between neighbours is given by one of five spatial cluster outcomes; in the context of this study it could be described as: (Tepanosyan et al., 2019):

1. High-High - A location with high quantities next to another high-quantity neighbour
2. Low-Low - A location with low quantities next to another low-quantity neighbour
3. High-Low - A location with high quantities next to a low-quantity neighbour
4. Low-High - A location with low quantities next to a high-quantity neighbour
5. Insignificant autocorrelation - No identifiable relationships

4.6 Spatial Linear Modelling

The spatial linear models implemented are defined in the following section. Spatial linear models are related to OLS but are used to address spatial dependencies in data. In the typical regression model, it is assumed that the elements of the dependent are uncorrelated but in the case of spatial data, there will be elements that are correlated. Observations that are closer together in space tend to be similar whereas those far apart are more dissimilar Dumelle et al. (2023). By not taking this spatial dependence into account models may not properly represent the true state of the scenario being modelled. The inclusion of a spatial random term τ results in a new spatial linear model.

$$Y_i = \beta X_i + \varepsilon_i + \tau_i \tag{4.13}$$

where
$$\begin{cases} Y_i : \text{Observed dependent variable at spatial unit } i \\ \beta : \text{Coefficient associated with the independent variables} \\ X_i : \text{Observed values for the independent variables at spatial unit } i \\ \varepsilon_i : \text{Independent error term for spatial unit } i, \\ \tau_i : \text{Spatially structured error term for spatial unit } i, \end{cases}$$

The spatially structured error term τ_i can be expressed in terms of a spatial covariance function which quantifies spatial dependence between observations; τ_i can be represented as

$$\tau_i = \epsilon_i \text{ x } C(h) \tag{4.14}$$

where
$$\begin{cases} \tau_i : \text{Spatially structured error term for spatial unit } i \\ \varepsilon_i : \text{Independent error term for spatial unit } i \\ C(h) : \text{Represents the covariance between two observations separated by distance } h \end{cases}$$

The covariance function measures how the values at one observation point compare to another as a function of their spatial distance from one another. There are many different covariance functions such as exponential, spherical and Gaussian covariance functions; exponential covariance functions model spatial dependence that

decays exponentially with distance, spherical covariance functions model spatial dependence that increases to a certain distance until a maximum point is reached and decreases to zero thereafter and Gaussian covariance functions model spatial dependence using a bell-shaped curve where closer observations are more strongly correlated than those that are farther apart. In Table 4.2, the covariance functions discussed are presented with their mathematical expressions.

Spatial Covariance Type	Mathematical Expressions
Exponential	$\mathrm{C(h)} = \exp\left(-\frac{h}{\phi}\right)$
Spherical	$\mathrm{C(h)} = \begin{cases} 1 - \frac{3}{2}\frac{h}{r} + \frac{1}{2}\left(\frac{h}{r}\right)^3 & \text{if } h \leq r \\ 0 & \text{if } h > r \end{cases}$
Gaussian	$\mathrm{C(h)} = \exp\left(-\frac{h^2}{\phi^2}\right)$

Table 4.2: Mathematical Expression for Common Covariance Functions

$$\text{where} \begin{cases} C(h) : \text{Represents the covariance between two observations separated by distance } h \\ \phi : \text{Represents the range parameter in the exponential and Gaussian covariance functions} \\ r : \text{Represents the distance at which spatial autocorrelation becomes 0 for spherical covariance functions} \end{cases}$$

For the spherical covariance type, the range parameter r is the parameter that controls the rate at which the correlation between two locations decreases as the distance h between the two locations increases. Many other covariance functions exist, but the three mentioned were chosen as they are widely used and were touched upon in the study by Muleia et al. (2020) which was discussed in Chapter 2.

4.7 Modelling and Data Considerations

HIV prevalence was estimated using the panel data. The panel consisted of N = 52 individual districts and T=3 years resulting in a panel with 156 observations. Using the first two years of 2020 and 2021 as training data, prevalence and PLHIV was estimated for 2022. The Naomi model estimations included estimated values for the mean HIV prevalence and people living with HIV (PLHIV) that were used as dependent variables for the models fitted. In addition, the ART number (residents) estimates were used for comparison in the drug consumption analysis.

Two approaches were followed in the modelling portion of the analysis. The first was to estimate prevalence directly from the quantities ordered per district, and the second approach was to calculate prevalence by first predicting PLHIV and using that estimate to calculate prevalence as prevalence is merely a ratio of PLHIV to the total population.

In cleaning the data and grouping the orders by district it was noted that 13 districts had no orders allocated. To ensure that all districts had some orders available, the orders were redistributed per province using population figures for people aged 15+. Taking Gauteng as an example, the total quantity of orders was summed up and distributed between the districts in the province on a population basis as seen in Table 4.3 showing the distribution for 2020.

The dataset with the redistributed order quantities was a logical choice to make as it has been shown in the data analysis that there is a clear relationship between population and quantities ordered.

If the imputation of the data had not been performed in this manner, then many smaller districts would have had 0 ARV orders measured which would result in inaccurate results as it would not be possible for a

District	Population	Percentage	Quantity	Distributed Quantity
Johannesburg MM	4592542.2	0.38	15447175	5927240.2
Tshwane MM	2881919.5	0.24	15447175	3719471.4
Ekurhuleni MM	3035256.3	0.25	15447175	3917371.4
Sedibeng DM	730244.7	0.06	15447175	942470.6
West Rand DM	728811.9	0.06	15447175	940621.4

Table 4.3: ARV Order Quantity Population-Based Distribution for 2020

district to have no ARV orders.

The implications of not redistributing the orders would mean that no drug utilisation estimations could be performed for many districts and that the models perform poorly. Redistributing the orders makes it such that each district receives an appropriate amount of ARV drugs based on their population.

4.8 Methodology Steps

Based on the content of this Chapter, the following process was followed to achieve the study's objectives.

1. The DDD per 1000 inhabitants per day was calculated for each district from 2020 to 2022.
2. The DDD per 1000 inhabitants per day was converted to population figures that could be compared to the number of residents on ART estimation from the Naomi model.
3. The Allocation Ratio was calculated for each district and province from 2020 to 2022 for analysis.
4. Pooled, random, and fixed effect models were fitted to the data.
5. The most appropriate modelling technique was selected using the Chow, Hausman, and LM tests.
6. Based on the results, a series of model fixed effect models were fitted using different model effects. The final models were chosen based on the RMSE score obtained.
7. The final models were used to predict prevalence and PLHIV. The predicted and calculated prevalence from PLHIV were compared to the Naomi model prevalence estimations.
8. Spatial autocorrelation was determined using Moran's I followed by the Local Morans I to determine if any clustering was present.
9. After confirming spatial autocorrelation a series of spatial linear models were fitted to predict prevalence and PLHIV by using different functional covariance types.
10. Using the AIC and RMSE, the most appropriate models were selected.
11. The final spatial models were then fitted and results were compared to the Naomi model estimates.
12. A final comparison between the panel data and spatial linear model results was performed.

Chapter 5

Results

5.1 Introduction

The results chapter presents the outcomes of the implemented methodology from Chapter 4. The chapter first presents the results of the drug allocation analysis of the two drugs studied, followed by the outcomes of the panel data and spatial linear HIV prevalence models. As initially stated, the study objectives were to establish if there is appropriate ordering of ARV drugs within the country on a district level and to determine if HIV prevalence may be predicted on a district level using medicine ordering data.

5.2 District and Provincial Level Drug Allocation Analysis

The drug allocation analysis was completed for the years 2020 to 2022 since the results of the consumption analysis were compared to the Noami model estimates for residents on antiretroviral therapy (ART) which was only available for this period. From an analysis perspective, the number of people on ART in each district would be the best estimation to compare drug consumption to, as those on ART are the People Living with HIV (PLHIV) and consuming the antiretroviral (ARV) drugs. The analysis was focused on the following:

1. Estimating district-level drug allocation and evaluating trends over the study period
2. Evaluating provincial-level drug allocation to better understand ARV drug distribution in the South African context.

5.2.1 District-Level Drug Allocation

The analysis of the district-level consumption of the efavirenz/tenofovir disoproxil/emtricitabine (TEE) and dolutegravir/lamivudine/ tenofovir disoproxil (TLD) fixed-dose ARV drugs are presented in this section. Figures 5.1 to 5.3 are plotted with district boundaries expressing the drug allocation through the ratio from Equation 4.3, which is the ratio of the estimated number of people consuming the drugs from the DDD per 1000 inhabitants per day calculation and the estimated number of residents on ART from the Naomi model estimates.

The DDD per 1000 inhabitants per day is an approximation of the proportion of a population that uses a drug. For example, in 2020 the City of Cape Town Metropolitan Municipality had a DDD value of 125.83 DDDs meaning that 12.58% of the population consumed these drugs in this period. This represents 432804.76

residents consuming one of the two ARV drugs; the full results of the DDD per 1000 inhabitants per day calculations for each year and district are provided in Appendix A. The Allocation Ratio expresses the ratio of the number of people consuming the ARV drugs to the number of residents on antiretroviral therapy (ART).

A ratio value close to 0 indicates an undersupply of ARV drugs in the district, suggesting that the available stock is insufficient. A ratio value close to 1 represents an ideal allocation, where there is adequate availability to meet the demands of PLHIV. A value greater than 1 indicates an oversupply of ARV drugs, meaning there is more stock than necessary for the district

2020 - District Analysis

The 2020 drug consumption analysis showed that apart from Mangaung, metropolitan municipalities had the largest ARV drug stock as the Allocation Ratio calculated for these regions exceeded or was close to 1 indicating that a sufficient amount of medication was available to those infected with HIV as seen in Figure 5.1. Table 5.1 indicates the calculated usage for these metropolitan municipalities.

District	**Allocation Ratio**
City of Cape Town Metropolitan Municipality	2.73
eThekwini Metropolitan Municipality	2.55
City of Johannesburg Metropolitan Municipality	1.65
City of Tshwane Metropolitan Municipality	1.41
Buffalo City Metropolitan Municipality	0.92
Ekurhuleni Metropolitan Municipality	0.91
Mangaung Metropolitan Municipality	0.27

Table 5.1: Metropolitan Allocation Ratios for 2020, Ranked from Highest to Lowest

This finding was not surprising as it aligns with Figure 3.5 that showed that metropolitan municipalities in the Eastern Cape, Western Cape, Gauteng, and KwaZulu-Natal would have higher quantities of ARV drugs available. Unlike, the metropolitan municipalities the district municipalities appeared to be under-supplied, this was evident by the Allocation Ratios calculated for these areas.

When excluding metropolitan municipalities, it could be determined that 30 of the remaining districts had ratios lower than 0.5 which implies a lack of resources. Districts falling in this group are the likes of the Central Karoo, Namakwa, Xhariep, and the JT Gaetsewe District Municipalities with populations of 52,410.37, 71,360.77, 83,455.45 and 152,113.28 which are all below 200,000.

Gauteng had the highest average Allocation Ratio of 0.85 because of the three metropolitan municipalities in the province. Gauteng was followed by the Eastern Cape, Limpopo, and Western Cape respectively with averages of 0.84, 0.76, and 0.68; the Free State had the lowest average of 0.14. Interestingly, the ARV supply available at a district level did not correlate with HIV prevalence as a district such as the City of Cape Town Metropolitan Municipality with a relatively low prevalence of 0.15 had a high Allocation Ratio of 1.65, but the Gert Sibande District Municipality which had a high prevalence of 0.28 had a calculated Allocation Ratio of 0.34. This implies that PLHIV in smaller districts must find other means to access ART, or ARV stock is delivered to health facilities in smaller districts from larger districts at a later time.

2021 - District Analysis

In 2021, there was an overall decrease in orders made by health facilities. It was determined that 42 (80.77%) of the 52 districts ordered less medication than in 2020 equating to 8,934,138 fewer units of TLD and TEE being ordered. As expected, the metropolitan municipalities accounted for a majority of the orders not made; with 4,290,998 (48,02%) fewer orders in 2021. The City of Cape Town, eThekweni, and the City of Johannesburg Metropolitan Municipalities saw the largest reduction of stock ordered with 1,690,959.22, 772,185.60, and 611,271.17 fewer units ordered respectively.

The impact of the reduction of orders could be seen in the decrease in the Allocation Ratio from Figure 5.2, with Nelson Mandela Bay Municipality experiencing the largest decrease of 1.08 from 2.39 in 2020 to 1.31 in 2021. This decrease may have been attributed to the COVID-19 pandemic which severely impacted health service delivery and the availability of resources for the period between 2020 to 2021.

A handful of districts such as the Nkangala, Ehlazeni, JT Gaetsewe, Gert Sibande, ZF Mgcawu, and Namakwa District Municipalities experienced an uptick in their Allocation Ratio of 0.31, 0.2, 0.15, 0.13, 0.13 and 0.11; indicating that more stock was available to treat PLHIV in those districts. The districts that experienced an increase in the Allocation Ratio were either part of the Northern Cape or Mpumalanga provinces.

2022 - District Analysis

2022 saw an increase of 58.21% of the quantity of stock ordered by health facilities. The increase in orders had the effect of driving up the Allocation Ratios in all districts, especially the metropolitan municipalities as seen in Figure 5.3. It is not clear why there was a substantial increase in orders during this period, but it is most likely attributed to healthcare services operating at normal levels post-COVID-19 pandemic. Compared to 2020 which had an average Allocation Ratio of 59.85%, this increased to 78.35% in 2022. In terms of the districts, the metropolitan municipalities were the biggest gainers in this regard with eThekweni Metropolitan Municipality, Nelson Mandela Bay Municipality, and the City of Cape Town Metropolitan Municipalities experiencing increases of 2.05,1.64 and 0.97 equating to 2,618,905, 2,342,495 and 2,296,208 more units ordered than the previous year.

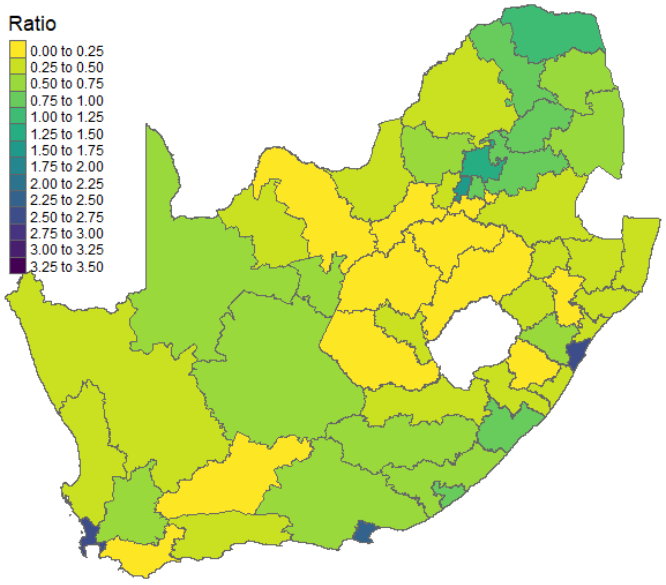


Figure 5.1: TEE and TLD Antiretroviral Drug Allocation for 2020 at a District Level

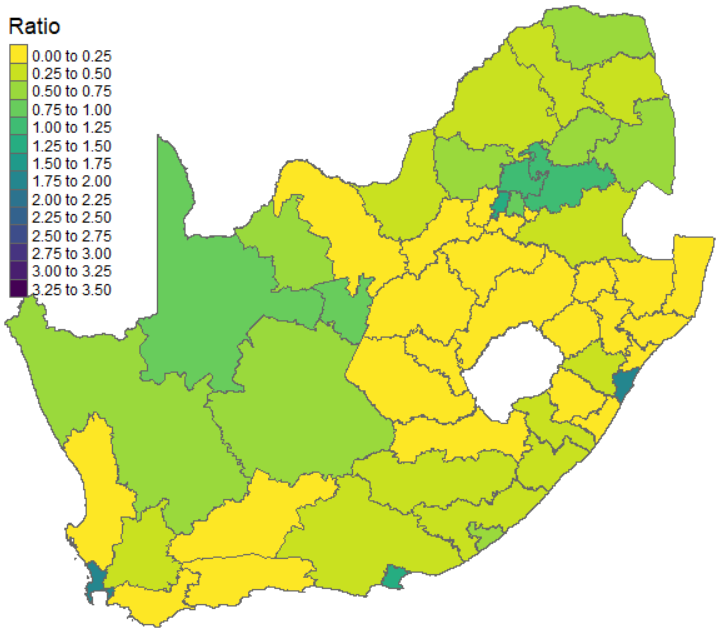


Figure 5.2: TEE and TLD Antiretroviral Drug Allocation for 2021 at a District Level

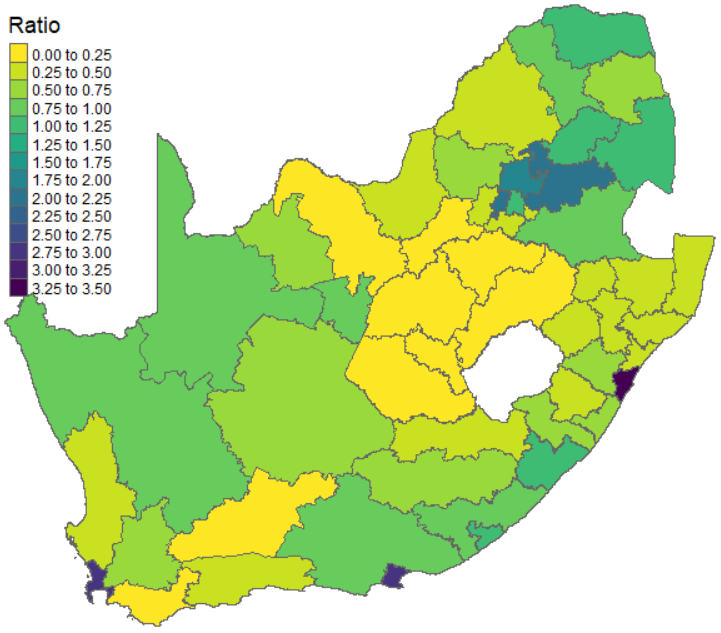


Figure 5.3: TEE and TLD Antiretroviral Drug Allocation for 2022 at a District Level

5.2.2 Provincial-Level Drug Allocation

Figures 5.1 to 5.3 are snapshots that provide a view of ARV drug availability through the Allocation Ratio for each district. These figures however do not explain how ARV drugs are distributed and obtained by those

infected with HIV. On a district level, there is a substantial difference between the quantity of ARV drugs ordered in the metropolitan and district municipalities to the extent that it appears that little ART therapy is available in the smaller less populated districts. The approximation of the estimates such as ART attendance and coverage at a district level remains a challenging task as PLHIV may access treatment in different districts other than their own for reasons such as the perceived closeness of facilities to their residences, or the quality of services from other health facilities (Estimates, 2024). Another reason for metropolitan municipalities holding large quantities of ARV drugs may be a result of them ordering and storing stock which may be distributed to small districts when required.

Table 5.2 presents the Allocation Ratios for 2020 to 2022 on a provincial level. As with the district-level results, there was a decrease in ARV availability from 2020 to 2021 before the increase in 2022. The three highest populated provinces Gauteng, KwaZulu-Natal, and the Western Cape have had sufficient ARV drug stock even with the dip in 2021 that only affected KwaZulu-Natal's ability to allocate medication to all its infected population.

Towards the end of the period studied in 2022 five provinces had Allocation Ratios above 1, meaning that they could supply every person in the district with medication with additional stock remaining. The Northern Cape and Limpopo had enough stock to treat 80% and 86% of their populations infected with HIV respectively, but the Free State and North West had insufficient ARV drug stock available to treat half of the population living with HIV.

Province	Allocation Ratio 2020	Allocation Ratio 2021	Allocation Ratio 2022
Free State	0.16	0.14	0.13
North West	0.42	0.36	0.39
Mpumalanga	0.52	0.71	1.20
Nothern Cape	0.58	0.70	0.80
Limpopo	0.76	0.46	0.86
Eastern Cape	0.86	0.49	1.06
KwaZulu-Natal	1.01	0.73	1.51
Gauteng	1.19	1.01	1.57
Western Cape	1.93	1.34	2.10

Table 5.2: Provincial Allocation of TEE and TLD Antiretroviral Drugs for 2020 to 2022

Even though on a district level there was a disparity in the amount of ARV drugs available to treat the population inflicted with HIV, there seemed to be an ample supply on a provincial level. This is due to the large quantities being held by metropolitan municipalities, which when viewed on a provincial level shows sufficient ARV drug stock to treat the provincial populations.

The assumption that the impact of the COVID-19 pandemic could explain the low ordering from 2020 to 2021 is plausible when looking at the provincial trends. A study by (Pillay et al., 2021) studied the impact of the pandemic on routine health care services and found that access to public health services was disrupted from March to December of 2020 and that the services most affected were antenatal visits, access to contraceptives, and HIV and TB testing in conjunction with reduced medical staff.

These factors could have impacted the HIV planning for 2020 and 2021 as the available resources would have been focused on fighting the COVID-19 pandemic and not necessarily focused on ART treatment programmes. If 2022 was considered a normal year, then the allocation of ARVs was sufficient on a provincial level; the unknowns relating to how and where patients receive treatment cannot be explained by the data used in this

study alone.

5.3 HIV Prevalence Modelling Results

Two different measures for prevalence were estimated. The first measure was the estimation of prevalence with quantity as the independent variable, and the second measure was the calculation of prevalence by estimating PLHIV from quantity and calculating prevalence using the ratio of PLHIV and total district population. In each model, the dependent variable prevalence was the mean prevalence estimation from the Noami model estimation outputs.

The panel data contained 156 total observations, split into 3 years with observations for each of the 52 districts. A 2:1 train-test split was used, with data from 2020 and 2021 used for training, and data from 2022 used for testing purposes.

5.3.1 Panel Model Statistical Tests

The first step was to find the most appropriate panel modelling approach, this was done through a series of tests evaluating pooled OLS, random Effect (RE), and fixed effect (FE) models against one another. First, by using the Lagrange multiplier test on a pooled model it could be determined if a basic linear regression model would be more suitable than a random effect model. The test was conducted for both instances where prevalence and PLHIV were predicted.

From Table 5.3 it can be seen that the null hypothesis was rejected in both cases when predicting prevalence and PLHIV with quantity as an independent variable. From this, using a pooled ordinary least squares regression model was rejected.

Lagrange Multiplier Test			
Dependent Variable	p-value	Hypothesis Result	Outcome
prevalence	<0.0001	H_1 : Heteroscedasticity is present	Select RE Model
PLHIV	<0.0001	H_1 : Heteroscedasticity is present	Select RE Model
Hausman's Test			
Dependent Variable	p-value	Hypothesis Result	Outcome
prevalence	0.0075	H_1 : One model is inconsistent	Select FE Model
PLHIV	<0.0001	H_1 : One model is inconsistent	Select FE Model

Table 5.3: Lagrange and Hausman Statistical Test Results for Panel Data Modelling

Using Hausman's test, the choice between the fixed and random effect modelling approaches was made. Four models were built, a fixed and random effect model for predicting both prevalence and PLHIV. Ultimately through the process of elimination, the fixed effect modelling approach was determined to be appropriate for this panel data as it would be more consistent compared to a random effect modelling approach for predicting prevalence and PLHIV.

5.3.2 Prevalence District-Level Fixed Effect Models

Three fixed effect models were fitted to predict prevalence and PLHIV for 2022. Each model made use of different fixed effects with the first taking individual effects into account, the second time effects, and the third two-way effects. Once fitted these models were compared to one another using the root mean square error (RMSE) as the metric of comparison. The three model equations are given below:

Model 1: Individual Effect

$$\text{prevalence}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \alpha_{u_i} + \mu_{u_{it}} \tag{5.1}$$

Model 2: Time Effect

$$\text{prevalence}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \gamma_t + \mu_{u_{it}} \tag{5.2}$$

Model 3: Two-way Effect

$$\text{prevalence}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \alpha_{u_i} + \gamma_t + \mu_{u_{it}} \tag{5.3}$$

The model parameter estimations in Table 5.4 show that a fixed effect model with time effects performed better than the individual and two-way effects models as it produced a lower RMSE of 0.012. In addition, Model 1 produced a large p-value indicating that there was no evidence that a correlation existed between the dependent and independent variables. It can be said that Model 2 using time effects could capture time-specific trends and changes over time, as well as reduce omitted variable bias that may influence prediction accuracy thus performing better.

	Model 1: Individual FE	Model 2: Time Effect FE	Model 3: Two-ways Effect FE
Coefficients	quantity	quantity	quantity
Estimate	-4.04 $\times 10^{-9}$	-1.83 $\times 10^{-8}$	-1.41 $\times 10^{-8}$
Std. Error	0.0001	4.7077 $\times 10^{-9}$	5.3330 $\times 10^{-9}$
Pr(>|t|)	0.482	<0.0001	0.011
RMSE	0.059	0.012	0.062

Table 5.4: Statistics Summaries for the Individual, Time, and Two-way Fixed Effect Models for Estimating Prevalence from Quantity

Examining Model 2, the model selected to predict district-level prevalence, it could be seen from the parameter estimations that when predicting prevalence with quantity as an independent variable there is a negative relationship between the two variables.

The Beta coefficient estimation for quantity was a small negative value of -1.828e-08. Given that ART is intended to extend the life span of PLHIV it would be expected that there would be a positive correlation between the variables since the mortality rate of PLHIV would decrease resulting in the proportion of PLHIV increasing resulting in higher prevalence.

The effectiveness of ART would indeed result in more PLHIV living longer, but other factors may explain the negative correlation observed due to effective ART:

- Mother-to-child vertical transmission is greatly reduced with the use of ARV drugs. With effective ART and choice of delivery route; the risk of transmission from mother to child is less than 2% compared to the range of 15 to 45% if untreated (Barral et al., 2014).
- ART leads to a reduction in the viral load of PLHIV. It has been shown that treatment reduces transmission; an individual with a viral load below 1000 copies per ml would have an almost zero chance of sexually transmitting the disease (Broyles et al., 2023).

These two points highlight the fact that accessibility to ARV drugs and treatment results in reduced HIV incidence and infections. If the rate of new infections decreases, while the general population increases then the prevalence would decrease as a smaller proportion of the population would be infected. Naomi model estimates for ART coverage showed an increase from 67.95% to 73.99% from 2020 to 2022 with a decrease in new HIV infections from 217,588 to 164,580; and a decrease in prevalence from 0.1784 to 0.1777 over the same period for people aged 15+.

An increase of 1 unit of quantity would result in a small decrease in prevalence of -1.828e-08 units and vice versa. The time effect term captures changes over time that are correlated with quantity, therefore when interpreting the Beta coefficient in this case the negative correlation may reflect the true relationship between prevalence and quantity when considering the above points and that new infections appear to be decreasing year-on-year.

In the context of this study, where thousands of individual packages of ARV drugs are ordered; although there is a small change in prevalence when quantity changes, the sheer magnitude of ARV drugs ordered could affect prevalence if enough PLHIV obtain treatment and adhere to it.

5.3.3 PLHIV District-Level Fixed Effect Models

Similar to the prevalence models, the PLHIV FE models were assessed. The addition of the PLHIV was included to determine if calculating prevalence using an estimation of those inflicted with HIV would provide a more accurate result. The PLHIV models are described by Equations 5.4 to 5.6.

Model 4: Individual Effect

$$\text{PLHIV}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \alpha_{u_i} + \mu_{u_{it}} \tag{5.4}$$

Model 5: Time Effect

$$\text{PLHIV}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \gamma_t + \mu_{u_{it}} \tag{5.5}$$

Model 6: Two-way Effect

$$\text{PLHIV}_{u_{it}} = \beta\text{quantity}_{u_{it}} + \alpha_{u_i} + \gamma_t + \mu_{u_{it}} \tag{5.6}$$

Table 5.5, showing the parameter estimates for Models 4,5, and 6 indicated that for all the models quantity had a relationship with with PLHIV as significant p-values(p-value < 0.05) were obtained. The model making use of time effects had the best performance based on the RMSE value of 16280.7700 which was lower than that of Models 4 and 6.

The Beta coefficient for Model 5 indicated that the relationship between quantity and PLHIV was negative, which would result in a 1 unit increase in quantity resulting in a decrease in PLHIV by 0.027 units. Similar to prevalence, ART would decrease the mortality rate of PLHIV but would also result in fewer new HIV infections and a decrease in PLHIV over time as the general population grows.

	Model 4: Individual FE	Model 5: Time Effect FE	Model 6: Two-ways Effect FE
Coefficients	quantity	quantity	quantity
Estimate	0.125	-0.027	-0.031
Std. Error	0.006	0.004	0.005
Pr(>|t|)	<0.0001	<0.0001	<0.0001
RMSE	88245.800	16280.770	188944.220

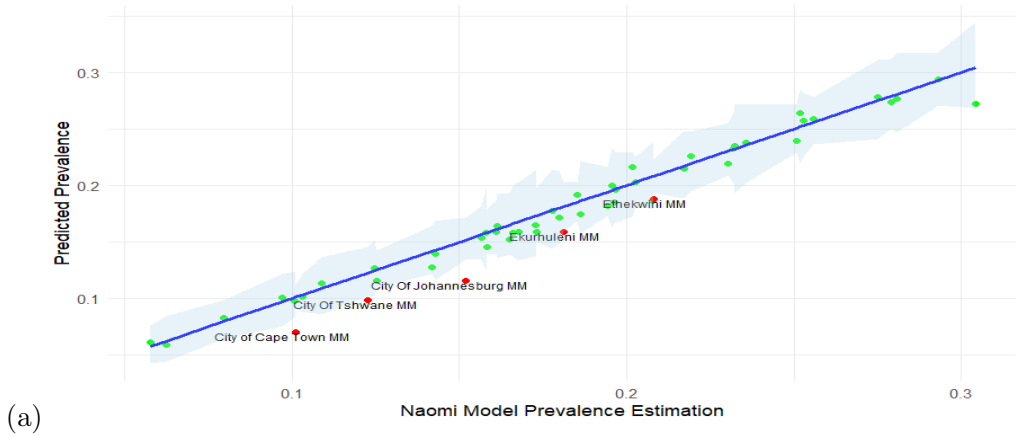
Table 5.5: Statistics Summaries for the Individual, Time and Twoways Fixed Effect Models for Estimating PLHIV from Quantity

5.4 Prevalence Results

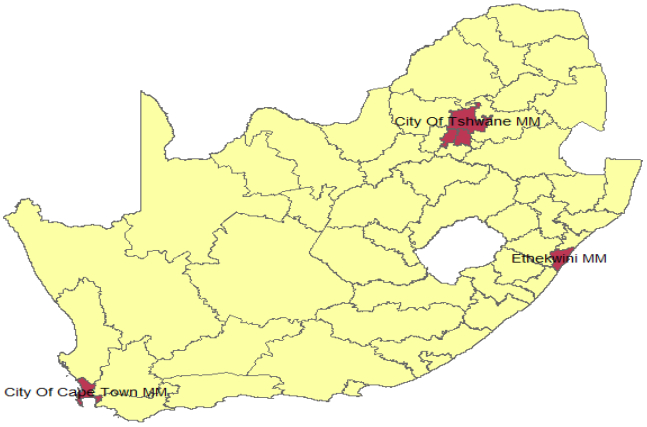
Figure 5.4 presents the predicted and calculated prevalence results; the complete results are available in Appendix B which provides the predicted and calculated prevalence for each district for 2022. The calculated prevalence was obtained from the ratio of PLHIV and the total population of each district for those aged 15+, this ratio is shown in Equation 5.7.

$$\text{Prevalence} = \frac{\texttt{PLHIV}}{\texttt{District Population}} \tag{5.7}$$

The Naomi model provides estimates for prevalence for each district with the mean, lower, and upper bound estimates. Figure 5.4a shows the predictions relative to the mean prevalence and estimation bounds for Model 2. The plot of the South African districts in Figure 5.4b shows that the model failed to accurately predict within the bounds for the City of Cape Town, City of Johannesburg, City of Tshwane, Ekuhurleni, and eThekwini Metropolitan Municipalities as all estimations fell below the lower bound of the Naomi estimates.



(a)

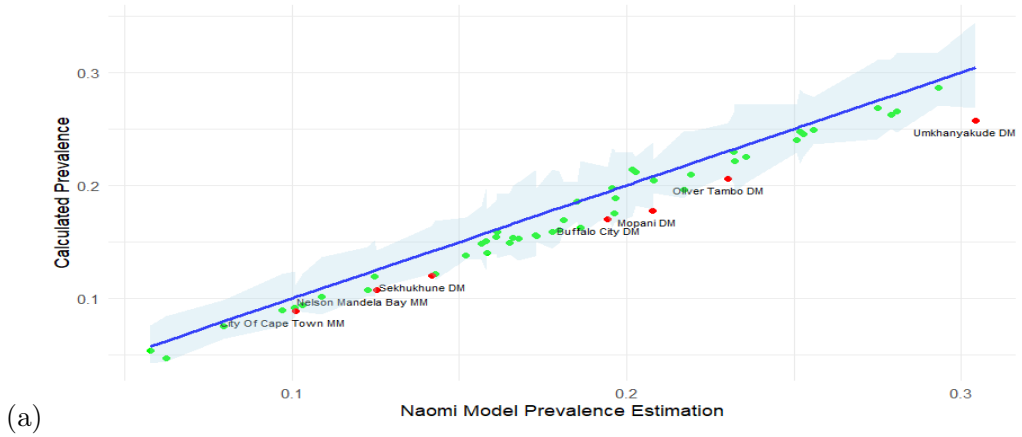


(b)

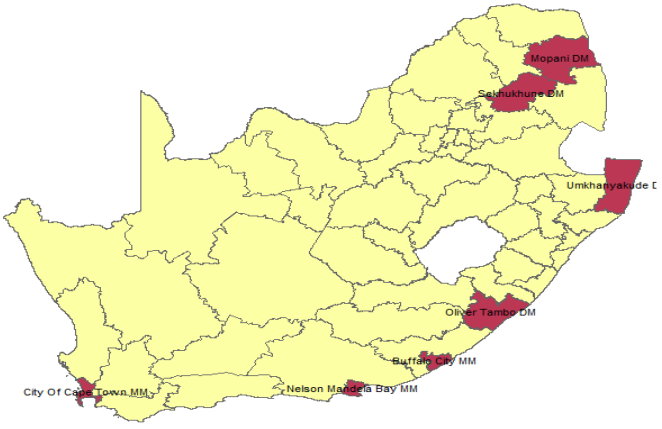
Figure 5.4: a) Plot Comparing Predicted Prevalence Results to the Naomi Model Prevalence Estimations b) District-Level Map of South Africa Indicating Prevalence Prediction Success

The poor performance of Model 2 in predicting prevalence in the Metropolitan districts could be attributed to the high quantities of ARV drugs ordered by these districts. Referring back to Table 5.4, the negative Beta coefficient for quantity would result in a decreasing prevalence as quantity increased. Due to the high quantities ordered the prevalence in these regions would result in lower estimations as the prevalence estimations were driven down by high quantity.

Examining the results of Model 5 with the calculated prevalence from PLHIV; the model failed to accurately predict seven district's prevalence within the acceptable bounds as seen in Figures 5.5a and 5.5b. Unlike Model 2, this model only failed to predict the prevalence for three metropolitan municipalities, the City of Cape Town, Buffalo City, and Nelson Mandela Bay Metropolitan Municipalities. A possible reason for this was that the estimations for the PLHIV were not impacted as much by the high quantities in the metropolitans since the number of PLHIV tends to increase with high population densities.



(a)



(b)

Figure 5.5: a) Plot Comparing Calculated Prevalence Results to the Naomi Model Prevalence Estimations b) District-Level Map of South Africa Indicating Prevalence Calculation Success

5.5 Spatial Auto Correlation - Moran's I

Before attempting to explore the use of spatial modelling techniques, the dataset had to be assessed for any spatial autocorrelation. A shapefile providing the administrative boundaries or districts of South Africa was used to provide the geographical information used in the spatial analysis performed. This was confirmed with the use of the Moran's I test; Table 5.6 contains the results of the Moran's I test for each year from 2020 to 2022.

Table 5.6: Moran's I Test Results

Year	Moran I Statistic	p-value
2020	0.333	0.000
2021	0.146	0.038
2022	0.328	0.000

With statistically significant p-values (p-value<0.05) it was confirmed that there was spatial correlation and that quantity was not randomly distributed. The Moran's I statistics from Table 5.6 suggested that for each year there was positive spatial autocorrelation. In 2021, Moran's I statistic was the lowest of the three years

indicating relatively weak spatial clustering in this period but there was still evidence of a spatial pattern.

In 2020 and 2022 the Moran's I statistics indicated a moderate degree of spatial clustering. Districts with high quantities of ARV drugs were somewhat close to other high-ordering areas, and districts with low quantities were somewhat close to other low-ordering areas. Figures 5.6 to 5.8 shows the plots of Moran's I for each district; as mentioned in the drug allocation analysis the plot in Figure 5.7 showed that in 2021 the ordering behavior differed significantly from 2020 and 2022. As seen by the p-values for each year, although significant, the p-value for the Morans's I test in 2021 is relatively close to the threshold of 0.05 which further suggests that in 2021 fewer resources were allocated to HIV-related efforts.

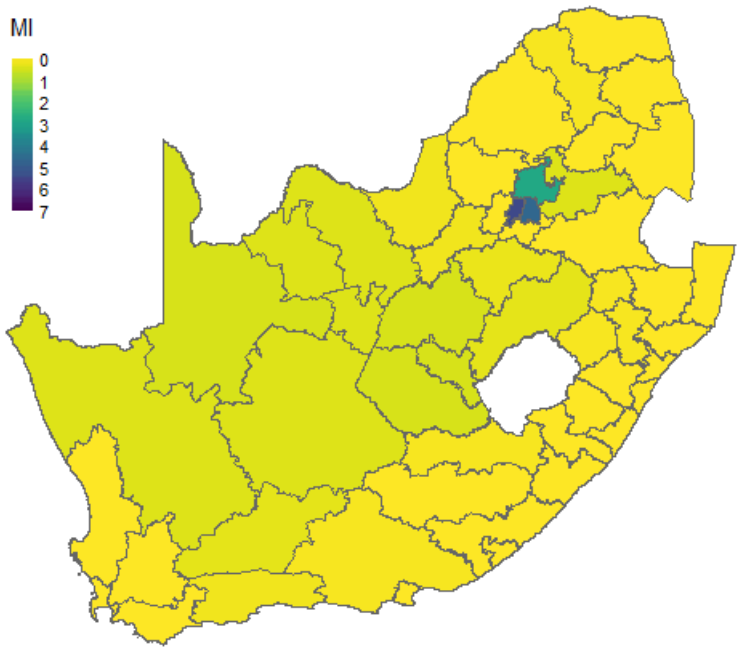


Figure 5.6: Moran's I Spatial Autocorrelation Analysis for 2020

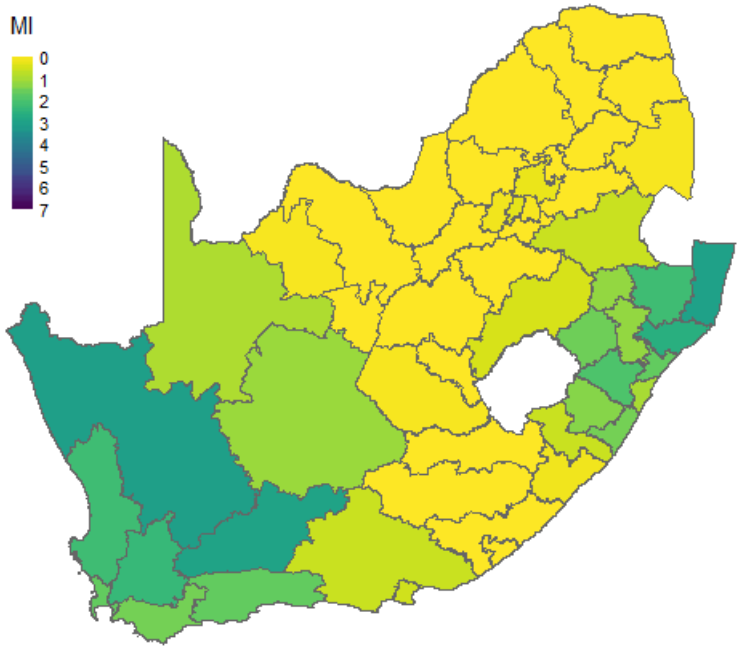


Figure 5.7: Moran's I Spatial Autocorrelation Analysis for 2021

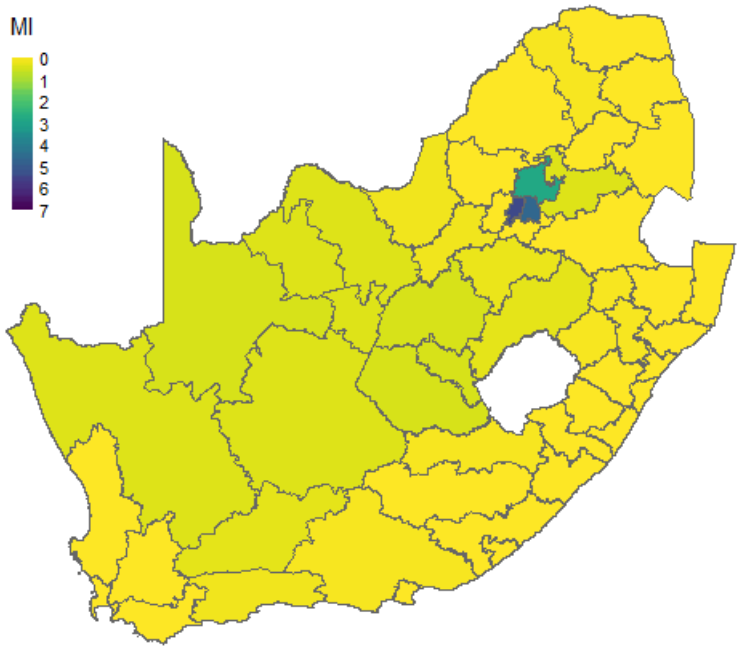


Figure 5.8: Moran's I Spatial Autocorrelation Analysis for 2022

5.5.1 Local Moran's I- Local Spatial Autocorrelation

Moran's I is a global view of the spatial correlation but does not give a granular description of spatial correlation. Using the local Moran's I, local patterns, clusters, and hotspots could be identified. More importantly, the spatial relationship between each district to its neighbor was assessed.

Figures 5.9 to 5.11 show the clustering of the districts based on the quantities of stock ordered to reveal neighbour relationships. For all three years, only High-High and Low-Low clusters were present with High-High clusters indicating districts with high quantities of ARV stock close to other districts with high stock and Low-Low clusters indicating neighbouring districts with low stock quantities grouped. The High-High and Low-Low clusters present occurred within the same regions of the country every year; the Low-Low clusters were focused around the Northern Cape, Free State, and parts of the Western Cape such as the Namakwa, Pixeley ka Seme and ZF Mgcawu District Municipalities.

The High-High clusters were focused around Gauteng and its districts. This made sense since Gauteng, more specifically, the City of Johannesburg has the highest population and therefore a high number of PLHIV requiring ART. The metropolitan municipalities stock large quantities of ARV drugs; three metropolitan municipalities exist within the borders of Gauteng which would increase the likelihood of high-high clusters occurring in this region.

The results of the Moran's I and LISA analysis indicated that spatial autocorrelation was indeed present. The implementation of spatial models would therefore be appropriate considering the spatial autocorrelation analysis outcomes.

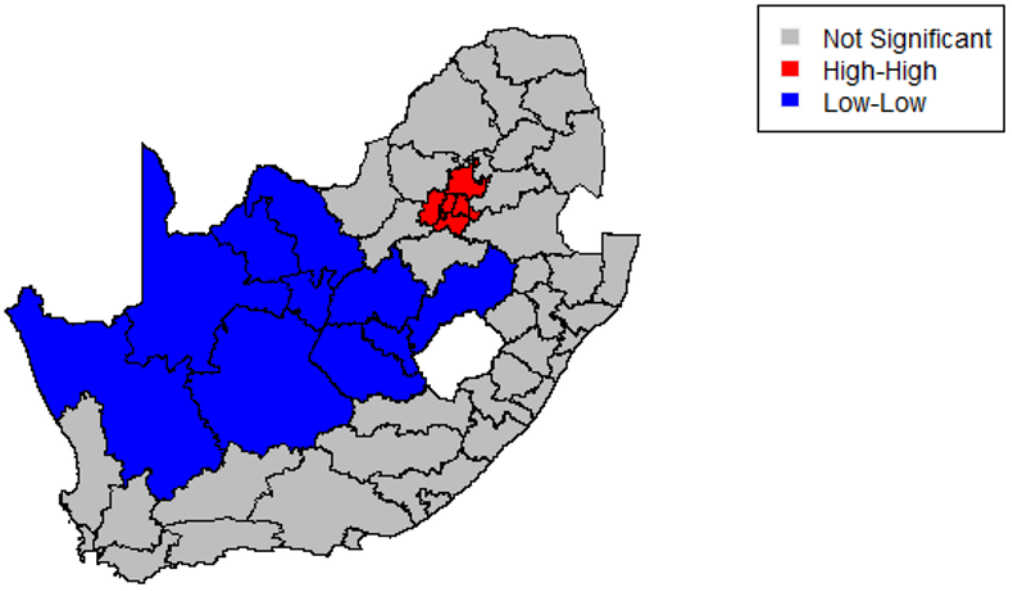


Figure 5.9: Local Indicators of Spatial Association (LISA) Cluster Map Showing Spatial Relationships Between Neighbouring Districts for 2020

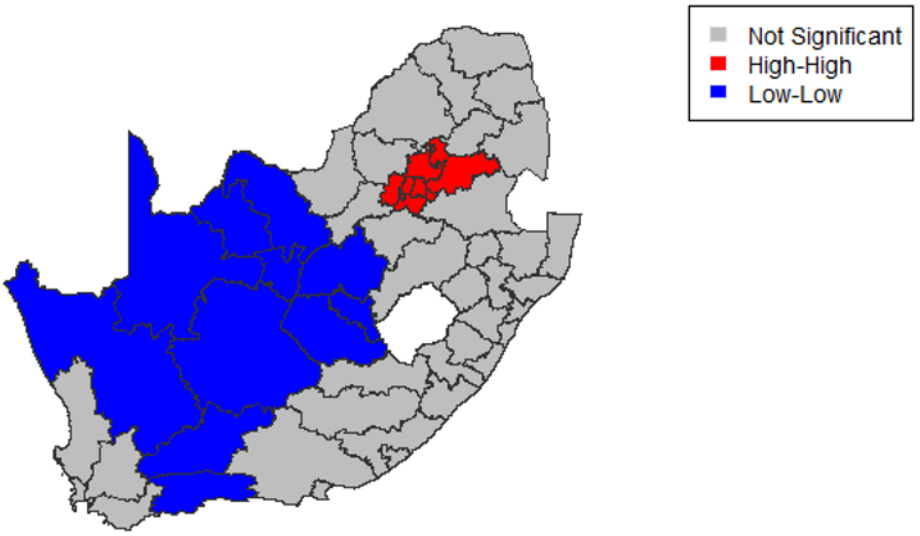


Figure 5.10: Local Indicators of Spatial Association (LISA) Cluster Map Showing Spatial Relationships Between Neighbouring Districts for 2021

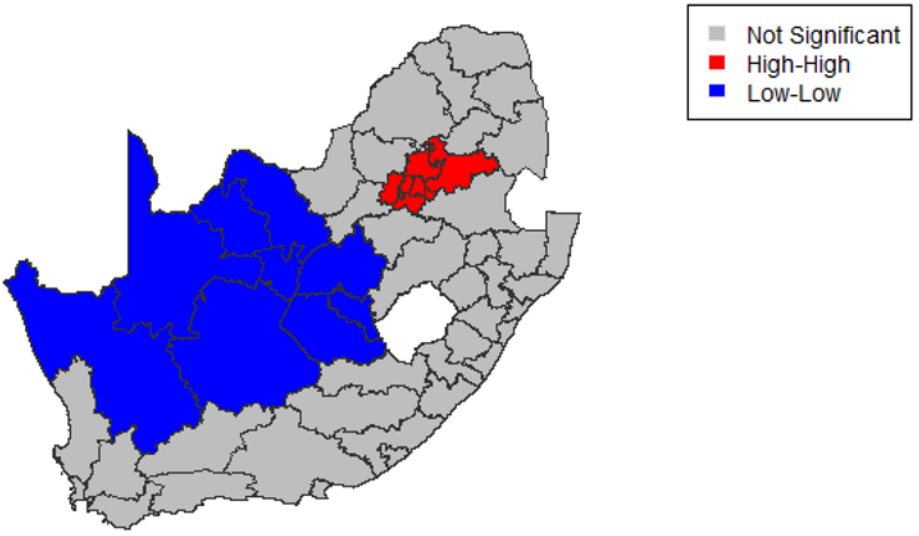


Figure 5.11: Local Indicators of Spatial Association (LISA) Cluster Map Showing Spatial Relationships Between Neighbouring Districts for 2022

5.6 Spatial Modelling

A similar approach was taken when implementing the spatial linear models. Prevalence was predicted with a spatial model, followed by a model used to predict PLHIV which was then used to calculate prevalence.

Prevalence Spatial Model

A series of spatial models were fitted and compared to each other to determine the best for estimation purposes. The parameter altered in each model was the spatial covariance type, which refers to how the

spatial relationships between observations are modeled. Table 5.7 shows the parameter estimates of the fitted models using exponential, spherical, and Gaussian functional covariance types for predicting prevalence.

Spatial Model 1 : Exponential Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.184	0.095	1.940	0.052
Quantity	-8.670 $\times 10^{-9}$	2.84 $\times 10^{-9}$	-3.052	0.002
Coefficients (exponential covariance):				
de	0.0131			
ie	0.000			
Range	20.640			
Metrics				
AIC	-468.475			
RMSE	0.009			
Spatial Model 2 : Spherical Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.186	0.068	2.716	0.007
Quantity	-9.10 $\times 10^{-9}$	2.86 $\times 10^{-9}$	-3.184	0.002
Coefficients (spherical covariance):				
de	0.001			
ie	0.000			
Range	21.870			
Metrics				
AIC	-468.208			
RMSE	0.009			
Spatial Model 3 : Gaussian Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.195	0.0362	5.391	0.000
Quantity	-9.48 $\times 10^{-9}$	2.43 $\times 10^{-9}$	-3.900	0.000
Coefficients (Gaussian covariance):				
de	0.004			
ie	0.001			
Range	6.029			
Metrics				
AIC	-408.921			
RMSE	0.020			

Table 5.7: Statistics Summaries for the Spatial Linear Models for Estimating Prevalence from Quantity

There was little difference in the Spatial Models 1 and 2 as an RMSE value of 0.009 was achieved by both which were the best-performing models. Ultimately Spatial Model 1 was selected as it had a better AIC score; this model used an exponential spatial covariance function that assumes that covariance between two observations decreases exponentially over distance. As with the fixed effect model, this model also indicated a negative correlation between prevalence and quantity.

From the dependent random error (de) term, and independent random error (ie) term it could be deduced that the spatial autocorrelation between districts holds over longer distances and that there is little variability at close spatial distances, which makes sense in the context of this study as from the local Moran's I in Figure 5.11 its seen that there is low variability across many districts in terms of their relationships with one another.

PLHIV Spatial Models

Spatial linear models were fitted to predict PLHIV. Table 5.8 provides the parameter estimates for each of the models fitted. As before different functional covariance types were used for each model fitted; from the AIC it could be seen that the spatial model using the exponential spatial covariance was the best-fit model; in addition, Spatial Model 4 produced the lowest RMSE.

Looking at the effect of the quantity on PLHIV, from the Beta coefficient estimate there was a negative relationship between PLHIV and quantity; an increase in quantity would result in a decrease in PLHIV. When examining the range, and the dependent and independent random error exponential spatial covariance coefficients; the high dependent random error term value suggested that the impact of quantity decreases more rapidly as the spatial distance increases. The independent error term's effect in terms of the PLHIV was high and indicated that there is a strong autocorrelation in PLHIV for neighbouring districts. This is apparent when considering Figure 5.8 which shows two distinct clusters of low-low and high-high neighbours.

The range value typically determines the spatial scale at which correlation becomes negligible; a small range would imply the effect of quantity on PLHIV is localised and reduces rapidly with distance, whereas a high range would suggest that the effect of quantity is widespread and persists with distance. From this, it could be said that the range is not large enough for quantity to have a significant effect on PLHIV. When considering the spatial autocorrelation between highly populated areas with large quantities of ARV drugs compared to smaller regions, it can be seen that there is a big difference in the quantity available.

Spatial Model 4 : Exponential Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.620×10^5	2.25×10^4	7.200	0.000
quantity	-0.020	0.004	-4.569	0.000
Coefficients (exponential spatial covariance):				
de	2.571×10^{10}			
ie	6.388×10^7			
range	0.0365			
Metrics				
AIC	2503.660			
RMSE	1.960908×10^5			
Spatial Model 5 : Spherical Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	8.087×10^4	1.750×10^5	0.462	0.644
quantity	0.116	0.007	15.763	0.000
Coefficients (spherical spatial covariance):				
de	4.752×10^{10}			
ie	1.622×10^9			
range	35.600			
Metrics				
AIC	2578.821			
RMSE	2.710463×10^5			
Spatial Model 6 : Gaussian Covariance				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	6.418×10^4	4.671×10^4	1.374	0.169
quantity	0.118	0.006	20.516	0.000
Coefficients (Gaussian spatial covariance):				
de	3.601×10^9			
ie	3.601×10^9			
range	11.690			
Metrics				
AIC	2584.146			
RMSE	2.678×10^5			

Table 5.8: Statistics Summaries for the Spatial Linear Models for Estimating PLHIV from Quantity

5.7 Spatial Prevalence Model Results

The spatial models yielded better results than the FE models fitted to predict and calculate prevalence. The RMSE obtained from the spatial models was 0.009 and 0.012 for the predicted and calculated prevalence compared to 0.012 and 0.015 for the fixed effect models. Note that the RMSE values for the calculated prevalence were obtained by treating the calculated prevalence values as predictions in the RMSE calculation.

The improvement in RMSE from the FE models resulted in the predicted and calculated prevalence estimations being more accurate with only two estimations for the predicted prevalence out of bounds and one estimation for the calculated prevalence out of the acceptable range dictated by the lower and upper bounds of the Naomi model.

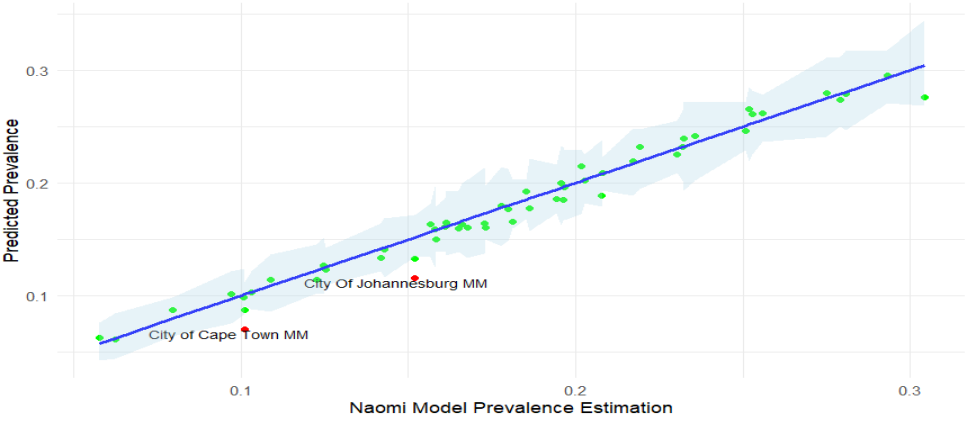
Figures 5.12 and 5.13 show the results of the model predictions. As seen in Figures 5.12 only two predictions were out of bounds being the City of Cape Town and the City of Johannesburg. Interestingly when comparing 5.12b and 5.13b the calculated prevalence produced more accurate predictions with only one inaccurate prediction being the uMkhanyakude District Municipality.

The errors for the prevalence model occurred when estimating for the City of Cape Town and Johannesburg Metropolitan Municipalities, showing the influence of the high quantities when predicting prevalence in these two areas which have relatively low HIV prevalence compared to the large amount of stock they hold.

The uMkhanyakude District was the only estimation out of bounds for the calculated prevalence. The FE model also failed to make a good estimate as both model estimates were below the lower limit; the spatial model calculation estimated prevalence at 0.262 (26.20%) while the FE model calculation was 0.257 (25.70%). The uMkhanyakude District Municipality had the highest average HIV prevalence sitting at 0.304 (30.40%) for 2022, this combined with the relatively low quantity of orders from this region could be the reason for both models failing to produce satisfactory predictions.

Examining the results as a whole, it appears that the models predict poorly under one of two circumstances. The first is low prevalence and high ARV stock, as seen in the City of Cape Town and Johannesburg Metropolitan Municipality estimations, and the second is when there is high prevalence and low ARV stock as in the case of the uMkhanyakude District Municipality.

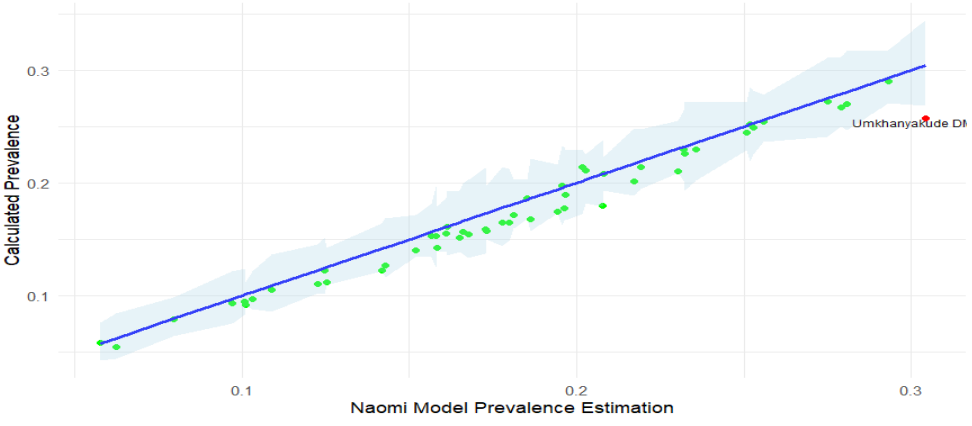
Comparing Figure 5.12a and Figure 5.13a which provide a view of the prevalence estimations for the predicted and calculated prevalence in comparison to the mean prevalence which is the Naomi prevalence estimation. It can be seen from Figure 5.13a, that the estimations of the calculated prevalence were below the mean but the predictions shown in Figure 5.12a were more evenly distributed. Although the calculated prevalence produced more accurate results that were within acceptable bounds, the overall estimations were not as close to the actual HIV prevalence from the Naomi model. Even though the calculated prevalence produced fewer prediction errors, the overall predictions were not as close to the actual HIV prevalence from the Naomi model meaning that consideration would have to be taken when deciding which estimates to use.



(a)

(b)

Figure 5.12: a) Plot Comparing Predicted Spatial Linear Model Prevalence Results to the Naomi Model Prevalence Estimations b) District-Level Map of South Africa Indicating Prevalence Prediction Success



(a)

(b)

Figure 5.13: a)Plot Comparing Calculated Spatial Linear Model Prevalence Results to the Naomi Model Prevalence Estimations b) District-Level Map of South Africa Indicating Prevalence Calculation Success

Chapter 6

Discussion and Conclusion

6.1 Conclusion

The study had two objectives, the first was using ordering data of antiretroviral (ARV) drugs to estimate the proportion of the population using ARV drugs on a district-level. This was done by calculating an Allocation Ratio which would indicate ARV availability within each district. The second objective was to determine if district-level Human Immunodeficiency Virus(HIV) prevalence could be predicted using panel data and spatial linear models since this is the administration level where HIV planning and resource allocation are typically undertaken. By analysing the two drugs selected, efavirenz/tenofovir disoproxil/emtricitabine (TEE) and dolutegravir/lamivudine/tenofovir disoproxil (TLD) these objective could be realised as they make up more than 70% of the ordered ARV drugs at 31.27% and 40.12% respectively.

The distribution of TEE and TLD was analysed by converting the quantities of ARV drugs ordered to a meaningful value, being the DDD per 1000 inhabitants per day (DDDs) which relates to the number of people in the population using a given drug. Through this exercise, the number of people using these drugs could be compared to the number of residents on antiretroviral treatment (ART) estimation from the Naomi model. The district-level analysis showed that in general, metropolitan municipalities held the most stock compared to surrounding district municipalities. The allocation of TEE and TLD in metropolitan districts far exceeded that which was needed by the populations in these regions whereas the district municipalities would generally not have enough stock to meet the needs of People Living with HIV(PLHIV).

Performing this analysis on a provincial level revealed that even though on a district-level many district municipalities would not be able to provide treatment to all of its residents living with HIV, when scaling up to a provincial level there was indeed enough ARV drugs available to treat PLHIV. The Western Cape, Eastern Cape, Gauteng, Mpumalanga and KwaZulu-Natal had allocation ratios above 1 by 2022, the Northern Cape and Limpopo were above 0.8 meaning at least 80% of PLHIV would have access to treatment while the Free State and North West had insufficient inventory over all three years.

Although it was not confirmed, it can be assumed that there is cross-district travel for treatment as there are enough ARV drugs available that people living in district municipalities could indeed receive treatment from one of the neighbouring districts with an oversupply of ARV drugs. A person in the West Rand District Municipality could easily receive treatment in the City of Johannesburg Metropolitan Municipality due to factors such as better health facilities, health professionals, or perceived quality of treatment. The year-on-year ordering showed that in 2022 a substantial amount of ARV drugs were ordered when compared to 2020 and 2021. This could be a result of the COVID-19 pandemic which resulted in a reallocation of resources

and less focus on HIV treatments and planning.

The second objective was to determine if ordering data could be used to predict district-level HIV prevalence. In terms of the modelling, it was found that models taking spatial effects into account perform better than those that do not. The fixed effect models used to predict prevalence and PLHIV on a district level produced satisfactory results, as the prevalence model could accurately predict 47 of the 52 districts within the acceptable bounds defined by the Naomi model's output. The prevalence calculated from the PLHIV predictions resulted in 45 acceptable prevalence estimations but was better in predicting prevalence for the metropolitan municipalities. A negative relationship between prevalence and quantity was discovered which was unexpected since antiretroviral therapy (ART) treatment would increase the lifespan of those living with HIV thus increasing prevalence. This assumption was true but didn't take into account that prevalence may be reduced when effective ART is provided resulting in a decrease in new HIV infections thus reducing prevalence over time as the proportion of PLHIV in the general population decreases.

It was found that there was spatial autocorrelation within the data. Low-low and high-high clusters were discovered, with the low-low clusters found in the southwest region of the country where many low-populated districts are located; this is expected as these districts with populations less than 200,000 would not require as many resources. The high-high cluster found in Gauteng and surrounding areas was expected as there are three metropolitan municipalities in this region of the country.

The results of the spatial linear models showed that incorporating spatial effects improved prediction results. The predicted prevalence from the spatial linear model produced two estimations out of the acceptable range but the calculated prevalence from the PLHIV estimation produced one out-of-bounds estimation.

In terms of the results of the models, it was apparent that the fixed effect models were more sensitive to the effects of the metropolitan municipalities as the increasingly high quantities of ARV drugs resulted in predictions lower than that of the Naomi model's lower bound for each metropolitan district. Unlike the fixed effect models, the spatial linear models performed better and were less influenced by the metropolitan municipalities due to the influence of the spatial effects included. It is apparent that the models struggle to predict accurately when a district has large stockpiles of ARV stock and low HIV prevalence as seen in the City of Cape Town and Johannesburg Municipalities, and when districts have low stock but high HIV prevalence as seen with the uMkhanyakude District Municipality.

6.1.1 Limitations

Several limitations may have impacted the study. The identified limitations are:

1. **Data Quality** - The RSA Pharma data used for this study had data quality issues related to how customer names were recorded, with many entries for the same customer or health facility having different spelling. This required cleaning the customer names for thousands of observations which could have resulted in errors and some customer orders not being included after data pre-processing.
2. **Methodology Limitations** - The DDD per 1000 inhabitants per day methodology used for the drug utilisation analysis has the shortcoming of not adequately accounting for factors that could affect drug utilisation such as health care access and distribution of medications. It was apparent from the analysis conducted that the data and methods used do not explain how and where PLHIV obtain their ARV medication as many districts appeared to be understocked and metropolitan municipalities were overstocked.
3. **COVID-19 Impact** - Disruptions in the healthcare system due to the pandemic had an impact on the data for 2021 as evidenced by the reduction of ARV drug orders during this period. To obtain

a clearer picture of the ARV procurement landscape in South Africa it would be advised to use data after the pandemic when the healthcare system recovered.

4. **Resource Limitations** - No R package was found that can effectively work with spatial panel data to take individual and time-specific effects into account. The splm package (Bivand et al., 2021) can be used to fit fixed and random effect spatial panel data models but lacks the functionality to predict on unseen data.

5. **Knowledge Gaps** - The results produced only make use of ordering data to provide a view of the South African HIV landscape in terms of ARV availability. There is a gap in the analysis relating to the societal factors relating to healthcare access, gender equality and stigma which could influence adherence to ART. This study assumes that all ARV drugs ordered are consumed in the same year when that may not be the case.

6.1.2 Recomendations

Recommendations that may improve the study or future studies are given:

1. **Methodological Improvements** - Although they represent more than 70% of the ARV drugs ordered, TEE and TLD consumption do not represent the entire population who are on ART. In future studies, focus could be placed on including other ARV drugs to improve the study's overall results.

2. **Data Collection** - To improve the quality of the models built, more data must be obtained from years following the COVID-19 pandemic as this would be a better representation of a normal operating health system.

3. **Future Research** - The study has shown that medicine ordering data has the potential to be used in estimating prevalence. By following a similar approach, other diseases of interest may be studied.

   Considerations that should be taken into account are:

   - If the appropriate medications used to treat the disease are identified. These medications would need to be those used by a majority of the population suffering from the disease or health condition to provide the best outcomes.
   - For drug utilisation analysis, it would be ideal to analyse diseases or conditions where medications are taken for prolonged periods such as diabetes or hypertension. Drugs that allow for consistent and reliable analysis are better for studying long-term trends in a population.

6.1.3 Concluding Remarks

The results of this study have shown that there is potential to build upon this work in future studies as this concept of using ordering data to estimate drug utilisation and disease prevalence is possible. This work could be refined and adapted to diseases that do not have established models available. The analysis results have highlighted the fact there is a gap in the study that cannot account for how and where PLHIV receive their treatment; the methods used in this study can only account for the quantities ordered by each district but not how the medication is distributed afterward.

Bibliography

Astawesegn, F. H., Stulz, V., Conroy, E., and Mannan, H. (2022). Trends and effects of antiretroviral therapy coverage during pregnancy on mother-to-child transmission of hiv in sub-saharan africa. evidence from panel data analysis. *BMC Infectious Diseases*, 22:1–13.

Baltagi, B. H., Feng, Q., and Kao, C. (2012). A lagrange multiplier test for cross-sectional dependence in a fixed effects panel data model. *Journal of Econometrics*, 170(1):164–177.

Barral, M. F. M., de Oliveira, G. R., Lobato, R. C., Mendoza-Sassi, R. A., Martínez, A. M. B., and Gonçalves, C. V. (2014). Risk factors of hiv-1 vertical transmission (vt) and the influence of antiretroviral therapy (art) in pregnancy outcome. *Revista do Instituto de Medicina Tropical de Sao Paulo*, 56(2):133—138.

Bivand, R., Millo, G., and Piras, G. (2021). A review of software for spatial econometrics in R. *Mathematics*, 9(11):1–40.

Biørn, E. (2017). *Econometrics of Panel Data: Methods and Applications*. Oxford University Press.

Broyles, L., Luo, R., Boeras, D., and Vojnov, L. (2023). The risk of sexual transmission of hiv in individuals with low-level hiv viraemia: a systematic review. *The Lancet*, 402:464–471.

Colonescu, C. (2017). *Using R for Principles of Econometrics*. Lulu.com.

Criado-Alvarez, J. J., Barrientos, C. R., González, J., Montero, J. C., and Moriano, A. M. (2017). Estimating the prevalence of attention deficit hyperactivity disorder in castile-la mancha, spain (1992-2020). *Journal of Childhood Developmental Disorders*, 3(1):1–4.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2):1–43.

Dooley, K. E., Kaplan, R., Mwelase, N., Grinsztejn, B., Ticona, E., Lacerda, M., Sued, O., Belonosova, E., Ait-Khaled, M., Angelis, K., et al. (2020). Dolutegravir-based antiretroviral therapy for patients coinfected with tuberculosis and human immunodeficiency virus: a multicenter, noncomparative, open-label, randomized trial. *Clinical Infectious Diseases*, 70(4):549–556.

Dumelle, M., Higham, M., and Ver Hoef, J. M. (2023). spmodel: Spatial statistical modeling and prediction in r. *PLOS ONE*, 18(3):1–32.

Eaton, J. W., Dwyer-Lindgren, L., Gutreuter, S., O'Driscoll, M., Stevens, O., Bajaj, S., Ashton, R., Hill, A., Russell, E., Esra, R., et al. (2021). Naomi: a new modelling tool for estimating hiv epidemic indicators at the district level in sub-saharan africa. *Journal of the International AIDS Society*, 24:27–39.

Elhorst, J. P. (2017). Spatial panel data analysis. *Encyclopedia of GIS*, 2:2050–2058.

Estimates, H. (2024). South africa district hiv estimates. https://www.hivdata.org.za/.

Government, S. A. (1998). Local government: Municipal structures act. Act No. 117 of 1998.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, 7th edition.

Gujarati, D. and Porter, D. (2009). *Basic Econometrics*. Economics series. McGraw-Hill Irwin.

Hollingworth, S. and Kairuz, T. (2021). Measuring medicine use: Applying atc/ddd methodology to real-world data. *Pharmacy*, 9(1):1–8.

Johnson, L. F. and Dorrington, R. E. (2021). Modelling the impact of hiv in south africa's provinces: 2020 update. *Cape Town: Centre for Infectious Disease Epidemiology and Research, University of Cape Town*, pages 1–136.

Kitenge, M. K., Fatti, G., Eshun-Wilson, I., Aluko, O., and Nyasulu, P. (2023). Prevalence and trends of advanced hiv disease among antiretroviral therapy-naïve and antiretroviral therapy-experienced patients in south africa between 2010-2021: a systematic review and meta-analysis. *BMC Infectious Diseases*, 23(1):1–16.

Laugesen, K., Jørgensen, J. O. L., Sørensen, H. T., and Petersen, I. (2017). Systemic glucocorticoid use in denmark: a population-based prevalence study. *BMJ open*, 7(5):1–5.

Lee, H. (2008). Using the chow test to analyze regression discontinuities. *Tutorials in Quantitative Methods for Psychology*, 4:46–50.

Maggiolo, F., Gulminetti, R., Pagnucco, L., Digaetano, M., Benatti, S., Valenti, D., Callegaro, A., Ripamonti, D., and Mussini, C. (2017). Lamivudine/dolutegravir dual therapy in hiv-infected, virologically suppressed patients. *BMC infectious diseases*, 17(1):1–7.

Mathur, M. (2015). Spatial autocorrelation analysis in plant population: An overview. *Journal of Applied and Natural Science*, 7:501–513.

Mukuna, D. M., Decroo, T., and Nyapokoto, C. M. (2024). Effect of dolutegravir-based versus efavirenz-based antiretroviral therapy on excessive weight gain in adult treatment-naïve hiv patients at matsanjeni health center, eswatini: a retrospective cohort study. *AIDS research and therapy*, 21(1):1–5.

Muleia, R., Boothe, M., Loquiha, O., Aerts, M., and Faes, C. (2020). Spatial distribution of hiv prevalence among young people in mozambique. *International Journal of Environmental Research and Public Health*, 17(3):1–20.

Olveira, G., Tapia, M. J., Colomo, N., Muñoz, A., Gonzalo, M., Federico, C., et al. (2009). Usefulness of the daily defined dose method to estimate trends in the consumption, costs and prevalence of the use of home enteral nutrition. *Clinical Nutrition*, 28(3):285–290.

Patel, R., Evitt, L., Mariolis, I., Di Giambenedetto, S., d'Arminio Monforte, A., Casado, J., Cabello Úbeda, A., Hocqueloux, L., Allavena, C., Barber, T., et al. (2021). Hiv treatment with the two-drug regimen dolutegravir plus lamivudine in real-world clinical practice: a systematic literature review. *Infectious Diseases and Therapy*, 10:2051–2070.

Pillay, Y., Pienaar, S., Barron, P., and Zondi, T. (2021). Impact of covid-19 on routine primary healthcare services in south africa. *South African Medical Journal*, 111(8):714–719.

Quercia, R., Perno, C. F., Koteff, J., Moore, K., McCoig, C., St Clair, M., and Kuritzkes, D. (2018). Twenty-five years of lamivudine: Current and future use for the treatment of hiv-1 infection. *J Acquir Immune Defic Syndr*, 78(2):125–135.

Rahman, S., Kesselheim, A. S., and Hollis, A. (2023). Persistence of resistance: a panel data analysis of the effect of antibiotic usage on the prevalence of resistance. *The Journal of Antibiotics*, 76:270–278.

Robinson, D. (2020). *fuzzyjoin: Join Data Frames on Fuzzy Matching*. R package version 0.1.6.

Simbayi, L., Zuma, K., Zungu, N., Moyo, S., Marinda, E., Jooste, S., Mabaso, M., Ramlagan, S., North, A., Van Zyl, J., Mohlabane, N., Dietrich, C., Naidoo, I., and the SABSSM V Team (2017). South africa national hiv prevalence, incidence, behavior and communication survey 2017, cape town.

Simelela, N. and Venter, W. D. F. (2014). A brief history of south africa's response to aids: history of hiv in sa-progress towards the millennium development goals. *South African Medical Journal*, 104(3):249–251.

Tepanosyan, G., Sahakyan, L., Zhang, C., and Saghatelyan, A. (2019). The application of local moran's i to identify spatial clusters and hot spots of pb, mo and ti in urban soils of yerevan. *Applied Geochemistry*, 104:116–123.

Truter, I. (2008). A review of drug utilization studies and methodologies. *Jordan Journal of Pharmaceutical Sciences*, 1(2):91–104.

UNAIDS (2023). Hiv sub-national estimates viewer. https://aidsinfo.unaids.org/.

Van Schalkwyk, C., Dorrington, R. E., Seatlhodi, T., Velasquez, C., Feizzadeh, A., and Johnson, L. F. (2021). Modelling of hiv prevention and treatment progress in five south african metropolitan districts. *Scientific reports*, 11(1):1–10.

Vella, S., Schwartländer, B., Sow, S. P., Eholie, S. P., and Murphy, R. L. (2012). The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. *Aids*, 26(10):1231–1241.

Venter, W., Kaiser, B., Pillay, Y., Conradie, F., Gomez, G., Clayden, P., Matsolo, M., Amole, C., Rutter, L., Abdullah, F., et al. (2017). Cutting the cost of south african antiretroviral therapy using newer, safer drugs. *SAMJ: South African Medical Journal*, 107(1):28–30.

WHO (2003). Introduction to drug utilization research, oslo: World health organization; 2003.

WHO (2021). Consolidated guidelines on hiv prevention, testing, treatment, service delivery and monitoring. *Recommendations for a public health approach. Geneva: World Health Organization*, pages 1–592.

WHOCC (2024). Structure and principles. https://www.whocc.no/atc/structure_and_principles/.

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1.

Zhang, L., Li, Y., Zeng, L., and Wang, Z. (2012). Applying "children defined daily dose" to assess appropriate dose in pediatrics. *Journal of evidence-based medicine*, 5(1):2–5.

Zulfikar, R. (2018). Estimation model and selection method of panel data regression : An overview of common effect, fixed effect, and random effect model. *Center for Open Science*, pages 1–10.

Appendix

.1 Appendix A

District	DDD	Population Consuming	Allocation Ratio
Alfred Nzo District Municipality	69.47	34653.62	0.45
Amajuba District Municipality	44.87	17157.96	0.30
Amathole District Municipality	73.79	39089.19	0.64
Bojanala Platinum District Municipality	64.88	89375.73	0.60
Buffalo City Metropolitan Municipality	84.79	51615.72	0.92
Cape Winelands District Municipality	25.21	17376.98	0.59
Capricorn District Municipality	74.53	64257.15	0.80
Central Karoo District Municipality	1.92	100.49	0.06
Chris Hani District Municipality	66.82	32059.76	0.61
City Of Cape Town Metropolitan Municipality	125.83	432804.76	2.73
City Of Johannesburg Metropolitan Municipality	119.57	549138.56	1.65
City Of Tshwane Metropolitan Municipality	75.03	216241.64	1.41
Dr Kenneth Kaunda District Municipality	26.22	14599.42	0.23
Dr Ruth Segomotsi Mompati District Municipality	14.09	4213.48	0.15
Eden District Municipality	16.63	7562.14	0.33
Ehlanzeni District Municipality	85.96	108283.80	0.50
Ekurhuleni Metropolitan Municipality	79.03	239864.72	0.91
Ethekwini Metropolitan Municipality	350.36	1046214.64	2.55
Fezile Dabi District Municipality	15.46	5526.20	0.13
Frances Baard District Municipality	52.38	12574.96	0.66
Gert Sibande District Municipality	61.74	55863.58	0.34
Harry Gwala District Municipality	37.41	11925.66	0.21
Ilembe District Municipality	54.72	25519.99	0.31
Joe Gqabi District Municipality	31.73	7230.32	0.30
John Taolo Gaetsewe District Municipality	33.19	5047.96	0.43
King Cetshwayo District Municipality	73.83	46451.19	0.33
Lejweleputswa District Municipality	19.99	9240.22	0.14
Mangaung Metropolitan Municipality	26.43	16150.63	0.27
Mopani District Municipality	66.96	51879.94	0.60
Namakwa District Municipality	15.57	1110.97	0.49
Nelson Mandela Bay Municipality	131.32	123819.20	2.39
Ngaka Modiri Molema District Municipality	28.37	17089.79	0.31
Nkangala District Municipality	82.01	98552.72	0.82
Oliver Tambo District Municipality	132.06	125206.30	0.85
Overberg District Municipality	8.02	1760.10	0.14
Pixley Ka Seme District Municipality	26.88	3312.73	0.51
Sarah Baartman District Municipality	50.10	18018.45	0.61
Sedibeng District Municipality	19.01	13883.93	0.21
Sekhukhune District Municipality	65.09	49019.40	0.94
Thabo Mofutsanyana District Municipality	21.98	11164.34	0.13
Ugu District Municipality	61.23	31950.13	0.33

Continued on next page

Table 1 – continued from previous page

District	DDD	Population Consuming	Allocation Ratio
Umgungundlovu District Municipality	94.45	76031.96	0.50
Umkhanyakude District Municipality	49.30	20718.39	0.26
Umzinyathi District Municipality	40.20	13776.51	0.24
Uthukela District Municipality	52.65	23623.81	0.26
Vhembe District Municipality	80.90	75720.38	1.03
Waterberg District Municipality	44.61	23021.62	0.43
West Coast District Municipality	12.18	4053.20	0.27
West Rand District Municipality	18.98	13829.50	0.25
Xhariep District Municipality	3.61	301.27	0.03
Zululand District Municipality	63.46	34318.89	0.32
Zwelentlanga Fatman Mgcawu District Municipality	37.14	6321.10	0.65

2020 Drug Consumption for all 52 Districts Showing Number of People Consuming ARV drugs Relative to the Number of People on ART Based on Naomi Model Estimates

District	DDD	Population Consuming	Allocation Ratio
Alfred Nzo District Municipality	40.64	20248.07	0.26
Amajuba District Municipality	33.75	12730.18	0.21
Amathole District Municipality	42.26	21895.92	0.35
Bojanala Platinum District Municipality	57.05	79178.30	0.50
Buffalo City Metropolitan Municipality	49.32	29815.22	0.50
Cape Winelands District Municipality	18.59	12776.39	0.39
Capricorn District Municipality	49.54	42919.74	0.47
Central Karoo District Municipality	1.39	71.52	0.04
Chris Hani District Municipality	38.43	18107.52	0.33
City Of Cape Town Metropolitan Municipality	92.83	318613.11	1.91
City Of Johannesburg Metropolitan Municipality	108.51	497051.75	1.34
City Of Tshwane Metropolitan Municipality	67.89	194586.36	1.16
Dr Kenneth Kaunda District Municipality	22.85	12703.89	0.20
Dr Ruth Segomotsi Mompati District Municipality	12.10	3561.14	0.12
Eden District Municipality	12.12	5430.25	0.23
Ehlanzeni District Municipality	112.49	138041.26	0.65
Ekurhuleni Metropolitan Municipality	70.75	211315.94	0.79
Ethekwini Metropolitan Municipality	267.14	797529.56	1.88
Fezile Dabi District Municipality	14.30	5017.37	0.12
Frances Baard District Municipality	73.91	17520.98	0.76
Gert Sibande District Municipality	82.10	73540.67	0.47
Harry Gwala District Municipality	28.01	8768.11	0.15
Ilembe District Municipality	41.11	18882.17	0.22
Joe Gqabi District Municipality	18.44	4169.22	0.16
John Taolo Gaetsewe District Municipality	47.56	7255.79	0.56
King Cetshwayo District Municipality	55.04	33849.29	0.24
Lejweleputswa District Municipality	18.42	8329.63	0.12
Mangaung Metropolitan Municipality	24.58	14824.41	0.23
Mopani District Municipality	44.69	34918.93	0.36
Namakwa District Municipality	22.18	1578.27	0.60
Nelson Mandela Bay Municipality	76.42	71584.19	1.31
Ngaka Modiri Molema District Municipality	24.57	14684.14	0.25
Nkangala District Municipality	109.98	131954.37	1.13
Oliver Tambo District Municipality	78.47	75482.98	0.49
Overberg District Municipality	5.90	1285.65	0.09
Pixley Ka Seme District Municipality	37.75	4569.60	0.60
Sarah Baartman District Municipality	29.41	10604.13	0.37
Sedibeng District Municipality	17.15	12420.50	0.21
Sekhukhune District Municipality	43.98	33815.28	0.56
Thabo Mofutsanyana District Municipality	20.36	10174.08	0.12
Ugu District Municipality	46.91	24590.58	0.24
Umgungundlovu District Municipality	71.59	57270.56	0.35
Umkhanyakude District Municipality	37.48	15699.16	0.17
Umzinyathi District Municipality	30.61	10471.45	0.17
Uthukela District Municipality	39.46	17404.11	0.19

Continued on next page

Table 2 – continued from previous page

District	DDD	Population Consuming	Allocation Ratio
Vhembe District Municipality	53.25	49583.95	0.65
Waterberg District Municipality	29.65	15372.98	0.26
West Coast District Municipality	9.00	2997.01	0.20
West Rand District Municipality	16.75	11844.97	0.21
Xhariep District Municipality	3.33	271.41	0.03
Zululand District Municipality	47.82	25554.29	0.23
Zwelentlanga Fatman Mgcawu District Municipality	52.84	8954.90	0.85

2021 Drug Consumption for all 52 Districts Showing Number of People Consuming ARV drugs Relative to the Number of People on ART Based on Naomi Model Estimates

District	DDD	Population Consuming	Allocation Ratio
Alfred Nzo District Municipality	93.38	47648.86	0.57
Amajuba District Municipality	70.10	26751.89	0.43
Amathole District Municipality	95.24	49574.95	0.75
Bojanala Platinum District Municipality	63.86	91043.85	0.56
Buffalo City Metropolitan Municipality	111.53	67983.25	1.08
Cape Winelands District Municipality	30.73	21226.00	0.71
Capricorn District Municipality	94.26	81413.05	0.88
Central Karoo District Municipality	2.27	116.07	0.06
Chris Hani District Municipality	86.34	40742.49	0.70
City Of Cape Town Metropolitan Municipality	153.98	532879.16	2.88
City Of Johannesburg Metropolitan Municipality	179.66	841124.82	2.19
City Of Tshwane Metropolitan Municipality	112.16	327838.37	1.81
Dr Kenneth Kaunda District Municipality	25.55	14577.67	0.21
Dr Ruth Segomotsi Mompati District Municipality	13.50	4068.74	0.13
Eden District Municipality	19.87	8874.40	0.35
Ehlanzeni District Municipality	192.97	236308.79	1.09
Ekurhuleni Metropolitan Municipality	116.31	352536.25	1.19
Ethekwini Metropolitan Municipality	553.82	1669626.90	3.93
Fezile Dabi District Municipality	13.10	4726.90	0.11
Frances Baard District Municipality	89.00	21015.94	0.90
Gert Sibande District Municipality	142.70	129217.66	0.79
Harry Gwala District Municipality	57.83	18205.23	0.30
Ilembe District Municipality	85.30	39611.46	0.46
Joe Gqabi District Municipality	41.83	9564.08	0.35
John Taolo Gaetsewe District Municipality	57.80	8863.02	0.59
King Cetshwayo District Municipality	113.66	70326.75	0.49
Lejweleputswa District Municipality	16.89	7860.00	0.12
Mangaung Metropolitan Municipality	21.32	12520.06	0.19
Mopani District Municipality	87.19	69662.70	0.68
Namakwa District Municipality	26.74	1896.28	0.80
Nelson Mandela Bay Municipality	173.25	164038.02	2.95
Ngaka Modiri Molema District Municipality	27.26	16589.37	0.27
Nkangala District Municipality	191.43	232548.05	2.01
Oliver Tambo District Municipality	179.79	176641.96	1.10
Overberg District Municipality	9.77	2144.40	0.16
Pixley Ka Seme District Municipality	45.33	5452.02	0.70
Sarah Baartman District Municipality	66.97	24508.93	0.77
Sedibeng District Municipality	27.52	19740.56	0.28
Sekhukhune District Municipality	85.68	67279.04	1.02
Thabo Mofutsanyana District Municipality	18.62	9557.69	0.11
Ugu District Municipality	97.97	52243.42	0.51
Umgungundlovu District Municipality	148.59	120187.03	0.73
Umkhanyakude District Municipality	77.90	33037.10	0.33
Umzinyathi District Municipality	63.79	22151.71	0.37
Uthukela District Municipality	81.56	36214.62	0.37

Continued on next page

Table 3 – continued from previous page

District	DDD	Population Consuming	Allocation Ratio
Vhembe District Municipality	103.70	98553.67	1.24
Waterberg District Municipality	57.63	30430.36	0.49
West Coast District Municipality	14.93	5011.48	0.31
West Rand District Municipality	27.37	19521.05	0.30
Xhariep District Municipality	3.17	277.10	0.03
Zululand District Municipality	98.81	53148.44	0.45
Zwelentlanga Fatman Mgcawu District Municipality	64.01	10869.98	0.94

2022 Drug Consumption for all 52 Districts Showing Number of People Consuming ARV drugs Relative to the Number of People on ART Based on Naomi Model Estimates

.2 Appendix B

District	Lower	Mean	Upper	Predicted	Calculated
Alfred Nzo DM	0,1886	0,2173	0,2483	0,2147	0,1961
Amajuba DM	0,1931	0,2323	0,2723	0,2353	0,2214
Amathole DM	0,1489	0,1799	0,2128	0,1719	0,1604
Bojanala Platinum DM	0,1676	0,1851	0,2033	0,1919	0,1859
Buffalo City MM	0,1728	0,1942	0,2166	0,1824	0,1698
Cape Winelands DM	0,0644	0,0797	0,0988	0,0833	0,0758
Capricorn DM	0,1503	0,1732	0,1979	0,1587	0,1555
Central Karoo DM	0,0444	0,0624	0,0842	0,0591	0,0474
Chris Hani DM	0,1575	0,1863	0,2216	0,1749	0,1629
City Of Cape Town MM	0,0913	0,1013	0,1121	0,0706	0,0886
City Of Johannesburg MM	0,1341	0,1519	0,1714	0,1161	0,1381
City Of Tshwane MM	0,1025	0,1227	0,1455	0,0984	0,1076
Dr K Kaunda DM	0,1673	0,1968	0,2290	0,1963	0,1887
Dr R S Mompati DM	0,1337	0,1677	0,2032	0,1592	0,1533
Eden DM	0,0882	0,1031	0,1216	0,1019	0,0939
Ehlanzeni DM	0,2292	0,2507	0,2717	0,2394	0,2402
Ekurhuleni MM	0,1597	0,1811	0,2033	0,1587	0,1693
Ethekwini MM	0,1930	0,2080	0,2234	0,1878	0,2041
Fezile Dabi DM	0,1626	0,1958	0,2331	0,2000	0,1979
Frances Baard DM	0,1357	0,1613	0,1925	0,1642	0,1588
Gert Sibande DM	0,2363	0,2560	0,2779	0,2586	0,2494
Harry Gwala DM	0,2021	0,2357	0,2718	0,2380	0,2256

Continued on next page

Table 4 – continued from previous page

District	Lower	Mean	Upper	Predicted	Calculated
Ilembe DM	0,2254	0,2528	0,2817	0,2576	0,2451
Joe Gqabi DM	0,1441	0,1780	0,2144	0,1777	0,1593
J T Gaetsewe DM	0,1375	0,1728	0,2131	0,1647	0,1558
King Cetshwayo DM	0,2706	0,2932	0,3179	0,2941	0,2861
Lejweleputswa DM	0,1727	0,2016	0,2300	0,2164	0,2141
Mangaung MM	0,1821	0,2027	0,2254	0,2033	0,2116
Mopani DM	0,1792	0,2079	0,2383	0,1867	0,1779
Namakwa DM	0,0425	0,0578	0,0759	0,0616	0,0538
Nelson Mandela Bay MM	0,1098	0,1254	0,1423	0,1155	0,1074
Ngaka Modiri Molema DM	0,1396	0,1609	0,1863	0,1589	0,1542
Nkangala DM	0,1345	0,1568	0,1815	0,1537	0,1484
Oliver Tambo DM	0,2086	0,2304	0,2552	0,2190	0,2057
Overberg DM	0,0860	0,1088	0,1366	0,1135	0,1015
Pixley Ka Seme DM	0,0828	0,1007	0,1236	0,0977	0,0916
Sarah Baartman DM	0,1165	0,1428	0,1690	0,1395	0,1219
Sedibeng DM	0,1390	0,1650	0,1931	0,1523	0,1493
Sekhukhune DM	0,1215	0,1418	0,1643	0,1279	0,1200
T Mofutsanyana DM	0,2025	0,2319	0,2643	0,2326	0,2298
Ugu DM	0,2196	0,2518	0,2848	0,2641	0,2478
Umgungundlovu DM	0,2411	0,2751	0,3109	0,2783	0,2684
Umkhanyakude DM	0,2687	0,3042	0,3445	0,2724	0,2571
Umzinyathi DM	0,1944	0,2193	0,2473	0,2262	0,2097

Continued on next page

Table 4 – continued from previous page

District	Lower	Mean	Upper	Predicted	Calculated
Uthukela DM	0,2498	0,2792	0,3118	0,2737	0,2624
Vhembe DM	0,1385	0,1583	0,1788	0,1459	0,1402
Waterberg DM	0,1667	0,1962	0,2321	0,1847	0,1755
West Coast DM	0,0755	0,0970	0,1220	0,1008	0,0894
West Rand DM	0,1376	0,1661	0,1995	0,1581	0,1539
Xhariep DM	0,1253	0,1580	0,1968	0,1581	0,1510
Zululand DM	0,2466	0,2806	0,3167	0,2767	0,2658
Z F Mgcawu DM	0,1017	0,1247	0,1518	0,1271	0,1195

2022 Fixed Effect Panel Model HIV Prevalence Results Showing the Predicted and Calculated Prevalence for the 52 South African Districts

.3 Appendix C

District	Lower	Mean	Upper	Predicted	Calculated
Alfred Nzo DM	0,1886	0,2173	0,2483	0,2197	0,2013
Amajuba DM	0,1931	0,2323	0,2723	0,2392	0,2259
Amathole DM	0,1489	0,1799	0,2128	0,1768	0,1653
Bojanala Platinum DM	0,1676	0,1851	0,2033	0,1924	0,1869
Buffalo City MM	0,1728	0,1942	0,2166	0,1861	0,1748
Cape Winelands DM	0,0644	0,0797	0,0988	0,0873	0,0793
Capricorn DM	0,1503	0,1732	0,1979	0,1607	0,1575
Central Karoo DM	0,0444	0,0624	0,0842	0,0610	0,0545
Chris Hani DM	0,1575	0,1863	0,2216	0,1773	0,1679
City Of Cape Town MM	0,0913	0,1013	0,1121	0,0872	0,0919
City Of Johannesburg MM	0,1341	0,1519	0,1714	0,1326	0,1407
City Of Tshwane MM	0,1025	0,1227	0,1455	0,1139	0,1102
Dr K Kaunda DM	0,1673	0,1968	0,2290	0,1963	0,1899
Dr R S Mompati DM	0,1337	0,1677	0,2032	0,1601	0,1548
Eden DM	0,0882	0,1031	0,1216	0,1033	0,0974
Ehlanzeni DM	0,2292	0,2507	0,2717	0,2461	0,2449
Ekurhuleni MM	0,1597	0,1811	0,2033	0,1658	0,1717
Ethekwini MM	0,1930	0,2080	0,2234	0,2092	0,2081
Fezile Dabi DM	0,1626	0,1958	0,2331	0,1997	0,1977
Frances Baard DM	0,1357	0,1613	0,1925	0,1647	0,1615
Gert Sibande DM	0,2363	0,2560	0,2779	0,2621	0,2542
Harry Gwala DM	0,2021	0,2357	0,2718	0,2415	0,2300

Continued on next page

Table 5 – continued from previous page

District	Lower	Mean	Upper	Predicted	Calculated
Ilembe DM	0,2254	0,2528	0,2817	0,2610	0,2494
Joe Gqabi DM	0,1441	0,1780	0,2144	0,1801	0,1649
J T Gaetsewe DM	0,1375	0,1728	0,2131	0,1645	0,1591
King Cetshwayo DM	0,2706	0,2932	0,3179	0,2952	0,2902
Lejweleputswa DM	0,1727	0,2016	0,2300	0,2149	0,2138
Mangaung MM	0,1821	0,2027	0,2254	0,2022	0,2111
Mopani DM	0,1792	0,2079	0,2383	0,1889	0,1800
Namakwa DM	0,0425	0,0578	0,0759	0,0629	0,0586
Nelson Mandela Bay MM	0,1098	0,1254	0,1423	0,1233	0,1124
Ngaka Modiri Molema DM	0,1396	0,1609	0,1863	0,1610	0,1554
Nkangala DM	0,1345	0,1568	0,1815	0,1635	0,1533
Oliver Tambo DM	0,2086	0,2304	0,2552	0,2254	0,2107
Overberg DM	0,0860	0,1088	0,1366	0,1143	0,1056
Pixley Ka Seme DM	0,0828	0,1007	0,1236	0,0986	0,0952
Sarah Baartman DM	0,1165	0,1428	0,1690	0,1413	0,1272
Sedibeng DM	0,1390	0,1650	0,1931	0,1598	0,1519
Sekhukhune DM	0,1215	0,1418	0,1643	0,1337	0,1222
T Mofutsanyana DM	0,2025	0,2319	0,2643	0,2321	0,2294
Ugu DM	0,2196	0,2518	0,2848	0,2658	0,2521
Umgungundlovu DM	0,2411	0,2751	0,3109	0,2799	0,2724
Umkhanyakude DM	0,2687	0,3042	0,3445	0,2757	0,2615
Umzinyathi DM	0,1944	0,2193	0,2473	0,2321	0,2142

Continued on next page

Table 5 – continued from previous page

District	Lower	Mean	Upper	Predicted	Calculated
Uthukela DM	0,2498	0,2792	0,3118	0,2741	0,2667
Vhembe DM	0,1385	0,1583	0,1788	0,1498	0,1423
Waterberg DM	0,1667	0,1962	0,2321	0,1852	0,1777
West Coast DM	0,0755	0,0970	0,1220	0,1018	0,0932
West Rand DM	0,1376	0,1661	0,1995	0,1632	0,1565
Xhariep DM	0,1253	0,1580	0,1968	0,1592	0,1527
Zululand DM	0,2466	0,2806	0,3167	0,2792	0,2699
Z F Mgcawu DM	0,1017	0,1247	0,1518	0,1273	0,1226

2022 Spatial Linear Model HIV Prevalence Results Showing the Predicted and Calculated Prevalence for the 52 South African Districts

.4 Appendix D

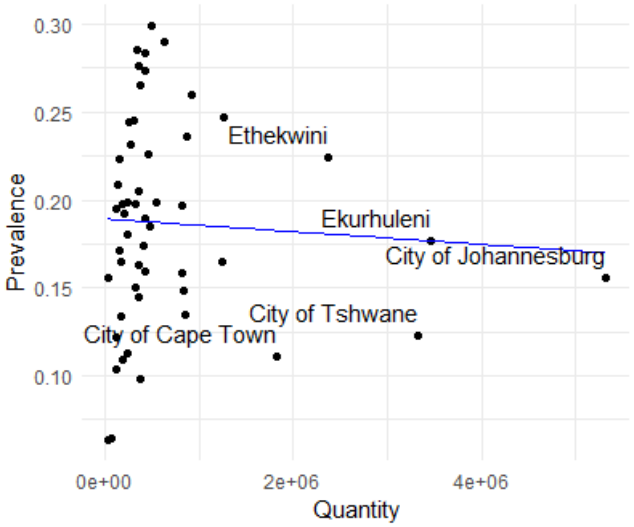


Figure 1: Scatterplot with a Fitted Regression Lines Showing the Relationship Between Quantity and Prevalence for 2021

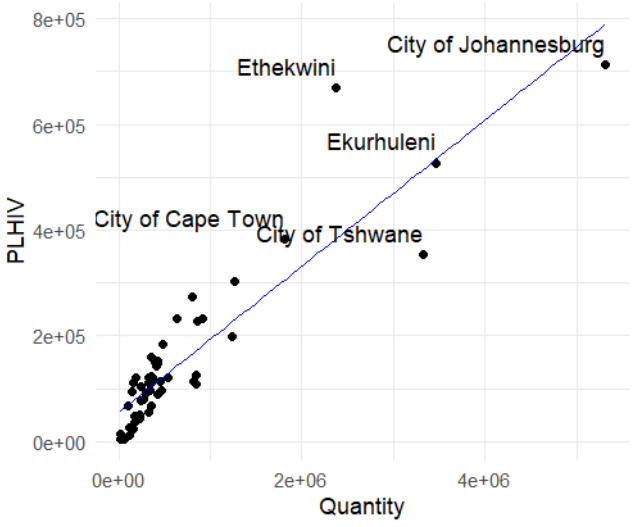


Figure 2: Scatterplot with a Fitted Regression Line Showing the Relationship Between Quantity and PLHIV for 2021

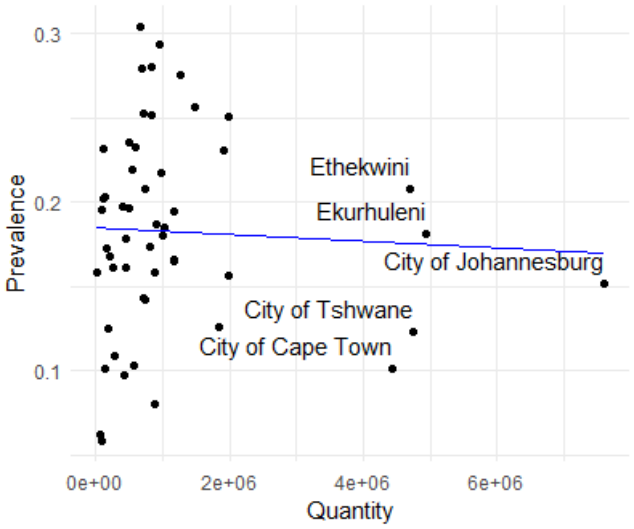


Figure 3: Scatterplot with a Fitted Regression Lines Showing the Relationship Between Quantity and Prevalence for 2022

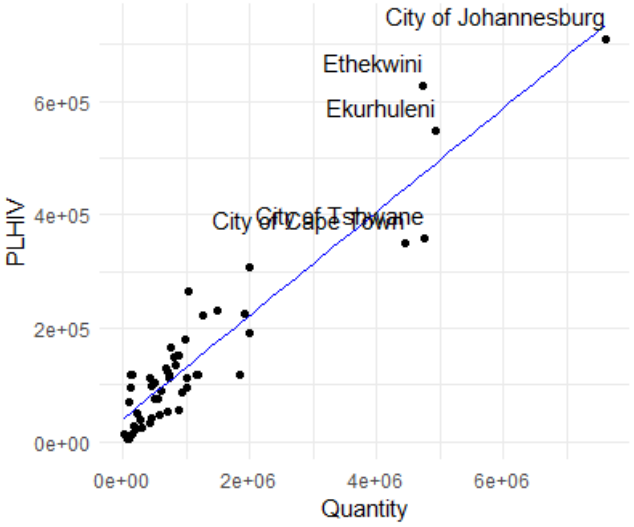


Figure 4: Scatterplot with a Fitted Regression Line Showing the Relationship Between Quantity and PLHIV for 2022