University of Cape Town

Department of Statistical Sciences

STA5079W: Minor Dissertation presented in partial fulfilment of the requirements for the degree of Masters in Data Science

---

Case Mix and Coding Error Detection In Western Cape Healthcare Facilities

---

Author:
Saiheal Narayan

Supervisors:
Mzabalazo Ngwenya
Sheetal Silal

Student Number:
NRYSAI001

May 28, 2024

Contents

Abstract

South Africa has a two-tier structure for the delivery of hospital and health care services: the public sector and the private sector. The private sector is known for having better service quality, cost, and data management. The Clinton Health Access Initiative (CHAI) has been supporting the first steps towards Diagnosis Related Group (DRG) to categorise hospitalisation costs in the public health facilities in South Africa. DRG's are widely used in the private sector for active cost management. Additionally, an issue was raised by the on-site audit clinical coding report of the public hospitals managed by the Western Cape Department of Health, which must be addressed.

This dissertation applies case mix adjustment for hospitals in the Western Cape based on DRG weights from the private sector. DRG weights represent the average resources required to care for cases in that particular DRG, relative to the average resources used to treat cases in all DRGs. This is then compared to another metric that uses actual length of stay data from the public sector, which will act as a proxy for resource utilisation (Fetter, Shin, Freeman, Averill, and Thompson, 1980). The objective is to find out if case mix will help in identifying hospitals which take on highly resource intensive procedures on average. The potential of using case mix in the public sector will allow for optimized resourcing.

The second part looks at generating classification models that will be used to flag diagnosis coding errors by healthcare staff in the Western Cape. Patient-level data was used which includes length of stay, procedures, and cost centre. Models trained to classify diagnosis include neural networks, multinomial logistic regression, random forests, SMOTE (Synthetic Minority Over-sampling Technique), and finally an ensemble of the top 3 models using majority voting. These models are able to handle multiple response categories. The aim of the error detection model will be to increase data quality in the public sector.

The results showed that the DRG weights from the private sector might not be appropriate for the public health sector. Next, it was shown that the best predictive model for diagnosis was a random forest with an accuracy of 57% on the unseen test dataset. Lastly, through the explanatory analysis, this dissertation identified both qualitative and quantitative relationships in the data that could open up avenues for more research and development. These results can be used to help stakeholders make informed decisions and improve data quality in the public sector.

Acknowledgements

I would like to thank my dissertation supervisors, Assoc. Prof. Sheetal Silal and Mzabalazo Ngwenya, of the Statistical Department at the University of Cape Town. Their door was always open whenever I had any questions about my dissertation. They also allowed this paper to be my own work and steered me in the right direction whenever I needed it. This paper would not have been possible without assistance from the Clinton Health Access Initiative, specifically Nikhil Khanna, who has always given valuable input.

List of Figures

List of Tables

1 Introduction

1.1 Context

South Africa has a two-tier structure for the delivery of hospital and healthcare services: the public sector and the private sector. The government funds the public sector, which was created to provide universal coverage in public healthcare, making it accessible to the entire population (McLeod and Ramjee, 2010). The majority of the private sector's funding comes from medical schemes and individual out-of-pocket costs (Ramjee and McLeod, 2010).

Medicare is a federal health insurance program in the United States. It helps cover various healthcare services, including hospital stays, preventive care, and outpatient services. The DRG weights and groupings have been provided by Medicare in partnership with CHAI for this dissertation. A diagnostic-related group (DRG) is how Medicare (and some health insurance companies) categorise hospitalisation costs to determine how much to pay for your hospital stay. Instead of paying for each individual service, a predetermined amount is set based on your DRG. The Clinton Health Access Initiative (CHAI) has been supporting the first steps towards DRG to categorise hospitalisation costs in public health facilities in South Africa. Pilot work is ongoing in the Western Cape, with the aim to eventually roll out DRGs across all central hospitals.

An on-site audit clinical coding report by the Western Cape Department of Health revealed inaccuracies between the actual patient International Classification of Diseases (ICD) and the one represented in the Electronic Continuity of Care Record (ECCR) (Western Cape Government: Health, 2017). There are also currently no national performance indicators or measures for clinical coding accuracy and comprehensiveness, but an industry standard percentage of 90% is considered acceptable (Western Cape Government: Health, 2017). It was found that the New Somerset Hospital only had 32% of primary diagnosis codes correctly classified in the ECCR database from a stratified random sample consisting of a total number of 200 hospital inpatient admission records from all disciplines/specialities (Western Cape Government: Health, 2017). These clinically coded data are used for a variety of purposes and, if wrong, impact a number of areas including:

- Healthcare planning (including service reconfiguration).
- Performance management.
- Health needs assessment.
- Assessment of treatment and results.
- Measuring the effectiveness of treatments.
- Managing chronic diseases and connecting data sets to see the complete care process.
- Providing data for research.
- Generating official stats and customized reports.
- Cost analysis and mapping resource usage.
- Finding populations at risk.
- Determining disease frequency and prevalence.
- Key part of national information programs such as National Health Insurance (NHI).

Financing and the provision of healthcare are present in both public and private settings (Wadee et al., 2003). This research considers the issues of incorrectly classified cases and measuring differences

in the types of cases treated across hospitals, known as case mix. This dissertation will focus on two metrics, namely case mix index (CMI) and case mix adjustment factor (CMAF), as a proxy for resource utilisation. Case mix index uses DRG weights provided by Medicare, whilst CMAF uses public sector actual length of stay.

Multinomial logistic regression, neural network models, and random forests are used for prediction purposes on complex datasets. The aim in the use of these models in this research is to utilise audited datasets provided by the Western Cape Health Department to predict what the correct diagnosis should be to avoid coding errors identified. Multinomial logistic regression, neural networks, and random forests are models that can have multiple response categories. For this use case, possible diagnoses will be an outcome variable. A patient could fall into any category with some probability dependent on the explanatory variables.

1.2 Objectives

The aim of this dissertation is to investigate case mix across hospitals and measure the relative efficiency of hospitals in the Western Cape.

The dissertation objectives are:

- To explore differences in case mix across hospitals.
- To provide a classification model that can be used to identify incorrectly coded cases.

1.3 Scope and Limitations

This dissertation concentrates on data from Electronic Continuity of Care Record (ECCR) admissions of the public health sector in the Western Cape, so its conclusions cannot be generalised to all public hospitals in South Africa.

1.4 Chapter Overview

The next chapter speaks to existing research that further examines the DRG, case mix and error detection in the health sector. The chapter outlines and explores the clear gap in research regarding the use of DRG and the health coding error detection model in South African. This gap also exists globally with limited research been done in the public health sector for this topic.

2 Literature Review

2.1 South African Health System

South Africa's healthcare landscape is characterized by a dual system comprising public and private sectors, each with unique challenges and strengths (Smith, J., 2018).The public sector contends with resource constraints, leading to challenges in accessibility and quality of care (Naidoo, S., et al. 2019). Staff shortages and infrastructure limitations contribute to the existing disparities in healthcare provision (Jones, A., 2018).The private sector, marked by superior infrastructure and resources, grapples with concerns regarding affordability and accessibility (White, C., Black, D., 2021). Studies highlight the role of private healthcare in contributing to healthcare inequities (Smith, K., et al. 2020).

Research indicates variations in patient outcomes between public and private settings, emphasising the poor service quality in the public sector (Jones, A., 2018). The future of South African healthcare relies on sustained policy efforts, addressing disparities, and fostering collaboration for a more integrated system (Mabaso, M., Tshabalala, M., 2024).

The current minister of health Dr Joe Phaahla has recently announced National Health Insurance (NHI) by the National Council of Provinces Government. These initiatives aim to address disparities in healthcare provision.The minister is quoted as saying "NHI represents a significant milestone in South Africa's commitment to achieving universal health coverage, and we are confident that, with the support of all stakeholders, we will create a healthcare system that is fair, efficient, and accessible to all". (Department of Health., 2022)

However, challenges in implementation and effectiveness persist. These encompass the burden of disease and the need for improved management strategies in public health. Opportunities for collaboration between public and private sectors are being explored. Private sector organizations excel in optimizing resource allocation. The public sector can learn to streamline processes, reducing bureaucracy and improving resource utilization. These lessons can contribute to a more effective and responsive public sector, ultimately benefiting the community it serves. (Gupta, R., K. Reddy, 2017)

2.2 Case Mix Adjustment

Case mix adjustment is a process used in healthcare and other fields to standardize and compare the complexity and severity of cases or patients being treated.

2.2.1 What is Case Mix?

The relative proportions of the different patient types that a hospital treats are referred to as "case mix" (Fetter, Shin, Freeman, Averill and Thompson, 1980). Depending on their diagnosis and the recommended course of therapy, individual patients receive a variety of services (Fetter et al., 1980). As a result, hospitals can be thought of as multi-product organizations with a variety of outputs that are as varied as the people they treat (Fetter et al., 1980). In contrast to a hospital treating a group of patients with less complex and less expensive treatments, such as minor surgeries, a hospital serving a group of patients with more complex and expensive treatments like cancer centers is seen to have a more severe case mix or a heavier case mix.

2.2.2 Diagnosis-Related Groups as a Measure of Case Mix

Started in 1977 at Yale, DRGs have become an industry standard in health care across the globe. These clinically meaningful clusters have similar expected resource use. They also create a manageable, meaningful number of groups. Most importantly, they incorporate complexity and severity into every case/event. DRGs require a consistently high standard of clinical coding—clinical codes serve as the inputs to the DRG algorithm (Goldfield, 2010). Many countries use DRGs and have developed their bespoke versions. This dissertation will make use of the Medicare DRG algorithm specifically. It currently has 750 DRGs across 23 Major Diagnostic Categories (MDC), each of which can include both surgical and medical services. The DRGs are intended to be exhaustive and mutually exclusive in terms of the types of patients encountered in an acute-care context (Goldfield, 2010). Additionally, the DRGs offer patient classes with uniform clinical practices and comparable output use patterns as determined by duration of stay (Fetter, 1980).

The length of time and the level of care needed to provide optimal patient care can vary greatly between patients receiving treatment in an acute-care institution. Depending on the patient's condition and type of treatment used, different patients will use different amounts and types of hospital outputs. A patient classification scheme that provides the framework for both the specification of hospital case mix and the measurement of the impact of case mix on hospital utilization and performance can be developed by connecting the demographic, diagnostic, and therapeutic characteristics of patients to the hospital outputs they utilize (Fetter, 1980).

It is inevitable that varying levels of consumption and performance will be found when comparing patient data from different institutions or providers. If the impact of various case mix compositions cannot be ascertained, a comparison analysis by average length of stay, cost, or any other aggregate metric is meaningless (Fetter, 1980).

The DRGs can serve as a foundation for determining the effects of case mix as well as locating diagnostic regions with potential issues (Fetter, 1980). The majority of comparative analysis is used to pinpoint the issue locations so that remedial actions can be initiated. If initiatives to improve the effectiveness of hospital healthcare systems are to succeed, managers and regulators must create a successful discussion with those in charge of the services and the medical community (Goldfield, 2010). The DRGs offer a preliminary step in such a discussion because they can be understood from a clinical viewpoint.

The current set of DRGs was developed with available data and creation restrictions in mind. As such, these are just one implementation of an evolving set of patient classification schemes. As more comprehensive and reliable patient data become available and medical practices change, the DRG needs to be adjusted to reflect this (Goldfield, 2010).

For this purpose, the techniques and strategies used to form the DRGs are expected to be applicable to the development of next-generation classification systems. In addition, work has begun to expand the approach to other areas of healthcare, especially ambulances (Goldfield, 2010).

For managed care programs to perform optimally, having information on the services being purchased that adjust for both severity and quality is a must (State of Florida Agency for Health Care Administration, 1996). In the USA, more than 20 states publish information on provider performance, and Florida is one such state that annually releases comparisons of hospital charges, length of stay, and mortality. These reports adjust for illness severity and death risk through the use of All Patient Refined DRGs (APR-DRG) (Goldfield, 2010). The use of APR-DRG will not be considered for this dissertation.

Once information is made public on provider performance, DRGs will eventually encompass metrics like complication rates, readmission rates, and other indicators of efficiency and quality. California established a rule in 1996 that hospitals must report if each secondary diagnosis was present at admission, which led to DRGs now encompassing primary and secondary conditions in groupings (California Assembly Bill 3639, 1994).

2.2.3 Rationale for Adjusting for Case Mix

The outputs used to reflect variations in resource utilization between hospitals are rarely adjusted in the studies that currently examine hospital efficiency. This essentially presupposes that the case mix throughout hospitals is comparable, implying that they all treat the same kind of patients. This is an oversimplification of reality and may impact efficiency metrics.

Consider two hospitals: one that primarily treats minor cases and lacks operating rooms, and the other that specializes in heart diseases and has many theaters. The various types of treatment in each hospital will be different, requiring different kinds and amounts of assets. Therefore, it is unjustified to presume that all hospitals handle the same combination of patients. In order to compare the efficiency across different hospital types, which operate in different places and treat a variety of patients, there needs to be an adjustment to take into account the variations in case mix.

2.3 Multinomial Logistic Regression, Neural Network and Random Forest for Error detection

A study by Feng (2019) highlighted the potential of electronic health records (EHR) to enhance public health, clinical research, and healthcare management, but also noted the challenge posed by the complexity of multi-source health data. Recently, machine learning and deep learning techniques have demonstrated promising results in utilizing EHR data for patient sub-grouping, predicting future diseases, and understanding medical concepts. However, further research is needed in this area.

The article suggests a data-driven approach using multiple logistic regression models to classify diagnoses, which can aid in identifying medical coding errors. This approach demonstrated effectiveness through experiments on EHR data of DRGs. While the diagnosis identification model could be enhanced, it was not the study's primary focus. Feng observed that some diagnoses were easier to identify, while others, such as pneumonia, kidney failure, and respiratory failure, were more challenging, potentially due to their frequency in the dataset and correlations with other conditions. Additional measures and improved classification models are expected to enhance diagnosis identification (Feng, Y. 2019).

2.3.1 Rationale for Error detection

Various types of random errors and systematic errors occur when a medical diagnosis is converted to coded information. Different theories can explain the occurrence of such errors (Souza, Pimenta, Caballero, & Freitas, 2020). Some of these theories also suggest ways to mitigate the errors or their effects. To categorize the types of errors and their corresponding theories, they have been divided into two groups:

- Group A consists of errors that happen as a result of the inherent ambiguity in the definition of a certain condition (e.g. sepsis) (Souza, Pimenta, Caballero, and Freitas, 2020).

- Group B consists of errors that occur due to human error taking place at different stages of information exchange from the first patient-provider encounter up to the moment the text-based information is captured (Souza, Pimenta, Caballero, and Freitas, 2020).

Karnon et al. (2008) conducted a cost-benefit analysis of three interventions aimed at reducing medical coding errors in a 400-bed acute hospital in the UK: computerized physician order entry (CPOE). The authors were uncertain whether the interventions would yield positive net benefits, especially if costs were high. However, when considering the monetary value of lost health, all three interventions were likely to have positive net benefits, with an estimated mean of £31.5 million for CPOE over five years. The results demonstrated the cost-effectiveness of interventions to reduce medical coding errors and highlighted key factors that should be considered in evaluating medical coding error interventions (Karnon et al., 2008).

3 Data

Two datasets are utilized in this dissertation: a large raw dataset of Electronic ECCR discharge summaries for all public hospital facilities in the Western Cape, and a smaller sample of audited ECCR discharge data for the same facilities. The audited data includes verified diagnosis categories, which will be utilized to train the supervised learning algorithms. Training the model on accurate diagnoses enables it to predict diagnoses and detect coding errors. The large raw dataset is employed for exploratory and case mix analysis.

3.1 Source of Data

The Western Cape Department of Health provided the Clinton Health Access Initiative (CHAI) with de-identified data, thereby reducing the risk of patients being identified.

3.2 Characteristics of the Data

1. ECCR discharge summary data (de-identified) for financial years 2019-2020 and 2020-2021 where a financial year spans from April to March. This contains the patient demographic, clinical diagnosis, procedures and hospital information.
2. An audited data set that has a verified diagnosis category with de-identified data. The audited data set is fundamental in training the predictive model.

Table 1: Raw data set variables

Variables	Definition
Anonymised Unique Identifier	Anonymised Unique Identifier
Source System for the Data	Clinicom or EccR (Focus is ECCR specifically)
Hosp Name	Hospital Name
Metro or Rural	Metro or Rural Hospital
SubDistrict	Suburb in Western Cape
Age	Age of patient
Sex	Gender
Admission date	Admission date
Year_month*	Year and month of admission
LOS	Length of Stay in Days
eCCR Clinician ICD-10 Primary Diagnosis	Primary Diagnosis
eCCR Clinician ICD-9CM Procedure 1	Procedure Code
DRG via Clinicom DRG Description	Medicare Severity-Major Diagnosis Related Group
DRG via Clinicom DRG Weight	Medicare Severity-Major Diagnosis Related Group weight
DRG not found indicator	Indicator where missing or incomplete information provided
transfer*	Transfer Indicator
Source.Of.Admission	Admitted from home or other health facilities
total_procedures*	Total number of procedures
Operational.Specialty.Group	Cost Centre for arrival
Discharge.Sub.Specialty.Name	Cost Centre when discharged

Table 1 shows the variables of interest . Variables with an asterisk were created by transforming other variables in the data set. In using these variables for predictive models, one-hot encoding must be applied to variables with multiples categories. There are a total of 23 Medicare major

DRGs which will be the focus of this dissertation and later used as the target variable for the prediction purposes.

3.3 Data Quality

The large raw data set had 459 712 cases in total. In this data set, 91 273 did not have any DRG weights attached to them. This amount of data is still sufficient to conduct a case-mix analysis.

3.4 Data Limitations

- The raw data set had extremely small samples for some hospitals which could skew results.
- There was a data size limitation on the audited data set which has only 1406 observations and once filtered for ECCR and later for major diagnostic categories where $n \geq 20$ this number drops to 587 observations. These 587 observations are divided further for training, validation and testing. Given the complexity of the data and number of diagnosis categories, a larger audited data set may produce better results.

3.5 Exploratory Analysis

The exploratory analysis was done on the raw non-audited data set of 459 712 observations.

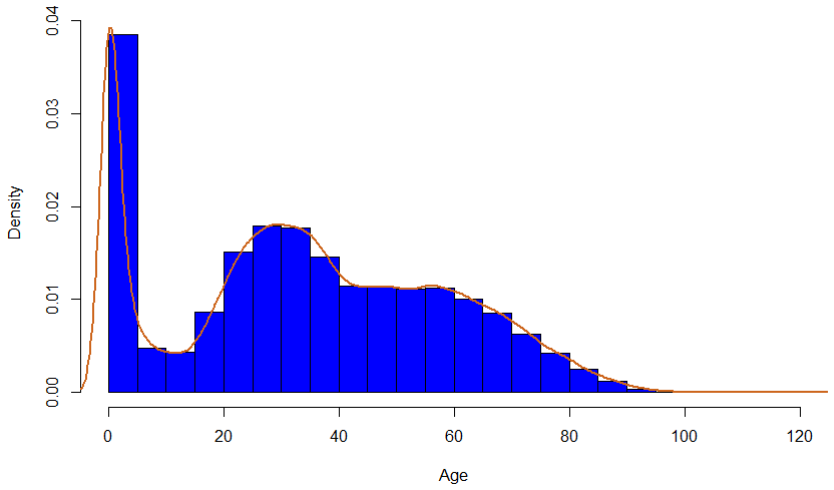

Figure 1: Age distribution 2019-2021

From Figure 1 it is clear that the patient files show a large number of newborns (age 0) that causes a peak to the left. There is a second peak between 30-35. The maximum patient age recorded is 120. This age is possible as it is in line with the oldest person in South Africa recorded at age 128 but will need to be investigated as a valid data point.

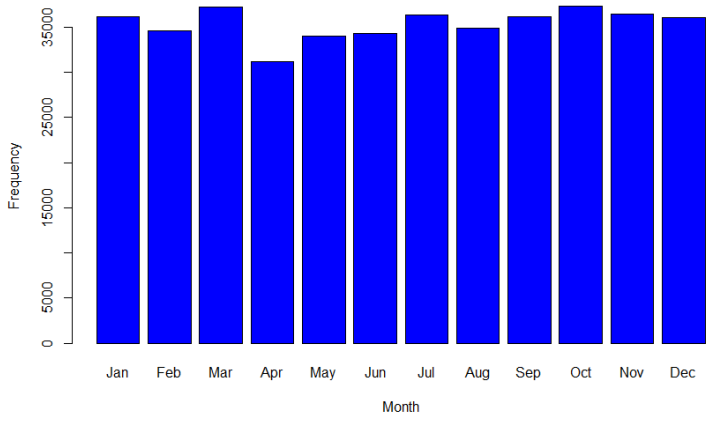

Figure 2: Volume of Cases by Month 2019-2021

Figure 2 shows that there is no clear indication of seasonality when looking at the volume of cases within the year for admissions. October has seen the highest admissions at 37 317 admissions.

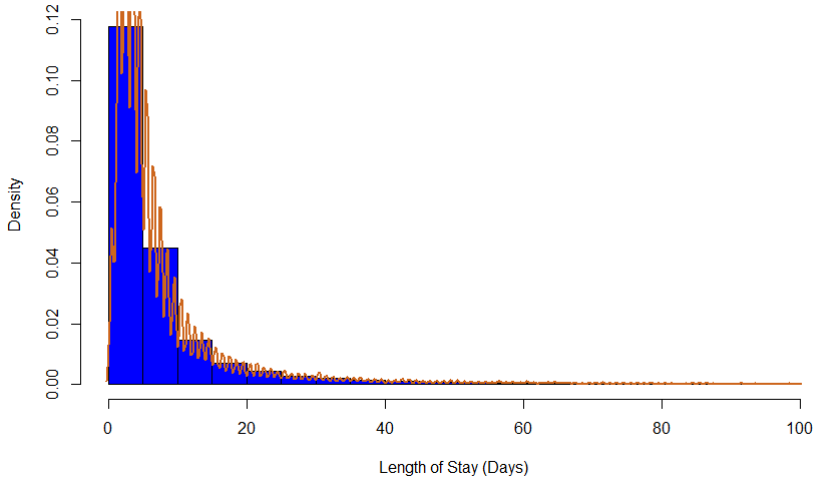

Figure 3: Length of stay (days)

Figure 3 shows that maximum days recorded for length of stay is 100. The distribution is skewed to the right with the majority of data between 0 and 15.

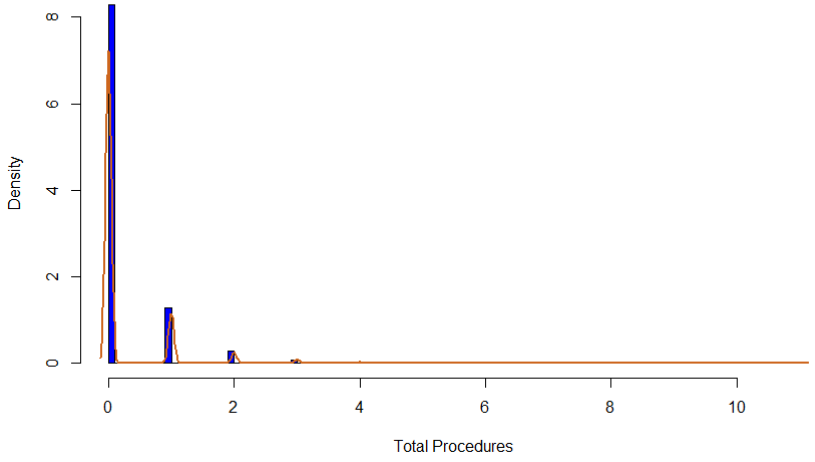

Figure 4: Total number of procedures

Figure 4 shows that majority of patients have no procedures. Maximum number of procedures recorded for a patient is 10. There are very few patients that underwent more than four procedures.

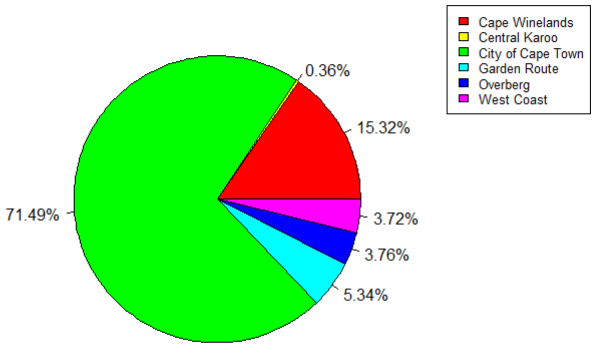

Figure 5: Hospital district

From Figure 5 we observe that most patients in the data are looked after in City of Cape Town area hospitals at 71.5%. This might be due to more specialised equipment and available beds. In addition the population density in the metro areas are higher. The Cape Winelands area is second at 15.32%.

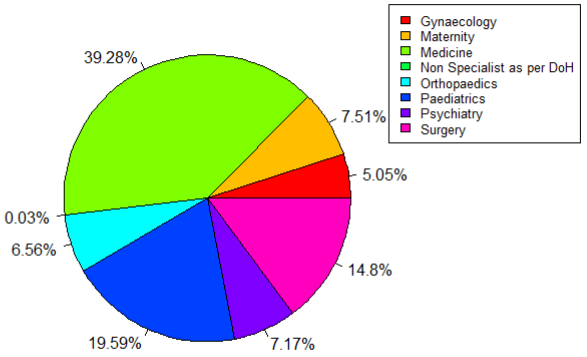

Figure 6: Specialty group

From Figure 6, it is evident the majority of admissions are made within general medicine at 39.29%. Second is paediatrics at 19.59% followed by surgery admissions at 14.8%.

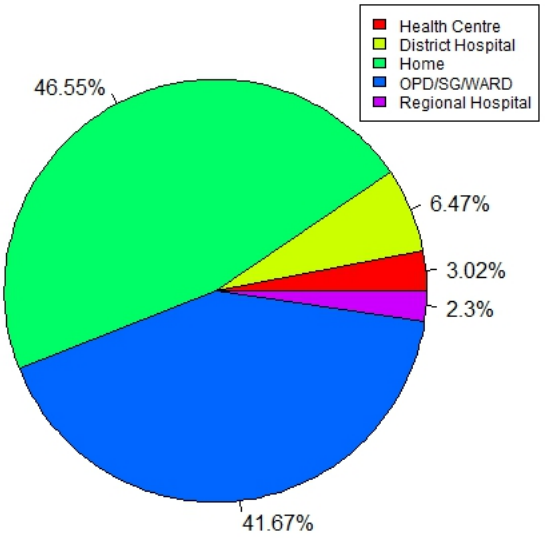

Figure 7: Source of admission

Looking at Figure 7 the largest source of patient admission is Home (normal admission where patient arrives by their own means from home) at 45.25%. Second is outpatient department admissions at 40.5% where there is no requirement to admit a patient in a hospital.

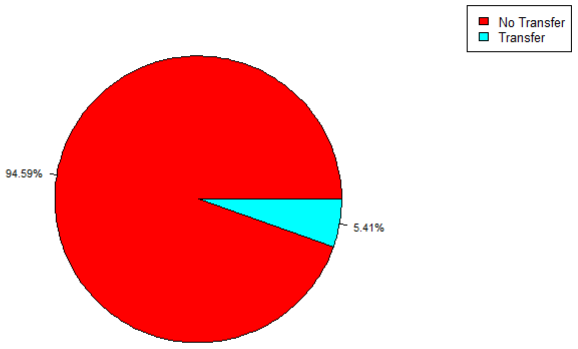

Figure 8: Transfers

From Figure 8 we see that 5.41% of patient admission are transfers from other hospitals. This happens when the hospital the patient was initially seen in did not have enough resources or equipment to adequately look after the patient.

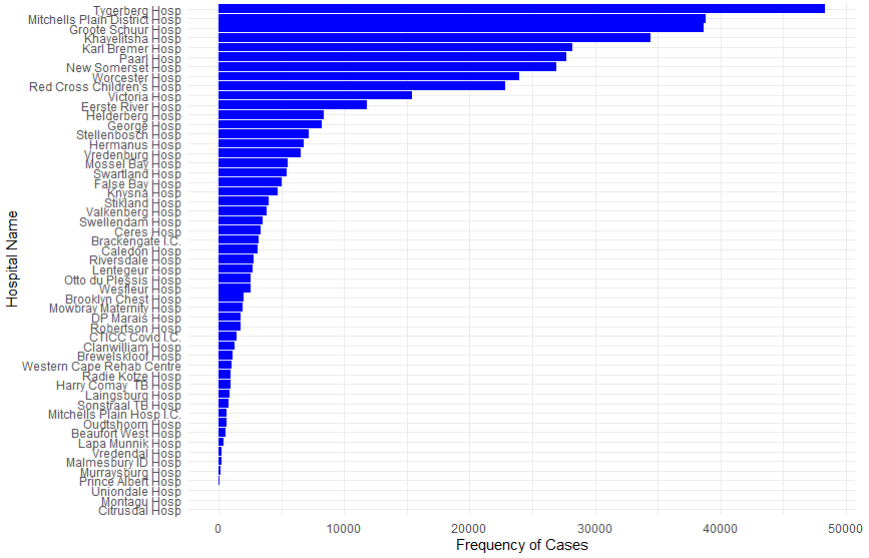

Figure 9: Volume of cases by hospital 2019-2021

Figure 9 shows that Tygerberg hospital has the highest number of cases at 48 336. Mitchells Plain District hospital has the second highest number of cases at 38 841. Groote Schuur hospital has the third highest number of cases at 38 682.

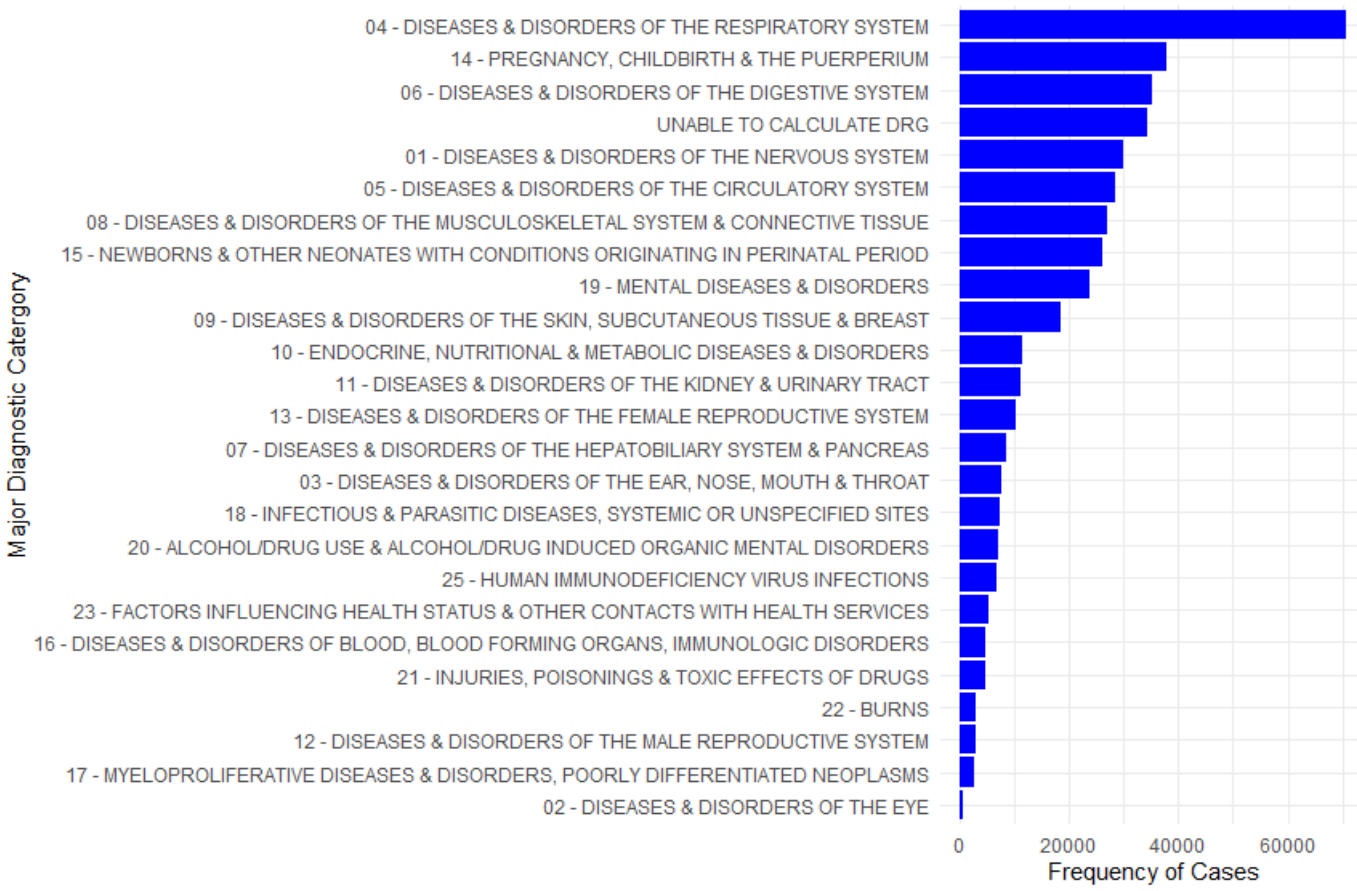

Figure 10: Volume of major diagnostic category

As mentioned earlier, the focus of this dissertation on the 23 Medicare Major DRG. Looking at Figure 10, respiratory system diseases and disorders made up the majority of the admissions at 70 742 cases. Second is pregnancy with 37 775 cases. Third is digestive system disorders with 35 085 cases. The lowest volume cases are diseases and disorders of the eye at 50 cases. The "unable to calculate DRG" category is where there is missing or incomplete information for the Medicare DRG algorithm to group the case. The details of this will be considered later in this section.

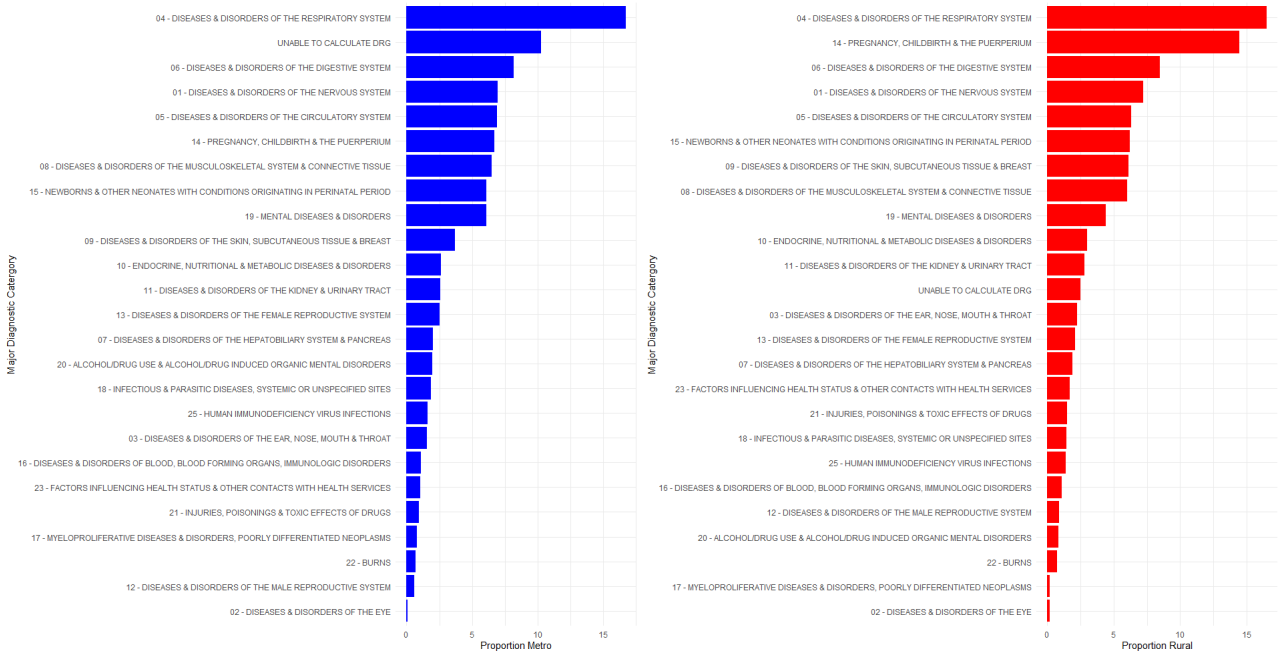

Figure 11: Metro and rural vs major diagnostic category

We observe from Figure 11 that there are different proportions of diseases experienced for the metro and rural hospitals. The biggest difference is that rural hospitals experience a larger proportion of childbirth's which is the second biggest admission category.

In metro hospitals, childbirths are the 5^{th} highest admission category. The metro hospitals also have a large proportion of "unable to calculate category" compared to rural hospitals which is their 12^{th} most prevalent in terms of proportions. The reason could be that metro hospitals receive much more complex cases.

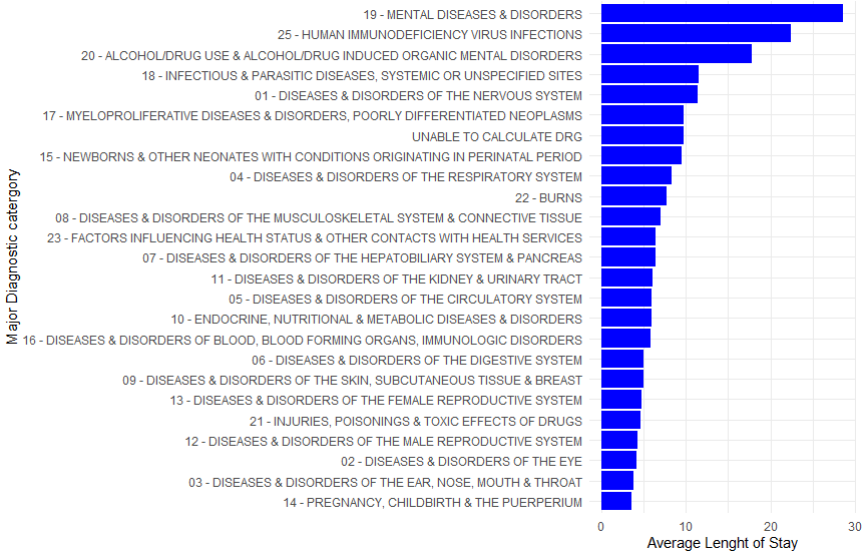

Figure 12: Average length of stay for major diagnostic category

Looking at Figure 12, mental disease and disorders have the highest average length of stay at 28.62 days. Second highest average length of stay is human immunodeficiency virus infections at 22.45 days . Third is alcohol/drug use at 17.75 days on average. Pregnancy has the lowest length of stay at 2 days on average.

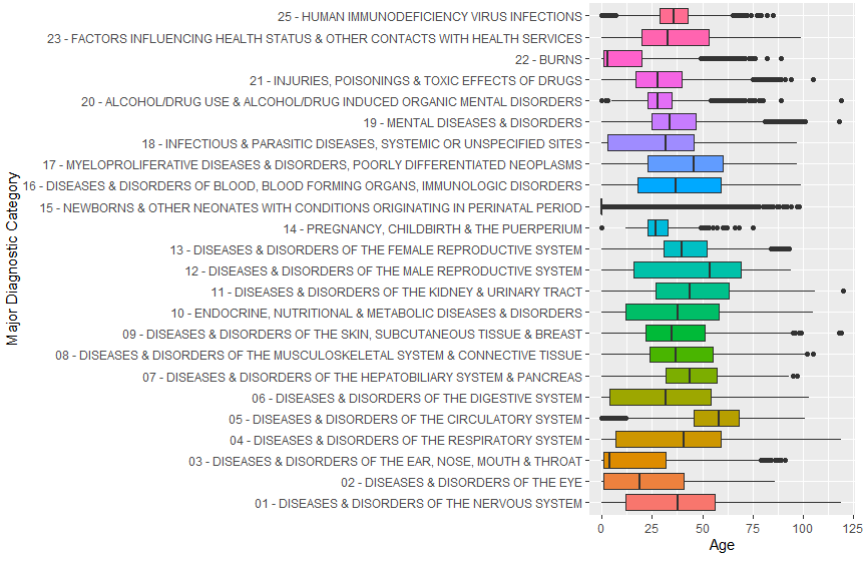

Figure 13: Age distribution for major diagnostic category

Looking at Figure 13, circulatory system disease and disorders cases have the highest mean age at 57 years. Burn and ear throat and nose disease and disorders cases have the lowest mean ages with 3 and 5 years old respectfully. It's observed that the age distribution between different DRG

cases varies.

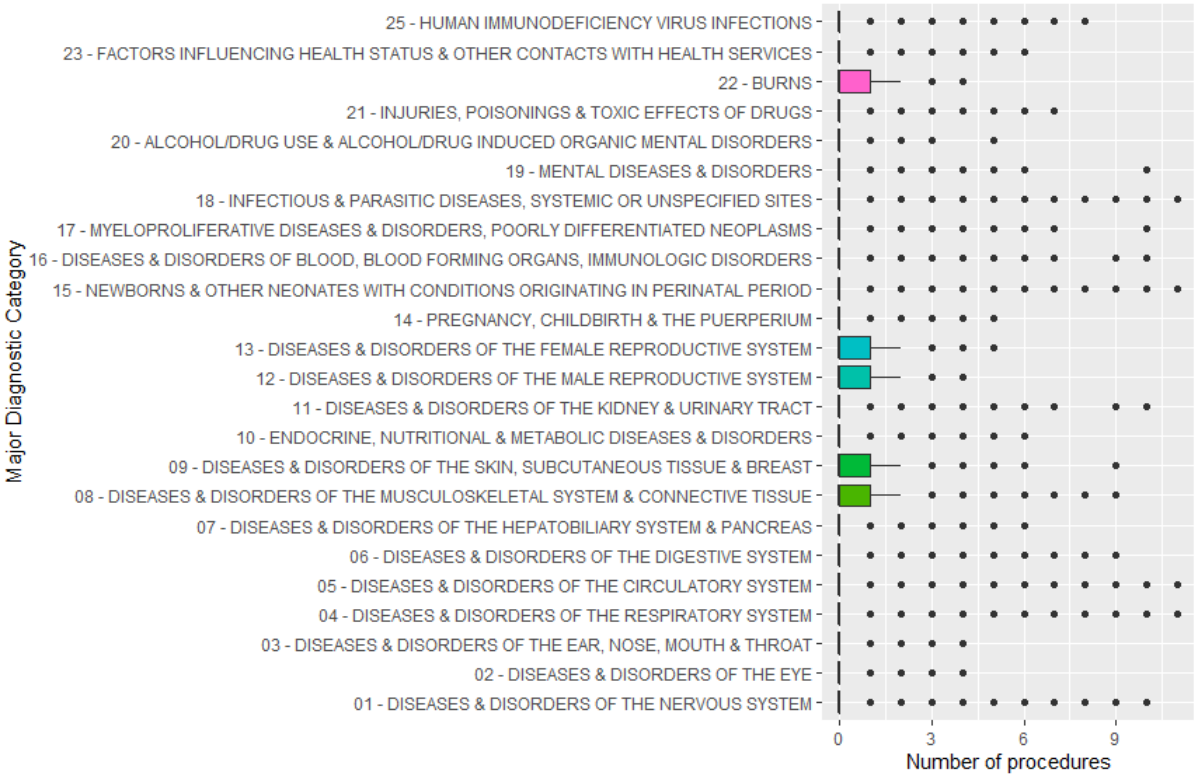

Figure 14: Total number of procedures distribution for major diagnostic category

Looking at Figure 14, most DRGs have the 25^{th} , 50^{th} and 75^{th} percentiles at 0 procedures. Burns, male and female reproductive system, skin subcutaneous tissue and musculoskeletal system disease and disorders have rightly-skewed distributions.

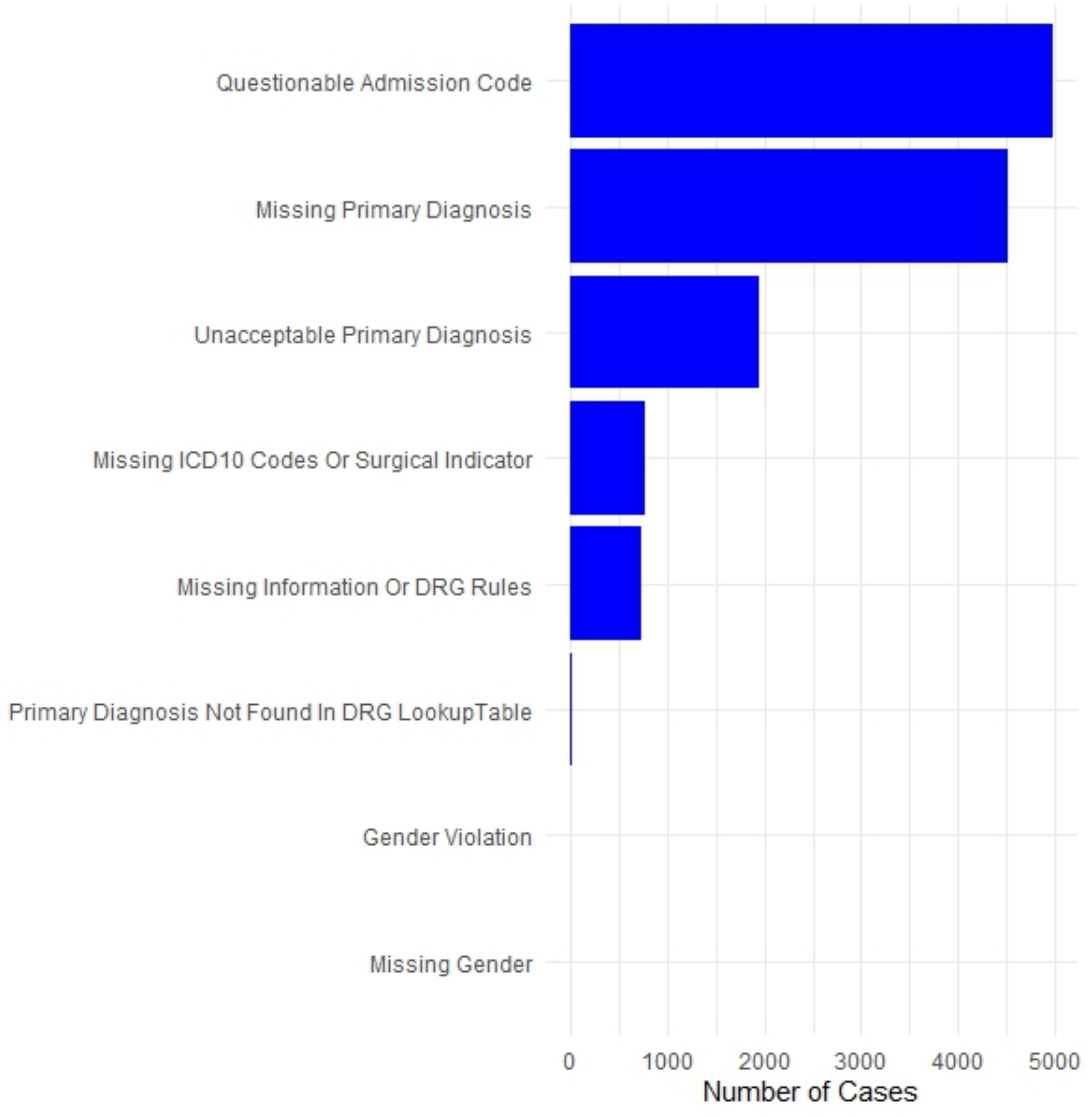

Figure 15: Unable to classify DRG reason

Figure 15 shows the reason for not being able to classify cases using the Medicare DRG algorithm. The highest error that occurs due to Questionable Admission Code at 4 984 cases. Second is Missing Primary Diagnosis at 4 508 cases. Third is due to Unacceptable Primary Diagnosis at 1 951 cases. The lowest three are Missing Gender, Gender Violation and Primary Diagnosis Not Found DRG.

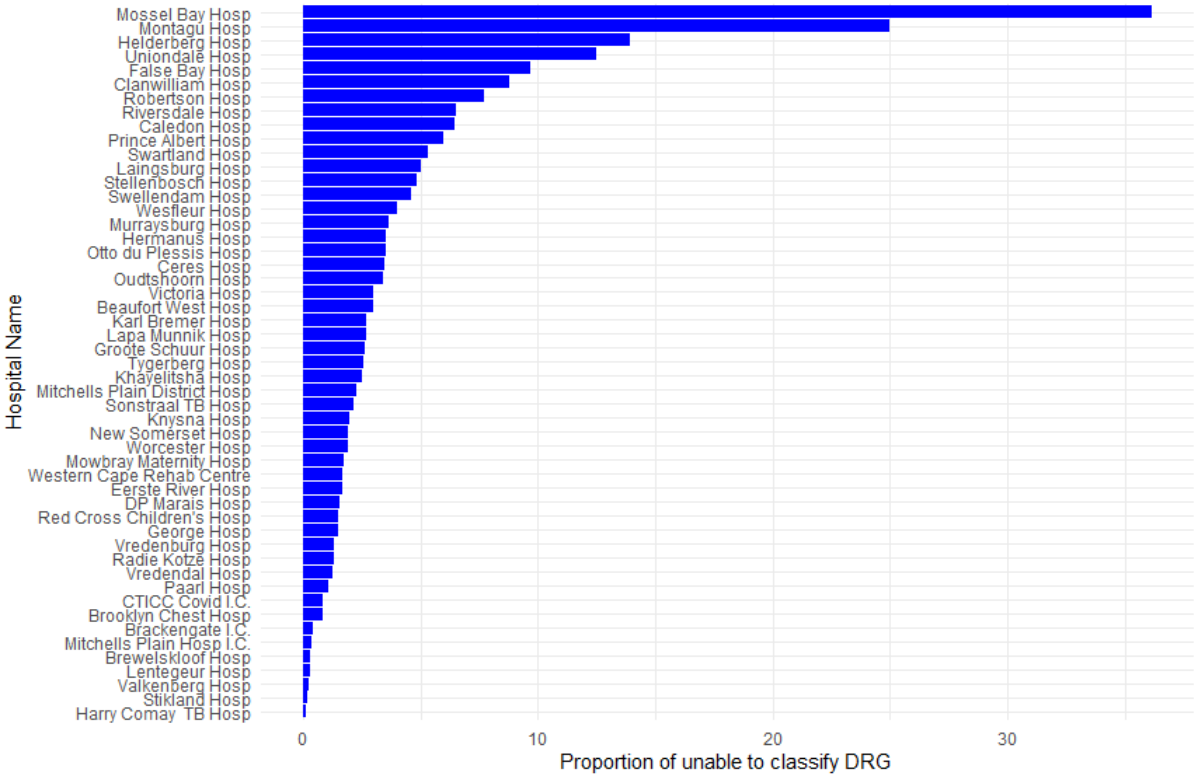

Figure 16: Proportion of unable to classify DRG by hospital

Looking at Figure 16, the top three hospitals with the highest DRG error proportions are Mossel Bay at 36.16%, Montagu at 25.00% and Helderberg Hospital at 13.94%. The lowest DRG error rates below 1% are at Valkenberg, Stikland, Harry Comay TB, CTICC Covid I.C., Brooklyn Chest, Brackengate I.C., Mitchells Plain I.C., Brewelskloof and Lentegeur Hospital. The lower errors are found at specialist hospitals, were diagnoses are limited to a specific specialty. For example, Harry Comay TB deals with tuberculosis cases exclusively.

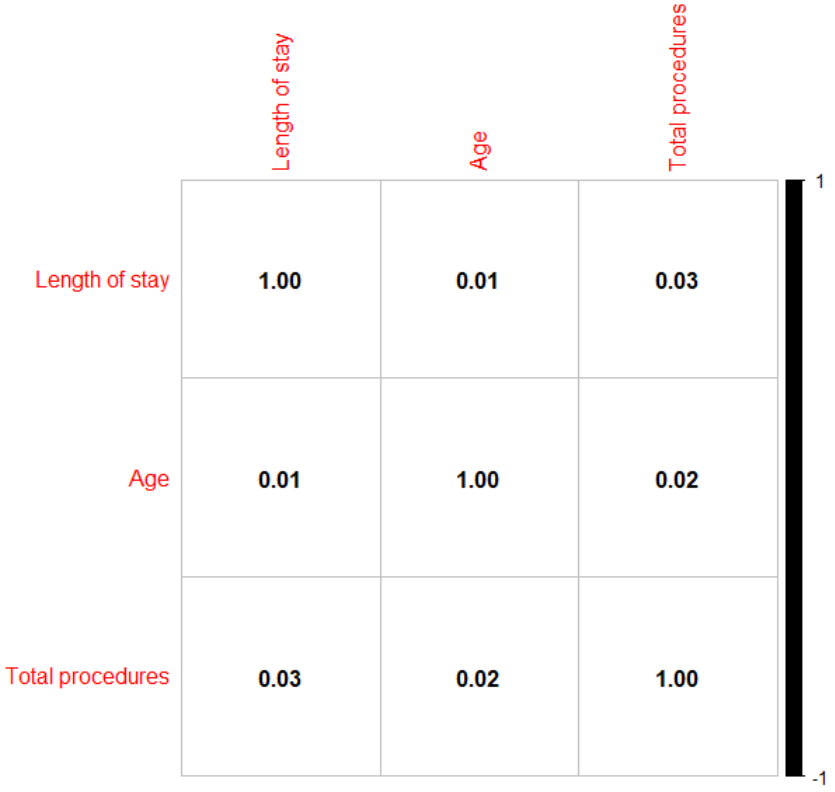

Figure 17: Correlation between continuous variables

Figure 17 above shows the Pearson correlation coefficient for all continuous variables. There is an extremely low correlation between age, length of stay and number of procedures.

4 Methodology

4.1 Case Mix

This section deals with the methodology used to process the Western Cape data and to calculate the CMI, CMAF and lastly compare the two methods. Case mix index is the more widely used metric however the Medicare weights are from the private sector and may not be appropriate for the public sector.

4.1.1 Overview of Methodology

DRG are formed through the use of a grouping algorithm. Every instance is placed in a single, final group, and its predicted value is the average of the group based on a weighted value using resource usage and clinical severity. Further partitioning is not possible if the sample sizes are insufficient or if the remaining variance cannot be reduced or explained through the variables in the database. In other words, if y_{kj} represents the value (based on a weighted value using resource usage and clinical severity) of the dependent variable for the j^{th} observation within the k^{th} group, then

$$y_{kj} = Y_k + \epsilon_{kj}$$

The procedure aims to minimize the sum of squared errors ($\sum \epsilon_{kj}^2$) between the predicted or estimated value Y_k for all members in the k^{th} group and the actual value y_{kj} of each observation. This results in the individual observations having values that tend to be close to the mean value of the group to which they belong. (Fetter, R., 1980)

The DRGs should be consistent in terms of their anatomy, physiology, or the way they are treated clinically. (Fetter, 1980):

- Must be uniform in terms of anatomy, physiology classification, or the clinical management approach.
- Must include a sufficient patient population.
- Must encompass all codes without any overlap or duplication.

The requirement for case mix enables the calculation of meaningful measures such as complication rates, admission severity, and risk of death, as well as the assessment of a patient's deterioration or improvement during their hospital stay. However, if multiple risk adjustment methodologies are required, it will become a burden for providers and hinder communication of results. (Goldfield, 2010)

4.1.2 Case Mix Index

The Case Mix Index (CMI) for the Western Cape is calculated as the average relative weight of a hospital's inpatient discharges, determined by the Medicare Diagnosis Related Group (DRG) weight. The Medicare DRG is a classification system employed by the Medicare program to categorize inpatient hospital stays into specific groups based on the patient's diagnosis and treatment. To compute the CMI, the DRG weight for each patient discharge is summed up, and the total is divided by the number of discharges.

As previously mentioned, the DRG weights signify the average resources required to care for cases in a particular DRG. The CMI reflects the range, clinical complexity, and resource needs of all patients in the hospital, with a higher CMI indicating a more complex and resource-intensive patient population on average. It's worth noting that although the DRG weights were developed for the Medicare population by the Centers for Medicare and Medicaid Services (CMS), they are utilized for all discharges in the dataset.

$$\text{Case Mix Index} = \frac{\sum_{j=1}^{total\ cases} DRG\ weight\ _j}{total\ cases} for j = 1, 2, 3...$$

where

- $DRG\ weight\ _j$ is the specific weight for the j^{th} case in a hospital.
- *totalcases* are the total number of patient admissions in a given hospital that we are calculating the CMI.

4.1.3 Case Mix Adjustment Factor

The purpose of a CMAF like the CMI is to account for variations in expected resource use for different illnesses and levels of severity. This allows for a more accurate comparison of efficiency scores by separating the effects of patient characteristics and treatment practices. Although the general approach is similar across studies, the methods used to develop the CMAF varies. Most studies construct the factors based on the distribution of cases across DRG. (Linna, 1998; Zuckerman et al., 1994; Fetter, 1991; Fetter et al., 1980).

Fetter et al. (1980) uses patient length of stay (LOS) as a proxy for resource consumption in their analyses. Using LOS is the most appropriate choice for this dissertation given the drawback of not having individual costs per case in the data set of the Western Cape (which is common in the public sector).

Consider:

$$CMAF_i = \frac{\sum_k A_k p_{ik}}{\sum_k A_k P_k} \text{ for k =1,2,3...}$$

where

- p_{ik} = the proportion of hospital $i's$ cases in the k_{th} DRG.
- P_k = the proportion cases across all hospitals in the k_{th} DRG.
- A_k = the average LOS for cases across all hospitals for the k_{th} DRG.

The disadvantage of using LOS is that it does not account for the fact that a day spent in an intensive care unit uses significantly more resources than is required for a day in a typical general ward. Billed amount in the standard measure of resource use in the private health sector. Using the billed amount has two drawbacks. First, the total amount billed is influenced by both patient-related traits and supply-side variables, such as prescription and referral practices of doctors. Therefore, the total amount billed does not include the extent of the illness or the entire amount of resources needed for the patient. Rather, it is complicated by supply-side issues instead of features. The second disadvantage is that costs per procedure may vary across hospitals and geographic locations.

4.2 Error Detection

This section dives into the methods for neural network, multinomial logistic regression and random forest. The aim of all these models is to predict the major DRGs using independent predictor variables. The audited data are used for training and testing on all models. A seed is set to 100 when randomly selecting a training and testing set for all models. This is to insure reproducing same splits when re-running the code. This process will also shuffle our data since our data was represented in order of category. There is further shuffling that happens when the models are built. The split for training and testing is 75% and 25% respectively for all models.

Max-Min Normalization is used to scale each variable between 0 and 1 for our independent variables. This is to ensure a common scale among the variables so that the models can interpret them fairly. A decision was made to focus on only 11 major diagnosis related groups for prediction purposes. This was due to low volumes present in some groups which caused an issue whereby these low volume diagnoses will appear in the testing set but not in the training set. An arbitrary value of $n \geq 20$ was chosen where n is the number of cases to determine which diagnoses to include. This insures the categories that remain are well represented in the data. After the removal the size of the remaining audited data set was 587 observations.

4.2.1 Neural Network

Overview of Methodology

The general modelling approach was taken which starts with a simple one layer neural network model as a benchmark for the approach and more complexity is then added keeping in mind both model fit and prediction error. The audited data are randomly assigned into a training and testing set at 75% and 25% respectfully. When tuning the neural network parameters for each model the training set is further broken down to validation subsets. The training-validation set consists of 30% of the training data. Inputs for the models include quantitative and qualitative data as seen in the exploratory analysis section 3.5. The Neural Network models were chosen as they do not rely on distributional assumptions and can accommodate nonlinear relations and interactions.

Structure

The number of parameters in a neural network is based on the size of the input vector, the size of the output vector and the number of units in each hidden layer. See figure 18 below:

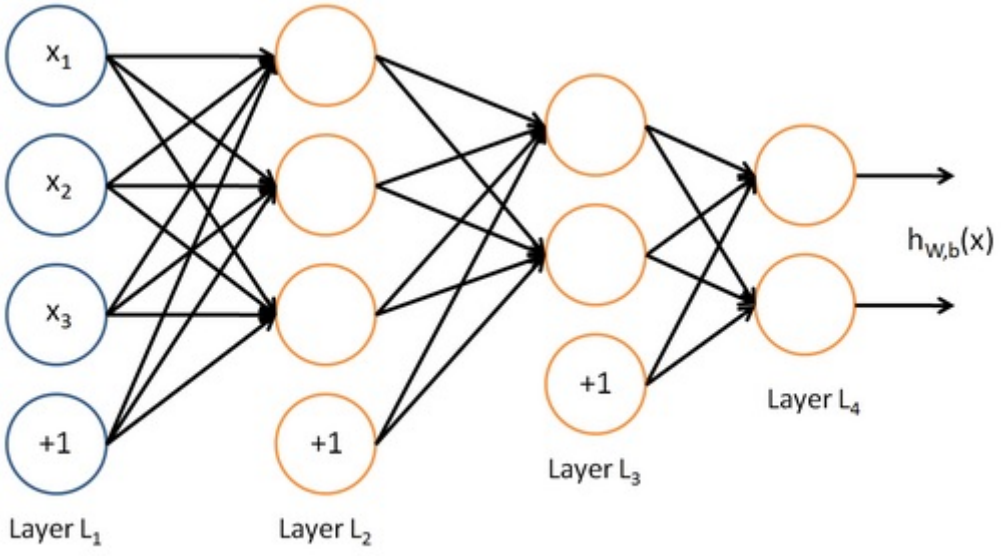

Figure 18: Structure of Neural Network

The figure will be used as an example to show how to calculate the number of parameters in a neural network. Here we have 3 input nodes + 1 input bias in L_1, 3 nodes and 1 bias in L_2 ,2 nodes and 1 bias in L_3, finally there are only two output nodes in L_4. So, in this case, we have (3x3) + 3 + (3x2) + 2 + (2x2) + 2 = 26 weight parameters. The model will always have 11 output nodes in our final layer since the predictor will have only 11 diagnosis categories after applying one hot encoding.

Cost Function

Since, our response variable is categorical, direct application of an appropriate distance measure is disadvantageous. Hence, for categorical response, it is better to formulate the objective function in terms of probabilities. This means that we are setting up the output layer to produce quantities that can fairly be interpreted as probabilities. Hence, on formulating an objective function in terms of the distribution for the target variable, we use cross entropy error. This error is formulated as:

$$C_{CE} = -\sum_{i=1}^{N} y_i log(\hat{y_i})$$

For a categorical response, the commonly used root mean squared error (RMSE) cannot be used as the cost function. Using RSME would saturate the output activations in order to minimise the distance between the observation data and the predictor.

Hidden Layer Activation Function

$$f(z) = max(0, z)$$

The activation function chosen for the hidden layers was Rectified Linear Unit (ReLU). One major benefit is ReLU helps alleviate the vanishing gradient problem, where the gradients become very

small and slow down training. Which is commonly encountered in networks with sigmoid or tanh activations. The other benefit of ReLU is that it encourages sparsity in the activations, meaning that many of the activations will be zero, which can be useful in some cases. ReLU is also computationally efficient compared to other activation functions such as sigmoid and tanh. ReLU provides a non-linear mapping, allowing neural networks to model complex relationships between inputs and outputs.

Data Transformation

Max-Min Normalization is used to scale each independent variable between 0 and 1. This makes the neural network more efficient when finding optimum weights. This also makes ReLU an appropriate choice of activation for our hidden layers. See figure 28 in appendix for results of transformation.

Final Layer Activation Function

$$softmax(z_i) = \frac{exp(z_i)}{\sum_i exp(z_i))}$$

The final layer activation function chosen was softmax because the response variable is a categorical variable. In the equation above $i = 1, 2, 3, 4..11$. It has the benefit of normalizing the output of a network to a probability distribution over predicted output classes. The components will add up to 1, so that they can be interpreted as a probability distribution. Furthermore, the larger input components will correspond to larger probabilities.

Handling Over-fitting

To avoid over-fitting dropout is used. This is a regularisation method that approximates training a large number of neural networks with different architectures in parallel.

Figure 19: Drop Out

By dropping a unit out, we mean temporarily removing it from the network, along with all its incoming and outgoing connections. Dropout has the effect of making the training process noisy, forcing nodes within a layer to take on more or less responsibility for the inputs.

Optimisation

The optimisation algorithm used will be Root Mean Square Propagation (RMSprop). It is an optimization algorithm for training deep learning models. It is an extension of the gradient descent

algorithm that uses a moving average of squared gradients to normalize the gradients, which helps reduce the amount of noise and oscillations in the gradients and allows for faster and more stable convergence. RMSprop also introduces a learning rate that adapts to the magnitude of the gradients, which helps to reduce the amount of time needed to train a model.

4.2.2 Multinomial Logistic Regression

Multinomial logistic regression is another model chosen because it can handle a dependent variable with more than two categories (i.e., multi-class classification). Unlike binary logistic regression, which deals with two categories, multinomial logistic regression can handle multiple classes simultaneously. Although the main objective is predictive accuracy, it is also valuable to note that it provides interpretable coefficients for each category. This allows for understanding of the the impact of independent variables on the different outcomes.

Multinomial logistic regression is a well-established and widely used method. In addition, unlike linear regression, it does not assume a linear relationship between the independent variables and the log-odds of the categories. It is important to note that like any statistical technique, multinomial logistic regression also has its limitations and assumptions. This will be discussed later in section 5.2.2.

Overview of Methodology

The Generalized Linear Model (GLM) is a statistical framework for modeling relationships between a response variable and one or more predictor variables.

It is characterized by three components:

- A random component that defines the conditional cummulative distribution function (CDF) of the response variable Y_i given X_i.
- A linear predictor η.
- A link function $g(\cdot)$ that maps the linear predictor to the mean of the response variable.

The multinomial (CDF) with probabilities $\pi_1, ..., \pi_J$, where $\sum \pi_r = 1$, is the random component in a GLM for a categorical response with J categories. The linear predictor, represented by $(\eta_1, ..., \eta_{J-1})$, is calculated by multiplying the design matrix $\boldsymbol{Z}$ and the unknown parameter vector $\boldsymbol{\beta}$.

The relationship between the linear predictor and the probabilities is defined by the link function $g(\pi) = \boldsymbol{Z\beta}$, with J$-$1 equations, where $g_j = \eta_j$. Peyhardi, Trottier, and Guédon (2015) suggested writing the link function as

$$g_j = F^{-1} \circ r_j \longleftrightarrow r_j = F(\eta_j) \quad \text{j = 1, 2, ..., J}-1$$

where F is a cumulative multinomial cdf and r=$(r_1, ..., r_{J-1})$ is a transformation of the expected value vector called the ratio.

The transformation that will be applied for this dissertation is the called the reference as seen below

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J}$$

The adjacent, cumulative, and sequential transformations all depend on the assumption that there is an ordering among the categories. Each component $r_j(\pi)$ can be considered a conditional prob-

ability. In the case of the reference ratio, each category j is evaluated in comparison to a reference category J. The reference ratio is specifically designed for nominal response data. The reference category J chosen for this dissertation was the diseases and disorders of the musculoskeletal system.

Forward selection

Forward selection is a stepwise regression method for selecting predictor variables in a multiple regression analysis. The idea behind forward selection is to start with an empty model and add one predictor variable at a time, based on the ability of each variable to improve the goodness of fit of the model. The variable that provides the largest improvement in the model fit (as measured by a chosen criterion such Akaike Information Criterion (AIC)) is selected and added to the model. This process is repeated until all predictor variables are included in the model or until no further improvement in the model fit is observed.

$$AIC = 2k - 2ln(\hat{L})$$

where:

- k is the number of parameters in the model
- L is the likelihood of the data, given the model.

AIC, or the Akaike Information Criterion, is a model selection criterion that balances the goodness of fit of a model with the number of parameters in the model. It is used to compare different statistical models and to choose the best model among them. (Akaike, 1974)

4.2.3 Random Forest

Overview of Methodology

Random forest is a popular machine learning algorithm that is widely used for classification, regression, and other predictive modeling tasks. It is an ensemble method that combines multiple decision trees to make predictions. Random forest has been applied in various fields such as finance, healthcare, and natural language processing. It has also been used for feature selection, outlier detection, and imbalanced data classification.(Ho, 1995)

The algorithm works by randomly selecting a subset of features and building a decision tree using only those features. This process is repeated many times, creating a collection of decision trees, each trained on a different subset of features.

The advantage of random forest is that it can handle a large number of input features, even when some of them are correlated or irrelevant. It also has the ability to detect interactions between features, making it well-suited for complex data-sets.(Ho, 1995)

First, you need to define your input data-set. Given a dataset D with n observations, where each observation i has p input features denoted by

$$X_i = (X_{i1}, X_{i2}, ..., X_{ip}),$$

where y_i is the corresponding class label for classification.

The random forest algorithm as follows for each tree t in the forest:

- Randomly select a subset of m features from the p input features.
- Create a bootstrap sample of n observations from the dataset D.
- Fit a decision tree to the bootstrap sample using the selected subset of m features.

For each observation i in the data-set, obtain the predicted class or response value from each decision tree in the forest. For classification, the class with the highest frequency is the predicted classes for each observation.

Overall, the random forest algorithm is designed to improve the accuracy and generalizability of decision trees by combining the predictions of multiple trees that are trained on different subsets of the input features and the data-set. This can help to reduce the variance and over-fitting of decision trees, and increase their robustness to noise and outliers in the data.(Ho, 1995)

4.2.4 Synthetic Minority Over-sampling Technique

Synthetic Minority Over-sampling Technique (SMOTE) is a statistical method used to address the problem of imbalanced data in classification problems. Imbalanced data refers to a situation where the number of observations in one class is much smaller than the number of observations in another class. This can lead to biased results and reduced accuracy of classification models.(Chawla, 2010)

SMOTE has been widely used in various fields such as finance, healthcare, and fraud detection. It has also been successfully applied in many machine learning algorithms such as decision trees.

Overview of Methodology

SMOTE is a technique that creates synthetic data points for the minority class by interpolating between existing minority class observations. The method works by randomly selecting a minority

class observation and then selecting one of its nearest neighbors. A new data point is then created along the line joining the selected observation and its neighbor. This process is repeated until the desired level of oversampling is achieved.(Chawla, 2010)

The advantage of SMOTE is that it increases the size of the minority class without introducing bias. It also reduces the risk of over-fitting by generating synthetic data that are similar to the original data.

The algorithm starts by defining the input data-set. Given a dataset D with n observations, where each observation i has p input features denoted by

$$X_i = (X_{i1}, X_{i2}, ..., X_{ip}),$$

and a corresponding class label y_i.

Next, you can define the SMOTE algorithm as follows:

1. Choose an observation i from the minority class.
2. Choose k nearest neighbors of i from the same class.
3. For each neighbor j, create a synthetic observation by interpolating between i and j:

$$\tilde{X}_{ij} = X_i + \lambda(X_j - X_i)$$

   where λ is a random number between 0 and 1.
4. Add the synthetic observations to the dataset.
5. Repeat steps 1-4 until the desired balance between the minority and majority classes is achieved.

The minority class is defined as the class with fewer observations, and the majority class as the class with more observations. Overall, the SMOTE algorithm is designed to increase the diversity of the minority class by creating synthetic observations that are similar to the existing minority observations, but also slightly different. This can help to improve the performance of classification models that are trained on imbalanced data-sets.(Chawla, 2010)

5 Results

5.1 Case Mix

This section looks at the results of CMI, CMAF and explores the relationship between the two.

5.1.1 Case Mix Index

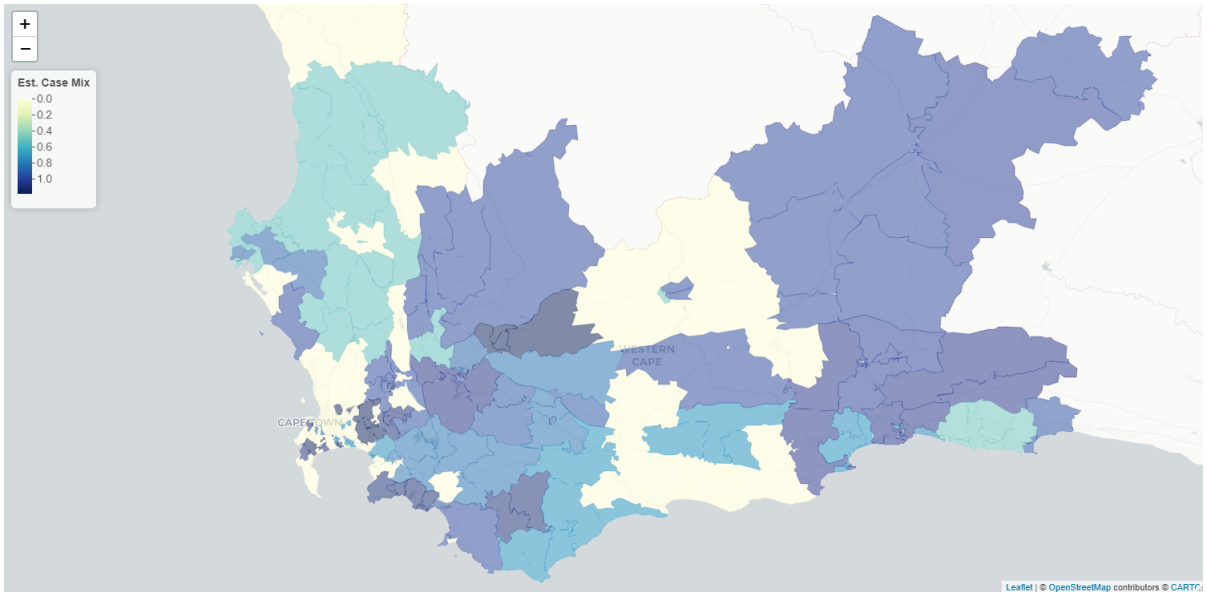

Figure 20: Choropleth Map of Western Cape based on CMI

Above figure 20 is a Choropleth Map of the CMI for the Western Cape. The map was created on a sub-district level. Darker blue hues represent a higher disease burden while white districts represent those without a hospital. The major metropolis areas have darker hues meaning that they have the highest disease burden in hospitals.

Hospitals situated in the metro areas have a higher average case mix index of 1.04 when compared to hospitals situated in the rural areas at 0.81. A reason for this could be that metro hospitals are better equipped to handle resource intensive diseases and that more severe cases are transferred to metro hospitals.

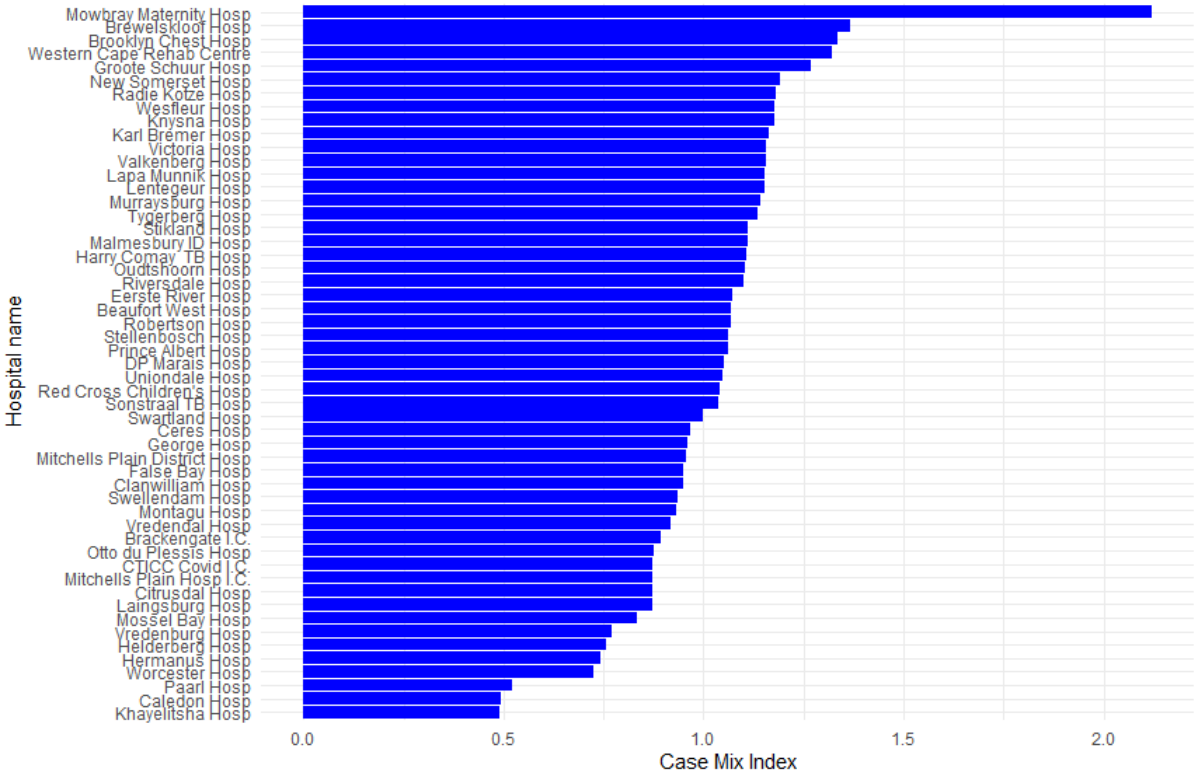

Figure 21: Case Mix Index for Individual Hospitals

As seen in Figure 21, Mowbray Maternity has the highest disease burden of 2.4. This may be due to the high proportion of neonates who require longer stays and a high level of care. The lowest disease burden was found in a rural hospital, Khayelitsha, at 0.51 and-as explained earlier-it may be the case that severe cases are transferred to other hospitals.

5.1.2 Case Mix Adjustment Factor

Using the methodology described in section 4.1.3 on the raw data set it's possible to calculate the hospital specific CMAF. Keep in mind that the difference between CMI and CMAF is that the CMAF uses actual LOS data instead of medicare DRG weights. A high CMAF value suggests that hospital i has a case mix that requires more resources compared to the average of all hospitals. A CMAF less than one means that the hospital is treating fewer severe cases compared to the average hospital.

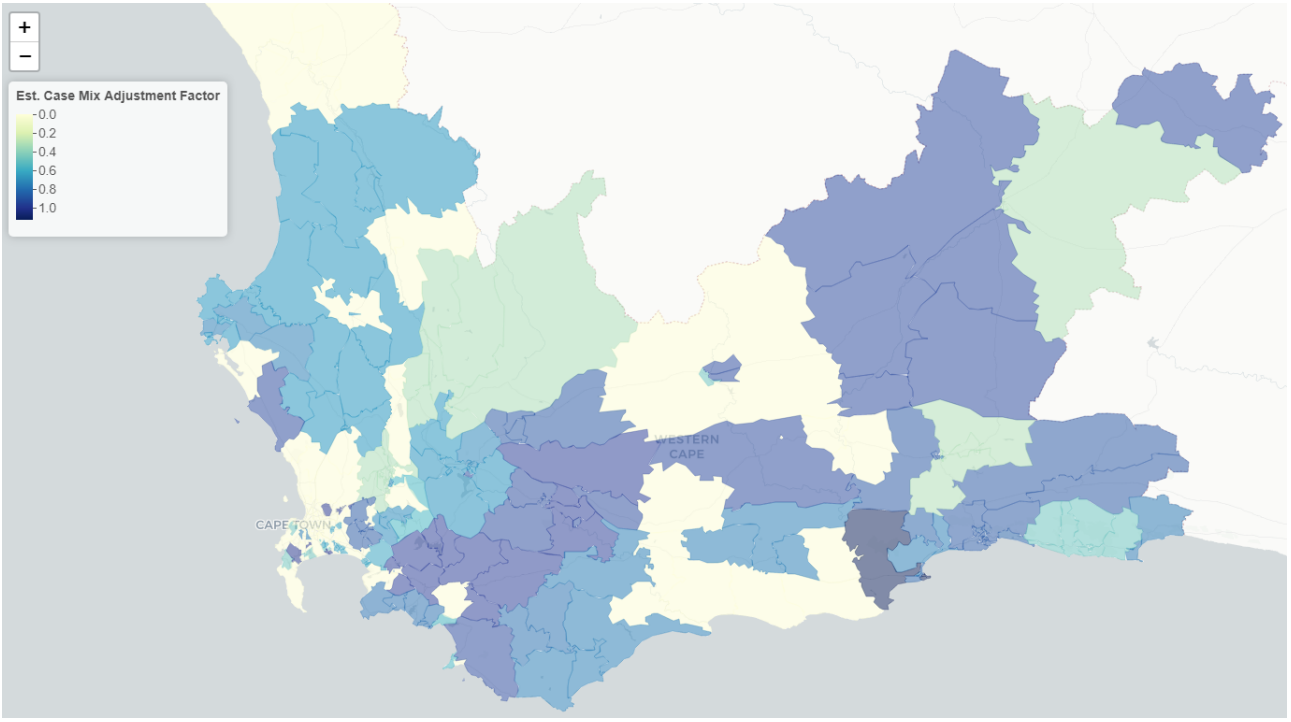

Figure 22: Choropleth Map of Western Cape based on CMAF

Figure 22 is a Choropleth Map based on the CMAF for the Western Cape. The major metropolis areas have darker blue hues indicating a higher disease burden. This is similar to what is seen in Figure 21 based on the case mix index.

Hospitals situated in the metro areas have a higher average CMAF of 1.52 when compared to hospitals situated in the rural areas at 1.02. As noted this may be due to more severe cases being transferred from resource-constrained rural hospitals to better-equipped metro hospitals.

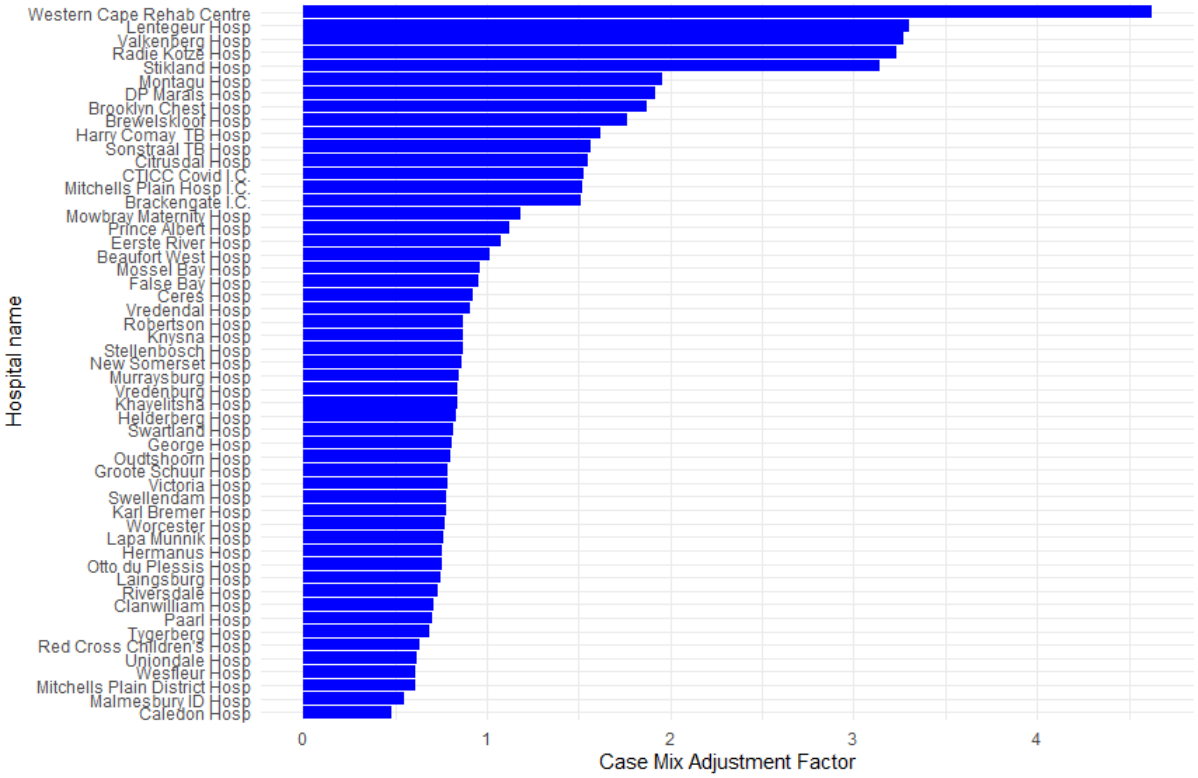

Figure 23: Case Mix Adjustment Factor for Individual Hospitals

As seen in Figure 23, the Western Cape Rehab centre has the highest CMAF of 5.4 and hence the highest disease burden. This is likely because mental health issues have the highest length of stay on average as seen in exploratory analysis section 3.5. This finding shows one of the drawbacks of this method, as looking at length of stay alone without accounting for costs and resource use can skew results. For example, a cancer patient who may have a length of stay for a single day might use more resources than a mental health patient whose length of stay is a week. Caledon, a rural hospital, has the lowest CMAF at 0.49.

5.1.3 Case Mix Adjustment Factor vs Case Mix Index

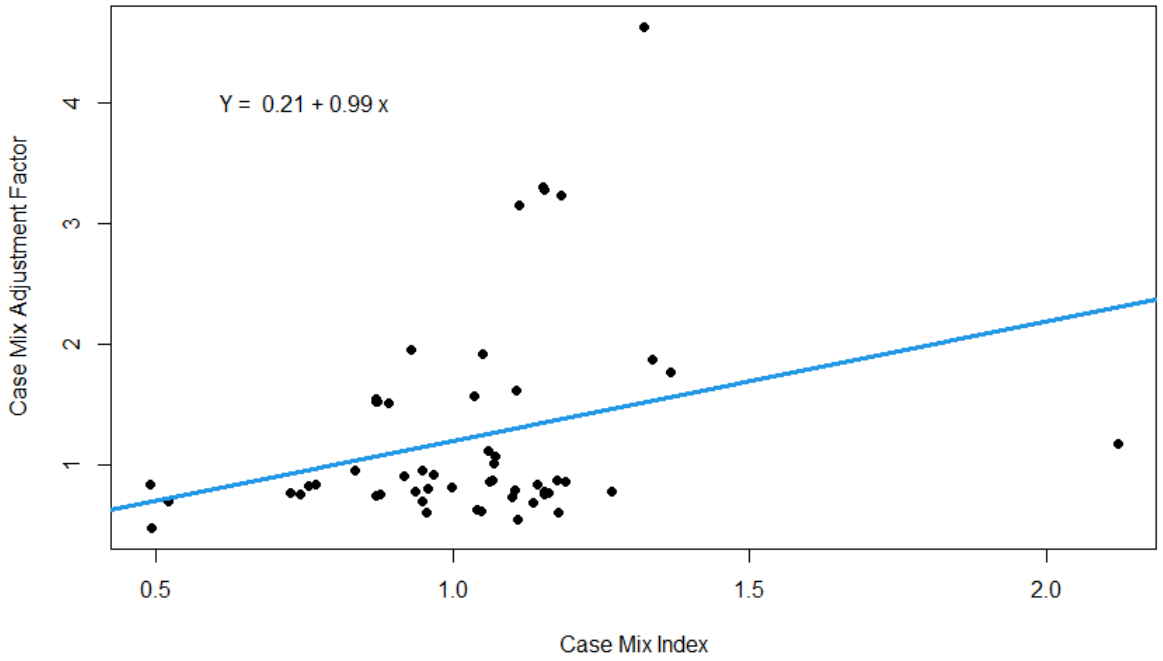

Figure 24: Case Mix Adjustment Factor vs Case Mix Index for Individual Hospitals

When looking at the scatter plot (Figure 24) it is observed that there is a weak positive relationship between the two metrics from the 53 hospitals. The trend line which is based on a simple linear regression, suggests the same showing an R-squared of 0.09. The correlation coefficient is 0.3, which is interpreted as weak positive correlation. As mentioned earlier looking at CMAF (which uses LOS as a proxy for resource utilisation) has disadvantages.

5.2 Results Error Detection

The models below are formulated using the methodology described in section 4.2. As explained one of the use cases identified was medical coding error detection. If the chosen model identifies a diagnosis different to the one reported with a high probability it can be used as an indicator for a possible error. This case can then be sent back to the hospital administration staff to be rectified. From the audit report mentioned in section 1.1, it is known that these errors happen frequently.

5.2.1 Neural Network

The first model was a simple model consisting of one hidden layer including 50 nodes and 11 output nodes. As discussed, our activation function is ReLU for the hidden layer and the output layers have a softmax function. The cost function that the neural network will try to minimise is cross-entropy error. This is used in optimising the weight parameters. A batch size of 20 was chosen. Small batch sizes can help to achieve a better trade-off between stability and generalization performance in deep learning training.

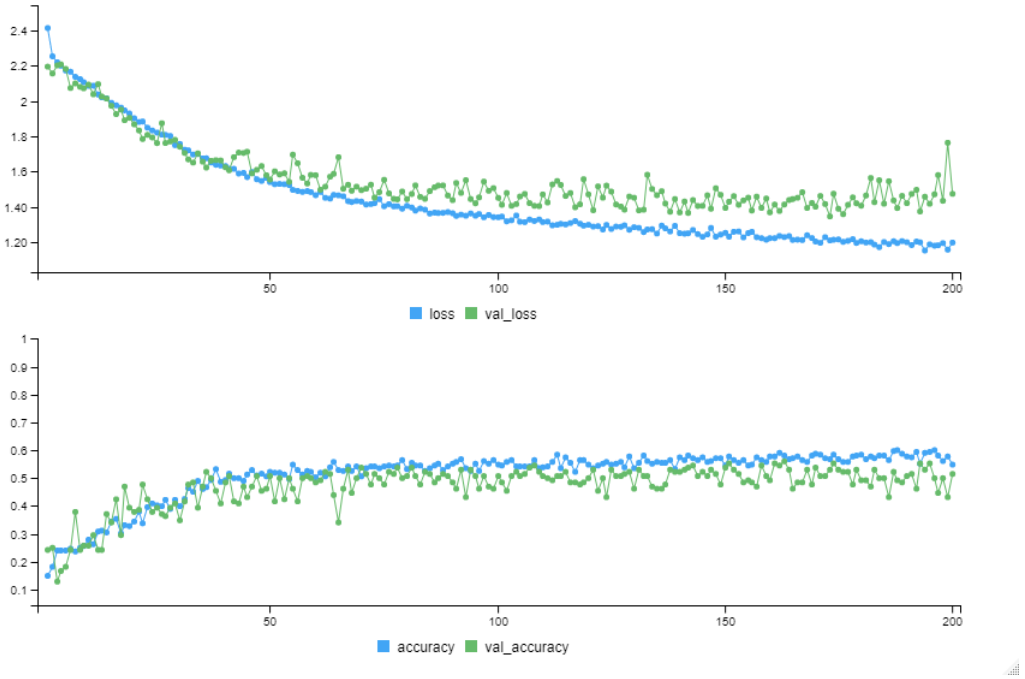

Figure 25: Simple Model 1

The green line in Figure 25 which represents the accuracy of the validation set and shows that over fitting becomes apparent at 50 epochs. The metric of interest for this task is accuracy. At around 100 epochs the accuracy for both the training and validation reaches a maximum and flattens. This means that accuracy does not improve significantly. This indicates that the number of epochs chosen is adequate. The RMSprop optimiser will automatically adjust the learning rate in order to increase the validation accuracy for this number of epochs. The validation accuracy was 51% .

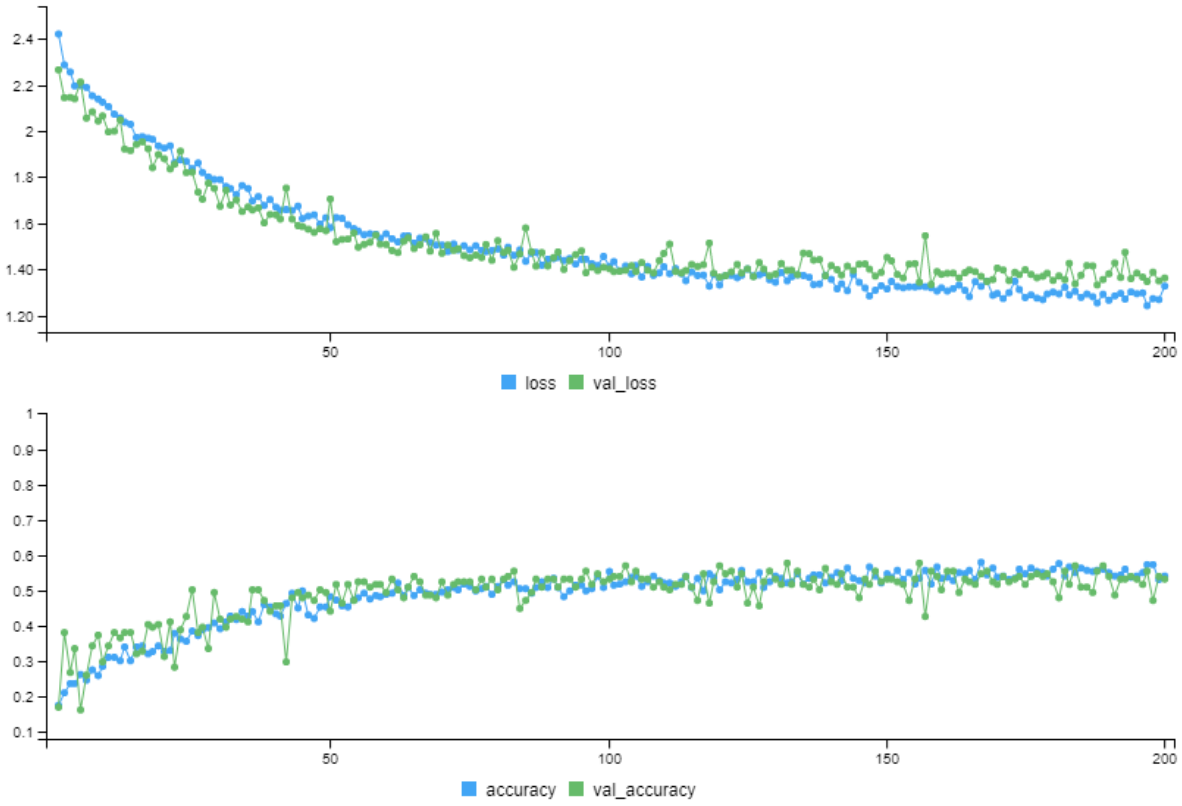

Figure 26: Model 1 with Drop Out rate=0.2

In Model 1 there are points where over-fitting does occur. Drop off was added to the hidden layer of Model 1 with rates at 0.1, 0.2 and 0.3. The model with drop out rate=0.2 performed the best for the same structure and epochs (Figure 26). The validation accuracy increased to 53%.

Table 2: Model Performance

Model	Hidden layers	Epochs	dropout rate	Validation Accuracy
1	[50]	200	[0.2]	53 %
2	[50, 50]	200	[0.2, 0.2]	53%
3	[50, 50, 50]	200	[0.2, 0.2, 0.2]	50 %
4	[100, 100, 100, 100]	200	[0.1, 0.1, 0.1, 0.1]	50%
5	[150, 150, 150, 150, 150]	200	[0.2, 0.2, 0.2, 0.2, 0.2]	49 %

Table 2 clearly shows that Model 1 and 2 are the best performing models of the five models because they have the highest validation accuracy. Model 2 is more complex and consists of a combination of hidden layers [50, 50] and drop out rates for each layer [0.2, 0.2]. Model 1 is a less complex structure [50] with a drop out rate of 0.2.

Test Performance

Model	Test Accuracy
1	52.7%
2	50 %
3	50 %
4	50 %
5	50 %

Model 1 also performs the best with the test data set with a accuracy of 52.7%. The other four models perform similarly at 50% on the test set.

Table 3: Detailed Test Performance

Diagnosis	Cases in Test set	Accurately Classified	Correct Classification Rate
Nervous System	20	1	5%
Respiratory System	20	17	85%
Circulatory System	20	8	40%
Digestive System	9	2	22%
Hepatobiliary System	6	3	50%
Skin, Subcutaneous Tissue & Breast	5	0	0%
Female Reproductive System	5	5	100%
Consitions Originating in Perinatal Period	12	10	83%
Immunological Disorders	4	0	0%
Mental Disease	22	12	55%
Musculoskeletal System	25	20	80%

Table 3 shows that Model 1 performs the best when classifying observations that belong to diseases and disorders of the female reproductive system at 100% and second best with diseases and disorders of the respiratory system at 85%. The worst performance is for diagnoses related to Skin, Subcutaneous Tissue & Breast and Nervous System disorders with 0% and 5% respectively.

5.2.2 Multinomial Logistic Regression

As there is an assumed distribution it is important to assess model assumptions before using multinomial logistic regression to ensure reliable. Several assumptions that must be adhered to when applying multinomial logistic regression:

- The dependent variables must be unordered categories. This holds as the training set was produced by selecting random rows.
- Each observation should be independent of all other observations in the data set which can be assumed to be true in most cases as patients are independent of each other. If a patient has a second hospitalisation that has also been audited this condition will not hold.
- There should be some variation in the dependent variable across the levels of the independent variables. There is variability looking at figures for diagnosis in section 3.5.
- The relationship between the independent variables and the log odds of the dependent variable should be linear.
- The sample size should be large enough to ensure reliable estimates of the model coefficients. This is a concern as there are only 587 observations and many parameter estimates, as seen in Table 6.
- The data should not contain extreme values or outliers that could have a significant impact on the model. All numeric data have been standardised between 0-1 to mitigate this.

Recapping section 4.2.3, the reference transformation was chosen to be applied to response variable diagnosis as seen below:

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J}$$

Hence the coefficient outputs for the multinomial logistic regression are in relation to the referent group. The chosen referent for this dissertation is Muscular Skeletal Disease.

Table 4: Forward Selection Model Build

Model	Variables	AIC
0	-	2100
1	Age	1772
2	Age + Sex	1730
3	Age + Sex + LOS	1711
4	Age + Sex + LOS + total_procedures	1698
5	Age + Sex + LOS + total_procedures + transfer	1602
6	Age + Sex + LOS + total_procedures+transfer + Operational Specialty Group	1680
7	Age + Sex + LOS + total_procedures + transfer + Discharge Sub Specialty Name	1720
8	Age + Sex + LOS+ total_procedures + transfer + Source Of Admission	1750

As discussed in Section 4.2.2, forward selection using AIC was chosen to find the champion model. Model 0 was derived using no variables in the model which produced a high AIC of 2100. Next, each variable is tested as it is added to the model. The variable that produces the lowest AIC is chosen to be added next. For example, the addition of age first lead to the largest reduction in AIC compared to other variables. Repeating the process indicated that Model 5 was the optimal model, as it has the lowest AIC at 1602. Further additions did not improve model fit.(see Table 4)

Source Of Admission, Discharge Sub-Specialty Name, and Operational Specialty Group have not been included in the model due to the fact that the Akaike Information Criterion (AIC) penalises variables with numerous levels. For instance, incorporating Operational Specialty Group, which has 8 levels and considering the 11 levels of the independent variable, would lead to 88 additional parameter estimates.

$$Relative\ Risk\ Ratio = e^{coef}$$

For ease of interpretation, the coefficient outputs will be transformed using the relative risk ratio seen above. A relative risk ratio (RRR) greater than 1 means that as the value of a variable increases, the likelihood of the outcome occurring in the comparison group compared to the referent group also increases. This means that the outcome is more likely to occur in the comparison group. On the other hand, an RRR less than 1 indicates that the risk of the outcome in the comparison group decreases as the variable increases, making it more likely to occur in the referent group. Generally speaking, if the RRR is less than 1, the outcome is more probable to occur in the referent group.

Table 5: Multinomial Logistic Regression Output

Intercept & X Variable	Estimate	Std Error	z-value	Pr(> |z|)	Relative Risk Ratio
(Intercept) - Nervous System	-0.50	0.63	-0.78	0.43	0.61
(Intercept) - Respiratory System	-0.59	0.59	-0.99	0.32	0.55
(Intercept) - Circulatory System	-1.76	0.74	-2.36	0.02	0.17
(Intercept) - Digestive System	-0.56	0.73	-0.77	0.44	0.57
(Intercept) - Hepatobiliary System	-1.21	0.92	-1.31	0.19	0.30
(Intercept) - Skin, Subcutaneous Tissue &Breast	-2.47	1.10	-2.24	0.02	0.09
(Intercept) - Female Reproductive System	-1.76	0.97	-1.82	0.07	0.17
(Intercept) - Conditions Originating in Perinatal Period	5.11	0.97	5.26	<0.001	166
(Intercept) - Immunological Disorders	1.78	0.83	2.16	0.03	5.93
(Intercept) - Mental Disease	0.17	0.61	0.28	0.78	1.19
Age - Nervous System	0.62	0.91	0.69	0.49	1.86
Age - Respiratory System	1.12	0.84	1.33	0.18	3.06
Age - Circulatory System	3.67	1.01	3.62	<0.001	39.25
Age - Digestive System	0.35	1.09	0.32	0.75	1.42
Age - Hepatobiliary System	1.39	1.41	0.99	0.32	4.01
Age - Skin, Subcutaneous Tissue &Breast	1.85	1.47	1.26	0.21	6.35
Age - Female Reproductive System	2.79	1.35	2.07	0.04	16.28
Age - Conditions Originating in Perinatal	-18.41	3.01	-6.12	<0.001	<0.001
Age - Immunological Disorders	-3.70	1.46	-2.53	0.01	0.03
Age - Mental Disease	-0.99	0.94	-1.06	0.29	0.37
Male - Nervous System	0.08	0.38	0.22	0.82	1.08
Male - Respiratory System	0.79	0.35	2.27	0.02	2.20
Male - Circulatory System	1.02	0.40	2.55	0.01	2.77
Male - Digestive System	0.66	0.45	1.48	0.14	1.93
Male - Hepatobiliary System	0.62	0.56	1.09	0.27	1.86
Male - Skin, Subcutaneous Tissue &Breast	0.38	0.60	0.63	0.53	1.46
Male - Female Reproductive System	-17.77	1380.02	-0.01	0.99	<0.001
Male - Conditions Originating in Perinatal	-1.38	0.74	-1.86	0.06	0.25
Male - Immunological Disorders	-0.78	0.65	-1.20	0.23	0.45
Male - Mental Disease	0.14	0.39	0.36	0.72	1.15
Length Of Stay - Nervous System	-1.05	2.50	-0.42	0.67	0.35
Length Of Stay - Respiratory System	-2.19	2.48	-0.88	0.38	0.11
Length Of Stay - Circulatory System	-9.52	4.17	-2.28	0.02	<0.001
Length Of Stay - Digestive System	-0.25	2.75	-0.09	0.93	0.78
Length Of Stay - Hepatobiliary System	-0.29	3.32	-0.09	0.93	0.75
Length Of Stay - Skin, Subcutaneous Tissue &Breast	-9.85	7.00	-1.41	0.16	<0.001
Length Of Stay - Female Reproductive System	-8.57	6.03	-1.42	0.16	<0.001
Length Of Stay - Conditions Originating in Perinatal	-4.47	4.91	-0.91	0.36	0.01
Length Of Stay - Immunological Disorders	-12.64	7.67	-1.65	0.10	<0.001
Total Procedure - Nervous System	1.11	3.18	-3.5	<0.001	3.03
Total Procedure - Respiratory System	-1.20	1.3	-4.2	<0.001	0.30
Total Procedure - Circulatory System	-1.77	5.00	-3.50	<0.001	0.17
Total Procedure - Digestive System	-1.20	3.08	-2.78	<0.001	0.30
Total Procedure - Hepatobiliary System	-1.74	8.04	-2.08	0.04	0.17
Total Procedure - Skin, Subcutaneous Tissue &Breast	2.08	2.87	0.07	0.94	8.00
Total Procedure - Female Reproductive System	-1.16	2.87	-0.40	0.69	0.31
Total Procedure - Conditions Originating in Perinatal	2.32	1.93	-1.20	0.23	10.18
Total Procedure - Conditions Originating in Perinatal	-1.53	2.17	-0.01	0.99	0.22
Total Procedure - Mental Disease	-6.07	2.50	-2.43	0.02	<0.001
Transfer - Nervous System	0.18	0.38	0.46	0.64	1.20
Transfer - Respiratory System	0.08	0.35	0.23	0.82	1.08
Transfer - Circulatory System	-0.72	0.39	-1.83	0.07	0.49
Transfer - Digestive System	-1.02	0.45	-2.25	0.02	0.36
Transfer - Hepatobiliary System	-3.22	1.06	-3.03	<0.001	0.04
Transfer - Skin, Subcutaneous Tissue &Breast	-0.06	0.61	-0.09	0.93	0.94
Transfer - Female Reproductive System	-0.60	0.52	-1.15	0.25	0.55
Transfer - Conditions Originating in Perinatal	-2.41	0.74	-3.24	<0.001	0.09
Transfer - Immunological Disorders	-0.71	0.57	-1.26	0.21	0.49
Transfer - Mental Disease	-0.41	0.38	-1.06	0.29	0.66

Table 5 provides an overview of the multinomial output. Here, the focus is on the most significant coefficients.

- For Nervous System Disease relative to Muscular Skeletal Disease the total procedures coefficient is significant at the 1% level. This is the only significant variable in relation to Nervous system disease. It means that if a patient is to increase total procedures by a single unit we would expect the relative risk of suffering from Nervous System Disease to Muscular Skeletal Disease to increase by a factor of three given that all other variables remain constant.
- For Circulatory System Disease relative to Muscular Skeletal Disease, age, gender, length of stay and total procedures coefficients are significant at the 5% level. The most significant variable is age. If a patient is to increase age by a single unit we would expect the relative risk for suffering from Circulatory System Disease relative to Muscular Skeletal Disease to increase by a factor of 39.2 given that all other variables remain constant.
- For Conditions Originating in Perinatal relative to Muscular Skeletal Disease, age and transfer are the only significant coefficients at the 5% level. The most significant variable is age. If a patient is to increase age by a single unit we would expect the relative risk for suffering from Conditions Originating in Perinatal to Muscular Skeletal Disease to decrease by a factor of $1.52e^{-8}$ given that all other variables remain constant.
- For Respiratory System Disease relative to Muscular Skeletal Disease, gender and total procedures are the only significant coefficients at the 5% level. The most significant variable is total procedures. If a patient is to increase total procedures by a single unit we would expect the risk for suffering from Respiratory System Disease relative to Muscular Skeletal Disease to decrease by a factor of 0.30 given that all other variables remain constant.

Table 6: Test Performance

Model	Test Accuracy
5	29%

Looking at Table 6 the accuracy of Model 5 for the multinomial logistic regression on the test data set is 29%.

5.2.3 Random Forest

The R package 'randomForest' was used to deploy the model. It uses the greedy algorithm to optimise the branching of each tree. As explained in Section 4.2.4, each tree containS a random subset of our data rows and explanatory variables p. The initial configuration Model 1 is a simple random forest consisting of 50 decision trees and all variables are considered for each split. The other models were built using an increasing number of trees and the number of variables considered at each split were varied. The ensemble of trees helped to reduce the model variability.

Out-of-bag is a method for estimating the performance of the model. Each decision tree is built on a bootstrap sample of the training data. This means that some samples are left out and not used to build the tree. These samples are known as the out-of-bag samples. The out-of-bag error is the average random forest error and is calculated on out-of-bag samples taken for validation. This allows the Random Forest classifier to be fit and validated whilst being trained.

Table 7: Model Performance

Model	Number of Trees	Number of variables randomly sampled	Out of Bag Error
1	50	8	53.6%
2	500	5	47.3%
3	1500	3	46.7%
4	2000	2	46.4%
5	3000	5	49.2%
6	5000	3	47.1%

Using various configurations of the random forest and evaluating the out-of-bag sample error the best performing configuration was achieved by using 2000 trees and randomly sampling two variables at each split. This led to an out-of-bag error rate of 46.4%. The better performing models have a higher number of trees and a smaller number of random variables at each split.

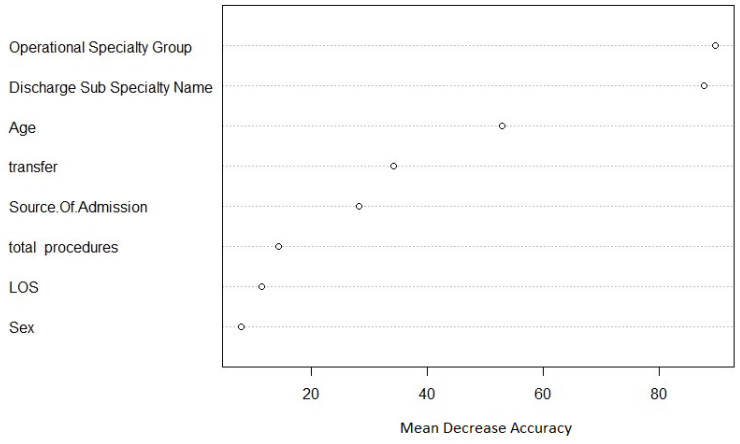

Figure 27: Variable Importance Model 4

Figure 27 is the variable importance of the best performing random forest Model 4 and shows the effect of removing specific variables from the model and on mean accuracy in relation to the out-

of-bag sample. The Operational Specialty group is the most impactful variable when predicting diagnosis group. We see the largest decline in accuracy when the Operational Specialty group is removed. Gender is the least important variable.

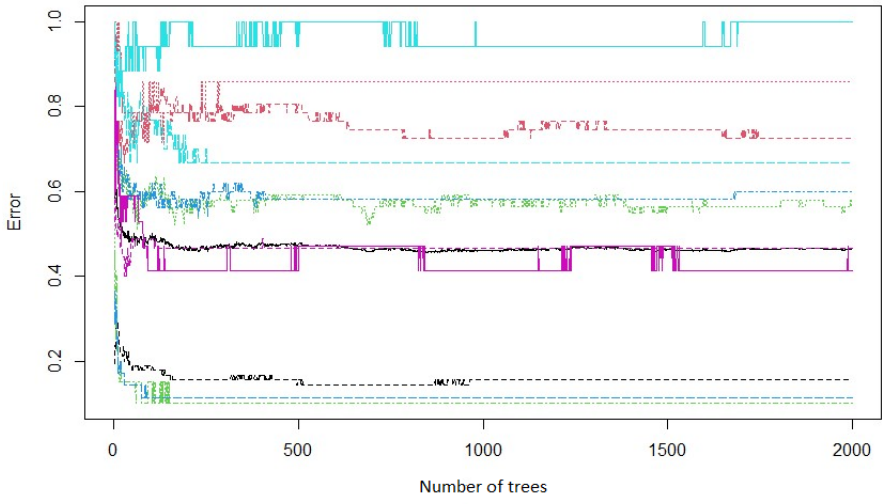

Figure 28: OOB error Model 4

Figure 28 shows various colours lines that represent the various diagnosis categories and changes in the out-of-bag error over the trees built in Model 4. The model will not perform better by adding more trees as the majority of the diagnosis categories' error rates have plateaued.

Table 8: Test Performance

Model	Test Accuracy
4	57%

Table 8 shows the performance of Model 4 on the test data is 57% which is better than both the neural network and multinomial logistic regression.

5.2.4 Random Forest with SMOTE

SMOTE can help to reduce the bias towards the majority class that can occur when using imbalanced datasets. The majority-to-minority frequencies are sampled up to have the same frequency as the most occurring level. Two datasets were created using SMOTE. The first applied SMOTE to the full dataset and then created a training and testing set using a 75% and 25% split, respectfully. The second dataset was created by splitting the original data set using a 75% and 25% split and then applying SMOTE to the training set only, leaving the test set unchanged.

Table 9: Test Performance

SMOTE Test Accuracy	**Actual Test Accuracy**
77%	30%

Table 9 shows that the test accuracy of applying SMOTE to the entire data set is 77%. Note this accuracy is based on a SMOTE generated test set. The actual test set performance (when SMOTE is not applied) was substantially lower at 30%. It can be inferred that the model is over-fitting to the synthetic data generated using SMOTE.

5.2.5 Ensemble

The last model was built using an ensemble of the two best performing models: the Neural Network and Random Forest. A third model was added to account for the minority classes. It will also create a third vote that will decide the overall winning class when there is a deadlock between the other two models. Random forest with SMOTE was chosen as the third model given it's good performance in dealing with minority classes.

Table 10: Majority Vote

Random Forest	Random Forest SMOTE	Neural Network
DISEASES MUSCULOSKELETAL SYSTEM	DISEASES MUSCULOSKELETAL SYSTEM	DISEASES MUSCULOSKELETAL SYSTEM
DISEASES OF THE RESPIRATORY SYSTEM	DISEASES OF THE NERVOUS SYSTEM	DISEASES MUSCULOSKELETAL SYSTEM
DISEASES OF THE SKIN	DISEASES OF THE SKIN	DISEASES OF THE SKIN
DISEASES OF THE RESPIRATORY SYSTEM	DISEASES OF THE RESPIRATORY SYSTEM	DISEASES OF THE NERVOUS SYSTEM

Majority voting was used to decide the predicted outcome. Looking at row one in table 10, the prediction would be Diseases of Musculoskeletal System as there are three votes for this case. In row two there is a three way split between the votes, hence this will be removed. In the fourth row there are two votes for Diseases of the Respiratory System by the Random Forest and Random Forest SMOTE, there is only one vote for Diseases of the Nervous System hence the predicted outcome would be Diseases of the Respiratory System.

Table 11: Test Performance

Model	**Test Accuracy**
Neural Network + Random Forest + Random Forest SMOTE	26%

The test accuracy is extremely low at 26% which is lower than the individual models. Cases with split votes were removed and could be the cause of weak performance. This is one drawback of using an ensemble.

6 Discussion

The aim of this dissertation was twofold. The first goal was to introduce Diagnosis Related Groupings (DRG) into the public sector and then explore the appropriateness of using DRG weights by examining differences between case mix index and CMAF across hospitals. The second objective was to address issues of errors in diagnosis coding by healthcare staff through the construction of a model that utilizes demographic and medical records to predict the correct diagnosis.

The CHAI partnered with the Western Cape government and was able to source Medicare DRG weights. These weights have been assigned to cases for all public hospitals in the province. A question arose regarding whether these weights are suitable for the public sector, given that they are based on private health sector costs and resourcing. This dissertation aimed to answer this question by using Length of Stay (LOS) as a proxy for resource utilization in the public sector (Fetter et al., 1980). Two measures were created. The first was CMI, which uses DRG Medicare weights to measure resource utilization at respective hospitals.

The second measure was CMAF, where this dissertation, limited to using LOS as a proxy for resource utilization, as public hospitals do not have financial records on a per case level. The respective measurements were calculated for every hospital, and the relationship between the two measures was assessed. The results showed a weak positive relationship between the two measures, with a correlation of 0.3. However, there are limitations in using LOS to calculate CMAF, which could have contributed to the weak relationship. Comparing these measures is important as there is no other way to validate whether the DRG Medicare weights are appropriate in the public health sector.

There have been major medical coding errors in the public sector, as seen in the onsite audit report quoted in Section 1.1. To address this issue, this dissertation looked into error detection models to mitigate these coding risks. Five models were built: a neural network, multinomial logistic regression, random forest, random forest using SMOTE, and an ensemble of three models using majority voting. These were built using a small amount of audited data in which some diagnosis categories had to be dropped due to low volume profiles.

By increasing the number of samples in the minority class, SMOTE helped to improve the performance of machine learning models on these minority classes. However, the overall accuracy decreased. SMOTE helped to reduce the bias towards the majority class that occurs when working with imbalanced datasets such as this. When used in this dissertation, the model tended to overfit to the synthetic data generated by SMOTE, and overall performance on the actual dataset was much lower than other models at 30%. The models built have a low accuracy rate on the test data, with the best performing model, the random forest, having a test accuracy of just 57%. However, this is better than the audit report mentioned in Section 1.1, which reported an accuracy of 32% of primary diagnosis codes correctly classified.

For neural networks, the model was able to identify certain disease categories with high accuracy. These include Diseases & Disorders of the Female Reproductive System at 100%, Diseases & Disorders of the Respiratory System at 85%, and a few others. Some diagnoses are inherently difficult to predict. For example, immunological disorders. These patients can be treated in multiple specialty groups, and the diagnosis can affect ages from neonates to geriatrics. This means that there are no specific factors/characteristics to distinguish this diagnosis from others. The accuracy for this specific diagnosis was 0% for this model. Given that there are 11 diagnosis categories to predict, the random probability of assigning a correct one would be 1/11 or 9%. The best performing model, random forest, had an accuracy of 57%, illustrating the power of this

predictive model.

The explanatory analysis found large differences in diagnosis groups between metro and rural areas, genders, as well as adults and pediatrics. It was also noted that the average LOS differs greatly between diagnosis groups.

Electronic Health Record (EHR) data has great potential to benefit public health, clinical research, and healthcare administration. Additional approaches not tried to improve the EHR data quality in this dissertation include using a Gradient Boosting Model (GBM) and Support Vector for the prediction of diagnosis.

The biggest limitations for the prediction models were the sample size of the audited dataset to train the model. In other studies like Feng, Y. (2019), natural language processing was used as an ensemble to the machine learning technique to account for the lower accuracy rate for some diagnoses. However, even after the ensemble, some diagnosis categories evaluated in this study were easier to identify, while pneumonia, kidney failure, and respiratory failure were harder to distinguish (Feng, Y. 2019).

This is possibly due to their frequencies being underrepresented in the dataset and the fact that these diagnoses are highly correlated with other diagnoses, making it difficult for the model to distinguish. An interesting observation is that in this dissertation, Diseases & Disorders of the Respiratory System had excellent accuracy at 85%. Natural language processing could have helped account for these cases in this dissertation. Doctor's notes, also known as medical notes or clinical notes, are the written records of a doctor's observations and assessments of a patient's symptoms, medical history, and physical examination findings. These notes are an essential tool for healthcare providers to track a patient's progress, make informed decisions about their care, and communicate important information to other healthcare professionals involved in the patient's treatment. It is much more detailed than the data found in the EHR and can account for complex cases.

Medical conditions can be complex, with multiple symptoms and causes, making it difficult to identify the root cause of the problem even for medical professionals. People can have different responses to the same disease, so predicting the progression and outcomes of a medical condition can be difficult. The EHR medical data is lacking in detail. For example, there are no clinical results present in the health record like blood pressure readings or HbA1c. These are used as an indicator for both hypertension and diabetes. Doctors could also elect to have multiple procedures just to rule out other possibilities which will still feed into the model as different procedures, making it more difficult to predict a diagnosis.

7 Conclusion

This dissertation has three main findings. The first is that there is a weak positive relationship between CMI and CMAF in the public sector. This indicates that the use of DRG weights from Medicare might not be appropriate in the public sector. However, CMAF, which used actual length of stay instead of costs per case as a proxy for resource utilization, might not be appropriate. The big limitation was that the public sector data does not track admission costs on a case level. There are currently projects in place to track costs on a cost center level which can be applied to CMAF for future work.

Next, it was shown that the best prediction model for diagnosis was a random forest with an accuracy of 57% on the unseen test data set. This is a reasonable accuracy rate given the number of diagnosis categories and the small audit data size that was used in the modeling process. In addition, if we consider the current error rate by staff of 68% as seen in the audit report in Section 1.1, there is definitely a value add. Future work can improve the accuracy of the model by incorporating natural language processing using the currently unavailable doctors' notes. Moreover, as more records are audited over time, the number of audited observations is expected to increase, which would allow all diagnosis categories to be included in the model.

Lastly, through the explanatory analysis, this dissertation identified both qualitative and quantitative relationships in the data that could open up avenues for more research and development. The biggest findings in this section were noting the difference between disease burdens seen in metro versus rural areas, the frequencies of diagnoses, and how the distributions for age and length of stay change for different diagnosis categories.

The predictive model will help identify cases where diagnostic coding errors exist, which is currently a major issue. The findings will also guide decision-makers on the appropriateness of using private sector (Medicare) DRG weights for the public sector.

Appendix A – Additional Tables

Table 12: Case Mix Adjustment Factor Calculation

	Hospital	$\sum_j A_j p_{ij}$	$\sum_j A_j P_j$	CMAF
1	Beaufort West Hosp	9.00	7.67	1.17
2	Brackengate I.C.	13.34	7.67	1.74
3	Brewelskloof Hosp	15.61	7.67	2.04
4	Brooklyn Chest Hosp	16.55	7.67	2.16
5	Caledon Hosp	4.23	7.67	0.55
6	Ceres Hosp	8.15	7.67	1.06
7	Citrusdal Hosp	13.73	7.67	1.79
8	Clanwilliam Hosp	6.25	7.67	0.81
9	CTICC Covid I.C.	13.50	7.67	1.76
10	DP Marais Hosp	16.95	7.67	2.21
11	Eerste River Hosp	9.54	7.67	1.24
12	False Bay Hosp	8.45	7.67	1.10
13	George Hosp	7.13	7.67	0.93
14	Groote Schuur Hosp	6.95	7.67	0.91
15	Harry Comay TB Hosp	14.29	7.67	1.86
16	Helderberg Hosp	7.36	7.67	0.96
17	Hermanus Hosp	6.69	7.67	0.87
18	Karl Bremer Hosp	6.85	7.67	0.89
19	Khayelitsha Hosp	7.41	7.67	0.97
20	Knysna Hosp	7.71	7.67	1.01
21	Laingsburg Hosp	6.58	7.67	0.86
22	Lapa Munnik Hosp	6.77	7.67	0.88
23	Lentegeur Hosp	29.21	7.67	3.81
24	Malmesbury ID Hosp	4.86	7.67	0.63
25	Mitchells Plain District Hosp	5.36	7.67	0.70
26	Mitchells Plain Hosp I.C.	13.46	7.67	1.76
27	Montagu Hosp	17.32	7.67	2.26
28	Mossel Bay Hosp	8.47	7.67	1.11
29	Mowbray Maternity Hosp	10.46	7.67	1.36
30	Murraysburg Hosp	7.48	7.67	0.98
31	New Somerset Hosp	7.63	7.67	1.00
32	Otto du Plessis Hosp	6.68	7.67	0.87
33	Oudtshoorn Hosp	7.06	7.67	0.92
34	Paarl Hosp	6.22	7.67	0.81
35	Prince Albert Hosp	9.89	7.67	1.29
36	Radie Kotze Hosp	28.56	7.67	3.73
37	Red Cross Children's Hosp	5.57	7.67	0.73
38	Riversdale Hosp	6.47	7.67	0.84
39	Robertson Hosp	7.72	7.67	1.01
40	Sonstraal TB Hosp	13.84	7.67	1.81
41	Stellenbosch Hosp	7.69	7.67	1.00
42	Stikland Hosp	27.80	7.67	3.63
43	Swartland Hosp	7.19	7.67	0.94
44	Swellendam Hosp	6.88	7.67	0.90
45	Tygerberg Hosp	6.07	7.67	0.79
46	Uniondale Hosp	5.46	7.67	0.71
47	Valkenberg Hosp	28.94	7.67	3.78
48	Victoria Hosp	6.93	7.67	0.90
49	Vredenburg Hosp	7.45	7.67	0.97
50	Vredendal Hosp	8.05	7.67	1.05
51	Wesfleur Hosp	5.40	7.67	0.70
52	Western Cape Rehab Centre	40.89	7.67	5.33
53	Worcester Hosp	6.81	7.67	0.89

Table 13: Case Mix Adjustment Factor and Case Mix Index

	Hospital	CMAF	CMI
1	Beaufort West Hosp	1.17	1.07
2	Brackengate I.C.	1.74	0.89
3	Brewelskloof Hosp	2.04	1.37
4	Brooklyn Chest Hosp	2.16	1.34
5	Caledon Hosp	0.55	0.49
6	Ceres Hosp	1.06	0.97
7	Citrusdal Hosp	1.79	0.87
8	Clanwilliam Hosp	0.81	0.95
9	CTICC Covid I.C.	1.76	0.87
10	DP Marais Hosp	2.21	1.05
11	Eerste River Hosp	1.24	1.07
12	False Bay Hosp	1.1	0.95
13	George Hosp	0.93	0.96
14	Groote Schuur Hosp	0.91	1.27
15	Harry Comay TB Hosp	1.86	1.11
16	Helderberg Hosp	0.96	0.76
17	Hermanus Hosp	0.87	0.74
18	Karl Bremer Hosp	0.89	1.16
19	Khayelitsha Hosp	0.97	0.49
20	Knysna Hosp	1.01	1.18
21	Laingsburg Hosp	0.86	0.87
22	Lapa Munnik Hosp	0.88	1.15
23	Lentegeur Hosp	3.81	1.15
24	Malmesbury ID Hosp	0.63	1.11
25	Mitchells Plain District Hosp	0.7	0.96
26	Mitchells Plain Hosp I.C.	1.76	0.87
27	Montagu Hosp	2.26	0.93
28	Mossel Bay Hosp	1.11	0.83
29	Mowbray Maternity Hosp	1.36	2.12
30	Murraysburg Hosp	0.98	1.14
31	New Somerset Hosp	1	1.19
32	Otto du Plessis Hosp	0.87	0.88
33	Oudtshoorn Hosp	0.92	1.11
34	Paarl Hosp	0.81	0.52
35	Prince Albert Hosp	1.29	1.06
36	Radie Kotze Hosp	3.73	1.18
37	Red Cross Children's Hosp	0.73	1.04
38	Riversdale Hosp	0.84	1.1
39	Robertson Hosp	1.01	1.07
40	Sonstraal TB Hosp	1.81	1.04
41	Stellenbosch Hosp	1	1.06
42	Stikland Hosp	3.63	1.11
43	Swartland Hosp	0.94	1
44	Swellendam Hosp	0.9	0.94
45	Tygerberg Hosp	0.79	1.13
46	Uniondale Hosp	0.71	1.05
47	Valkenberg Hosp	3.78	1.15
48	Victoria Hosp	0.9	1.15
49	Vredenburg Hosp	0.97	0.77
50	Vredendal Hosp	1.05	0.92
51	Wesfleur Hosp	0.7	1.18
52	Western Cape Rehab Centre	5.33	1.32
53	Worcester Hosp	0.89	0.73

Appendix B – Transformed variables

```
> summary(data[,1:40])
      Age              Sex              LOS          total_procedures     transfer      Source.Of.AdmissionCentral.Hospital
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000000
 1st Qu.:0.3083   1st Qu.:0.0000   1st Qu.:0.01875   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000000
 Median :0.4889   Median :0.0000   Median :0.03750   Median :0.00000   Median :1.0000   Median :0.000000
 Mean   :0.4729   Mean   :0.4302   Mean   :0.06052   Mean   :0.03684   Mean   :0.5209   Mean   :0.009777
 3rd Qu.:0.6444   3rd Qu.:1.0000   3rd Qu.:0.07500   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.000000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.000000
 Source.Of.AdmissionComm.Health.Centre Source.Of.AdmissionDistrict.Hospital Source.Of.AdmissionHome Source.Of.AdmissionMat.Obs.Unit.MOU.
 Min.   :0.00000                       Min.   :0.00000                      Min.   :0.0000          Min.   :0.000000
 1st Qu.:0.00000                       1st Qu.:0.00000                      1st Qu.:0.0000          1st Qu.:0.000000
 Median :0.00000                       Median :0.00000                      Median :0.0000          Median :0.000000
 Mean   :0.02933                       Mean   :0.06285                      Mean   :0.4525          Mean   :0.002793
 3rd Qu.:0.00000                       3rd Qu.:0.00000                      3rd Qu.:1.0000          3rd Qu.:0.000000
 Max.   :1.00000                       Max.   :1.00000                      Max.   :1.0000          Max.   :1.000000
 Source.Of.AdmissionMaternity.hospital Source.Of.AdmissionOld.Age.Home Source.Of.AdmissionOPD.SG.WARD...Inter.
 Min.   :0.000000                      Min.   :0.000000               Min.   :0.000
 1st Qu.:0.000000                      1st Qu.:0.000000               1st Qu.:0.000
 Median :0.000000                      Median :0.000000               Median :0.000
 Mean   :0.005587                      Mean   :0.002793               Mean   :0.405
 3rd Qu.:0.000000                      3rd Qu.:0.000000               3rd Qu.:1.000
 Max.   :1.000000                      Max.   :1.000000               Max.   :1.000
 Source.Of.AdmissionPrivate.Institution Source.Of.AdmissionPsychiatric.Hospital Source.Of.AdmissionRegional.Hospital
 Min.   :0.000000                       Min.   :0.000000                        Min.   :0.00000
 1st Qu.:0.000000                       1st Qu.:0.000000                        1st Qu.:0.00000
 Median :0.000000                       Median :0.000000                        Median :0.00000
 Mean   :0.002793                       Mean   :0.001397                        Mean   :0.02235
 3rd Qu.:0.000000                       3rd Qu.:0.000000                        3rd Qu.:0.00000
 Max.   :1.000000                       Max.   :1.000000                        Max.   :1.00000
 Source.Of.AdmissionRehab.facility Source.Of.AdmissionUndefined.for.Takeon Operational.Specialty.GroupGynaecology
 Min.   :0.000000                  Min.   :0.000000                        Min.   :0.00000
 1st Qu.:0.000000                  1st Qu.:0.000000                        1st Qu.:0.00000
 Median :0.000000                  Median :0.000000                        Median :0.00000
 Mean   :0.001397                  Mean   :0.001397                        Mean   :0.08939
 3rd Qu.:0.000000                  3rd Qu.:0.000000                        3rd Qu.:0.00000
 Max.   :1.000000                  Max.   :1.000000                        Max.   :1.00000
 Operational.Specialty.GroupMaternity Operational.Specialty.GroupMedicine Operational.Specialty.GroupOrthopaedics
 Min.   :0.000000                     Min.   :0.0000                      Min.   :0.0000
 1st Qu.:0.000000                     1st Qu.:0.0000                      1st Qu.:0.0000
 Median :0.000000                     Median :1.0000                      Median :0.0000
 Mean   :0.005587                     Mean   :0.5126                      Mean   :0.1662
 3rd Qu.:0.000000                     3rd Qu.:1.0000                      3rd Qu.:0.0000
 Max.   :1.000000                     Max.   :1.0000                      Max.   :1.0000
 Operational.Specialty.GroupPaediatrics Operational.Specialty.GroupPsychiatry Operational.Specialty.GroupSurgery
 Min.   :0.00000                        Min.   :0.00000                       Min.   :0.00000
 1st Qu.:0.00000                        1st Qu.:0.00000                       1st Qu.:0.00000
 Median :0.00000                        Median :0.00000                       Median :0.00000
 Mean   :0.07961                        Mean   :0.05587                       Mean   :0.09078
 3rd Qu.:0.00000                        3rd Qu.:0.00000                       3rd Qu.:0.00000
 Max.   :1.00000                        Max.   :1.00000                       Max.   :1.00000
```

Figure 29: Independent scaled variables

References

Akaike, H., 1974. A new look at the statistical model identification, 19(6), pp. 716-723.

Chawla, N.V., 2010. Data Mining for Imbalanced Datasets: An Overview, pp.1-289.

Department of Health., 2022. National Health Insurance in South Africa: Progress and Challenges.

Dreyer, K.A., 2013. The evaluation of case-mix adjusted efficiency scores the case of the South African private hospital industry (Master's thesis, University of Cape Town).

Feng, Y., 2019. Identification of Medical Coding Errors and Evaluation of Representation Methods for Clinical Notes Using Machine Learning (Doctoral dissertation, Ohio University).

Fetter, R.B., Shin, Y., Freeman, J.L., Averill, R.F. and Thompson, J.D., 1980. Case mix definition by diagnosis-related groups. Medical care, 18(2), pp.1-53.

Fetter, R.B., 1991. Diagnosis related groups: understanding hospital performance. Interfaces, 21(1), pp.6-26.

Goldfield, N., 2010. The evolution of diagnosis-related groups (DRGs): from its beginnings in case-mix and resource use theory, to its implementation for payment and now for its current utilization for quality within and outside the hospital. Quality Management in Healthcare, 19(1), pp.3-16.

Gupta, R., K. Reddy, 2017. Challenges to healthcare in South Africa.

Ho, T.K., 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 20(3) , pp. 278–282.

J. Peyhardi, C. Trottier, Y. Guédon 2015. A new specification of generalized linear models for categorical responses, Biometrika, 102(4) , pp.889–906.

Jones, A., 2018. Patient outcomes in public versus private healthcare settings in South Africa.

Johnson, M., 2020. Assessing the effectiveness of the National Health Insurance program.

Karnon, J., McIntosh, A., Dean, J., Bath, P., Hutchinson, A., Oakley, J., Thomas, N., Pratt, P., Freeman-Parry, L., Karsh, B.T. and Gandhi, T., 2008. Modelling the expected net benefits of interventions to reduce the burden of medication errors. Journal of health services research & policy, 13(2), pp.85-91.

Mabaso, M., Tshabalala, M., 2024. Future prospects for the South African healthcare system.

Naidoo, S., et al. 2019. Challenges in the South African public healthcare system.

Phillips, S., Rond, P.C., Kelly, S.M. and Swartz, P.D., 1996. The failure of triage criteria to identify geriatric patients with trauma: results from the Florida Trauma Triage Study. Journal of Trauma and Acute Care Surgery, 40(2), pp.278-283.

Shahraz, S., 2014. Accuracy of Medical Coding Algorithms to Identify Complex Conditions in United States Hospitals: The Case of Sepsis. Brandeis University, The Heller School for Social Policy and Management.

Smith, J., 2018. Overview of the dual healthcare system in South Africa.

Smith, K., et al. 2020. Private healthcare and its impact on healthcare equity in South Africa.

Souza, J., Pimenta, D., Caballero, I. and Freitas, A., 2020. Measuring data credibility and medical coding: a case study using a nationwide Portuguese inpatient database. Software Quality Journal, 28(3), pp.1043-1061.

State of Florida Agency for Health Care Administration, 1996 Guide to Hospitals in Florida, Tallahassee, Florida, 1996.

Wadee, H., Gilson, L., Thiede, M., Okorafor, O. and McIntyre, D., 2003. Health care inequity in South Africa and the public-private mix.

Western Cape Government: Health ., 2017. On-site audit Clinical Coding Report. rep. Western Cape: New Somerset Hospital, pp. 1–35.

White, C., Black, D., 2021. Cost-effectiveness analysis of public and private healthcare in South Africa.

White, D., 2018. Performance Measurement and Accountability: Private Sector Best Practices.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80–83.

Zuckerman, S., Hadley, J. and Iezzoni, L., 1994. Measuring hospital efficiency with frontier cost functions. Journal of health economics, 13(3), pp.255-280.