DETECTION OF METABOLIC DISORDERS FOR AFRICAN WOMEN IN A RURAL SOUTH AFRICAN SETTING: A CASE OF GA-DIKGALE IN LIMPOPO PROVINCE, SOUTH AFRICA

by

MMALEHU FRANCIN PHAHO

[redacted]

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Prof. D Maposa

CO-SUPERVISOR: Dr. T.B Darikwa

AUGUST 2024

Declaration

I, **Mmalehu Francin Phaho**, the undersigned, hereby declare that the dissertation submitted to the University of Limpopo, for the Master of Science degree in Statistics has not been previously submitted by me for a degree at this or any university; that it is my work in design and execution and that all material contained herein has been duly acknowledged and referenced accordingly.

Signature:.....*MFP*.....Date:... August 2024.....

Phaho, M F.

Abstract

The waist circumference cut-off point for diagnosing metabolic syndrome in Sub-Saharan Africa is based on standards established for European populations. The purpose of this study was to determine the prevalence of metabolic syndrome and other related disorders and to determine the waist circumference cut-off point that effectively discriminates between African women with and without metabolic syndrome. Initially, the study participants with metabolic syndrome were identified using the National Cholesterol Education Program - Third Adult Treatment Panel criteria, which was subsequently adapted to the International Diabetes Federation definition. According to the National Cholesterol Education Program - Third Adult Treatment Panel definition, metabolic syndrome is present if at least three of the following criteria are met: Triglycerides $\geq$1.7 mmol/L, High-density lipoprotein cholesterol $<$1.29 mmol/L, Glucose $\geq$5.6 mmol/L, Systolic Blood Pressure $\geq$ 130 mmHg, or Diastolic Blood Pressure $\geq$ 85 mmHg. The prevalence of metabolic syndrome and obesity (body mass index $\geq$30 kg) were 19% and 30%, respectively.

The optimal waist circumference for diagnosing metabolic syndrome was obtained using receiver operating characteristic curve analysis and was found to be 88 cm. Machine learning methods, including logistic regression, linear discriminant analysis and random forest were employed to further validate the cut-off point. The 88 cm cut-off point demonstrated superior performance compared to the European 80 cm cut-off pint, based on prediction accuracy, specificity and positive predictive value. The study highlights how important

it is to have population-specific cut-off points for correctly diagnosing metabolic syndrome in order to reduce the risk of misdiagnosis and related complications. The findings advocate for using the 88 cm cut-off point, which differs with the recommended cut-off point of 80 cm. This is a as a quick and cost-effective measure for identifying obesity, potentially improving public health interventions for African populations.

Keywords: obesity, waist circumference, metabolic syndrome, machine learning.

Dedication

This work is dedicated to my mother, Kgagogelo Mercy Phaho, my daughter, Silindile Consia Phaho and my son, Ditheto Phaho.

Acknowledgments

I thank God the almighty for giving me strength and guidance especially when I was about to give up. Special thanks go to my supervisor, Professor Daniel Maposa and my co-supervisor, Doctor Timotheus Brain Darikwa for the support and guidance throughout the process. To all the colleagues who opened their doors for me when I needed help, I trully appreciate. I am also grateful for the University of Limpopo for granting me the opprtunity to further my studies. I would like to extend my gratitude to my family and friends for their continuous support, encouragement and believing in me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and background

1.1 Introduction

The concept of metabolic disorder has been in existence for almost 10 decades. It was first defined by a Swedish physician in the 1920s as a cluster of metabolic abnormalities such as hypertension, hyperglycemia and hyperuricemia (Kylin, 1923). Two decades later, Vague defined it as the upper body adiposity mostly associated with metabolic abnormalities such as diabetes and cardiovascular diseases (CVDs) (Vague, 1947). Four decades later, Reaven came up with the term "syndrome X" that also defines a collection of metabolic abnormalities which lead to CVDs (Reaven, 1988). However, the formalised and widely accepted definitions for diagnosing metabolic syndrome (MS) were not established.

According to the latest definition by the International Diabetes Federation (IDF), obesity is the major component of MS (Alberti et al., 2006). Obesity increases the risk of CVDs (Grundy, 2004b). It is among the leading causes

of deaths globally, with at least 2.8 million people dying each year. The global prevalence of obesity has been increasing considerably since 1980 (Seidell and Halberstadt, 2015) and has doubled in more than 70 countries (Chooi et al., 2019; Collaborators, 2017). The IDF definition further includes gender and ethnic specific cut-off points for waist circumference. However, cut-off values for African populations have not been defined due to lack of data.

The Sub-Saharan Africa (SSA) and other low- or middle-income countries are facing a serious public health issue with obesity and its associated diseases (Lopez et al., 2006; Seidell and Halberstadt, 2015). South Africa is one of the countries in the SSA. (Ajayi et al., 2016) conducted a study to determining the prevalence of obesity for four countries in the SSA: Uganda, Nigeria, Tanzania and South Africa. The prevalence of obesity in rural Uganda, peri-urban Uganda, Nigeria, Tanzania and South Africa were 10%, 14%, 31%, 40% and 54% respectively, with South Africa having the highest prevalence. South Africa was among the top five African countries with the highest prevalence of overweight and obesity in 2010 (Agyemang et al., 2015). The increasing prevalence of obesity leads to an increasing prevalence of MS (Grundy, 2008).

There are disparities in obesity and MS rates based on ethnicity, age, gender, and socioeconomic status (Kanter and Caballero, 2012; Anekwe et al., 2020). The prevalence of obesity in the United States between 2011 and 2014 was higher in women (Ogden et al., 2015). In the African nations, epidemiological studies have also revealed a clear gender difference in the incidence of obesity with levels substantially greater in females than in males (Van Der Merwe and Pepper, 2006; Abubakari et al., 2008). In South Africa, the 1998 National Demographic Health Survey (NDHS) reported a high prevalence of obesity in women than men, with Black women being the most affected than the White and Indian women (Puoane et al., 2002).

1.2 Background

MS is not rare in African countries, especially in developing or middle-income countries like South Africa (Sekgala et al., 2018a). South Africa is one of the countries in economic transition from undeveloped to developed, which is affected by the increasing rate of obesity (Kruger et al., 2005). Obesity is closely related to numerous health conditions, including but not limited to hypertension, coronary heart diseases, diabetes, stroke and cancer (Jebb, 2004; Mehrzad, 2020). Obesity and its related conditions shorten the quality of life and lead to premature death, this is a concern to the economical and public health system. The determinants of obesity range from sociodemographic factors, socioeconomic status and lifestyle factors (Agyemang et al., 2015). Obesity can be considered a chronic relapsing and a progressive disease (Bray et al., 2017). The global epidemic of obesity affects about 500 million people worldwide and it is anticipated to increase to one billion by 2030 (Micklesfield et al., 2013).

MS is a cluster of metabolic abnormalities associated with the risk of CVDs and Type 2 diabetes (Cornier et al., 2008; Han and Lean, 2015). These clustering of different conditions led to a development of different criteria of diagnosing MS by several expert groups and institutions (Fezeu et al., 2007). However, it has only been able to compare the prevalence among populations around the world after some consensus on definitions of MS was reached (Cameron et al., 2004). The four widely recognised definitions of MS include the World Health Organisation (WHO) in 1999 (WHO, 1999), the European group for the Study of Insulin Resistance (EGIR) also in 1999 (Balkau and Charles, 1999), the National Cholesterol Education Program - Adult Treatment Panel III (NCEP-ATP III) in 2001 (NCEP, 2001) and the IDF in 2006 (Alberti et al., 2006).

1.3 Problem statement

The prevalence of MS is increasing worldwide (Eckel et al., 2005; Grundy, 2008). The prevalence estimates of MS depend on the definition used (Cornier et al., 2008). Initially, there were three most widely accepted definitions of MS and each of these definitions used a different set of criteria. The use of these definitions resulted in different prevalence rates which makes it difficult to compare the prevalence of MS for diverse populations. In response to this controversy, the IDF proposed a unified definition that can be used in both epidemiology and clinical practice worldwide (Alberti et al., 2006). This unified definition emphasised obesity as a major component for diagnosing MS. Furthermore, obesity is measured by waist circumference that is gender and ethnic specific. The specific cut-off values for various ethnic groups were given. However, the European cut-off points were recommended for African populations and other population groups where specific data are not yet available.

According to Grundy (2008), the prevalence of MS is increasing throughout the world. The prevalence of obesity in the SSA is high and South Africa being the most affected (Agyemang et al., 2015). A study by Van Zyl et al. (2012) reported the prevalence of overweight and obesity as: 46%, 48%, 68%, 75% and 85% for rural Uganda, peri-urban Uganda, Nigeria, Tanzania and South Africa, respectively. Obesity is the key determinant for the prevalence of MS in the SSA. (Fezeu et al., 2007). Obesity is accompanied by an increased risk of CVDs and it affects metabolism of lipids and glucose, regulation of blood pressure, thrombotic and fibrinolytic processes, and inflammatory reactions (Grundy, 2004a).

The European cut-off points for diagnosing MS is not appropriate for the SSA populations (Crowther and Norris, 2012). There are several studies in the SSA populations that found the waist circumference cut-off points that are different from the European based, 94 cm and 80 cm for men and women, respectively.

For example, a study by Prinsloo et al. (2011) in an urban African population found that the waist circumference cut-off points are 90 cm and 98 cm for men and women, respectively. Magalhães et al. (2014) also did a study in Angola and found the waist circumference cut-off points of 87.5 cm and 80.5 cm for men and women, respectively. In a separate study in Ethiopia, Sinaga et al. (2018) revealed waist circumference cut-off points of 83.7 cm and 78 cm for men and women, respectively.

Therefore, this study determines the prevalence of MS and derives the appropriate waist circumference cut-off point for women in the rural community of Limpopo, South Africa.

1.4 Rationale

The rising prevalence of obesity poses a challenge to the public health worldwide. Obesity is the key determinant of MS in the SSA (Fezeu et al., 2007). South Africa is among the top countries with the highest prevalence of obesity. The cost of treating obesity comorbidities pose a challenge to the health care system (Kramer et al., 2013). Epidemiological research has shown that obesity leads to non-communicable diseases such as Type 2 diabetes, CVDs and hypertension (Gbadamosi and Tlou, 2020). Obesity is also associated with an increased risk of long-term adverse health outcomes and a higher chance of dying.

Despite the intervention by the National Department of Health putting measures in place to reduce the rate of obesity in South Africa, the rate is continuously increasing. The prevention strategies were documented in the Strategy for Prevention and Control of Obesity document (2015-2020), with the aim of

preventing and reducing obesity by 10% by 2020 (Health and Africa, 2015) . However, there seems to be limited published evidence about the effectiveness of measures adopted to reduce obesity. Therefore, it is hard to evaluate the effectiveness of the 2015–2020 national obesity strategy because there are no interim monitoring data available (Manafe et al., 2022). According to Owolabi (2017) there is a substantial increase in the prevalence of obesity in South Africa. The prevalence is projected to increase by 47.7% in females and 23.3% in males by 2025 (Van Vollenstee and Van der Merwe, 2021).

Faijer-Westerink et al. (2020) revealed that the prevalence of MS was higher among women than men in SSA. For example, Gebreyes et al. (2018) did a study on the prevalence of MS in Ethiopia and found that the prevalence of MS is 8.6% and 1.8% for females and males, respectively. Another evidence by Van Der Linden et al. (2019) in Ghana found the prevalence of MS to be 25.6% and 7.3% for females and males, respectively. Van Zyl et al. (2012) did a study on the risk-factor profiles for chronic diseases of lifestyle and MS in an urban and rural setting in South Africa and concluded that the rural study population had a higher prevalence of risk factors for MS.

According to Crowther and Norris (2012), the recommended cut-off point of 80 cm for women in diagnosing MS is not appropriate for women in the SSA. In South Africa, Owolabi et al. (2018) did a study in Eastern Cape using the IDF definition and found that the prevalence of MS is 15.6% and 24.8% and the waist circumference cut-off points of 95.25 cm and 89.45 cm for men and women, respectively. In a separate study in Ellisrus, Sekgala et al. (2018b) also used the IDF definition and found that the prevalence of MS is8.6% and 36.8% and the waist circumference cut-off points of 75.07 cm and 82.20 cm for men and women, respectively. Therefore, it is important to find the appropriate waist circumference so that the populations in the SSA can be correctly

diagnosed.

1.5 Aim and objectives

1.5.1 Aim

The aim of this study is to determine the prevalence of obesity and other related metabolic disorders and to derive a suitable waist circumference cut-off point for diagnosing the metabolic syndrome for African women in a rural South African setting.

1.5.2 Objectives

The objectives of the study are to determine the:

1. Prevalence of metabolic disorders among black female respondents.
2. Suitable cut-off point for diagnosis of metabolic syndrome.
3. Prevalence of obesity in the female population based on the new cut-off point and compare with the European standard.
4. Similarities and dissimilarities in classification between different cut-off points.

1.6 Contribution of the study

In determining the prevalence of obesity that is increasing, we will be able to recommend to the Department of Health that in developing strategies that will reduce the rate of obesity, they must also make sure that they are effective and monitored. This will reduce the burden of obesity on the health and economical sectors. Obesity is a key determinant of MS, therefore, an increase in the

prevalence of obesity will result in a parallel increase in the prevalence of MS. Research has shown that the European cut-off point is not appropriate for the SSA population. An appropriate waist circumference cut-off point is needed to correctly estimate the prevalence of MS. The findings will contribute to the data needed to determine the appropriate circumference cut-off point for the SSA populations.

1.7 Structure of the dissertation

This dissertation is divided into five chapters. Chapter 1 gives the background information about the prevalence of MS worldwide. It also provides information about the waist circumference cut-off point for diagnosing MS in SSA, problem statement, rationale, aim and objectives and significance of the study. Chapter 2 is the literature review on the previous related studies that have been conducted. Chapter 3 presents the methodology. Chapter 4 presents the results and discussion and Chapter 5 provides conclusion and recommendations that address the objectives set in Chapter 1.

Chapter 2

Literature review

2.1 Introduction

This chapter outlines the literature on metabolic syndrome (MS) in the Sub-Saharan Africa (SSA) and worlwide. These include various definitions used to estimate the prevalence of MS, the prevalence of obesity, and the optimal cut-off points for diagnosing MS in SSA.

2.2 Definitions of metabolic syndrome

Several organisations and expert groups proposed definitions of MS. However there were three formalised and widely recognised definitions from 1999 to 2001, the World Health Organisation in 1999 (WHO, 1999), the European Group for the study of Insulin Resistance also in 1999 (Balkau and Charles, 1999) and the National Cholesterol Education Program - Third Adult Treatment Panel in 2001 (NCEP, 2001). Then in 2004, the IDF established the

fourth unified definition of MS (Alberti et al., 2006). The definitions by each organisation are outlined in the next subsections.

2.2.1 The World Health Organisation

The World Health Organization (WHO) is an agency of the United Nations responsible for international public health. One of the key fuctions of this orgainsation is disease prevention and control that includeS monitoring, prevention, and control of infectious diseases, such as malaria, COVID-19, and HIV/AIDS and non-communicable diseases,such as diabetes and heart related diseases. MS is one of the non-communicable disease and the first formalised and widely accepted definition of MS was formulated by WHO. They defined MS as glucose intolerance, impaired glucose tolerance (IGT) or diabetes mellitus (DM), and/or insulin resistance together with two or more of the following components: blood pressure: $\geq 140/90$mmHg, triglycerides: $\geq$150 mg/dl (1.7 mmol/l) and/or high-density lipoprotein (HDL) cholesterol: < 35 mg/dl (0.9 mmol/l) in men and $<$39 mg/dl (1.0 mmol/l) in women, obesity (waist/hip ratio (WHR): $>$0.9 in men and $>$0.85 in women and/or body mass index (BMI): $>$30 kg/m^2 and microabuminuria :urinary albumin excretion rate $\geq$20 μg/minute or albumin/creatine ratio $\geq$30 gm/g. Their suggestions for this description were made as a working guideline to be modified in the future, not as an exact definition.

2.2.2 The European Group for the study of Insulin Resistance

After the publication of the WHO definition, the European Group for the study of Insulin Resistance (EGIR) proposed a modification and used it for non-diabetic subjects only. The definition was set to be much simpler to use in epidemiological studies since it does not need a euglycaemic clamp to measure insulin resistance. They also proposed a slight modification of the cut-off points for

hypertension, HDL cholesterol and triglycerides, and altered measures for cut-off points for obesity. The EGIR then defined MS as insulin resistance: 75 percentile of fasting insulin values for non-diabetic population plus two of the following components: fasting plasma glucose: $\geq$110 mg/dl (6.1 mmol/l), but non-diabetic, blood pressure: $\geq$140/90 mmHg or treatment, triglycerides: $>$178 mg/dl (2.0 mmol/l) and/or HDL cholesterol: $<$39 mg/dl (1.0 mmol/l) and obesity: waist circumference $\geq$94 cm for men and $\geq$80 cm for women.

2.2.3 The National Cholesterol Education Program - Third Adult Treatment Panel

In 2001, the National Cholesterol Education Program - Third Adult Treatment Panel(NCEP-ATP III) also designed a definition that was to be manageable as a measurement in clinical practice (NCEP, 2001). It differed with the WHO and EGIR definitions in two ways. Firstly, it does not include insulin resistance as a component. Secondly, it treats all the other components as having equal weights to the diagnosis of MS. The NCEP ATP III defined MS as the presence of three of the following five components: fasting plasma glucose: $\geq$100 mg/dl (5.6 mmol/l), blood pressure: $\geq$130/85 mmHg, triglycerides: $\geq$150 mg/dl (1.7 mmol/l), HDL-C: $<$40 mg/dl (1.03 mmol/l) in men and $<$50 mg/dl (1.29 mmol/l) in women and obesity (waist circumference): $>$102 cm for men and $>$88 cm for women.

2.2.4 The International Diabetes Federation

In 2004, the International Diabetes Federation (IDF) established a unified definition to be used worldwide so that results from different populations can be comparable (Alberti et al., 2006). The IDF defined obesity as the major component of MS. Taking into consideration the fact that different populations, nationalities, and ethnicities have different distributions of norms for body weight

and waist circumferences, the IDF then used ethnic-specific cut-off points (Table 2.1) as a requirement for measuring obesity. The IDF defined MS as the presence of obesity being the major component together with two of the following components: raised plasma triglyceride: $\geq$150 mg/dl (1.7 mmol/l), HDL-C: $<$40 mg/dl (1.03 mmol/l) in men and $<$50 mg/dl (1.29 mmol/l) in women or specific treatment for these lipids abnormalities, fasting plasma glucose: $\geq$100 mg/dl (5.6 mmol/l) or previously diagnosed with Type 2 diabetes, blood pressure: $\geq$130/85 mmHg.

Table 2.1: Ethnic-specific value for waist circumference

Ethnic group	Waist circumference
Europids	≥ 94 cm (male) and ≥ 80 cm (female)
South Asians	≥ 90 cm (male) and ≥ 80 cm (female)
Chinese	≥ 90 cm (male) and ≥ 80 cm (female)
Japanese	≥ 85 cm (male) and ≥ 90 cm (female)
South and Central Americans	Use South Asian data until more specific data are available.
Sub-Saharan Africans	Use European data until more specific data are available.
Eastern Mediterranean and Arabs	Use European data until more specific data are available.

2.3 Prevalence of metabolic syndrome.

The prevalence of MS is increasing globally (Grundy, 2008; Eckel et al., 2005) and according to Grundy (2008) this is due to an increase in the prevalence of obesity. Prevalence estimates of MS differ depending on the definition used and the population being studied (Cornier et al., 2008). The population being studied may be influenced by ethnicity, age, gender, physical activity, socioeconomic factor and others.

Cameron et al. (2007) conducted a study using all widely accepted definitions of MS in estimating the prevalence in Australia. They found a higher prevalence when using the IDF (34.4%), followed by WHO (25.4%), then NCEP-ATP III (24.4%) and the lowest with EGIR (15.6%) for men. However, it was not the case with women. The prevalence estimates of MS was higher when using the IDF (27.2%), followed by NCEP-ATP III (19.9%), then WHO (18.2%) and lowest with EGIR (11.3%). A separate study was conducted by Chien et al. (2008) in China also using all definitions. They found a higher prevalence when using the NCEP-ATP III (18.5%), followed by WHO (18.1%), then IDF (16.2%) and the lowest with EGIR (11.8%) for men. However, it was still not the case with the Chinese women. The prevalence estimates of MS for women was higher when using the WHO (22.4%), followed by IDF (19.0%), then NCEP-ATP III (15.7%) and lowest with EGIR (12.2%).

In SSA, Sumner et al. (2010) conducted a study in Nigeria using two definitions of MS. They found the highest prevalence when using IDF (26%) and the lowest using NCEP-ATP III (19.0%) for men. The prevalence was also higher when using IDF (48%) and lower using NCEP-ATP III (42%) for women. Another study in Ghana used the same definitions as in the Nigerian study and found the highest prevalence when using IDF (15.7%) and the lowest using NCEP-ATP III (5.9%) for men. Furthermore, the prevalence was higher when using IDF (55%) and lower when using NCEP-ATP III (42%) for women (Gyakobo et al., 2012).

In South Africa, a study conducted by Motala et al. (2011) in KwaZulu-Natal using two definitions found the highest prevalence when using IDF (11.6%) and the lowest using NCEP-ATP III (9%) for men. The prevalence was also higher when using IDF (26.3%) and lower using NCEP-ATP III (20.9%) for women. Another study by Erasmus et al. (2012) in Bellville, Cape Town used the same

definitions as in the KwaZulu-Natal study and found the highest prevalance when using IDF (30.6%) and the lowest using NCEP-ATP III (26.8%) for men. Furthermore, the prevalence was higher when using IDF (67.8%) and lower using NCEP-ATP III (62.2%) for women.

2.4 Risk factors of metabolic syndrome.

According to Cassells and Haffner (2006), the risk factors of MS include obesity, high triglycerides level, HDL cholesterol, hypertension, fasting plasma glucose amongst others. However, obesity seems to be the major determinant or risk factor of MS (Fezeu et al., 2007; Carr et al., 2004). Furthermore, the IDF defines obesity as the major component of MS (Alberti et al., 2006).

2.4.1 Hypertension

Hypertension, also called high blood pressure, is defined as a condition whereby long-term force of the blood against the arteries is high enough to cause health problems. It is determined by the amount of blood the heart pumps and the resistance to the blood flow in the arteries. The blood pressure reading has got two numbers. The first or the upper number is for measuring the pressure in the arteries when the heart beats and is called systolic pressure. The second or the bottom number is for measuring the pressure in the arteries between beats and is called diastolic pressure. The more blood the heart pumps and the narrower the arteries, the higher the blood pressure.

High blood pressure has no symptoms, thus every person visiting a health facility is taken a blood pressure reading regardless of the problem they are consulting about. Hypertension can be caused by the following factors: (1) Tobacco use: the chemical in tobacco gradually damages the lining of the artery walls, causing the arteries to narrow, (2) Age: as a person ages, the risk of high blood

pressure also increases, (3) Consuming too much salt: too much salt can cause the body to retain fluids, which increases blood pressure, (4) Lack of physical activities: physical inactivity can lead to higher heart rates and the higher the heart rate, the harder the heart must work with each contraction and the stronger the forces on the arteries and many other related factors. The blood pressure is measured by an upper arm blood pressure monitor or sphygmomanometer.

The prevalence of high blood pressure is increasing with age and it is most common in low-income and middle-income economies. About 13.5% of the global premature death were related to high blood pressure (Lawes et al., 2008). There is a strong positive association between body weight and high blood pressure. This is observed in people of different ethnic backgrounds, different ages and gender (Pausova, 2006).

2.4.2 Fasting plamsa glucose

Fasting plasma glucose is the amount of sugar/glucose in the liquid part of the blood rather than the whole body. In testing the glucose level, it is required that the patient be in a fasting state, meaning having not eaten for at least eight hours. Hence this test is usually scheduled for the morning. A high plasma glucose level may lead to diabetes (Nichols et al., 2008). A healthy plasma glucose level is $<$5.6 mmol/l, between 5.5 mmol/l and 6.9 mmol/l means pre-diabetic and $\geq$7.0 mmol/l means diabetic. The glucose level can simply be measured by applying a drop of blood to a chemically disposable test strip, and then inserted into an electronic blood glucose meter.

2.4.3 Triglycerides

Triglycerides are a type of fat in the blood that result from any unwanted calories in the body. They are stored in the fat cells and released for energy between meals. Elevated triglycerides level may lead to atherosclerosis which causes arteries to narrow hence blocking the flow of blood. Highly elevated triglycerides level may lead to fatty liver diseases, pancreatitis and kidney diseases. The normal healthy triglycerides level is less than 150 mg/dl, anything above that is a health risk. High triglycerides level is a result of the food consumed such as sugary foods and foods high in saturated fats like cheese and red meat. Triglycerides levels are measured using blood samples.

2.4.4 High-Density Lipoprotein (HDL) Cholesterol

HDL cholesterol is a lipid protein that absorbs cholesterol and carries it back to the liver so that it can be flushed out of the body. If the amount of HDL cholesterol is not enough in the body, then cholesterol remain in the bloodstream and build up in the arteries which will cause difficulties in the flow of blood. High level of HDL cholesterol lowers the risk of heart diseases, blood clots and stroke. A low HDL cholesterol can be caused by the use of tobacco, excess weight, poor diet and many others. HDLC level is also measured using blood samples.

2.4.5 Obesity

Obesity is defined as an excessive amount of fat in the body and has become a major public health problem worldwide (Vazquez et al., 2007). The distribution of fat in the body is an important factor of obesity. Studies have shown that fat in the upper or central body is mostly related to health risk (Samuel et al., 2007; Goh et al., 2014; Barroso et al., 2017). Central obesity increases the risk of heart diseases, cancer, Type 2 diabetes and other diseases. This condition

can affect people from childhood, adolescent stage and adulthood. Adult obesity is associated with increased morbidity and mortality (Ness-Abramof and Apovian, 2008). Causes of obesity include: (1) Diet: energy imbalance between calories consumed and calories expanded, (2) Lack of physical activity: today children seem less likely to walk to school and spend more time on computers and television, (3) Heredity: genetics play a role in obesityAS obese parents are more likely to have obese children, and many other causes.

Central obesity can be accurately measured by a computer tomography (CT) scan or a magnetic resonance imaging (MRI) machine, but this approach is way too expensive for routine clinical use. As a result, simple arthropometric measurements can be used to measure central obesity. The most common and standard arthropometric measures are body mass index (BMI) and waist circumference (WC). These two measures are highly correlated (Koster et al., 2008; Ness-Abramof and Apovian, 2008). The BMI is the relative weight for height that correlates significantly with the total body fats. Pi-Sunyer (2000) defines a BMI of $<$18.5 kg/m^2 as underweight, 18.5 - 24.9 kg/m^2 as normal weight, 24.9 - 29.9 kg/m^2 as overweight and $\geq$30 kg/m^2 as obese.

WC is the presence of excess fat in the abdomen out of proportion to the total fat in the whole body. It is a perimeter which provides an estimate of body girth at the abdominal level. According to Samuel et al. (2007), WC measurements are taken three times and recorded to the nearest 0.1 cm. Although specific techniques have been recommended for measuring WC in clinical setting, there is still no common recognised approach (Heart et al., 1998). The IDF definition of MS emphasise the WC cut points that are gender and ethnic-specific. Therefore when estimating the prevalence of MS and diagnosing people with MS, gender and ethnicity should be taken into consideration. WC is measured at the midpoint between the low rib and the iliac crest using a measuring tape

that cannot be easily stretched.

Obesity is a global health burden and its prevalence is increasing with the rates having tripled in the past four decades (Blüher, 2019). According to the global burden of disease, over four milloin people are dying each year as a result of being overweight or obese. In Africa, a complex coexistence of under and over nutrition has been reported. The prevalence of overweight and obesity has also tripled between 1992 and 2005 (Ziraba et al., 2009). In the SSA populations, the prevalence of obesity is higher in women than men and also higher in urban than rural populations. A study by Pereko et al. (2013) in Ghana revealed that women are about eight times more likely to be overweight or obese than men. Prentice (2006) also conducted a study in South Africa and revealed that women are about three times more likely to be overweight or obese than men. In addition, Malhotra et al. (2008) did a study which revealed that the prevalence was high in South African women than men. Obesity increases with age, it increases up to a certain age and then decreases afterwards. Some studies in Africa reported a positive association between age and obesity (Dake et al., 2011; Muhihi et al., 2012; Masibo et al., 2013).

2.5 Prevalence of obesity

Over the past few decades, there has been an increase in obesity prevalence, according to several international surveys (Popkin, 1994; Matta et al., 2018). Obesity rates are rising across all economic levels and age groups in developing nations in particular, like China, Brazil, and South Africa, which are in the process of making the economic transition from underdeveloped to developed (Region, 2000). In South Africa, the prevalence of obesity has steadily increased, especially among black women (Steyn et al., 1998; Van Der Merwe and Pepper, 2006). Obesity is influenced by multiple factors such as socio-economic,

socio-demographic, ethnicity or culture, perceptions and beliefs, genetic among others.

2.5.1 Socio-economic factors

Despite being a developing nation, South Africa still has a large number of impoverished households (May, 2000). Non-communicable disease cases are increasing as a result of urbanisation, poverty, violence, rapid social and economic change, lack of knowledge, and insufficient services (Vorster et al., 2000; Ball et al., 2002). A group of economically engaged South Africans from four different ethnic groups found a correlation between low education level and having a higher BMI (Senekal et al., 2003). According to the 1998 national survey, obesity was associated with level of education, age, ethnicity and geographic location (Puoane et al., 2002).

2.5.2 Perceptions and beliefs

Cultural standards and views contribute to the rise in overweight and obesity, especially in Africa. For instance, it is widely accepted that thin people are not healthy because it represents ill-health and poverty (Simfukwe et al., 2017). Another widely held misconception that fuels the need for a bigger body structure is that thinness is associated with HIV/AIDS which is still stigmatised (Ofori and Angmorterh, 2019). South Africa, as well as nations like Morocco and the USA, accept the idea that being overweight is good, healthy, and a sign of wealth (Okop et al., 2016).

2.5.3 Globalisation and urbanisation

The majority of the SSA countries still have this rural-urban gradient, while in South Africa, the differences between rural and urban areas are less pronounced (Delisle et al., 2012). The obesity epidemic among South Africans

is a reflection of globalisation, which is the main force behind the nutritional change (Kruger et al., 2005). The consumption of obesogenic foods is more likely to be seen in urban locations where energy-dense foods linked to a western lifestyle have been adopted (Neupane et al., 2015; Motadi et al., 2018). Making healthy food choices is getting harder as South Africa undergo a nutrition revolution, as energy-dense foods are more readily available and more reasonably priced (Tugendhaft et al., 2016). Furthermore, by fostering a culture that encourages the consumption of foods high in fat and sugar, globalisation thereby raises the risk among the urban population (Khan and Bowman, 1999). However, according to the 2003 South African Demographic and Health Survey (SADHS), there is a 21% prevalence in rural black women and a 31% prevalence in urban black women (Micklesfield et al., 2013).

2.5.4 Gender

There has been a consistent report from studies in both SSA and South Africa indicating that the prevalence of obesity is greater in women than men (Dake et al., 2011; Shayo and Mugusi, 2011; Kruger et al., 2012). In most nations, women are more likely than men to be obese, although in other nations and population subgroups, this disparity is more pronounced (Cooper et al., 2021). According to the available data, being a woman may increase the risk of becoming obese (Wardle et al., 2002; Cooper et al., 2021). Wells et al. (2012) conducted a study analysing the association of economic and gender inequalities with the global obesity prevalence of 68 nations and found that gender disparities were linked to obesity in both developed and developing European nations. Women are more likely than men to receive an obesity diagnosis, seek and receive all forms of therapy for obesity, including behavioral, pharmacological, and bariatric surgery (Cooper et al., 2021).

2.6 Optimal cut-off points in Sub-Saharan Africa (SSA)

According to Crowther and Norris (2012), the recommended cut-off points for diagnosing MS is not appropriate for population groups in SSA. This can lead to under or over estimating the prevalence of MS in these population groups. If underestimation occurs, it means some people will be at high risk of MS and not getting treatment and vice versa. Therefore studies were done in the SSA to determine the appropriate optimal cut-off points.

Some related studies were conducted in some parts of SSA and the following cut-off points were obtained: (1) Magalhães et al. (2014) in Angola found optimal cut-off point of 87.5 cm for men and 80.5 cm for women. (2) Ekoru et al. (2018) used data from various countries in the SSA including: Kenya, Uganda, Tanzania, Nigeria, Benin and South Africa. They found optimal cut-off points of of 81.2 cm for men and 81 cm for women. (3) Sinaga et al. (2018) in Ethiopia found the optimal cut-off points of 83.7 cm for men and 78 cm for women. (4) Tladi et al. (2020) in Botswana found optimal cut-off points of 91 cm for men and 82.3 cm for women.

According to the studies conducted in South Africa,the following cut-off points were obtained: (1) Motala et al. (2011) in KwaZulu-Natal found optimal cut-off points of 86 cm for men and 92 cm for women. (2) Crowther and Norris (2012) in Soweto, Johannesburg found optimal cut-off point of 91.5 cm for women. (3) Prinsloo et al. (2011) in North West found the optimal cut-off points of 90 cm for men and 98 cm for women. (4) Owolabi et al. (2018) in Eastern Cape found optimal cut-off points of 92.25 cm for men and 89.45 cm for women.

2.7 Summary of the chapter

In this chapter we reviewed researches on the prevalence of metabolic syndrome and obesity in various countries, specifically focusing on South Africa. The determinants of metabolic syndrome included: triglycerides, high blood pressure, fasting plasma glucose, high-density lipoprotein cholesterol and obesity. We also reviewed the prevalence and factors of obesity. Lastly, we reviewed the optimal cut-off points in the SSA and South Africa.

Chapter 3

Methodology

3.1 Introduction

This chapter gives an explanation of the sources of data utilised and study area. It explains the statistical tests and methods used to analyse the data.

3.2 Research methodology

3.2.1 Data source and study area

The study uses secondary data obtained from the Dikgale Health and Demographic Surveillance System (DHDSS) in the Capricorn District of the Limpopo Province, South Africa. The DHDSS was intended as a sample framework for monitoring non-communicable diseases (NCDs) and their risk factors. Dikgale is a rural area consisting of 15 villages with approximately 7200 households and an estimated population size of 50 000.

3.2.2 Receiver operating characteristic analysis

Receiver Operating Characteristic (ROC) analysis was formulated in the 1950s and was first applied during World War II for classification accuracy in differentiating signal from noise in radar detection (Lusted, 1971). It is commonly used in medical or clinical areas that depend on screening and diagnostic tests such as epidemiology, laboratory testing, bioinformatics, and many others (Tal et al., 2019). ROC analysis is a tool used to evaluate the performance of a diagnostic test and the accuracy of a statistical model that classifies the test as a dichotomous, ordinal or continuous outcome. In the case where an outcome is dichotomous, the approach of the diagnostic test evaluation uses sensitivity and specificity as a measure of accuracy in comparison with the gold standard status (Swets, 1979).

The ROC curve, a plot of sensitivity versus 1-specificity, displays all possible cutoff points and gives estimates of the frequency of various outcomes at each threshold. The ROC curve does not require the selection of a particular threshold because all possible threshold are included. It plays a role in evaluating diagnostic ability of tests to discriminate the true state of the subjects and finding the optimal cutoff values. This ROC curve and the area under the curve (AUC) are the effective measure of accuracy considered with meaningful interpretations (Hanley and McNeil, 1982).

3.2.3 Area under the curve

Area under the curve (AUC) is a measure of the overall performance of a diagnostic test (Zhou et al., 2009; Obuchowski, 2003). It is an effective and combined measure of sensitivity and specificity that describes the inherent of diagnostic test (Kumar and Indrayan, 2011). AUC $= 0$, which is a rare occasion,

means the test incorrectly classifies all positive subjects as negative and all negative subjects as positive. AUC $= 0.5$ means the chance discrimination that the curve is located on the diagonal line in the ROC space. AUC closer to 1 means a better overall diagnostic performance of the test. AUC $= 1$ means that the diagnostic test is perfectly accurate, meaning there is no overlapping.

3.2.4 Confusion matrix

A confusion matrix is a tabular summary that shows how well the model is performing. The standard way of describing the accuracy of a diagnostic test when the outcomes are dichotomous (positive and negative or present and absent) is through a two-by-two table/confusion matrix shown in Table 3.1.

Table 3.1: 2x2 confusion matrix.

		Actual value	
		Yes	No
Model prediction	Yes	TP	FP
	No	FN	TN

In machine learning applications, a true positive (TP) is a positive sample that is correctly classified as positive by the machine learning algorithm. True negative (TN) is a negative sample that is correctly classified as negative by the machine learning algorithm. False positive (FP) is a negative sample that is incorrectly classified as positive by the machine learning algorithm. False negative (FN) is a positive sample that is incorrectly classified as negative by the machine learning algorithm.

3.2.5 Sensitivity and specificity

Sensitivity is the probability of a diagnostic test to correctly classify the positive class. Specificity is the probability of a diagnostic test to correctly classify the

negative class. The ROC curve is plotted using the true positive rate (TPR) also called sensitivity on the y-axis versus the false positive rate (FPR) also called $1-$ specificity on the x-axis. Each point on the ROC curve represents TPR and FPR at a different threshold. Then sensitivity or TPR is given by:

$$\text{TPR} = \frac{TP}{TP + FN}$$

and $1-$ specificity or FPR is given by:

$$\text{FPR} = \frac{FP}{FP + TN}$$

3.3 Generalised linear models

The Generalised Linear Model (GLM) was formulated by Nelder and Wedderburn (1972) to provide a unified framework to study various statistical regression models. The GLM is an extension of the ordinary regression models that encompasses non-normal response distribution as well as modelling functions of the mean. The general linear model is given by:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{3.1}$$

where $\mathbf{X}$ represents the covariates matrix, β represents the parameters and ϵ denotes the vector of the error term. The GLM extends to include the response variable that follows any probability distribution in the exponential family of distributions. Furthermore, it allows the response variables to take the form of an exponential distribution. Therefore, it is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family.

The GLM has the following assumptions:

- The response variables $Y_1, Y_2, ..., Y_n$ are independently distributed.
- The error terms are not correlated.
- The relationship between the outcome variable and each explanatory variable is approximately linear.
- The residuals are independent with mean zero and a constant variance

The GLM is characterised by three components:

- The random component: that specifies the conditional distribution of the response variable Y_i, given the explanatory variables. The response variable $Y_1, ..., Y_n$ are the random components and they are assumed to be independent random variables, each with a distribution from a specified exponential family. These $Y_i's$ are not identically distributed, but each have a distribution from the same family: Binomial, Normal, Poisson, etc. With a binary outcome, the random component has a binomial distribution that leads to logistic regression model.
- The systematic component: that is a model with a function of the predictor variable x_i, linear in the parameter that is related to the mean of Y_i. It relates a vector $(\eta_i, ..., \eta_N)$ to the explanatory variables $(x_1, ..., x_k)$ through a linear model by creating a linear predictor as:

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2,} + ... = \sum \beta_j \mathbf{X}_{ij}, \tag{3.2}$$

- Lastly, the link function: that specifies the link between the random and systematic components. It determines the relationship between the expected value of the response variable and the linear predictor of the explanatory variable. If $\mu = E(Y_i)$ is the expected value of Y_i, then the link

function specifies the function $g(.)$ relates μ to the linear predictor as:

$$g(\mu_i) = \alpha + \beta_k x_{ik}. \tag{3.3}$$

The ordinary regression model is the common form of the continuous responses. Other link functions allow the mean μ to be related to the predictors non-linearly. For example, suppose the link function $g(\mu) = log(\mu)$ models the log of the mean and the log function is only appropriate to the situation of modelling non-negative numbers such as count data. However, the log function of a GLM has the form:

$$log(\mu) = \alpha + \beta_k x_{ik}. \tag{3.4}$$

The link function $log(\mu) = log(\frac{\mu}{1-\mu})$ models the log of odds. This is called the logit link and it is appropriate to model μ when it lies between zero and one. This GLM is called the logistic regression model.

3.3.1 Exponential family of distributions

Exponential family of distributions is a parametric set or probability distributions for continuous, discrete of both variables. Some distributions like Bernoulli, Binomial, Normal, Poisson, Exponential, Gamma and Weibull belong to this family of distributions. The natural form of the exponential family of distributions is given by:

$$f_y(y, \theta, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \tag{3.5}$$

where $a(\phi)$ and $b(\theta)$ are known functions and $c(y, \phi)$ is a function of y_i and θ. θ is called a canonical parameter and ϕ is the dispersion parameter. If Y_i has a distribution in the exponential family, the mean and variance of y can be

derived from the well known relations and are indicated as follows:

$$E\left(\frac{\partial(l)}{\partial\theta}\right) = 0 \tag{3.6}$$

and

$$E\left(\frac{\partial^2(l)}{\partial\theta^2}\right) + \left[E\left(\frac{\partial(l)}{\partial\theta}\right)\right]^2 = 0 \tag{3.7}$$

$$l(\theta, y) = \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{3.8}$$

$$\frac{\partial(l)}{\partial\theta} = \frac{1}{a(\phi)}[y - b'(\theta)] \tag{3.9}$$

$$\frac{\partial^2(l)}{\partial\theta^2} = \frac{-b''(\theta)}{a(\phi)} \tag{3.10}$$

where primes are the second derivative with respect to θ.

$$E\left(\frac{\partial(l)}{\partial\theta}\right) = E\left(\frac{1}{a(\phi)}[y - b\prime(\theta)\right) = \frac{\mu_i - b'(\theta)}{a(\phi)} = 0 \tag{3.11}$$

Therefore

$$E(Y_i) = \mu_i = b'(\theta_i), \tag{3.12}$$

$$\frac{-b''(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)} = 0, \tag{3.13}$$

which implies that

$$Var(Y_i) = \sigma^2 = b''(\theta_i)a(\phi). \tag{3.14}$$

Hence the mean and the variance are given as $E(Y_i) = \mu_i = b'(\theta_i)$ and $mVar(Y_i) = \sigma^2 = b''(\theta_i)a(\phi)$ respectively.

3.3.2 Binomial distribution

Binomial distribution is a member of the exponential family, and for the binary outcome variable Y, the probability function is given by:

$$f(y_i, \pi_i) = \binom{k}{y_i} (\pi_i)^{y_i} (1 - \pi_i)^{k - y_i} \tag{3.15}$$

where π is the probability of success or parameter of interest and
k is the sample size.
The probability function is written in an exponential form as:

$$f(y_i, \pi_i) = exp\left[y_i log\pi_i - y_i log(1 - \pi_i) + klog(1 - \pi_i) + log\binom{k}{y_i} \right] \tag{3.16}$$

$$= exp\left[y_i log\pi_i - y_i log(1 - \pi_i) + klog(1 - \pi_i) + log\binom{k}{y_i} \right]$$

$$= exp\left[y_i log\left(\frac{\pi_i}{1 - \pi_i} \right) + klog(1 - \pi_i) + log\binom{k}{y_i} \right]. \tag{3.17}$$

Suppose $\frac{\pi_i}{1 - \pi_i} = \theta$
then:
$b(\theta) = -nlog(1 = \pi_i)$,
$c(y_i, \phi) = log\binom{k}{y_i}$,
$a(\phi) = 1$.
The mean and variance of a binomial distribution are $E(Y_i) = \mu_i = b'(\theta_i)$ and $Var(Y_i) = a(\phi)b''(\theta_i)$, respectively.

3.4 Special case of GLM - logistic regression

Logistic regression is a commonly used statistical method for modelling dichotomous response variable in which the log odds of the outcomes are mod-

elled as a linear combination of the regression variables and it is a family of the gereralised linear models. Logistic regression has the following assumptions:

- The dependent variable should be binary or dichotomous.
- The relationship between the independent variables and the log-odds of the dependent variable should be linear.
- The observations used in the logistic regression model should be independent of each other.
- Logistic regression assumes a reasonably large sample size to ensure reliable parameter estimation and robust statistical inference.
- The independent variables should not exhibit high levels of multicolliearity.
- There should be no influential outliers because logistic regression can be sensitive to outliers.

3.4.1 Binomial logit model for binary data

Many response variables are binary and have outcomes 0 and 1 which represent failure and success. In our case is the presence and absence of metabolic syndrome. The response variables $Y_1, ..., Y_n$ are independent and Y_i follows a Bernoulli distribution with parameter π_i. The Bernoulli family is an exponential family. The Bernoulli distribution is a special case of the Binomial distribution with $n = 1$. The Bernoulli distribution specifies probabilities $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$ for which $E(Y_i) = \pi_i$, hence for this model π_i is assumed to be related to x_i by:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_k x_i k \tag{3.18}$$

where $log\left(\frac{\pi_i}{1-\pi_i}\right)$ is the log odds of success for response Y_i which is the logit of π_i and the natural parameter of the exponential family. The model assumes that this logit is a linear function of the predictor x. The Bernoulli probability mass function can be written as:

$$\pi^y(1-\pi)^{1-y} = (1-\pi)exp\left[ylog\left(\frac{\pi_i}{1-\pi_i}\right)\right]. \tag{3.19}$$

The logit transformation is used to link the response variable to the set of explanatory variables denoted by $(x_1, x_2, ..., X_n)$. The logit link is given by:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = \sum_{j=1}^{k} \beta_j X_j. \tag{3.20}$$

It can be deduced from Eq. 3.20 that:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \sum_{j=1}^{k} \beta_j X_j \tag{3.21}$$

$$\frac{\pi(x_i)}{1-\pi(x_i)} = \exp\left(\sum_{j=1}^{k} \beta_j X_j\right)$$

$$\pi(x_i) = \left(\sum_{j=1}^{k} \beta_j X_j\right)\left(1-\pi(x_i)\right)$$

$$\pi(x_i) = \exp\left(\sum_{j=1}^{k} \beta_j X_j\right) - \pi(x_i)\exp\left(\sum_{j=1}^{k} \beta_j X_j\right)$$

$$\pi(x_i)\left(1+\exp\left(\sum_{j=1}^{k} \beta_j X_j\right)\right) = \exp\left(\sum_{j=1}^{k} \beta_j X_j\right)$$

Therefore,

$$\pi(x_i) = \frac{\exp\left(\sum_{j=1}^{k} \beta_j X_j\right)}{1 + \exp\left(\sum_{j=1}^{k} \beta_j X_j\right)}. \tag{3.22}$$

Thus, Eq. 3.22 is the logistic function for describing the relationship between $\pi(x_i)$ and the explanatory variables.

3.4.2 Maximum likelihood estimation

The goal of logistic regression is to estimate the parameters and this is done through maximum likelihood estimation (MLE). MLE entails finding a set of parameters for which the probability of the observed data is greatest. The MLE is derived from the probability distribution of the dependent variable. Let each y_i repesent a binomial count of the i^{th} population, then the probability density function is defined as:

$$f_{y_i}(y_i, \theta_i, \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]. \tag{3.23}$$

The likelihood function of the probability density function is given by:

$$L(y_i, \theta_i) = \prod_{i=1}^{n} f(y_i, \theta_i, \phi) = \prod_{i=1}^{n} exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]. \tag{3.24}$$

By taking the logarithm of Eq. 3.24, the log-likelihood function will be:

$$log[L(y_i, \theta_i)] = log \prod_{i=1}^{n} exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right] \tag{3.25}$$

which implies,

$$l(y_i, \theta_i) = \sum_{i=1}^{n} \log \exp \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]$$

$$= \sum_{i=1}^{n} a(\phi)^{-1}()y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi). \tag{3.26}$$

The estimates are obtained by taking the derivative of the log-likelihood for the i^{th} observation with respect to the regression coefficient β_i and equating them to zero. The score function given by $(U_{\beta_1}, U_{\beta_2}, ..., U_{\beta_p})'$ will be obtained, where p is the number of parameters and U_{β_j} is given by:

$$\frac{\partial(l_i)}{\partial\beta_j} = \frac{\partial\left(\sum_{i=1}^{n} a(\phi)^{-1}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi) \right)}{\partial\beta_j} = 0. \tag{3.27}$$

Using the chain rule of differentiation, $\frac{\partial(l_i)}{\partial\beta_j}$ can be written as:

$$\frac{\partial(l_i)}{\partial\beta_j} = \frac{\partial(l_i)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_j}\frac{\partial\eta_i}{\partial\beta_j}. \tag{3.28}$$

$E(Y_i) = \mu_i = b'(\theta_i)$, $Var(Y_i) = a(\phi)b''(\theta_i)$.

Since $\frac{\partial(l_i)}{\partial\theta_i} = \frac{y_i - b(\theta_i)}{a(\phi)}$, then $\frac{\partial(l_i)}{\partial\theta_i} = \frac{y_i - \mu_i}{a(\phi)}$

and $\frac{\partial\mu_i}{\partial\theta_i} = b''(\theta_i) = \frac{Var(Y_i)}{a(\phi)}$

which implies that,

$\frac{\partial\theta_i}{\partial\mu_i} = \frac{a(\phi)}{Var(Y_i)}$.

$\eta_i = \sum \beta_j X_{ij}$, then $\frac{\partial\eta_i}{\partial\beta_j} = x_{ij}$.

Substituting all these in Eq. 3.28, we get:

$$\frac{\partial(l_i)}{\partial\beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{\partial\mu_i}{\partial\eta_j} x_{ij} = \frac{y_i - \mu_i}{Var(Y_i)} x_{ij} \frac{\partial\mu_i}{\partial\eta_j}. \tag{3.29}$$

Equating Eq. 3.29 to zero, the maximum likelihood of β can be obtained using Newton-Raphson and Fischer scoring methods. Newton-Raphson is an iterative method for solving non-linear equations, whereas Fischer scoring is an iterative method for solving likelihood equations.

3.4.3 Fitting the logistic regression model

The goal of logistic regression is to estimate $K+1$ parameters using the method of maximum likelihood in:

$$ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \sum_{k=1}^{K} \beta_i X_j. \tag{3.30}$$

The maximum likelihood equation is derived from the probability distribution. Each y_i represents the count in the i^{th} population, and the joint probability density function of y_i is given by:

$$L(y|\pi) = \prod_{i=1}^{N} f(y_i|\pi_i) = \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} \left(1-\pi_i\right)^{n_i}. \tag{3.31}$$

Since $\pi(x_i) = \dfrac{\exp\left(\sum_{k=1}^{K} \beta_k X_k\right)}{1 + \exp\left(\sum_{k=1}^{K} \beta_k X_k\right)}$, the joint density function can be written as:

$$L(y|\pi) = \prod_{i=1}^{N} \binom{n_i}{y_i} \left[\exp\left(\sum_{k=1}^{K} \beta_k X_k\right)\right]^{y_i} \left[1 + \exp\left(\sum_{k=1}^{K} \beta_k X_k\right)\right]^{-n_i}. \tag{3.32}$$

Taking the logarithm of Eq.3.32 and since some terms in the summation do not depend on β_k, we will treat them as constants, then the log-likelihood function will be:

$$l(\pi) = \sum_{i=1}^{N} y_i \left[\exp\left(\sum_{k=1}^{K} \beta_k X_k \right) \right] - \sum_{i=1}^{N} n_i log \left[\exp\left(1 + \sum_{k=1}^{K} \beta_k X_k \right) \right]. \tag{3.33}$$

To find the critical value of the log-likelihood function, we set the first derivative with respect to β equal to zero, then we get:

$$\frac{\partial}{\partial \beta_k} \sum_{k=1}^{K} X_{ik} \beta_k = x_{ik}. \tag{3.34}$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \frac{1}{1 + \exp\left(\sum_{j=1}^{k} \beta_j X_j \right)} \frac{\partial}{\partial \beta_k} \left[1 + \exp\left(\sum_{j=1}^{k} \beta_j X_j \right) \right] \tag{3.35}$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \left(\frac{\exp(\sum_{j=1}^{k} \beta_j X_j)}{1 + \exp(\sum_{j=1}^{k} \beta_j X_j)} \right)$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ij} \tag{3.36}$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} x_{ik} (y_i - n_i \pi_i). \tag{3.37}$$

The maximum likelihood estimates for β is obtained by letting each $K+1$ equations in Eq. 3.37 to zero and solving for each β_k. Then, the maximum likelihood estimates will be given by:

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} x_{ik} (y_i - n_i \pi_i) = 0. \tag{3.38}$$

Furthermore, the equations are non-linear and can be solved using iteration algorithms such as Newton-Raphson and Fischer scoring methods.

3.4.4 Confidence interval estimation

The confidence interval estimate for the slope coefficient is given by:

$$\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_i), \tag{3.39}$$

and the confidence interval estimate for the intercept is given by:

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_0). \tag{3.40}$$

3.4.5 The odds ratio

The logistic regression coefficients are usually interpreted using the odds ratios. The odds of the outcome being a success, where $x = 1$ is given by:

$$\frac{\pi(1)}{1 - \pi(1)}, \tag{3.41}$$

and the odds of the outcome being a success, where $x = 0$ is given by:

$$\frac{\pi(0)}{1 - \pi(0)}. \tag{3.42}$$

According to Hosmer and Lemeshow (2013) , the odds ratio (OR) is defined by:

$$OR = \frac{\frac{\pi(1)}{1 - \pi(1)}}{\frac{\pi(0)}{1 - \pi(0)}}. \tag{3.43}$$

Consider the relationship between an outcome variable and one explanatory variable X, then,

$$logit(\pi) = \beta_0 + \beta_1 X_1 \tag{3.44}$$

and the logit function can be written as:

$$logit(\pi) = log\frac{\pi}{1-\pi} = log(odds) \tag{3.45}$$

$$log(odds) = \beta_0 + \beta_1 X_1 = \beta_0 + \beta_1$$

which implies,

$$\pi = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}.$$

let $\pi(1) = \pi = \dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$

and $\pi(0) = \dfrac{\exp \beta_0}{1 + \exp(\beta_0)}$

then, the odds ratio can be written as:

$$OR = \frac{\dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \Big/ \dfrac{1}{1 + \exp(\beta_0 + \beta_1)}}{\dfrac{\exp \beta_0}{1 + \exp(\beta_0)} \Big/ \dfrac{1}{1 + \exp(\beta_0)}}$$

$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)}.$$

Therefore,

$$OR = \exp(\beta_1). \tag{3.46}$$

The parameter β_1 which is associated with X represents the change in the log odd from zero to one. Therefore, when

$OR = 1$, it means that the odds for $x = 1$ and $x = 0$ are equal,

$OR > 1$, it means that the odds for $x = 1$ is greater than the odds for $x = 0$ and

$OR < 1$, it means that the odds for $x = 1$ is less than the odds for $x = 0$.

3.4.6 Confidence interval estimation for the odds ratio

The confidence interval estimate for the intercept is given by:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_j). \tag{3.47}$$

The confidence interval is transformed by introducing the exponent so that the corresponding confidence interval for the odds ratio is given by:

$$\exp(\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_j)). \tag{3.48}$$

3.5 Random forest

The first algorithm of random decision forest was proposed using the random sample space in 1995 by Ho (1995). Breiman (1996), Amit and Geman (1997), Dietterich (1998) and many other researchers developed other machine learning techniques of random selection until in 2001 when Leo (2001) was influenced by Amit and Geman (1997)'s idea and developed an extension of the algorithm. Random forest is a learning machine techniques introduced by Leo (2001) as a combination of tree predictors such that each tree depends on the value of a random vector sampled independently and with the same distribution for all trees in the forest. Random forest is a popular algorithm for both classification and regression and has the following assumptions:

- Random forest assumes that the observations used to train the model are independent of each other.
- Random forest typically requires enough training samples to learn meaningful patterns and generalise well to the unseen data.
- Random forest generally does not handle missing data internally. Therefore, it is recommended to handle missing values appropriately before

training the model, either by imputing missing values or removing instances with missing data.

- While random forest can handle imbalanced class distributions, it tends to perform better when the classes are roughly balanced.
- Random forest is robust to multicollinearity, as it considers only a subset of features at each split.
- Random forest does not require standardising or scaling the input features. It can handle variables with different scales and units effectively.

Definition 3.1. *Random forest (Leo, 2001).*
A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(\boldsymbol{x}, \Theta_k), k = 1, ...\}$ *where* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\boldsymbol{x}$.

A random vector Θ_k is generated for the k^{th} tree where Θ_k is independent of the past random vectors $\Theta_1, ..., \Theta_{k-1}$ but with the same distribution. A tree is grown using the training set and Θ_k, resulting in a classifier $h(\mathbf{x}, \Theta_k)$ where $\mathbf{x}$ is an input vector. In random split selection Θ consists of several independent random integers between 1 and K. The nature and dimensionality of Θ depends on its use in tree construction. Therefore, a large number of trees is generated and they vote for the most popular class, then the decision function is given by:

$$H(x) = argmax_Y \sum_{i=1}^{k} I(h_i(x) = Y), \tag{3.49}$$

where $H(x)$ is the combination of the classification model, h_i is a single decision tree model, Y is the output variable, and I is the indicator function.

3.5.1 Characters of random forest

In random forest, the margin function is used to measure the extend to which the average number of votes at X, Y for the right class exceeds that for the wrong class. We define the margin function as:

$$mg(X, Y) = av_k I(h_k(X) = Y) - max_{j \neq Y} av_k I(h_k(X) = j). \tag{3.50}$$

The larger the margin, the more confidence in classification and the higher accuracy of classification prediction. The generalisation error of the classifiers is defined as:

$$PE^* = P_{X,Y}(mg(X, Y) < 0), \tag{3.51}$$

where the subscripts X,Y indicate that the probability is over the X,Y space. When the number of decision tree is big enough, $h_k(X) = h(X, \Theta_k)$ follows from the strong law of large numbers and the tree structure. As the number of decision trees increase, for almost all sequences Θ_i, PE^* converges to

$$P_{X,Y}(P_\Theta(h(X, , \Theta = Y)) - max_{j \neq Y} P_\Theta(h(X, \Theta) = j) < 0). \tag{3.52}$$

3.6 Linear discriminant analysis

Linear discriminant analysis (LDA) is a statistical technique used for classification problems that find a linear combination of features that separates different classes. The assumptions of LDA include:

- LDA assumes that the predictors or independent variables follow a multivariate normal distribution within each class.
- Homoscedasticitya which assumes that the variance of the predictor variables is equal across all classes.

- LDA assumes that the predictor variables are independent of each other within each class. This assumption suggests that the correlations between the predictor variables should be minimal within each class.
- LDA assumes that the covariance matrices of the predictor variables are equal across all classes.
- LDA assumes that the predictor variables can be combined linearly to create a discriminant function that can separate the classes effectively.

The LDA method identifies a projection vector to maximise between-class scatter matrix while minimising within-class scatter matrix in the future space. The linear function is given by:

$$y = a_1 x_{i_1} + a_2 x_{i_2} + ... + a_q x_{i_q} \tag{3.53}$$

where

$$a^T = \left[\{a_1, a_2, ..., a_q\}\right] \tag{3.54}$$

is a vector of coefficients that need to be determined, while

$$x_i = \left[x_{i_1}, x_{i_2}, ..., x_{i_q}\right] \tag{3.55}$$

are the subjects and

$$x_j = \left[x_{j_1}, x_{j_2}, ..., x_{j_q}\right] \tag{3.56}$$

are the features.

3.7 Model building

Model building is a phase which includes model building for prediction of the disease. In this section, we will provide an outline various machine learning algorithms for MS prediction. These algorithms include: Logistic regression,

Random forest and Linear discriminant analysis. The following steps are followed during prediction using various machine learning algorithms:

Algorithm 1:

1. Generate training set and test set randomly.
2. Specify algorithms that are used in model

Algorithm 2:

1. Import required libraries.
2. Import MS dataset.
3. Create pipeline for algorithms giving highest accuracy.
4. Add theses pipeline to a dictionary where all pipelines will be stored.
5. Fit the pipelines in training dataset.
6. Compare accuracies of all pipelines added.
7. Prediction and identification of the most accurate model will be done on test data.

3.8 Model evaluation

This is a phase where we evaluate the prediction results using various evaluation performance measures such as sensitivity (recall), specificity, positive predictive value(PPV), negative predictive value(NPV), F1 score and accuracy which can all be obtained through a confusion matrix.

Table 3.2: The confusion matrix

		Actual	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

A confusion matrix is a tabular representation of the model's performance that displays the quantity of true positives, true negatives, false positives, and false negatives . Confusion matrix analysis offers a more profound comprehension of the model's accuracy in instance classification.

3.8.1 Sensitivity / Recall

Sensitivity quantifies the proportion of true positives that the test accurately identifies. It provides information on how successfully the test detects true positives. It is given by:

$$\text{Sensitivity / Recall} = \frac{TP}{TP + FN}. \tag{3.57}$$

3.8.2 Specificity

Specificity quantifies the proportion of true negatives that the test accurately identifies. It provides information on how successfully the test detects true negatives. It is given by:

$$\text{Specificity} = \frac{TN}{FP + TN}. \tag{3.58}$$

3.8.3 Positive predictive value / Precision

Positive predictive value is the proportion of positive test results that are true positives. It shows the likelihood that those who have positive test results for the condition are infact in the group. It is given by:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3.59}$$

3.8.4 Negative predictive value

Negative predictive value is the proportion of negative test results that are true negatives. It shows the likelihood that those who have negative test results for the condition are infact not in the group. It is given by:

$$\text{NPV} = \frac{TN}{FN + TN}. \tag{3.60}$$

3.8.5 Accuracy

Accuracy is the proportion of correctly classified instances, both true negative and true negatives. It is given by:

$$\text{Accuracy} = \frac{TP + FN}{TP + FN + FN + TN} \tag{3.61}$$

3.8.6 F1 score

It is used to measure a test's accuracy. F1 score is the harmonic mean between precision and recall. The range for F1 score is $[0, 1]$. It tells us how precise the classifier is as well as how robust it is. It is given by:

$$\text{F1} = \frac{2}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right)}. \tag{3.62}$$

3.9 Summary of the chapter

This chapter outlines the approach, techniques, and procedures used to collect data. It also outlines the various statistical and machine learning techniques employed to analyze the data and predict outcomes.

Chapter 4

Results and discussion

4.1 Introduction

This chapter presents data analysis, interpretation and discussion of the results in line with the research objectives and methodology of the study. The first part consists of the prevalence of metabolic syndrome (MS) and related disorders, the summary statistics, determination of the waist circumference optimal cut-off point and direct evaluation of the newly derived optimal cut-off point. The second part is the application of machine learning tools to discriminate participants with and without MS. The last part is sensitivity analysis.

4.2 Data description

There are three female datasets in this study, the 1997, 2005 and 2012. The 1997, 2005 and 2012 datasets consist of 1082, 480 and 521 participants, respectively, all with 9 variables. The variables include body mass index (BMI),

age, trigycerides (Trig), total cholesterol (TotChol), glucose (Gluc), high-density lipoprotein (Hdl), waist circumferenceWC (), systolic blood pressure (Sbp) and diastolic blood pressure (Dbp).

4.3 Descriptive statistics

This section presents characteristics of the study population. We first assumed an waist circumference cut-off point for MS. In this case, participants were diagnosed with metabolic syndrome when they presented at least three of the following components: Trig $\geq$1.7 mmol/l, Hdl $<$1.29 mmol/l, Gluc $\geq$ 5.6 mmol/l and Sbp $\geq$ 130 mmHg or Dbp $\geq$ 85 mmHg. This is inline with the derivations done by Crowther and Norris (2012). The prevalence of MS and associated disorders are shown in Table 4.1.

Table 4.1: Prevalence of MS and other related disorders

Variable	Number (%)
MS (at least 3 disorders)	209 (19)
Trig $\geq$ 1.7 mmol/l	170 (16)
Hdl $<$1.29 mmol/l	689 (64)
Gluc $\geq$ 5.6 mmol/l	275 (25)
Sbp $\geq$ 130 mmHg	582 (54)
Dbp $\geq$ 85 mmHg	551 (51)
BP $\geq$ 130/85 mmHg	673 (62)
BMI $\geq$ 30 kg	328 (30)
WC $\geq$ 80 cm	744 (69)
WC $\geq$ 80 cm & at least 2 disorders	457 (42)

In Table 4.1, the number of participants with MS were found to be high at 209 in number (19%). It can also be seen that the prevalences are quite high for

high-density lipoprotein cholesterol levels less than 1.29 mmol/L at 64 % and those with high levels of blood pressure at 62%. The less frequent components were fasting plasma glucose greater than 5.6 mmol/L and triglycerides greater than 1.7 mmol/L with prevalence of 25% and 16%, respectively.

Table 4.2: Summary statistics of characteristics of metabolic syndrome

	min	max	mean	median	Q_1	Q_3	kurtosis	skewness
Age	30.00	89.00	53.34	53.00	41.00	63.00	2.19	0.18
BMI	13.09	72.56	27.47	26.57	22.97	31.34	7.17	1.24
WC	55.00	150.00	86.84	86.00	77.00	95.00	3.56	0.57
Trig	0.20	7.10	1.19	1.00	0.80	1.40	18.68	2.64
Hdl	0.40	3.30	1.22	1.20	1.10	1.30	14.14	1.97
Gluc	1.00	27.80	5.52	4.90	4.30	5.60	21.60	4.00
Sbp	96.00	250.00	135.80	130.00	120.00	149.30	4.48	1.09
Dbp	48.00	140.00	86.68	86.00	80.00	93.92	3.81	0.66

Table 4.2 presents the summary statistics for the characteristics of metabolic syndrome. The results revealed that the ages of participants range from 30 to 89 years with a median of 53 years. The data for age variable is positively skewed since the mean is greater than median. A kurtosis value less than 3 suggests a flat-peaked distribution. The data for the other variables; BMI, WC, Trig, Hdl, Gluc, Sbp and Dbp have a mean greater than the median, therefore positively skewed. Kurtosis values greater than 3 suggests that the data follows a sharp-peaked and heavy-tailed distribution and therefore non-normal.

The next section presents the anthropometric and fasting biochemical parameters by the disease status. Shapiro-wilk test was used to test for normality and a test was considered significant if p-value is less than 0.05. The data that was not normally distributed was then transformed using the log function.

Table 4.3: Shapiro-Wilk test for normality

	MS		NoMS	
	Statistic	P-value	Statistic	P-value
Number (%)	209 (19)		873(81)	
Age	0.9849	0.0250	0.9668	<0.001
BMI	0.9796	0.0040	0.9213	<0.001
WC	0.9934	0.4749	0.9744	<0.001
Trig	0.9545	<0.001	0.7049	<0.001
Hdl	0.7318	<0.001	0.8638	<0.001
Gluc	0.7019	<0.001	0.5401	<0.001
Sbp	0.9584	<0.001	0.9022	<0.001
Dbp	0.9506	<0.001	0.9584	<0.001

Table 4.3 shows the test statistics and the corresponding p-values. It can be seen that all p-values are very small (<0.05), except for waist circumference of participants suffering from MS with a p-value of 0.4749 which is not less than 0.05. However, all the other variables resulted in a p-value less than 0.05, so the null hypothesis was rejected. This implies that the normal distribution is not assumed, therefore, the data must be transformed for subsequent analysis. The reason for log-transformation of the data is to approximate normality which is assumed by the t-test done in Table 4.4. The test is done specifically for the difference between two means for the anthropometric and fasting biochemical parameters with regards to participants identified with MS and those without.

Table 4.4: Test for the difference betweern two means for anthropometric and fasting biochemical parameters (at least three components of MS)

	MS	NoMS	P-value
Number (%)	209(19)	873(81)	-
Age	4.0246±0.2184	3.9225±0.2724	<0.001
BMI	3.3553±0.2236	3.2710±0.2222	<0.001
WC	4.5287±0.1408	4.4354±0.1412	<0.001
Trig	0.4650±0.5214	0.0506±0.4155	<0.001
Hdl	0.1182±0.1380	0.1950±0.1811	<0.001
Gluc	1.9869±0.4514	1.5455±0.2570	<0.001
Sbp	4.9924±0.1353	4.8746±0.1617	<0.001
Dbp	4.5340±0.1202	4.4321±0.1422	<0.001

The study group for this analysis consists of 1082 participants with 209 suffering from MS and 873 not suffering from MS. The data were transformed using the natural logarithm and presented as mean ± standard deviation (SD). It can be seen in Table 4.4 that participants who are suffering from MS had significantly higher mean values than those who are not suffering from MS for age, BMI, WC, Trig, Gluc, Sbp and Dbp (all with p-values < 0.001). However, participants who are suffering from MS had a significantly lower mean value than those who are not suffering from MS for Hdl (also with p-value < 0.001).

4.4 Receiver Operating Characteristic (ROC)

In this section, we determine the waist cicumference optimal cut-off point. The optimal cut-off point was determined using the Youden J index approach.

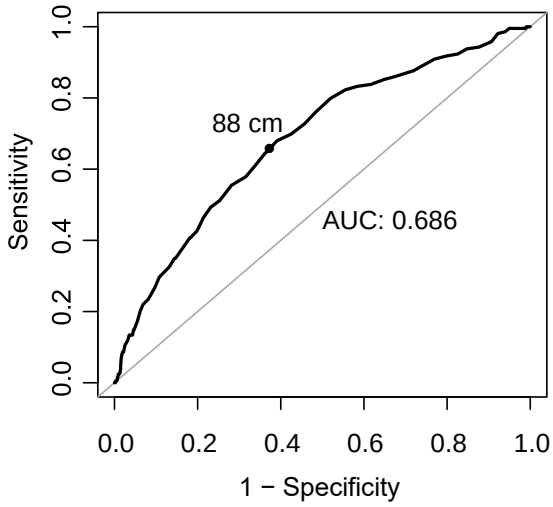

Figure 4.1: ROC curve for waist circumference. Area under the curve: 0.686, Waist cut-off point: 88 cm, Sensitivity: 68% and Specificity: 61%

Table 4.5 presents the summary of the ROC curve in Fig. 4.1. The optimal waist circumference cut-off point to detect MS with the maximum sensitivity and specificity was found to be 88 cm. This is a value above which a participant is most probably having MS and below which a participant is not having MS. Sensitivity was found to be 67.94% with a 95% confidence interval of (61.09 , 74.34) and specificity of 60.93% with a 95% confidence interval of (57.43 , 64.10). The Youden J statistics for optimising sensitivity and specificity values resulted in the Youden index = max (sensitivity + specificity – 1) of 0.2783. The area under the curve, AUC=0.686 with a 95% confidence interval (0.645,0.727). This AUC of 0.69 is not far from 0.7 and the confidence interval for AUC (0.645,0.727) contains 0.70. Therefore, one can conclude that this is good a AUC since it is close to the acceptable AUC value and can better discriminate participants suffering from MS and those not suffering.

Table 4.5: ROC curve analysis

	Estimates	95% Lower limit	95% Upper limit
Cut-off point	88.0000	-	-
Sensitivity	0.6794	0.6115	0.7421
Specificity	0.6094	0.5743	0.6410
PPV	0.2940	0.2662	0.3612
NPV	0.8881	0.8550	0.9012
DLR Positive	1.7394	1.5356	1.9703
DLR Negative	0.5260	0.4288	0.6454
AUC	0.686	0.645	0.726
Youden statistics	0.2888	-	-

Table 4.6: The confusion matrix

		MS	
		Yes	No
WC≥ 88	Yes	142	341
	No	67	532

Table 4.6 presents a confusion matrix of the presence of at least three components of MS and the waist circumference greater or equal 88 cm. 209 participants had MS and 483 had a waist circumference is greater or equal to 88 cm. From this 483 only 142 had MS, thus the positive predictive value (PPV) of 0.2940 from Table 4.5. The proportion of correctly classifyng participants with MS, sensitivity is 142 out of 209 (0.6794). 873 participants had no MS and 599 had a waist circumference is not greater or equal to 88 cm. From this 599 only 532 had no MS, thus the negative predictive value (NPV) of 0.8881. The proportion of correctly classifyng participants with no MS, specificity is 532 out of 873 (0.6094) The confusion matrix also shows the false negative value of 341, which is the number of participants who infact do not have MS but have a waist circumference greater or equal to 88 cm and a false positive of 67 which is the

number of participants who infact have metabolic syndrome but have a waist circumference not greater or equal to 88 cm.

In the next section we present the anthropometric and fasting biochemical parameters for the three dataset that will be used for analysis. The datasets include the 1997, 2005 and 2012 data. The models will be fitted using the 1997 dataset and evaluated using the 2005 and 2012 data.

Table 4.7: Normality test using Shapiro-Wilk test

	WC$\geq$88cm		WC$<$88cm	
	Statistic	P-value	Statistic	P-value
Number (%)	483 (45)		599(55)	
Age	0.9108	$<$0.001	0.9639	$<$0.001
BMI	0.9229	$<$0.001	0.9266	$<$0.001
WC	0.8821	$<$0.001	0.9598	$<$0.001
Trig	0.8016	$<$0.001	0.7972	$<$0.001
Hdl	0.7975	$<$0.001	0.8930	$<$0.001
Gluc	0.6047	$<$0.001	0.4890	$<$0.001
Sbp	0.9373	$<$0.001	0.9110	$<$0.001
Dbp	0.9572	$<$0.001	0.9273	$<$0.001

The variables in Table 4.7 are grouped according to the newly derived cut-off point and all of these variables have a p-value < 0.05. The distribution is assummed to be non-normal, therefore, the data must be transformed.

Table 4.8: Test for the difference betweern two means for anthropometric and fasting biochemical parameters (newly derived cut-off point)

	WC≥88cm	WC<88cm	P-value
Number (%)	483(45)	599(55)	-
Age	3.9780±0.2533	3.9134±0.2723	<0.001
WC	3.4466±0.1823	3.1587±0.1661	<0.001
BMI	4.5844±0.0863	4.3478±0.0860	< 0.001
Trig	0.1627±0.5079	-0.0427±0.4410	<0.001
Hdl	0.1723±0.1726	0.1865±0.1790	0.021
Gluc	1.6892±0.3745	1.5836±0.3227	<0.001
Sbp	4.9462±0.1586	4.8580±0.1570	<0.001
Dbp	4.4947±0.1422	4.4173±0.1359	<0.001

The data in Table 4.8 is presented as mean ± standard deviation. There are 483 participants with waist circumference greater or equal to 88cm. Participants with waist circumference greater or equal to 88cm had higher mean values than those with waist circumference less than 88cm for age, BMI, WC, Trig, Gluc, Sbp and Dbp (all with p-values < 0.001). However, participants with waist circumference greater or equal to 88cm had a lower mean value than those with waist circumference less than 88cm for Hdl (with p-value 0.021).

Table 4.9: Normality test using Shapiro-Wilk test

	WC≥80cm		WC<80cm	
Variable	Statistic	P-value	Statistic	P-value
Number (%)	744 (69)		338 (31)	
Age	0.9784	<0.001	0.9528	<0.001
BMI	0.9246	<0.001	0.9228	<0.001
WC	0.9185	<0.001	0.9392	<0.001
Trig	0.7944	<0.001	0.7672	<0.001
Hdl	0.8325	<0.001	0.8880	<0.001
Gluc	0.5729	<0.001	0.4786	<0.001
Sbp	0.9426	<0.001	0.8935	<0.001
Dbp	0.9637	<0.001	0.9674	<0.001

The variables in Table 4.9 are grouped according to the European standard and all of these variables have a p-value < 0.05. The distribution is assummed to be non-normal, therefore, the data must be transformed.

Table 4.10: Test for the difference betweern two means for anthropometric and fasting biochemical parameters (European standard)

	WC≥80cm	WC<80cm	P-value
Number (%)	744(69)	338(31)	-
Age	3.9683±0.2551	3.8848±0.2801	<0.001
WC	3.3776±0.1938	3.0885±0.1475	<0.001
BMI	4.5288±0.1040	4.2876±0.0649	<0.001
Trig	0.1085±0.4880	-0.0820±0.4443	<0.001
Hdl	0.1721±0.1773	0.1979±0.1727	0.04
Gluc	1.6629±0.3613	1.5600±0.3146	<0.001
Sbp	4.9275±0.1637	4.8310±0.1428	0.001
Dbp	4.4732±0.1434	4.4046±0.1336	<0.001

The data in Table 4.10 is presented as mean ± standard deviation. There are 744 participants with waist circumference greater or equal to 80cm. Participants with waist circumference greater or equal to 80cm had higher mean

values than those who are not suffering from metabolic syndrome for age, BMI, WC, Trig, Gluc, Sbp and Dbp (all with p-values < 0.001). However, participants with waist circumference greater or equal to 80cm had a lower mean value than those with waist circumference less than 80cm for Hdl (with p-value $= 0.04$).

4.5 Direct evaluation

In this section, we want to evaluate which cut-off point best discriminate metabolic abnormalities among the female population of Dikgale area. This is achieved using the 2005 and 2012 data.

4.5.1 Descriptive statistics for the evaluation data

Table 4.11: Test for the difference betweern two means for anthropometric and fasting biochemical parameters (at least three components of MS) for 2005 dataset

	MS	NoMS	P-value
Number (%)	62(13)	418(87)	-
Age	58.53±16.76	37.58±19.83	<0.001
BMI	29.74±6.49	25.83±5.84	<0.001
WC	95.24±12.78	82.07±13.97	<0.001
Trig	2.59±1.23	1.09±0.54	<0.001
Hdl	1.02±0.24	1.37±0.68	<0.001
Gluc	7.76±5.29	4.74±1.44	< 0.001
Sbp	144.26±23.38	115.33±20.22	< 0.001
Dbp	93.50±12.02	76.79±12.26	<0.001

The data in Table 4.11 was collected from 480 female participants in 2005 and presented as mean $\pm$ standard deviation. There are 62 participants suffering from metabolic syndrome. Participants who are suffering from metabolic syn-

drome had higher mean values than those who are not suffering from metabolic syndrome for age, BMI, WC, Trig, Gluc, Sbp and Dbp (all with p-values $<$ 0.001). However, participants who are suffering from metabolic syndrome had a lower mean value than those who are not suffering from metabolic syndrome for Hdl (also with p-value $<$ 0.001).

Table 4.12: Prevalences (2005 dataset)

Variable	Number (%)
MS	62 (13)
Trig	101 (21)
Hdl	263 (55)
Gluc	59 (12)
Sbp	110 (23)
Dbp	147 (31)
BP	163 (34)

Table 4.12 presents the prevalence of metabolic syndrome and the individual components of metabolic syndrome. The prevalence of metabolic syndrome is 13%. The most prevalent component is high-density lipoprotein with 55%, followed by blood pressure with 34%, then triglycerides with 21% and lastly, glucose with 12%.

Table 4.13: Test for the difference betweern two means for anthropometric and fasting biochemical parameters (at least three components of MS) for 2012 dataset

	MS	NoMS	P-value
Number (%)	115 (22)	406 (78)	-
Age	58.40±16.42	48.22±19.99	<0.001
BMI	29.95±8.63	26.11±7.29	<0.001
WC	89.73±31.89	79.98±26.82	<0.001
Trig	2.20±0.96	1.11±0.58	<0.001
Hdl	1.12±0.21	1.62±4.98	<0.001
Gluc	7.44±3.79	5.29±1.91	< 0.001
Sbp	144.29±23.56	124.85±24.08	< 0.001
Dbp	89.14±13.24	79.08±15.59	<0.001

The data in Table 4.13 was collected from 521 female participants in 2012 and presented as mean ± standard deviation. There are 115 participants suffering from metabolic syndrome. Participants who are suffering from metabolic syndrome had higher mean values than those who are not suffering from metabolic syndrome for age, BMI, WC, Trig, Gluc, Sbp and Dbp (all with p-values $<$ 0.001). However, participants who are suffering from metabolic syndrome had a lower mean value than those who are not suffering from metabolic syndrome for Hdl (also with p-value $<$ 0.001).

Table 4.14: Prevalences (2012 dataset)

Variable	Number (%)
MS	115 (22)
Trig	130 (25)
Hdl	277 (53)
Gluc	206 (40)
Sbp	206 (40)
Dbp	205 (39)
BP	266 (51)

Table 4.14 presents the prevalence of metabolic syndrome and the individual components of metabolic syndrome. The prevalence of metabolic syndrome is 22%. The most prevalent component is high-density lipoprotein with 53%, followed by blood pressure with 51%, then glucose with 40% and lastly, triglycerides with 25%.

Table 4.15: Predictions for WC≥88cm and WC≥80cm

Year	2005		2012	
Predictor	WC≥88	WC≥80	WC≥88	WC≥80
Sensitivity	0.7419	0.8710	0.7130	0.8174
Specificity	0.6794	0.4928	0.6133	0.3990
Accuracy	0.6875	0.5417	0.6353	0.4914
Kappa	0.2327	0.1515	0.2354	0.1277
Balanced Accuracy	0.7107	0.6819	0.6632	0.6082

Table 4.15 presents data classification for the newly derived cut-off point and the European standard using the 2005 and 2012 datasets. Sensitivity is lower for the newly derived cut-off point than the European standard and specificity is the opposite, it is higher for the newly derived cut-off point than the European standard. However, accuracy which is the overall classification and

balanced accuracy which is the average sensitivity and specificity are higher for the newly derived cut-off point than the European standard. Therefore we can conclude that the newly derived cut-off point best discriminates metabolic abnormalities among adults female population of the Dikgale area.

4.6 Application and evaluation of the models

In this section we will apply three models, logistic regression (LR), linear discriminant analysis (LDA) and the random forest (RF) to predict the classification of participants with and without MS. This section is divided into two subsections. In the first part, we fit the models to determine the cut-off point (between the European standard and the newly derived) that best discrimanate participants with and without MS. The evaluation of the fitted models will be done using the 20% from data partitioning. Further evaluation will be done using the independent datasets for 2005 and 2012. In the second subsection, we assess the performance of these models and compare them. In fitting these models, we first partition the 1997 dataset into 80:20 ratio, where 80 is the training data and 20 the testing data. Secondly, we apply the cross-validation method with k=5 (k is the number of folds). Each model will be fitted twice, using the newly derived cut-off point and the European standard.

4.6.1 Logistic regression model

This section presents the results of the binary logistic regression model fitted to correctly predict participants with (y=1) and without (y=0) MS. For the newly derived cut-off point, MS is defined as having WC$\geq$88cm and any two components of MS, whereas for the European standard MS is defined as having WC$\geq$80cm and any two components of MS.

Table 4.16: Fitting logistic regression model (Newly derived)

Coefficients	Estimate	SE	P-value
Intercept	5.0762	0.8469	<0.0001
Age	0.0002	0.0072	0.9796
Trig	-0.7150	0.1386	<0.0001
Gluc	-0.1002	0.0295	<0.0001
Hdl	3.1663	0.4911	<0.0001
Sbp	-0.0158	0.0060	<0.0001
Dbp	-0.0494	0.0111	<0.0001

Table 4.16 presents the y-intercept and the coefficients of variables used to fit binary logistic regression model for the newly derived cut-off point. The fitted logistic regression model can be expressed as:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

$$= \beta_0 + \beta_1 Age + \beta_2 Trig + \beta_3 Gluc + \beta_4 Hdl + \beta_5 Sbp + \beta_6 Dbp$$

$$= 5.0762 + 0.0002 Age - 0.7150 Trig - 0.1002 Gluc + 3.1663 Hdl - 0.0158 Sbp - 0.0494 Dbp$$

All variables were significant except for age with a p-value larger than 0.05. Metabolic syndome and triglycerides, glucose, systolic and diastolic blood pressure have a negative relationship. According to the IDF, a person is at risk of metabolic syndrome if their triglycerides is $\geq$ 1.7 mmol/l, glucose is $\geq$ 5.6 mmol/l, high-density lipoprotein cholesterol is $<$ 1.29 mmol/l, systolic blood pressure is $\geq$ 130 mmHg, and diastolic blood pressure is $\geq$ 85 mmHg (Alberti et al., 2006). Therefore, the logistic regression coefficients mean that:

- If the participant's age increases by 1 year then the likelihood that they will have MS increases by the log odds of 0.0002.
- If the participant's level of triglycerides decreases by 0.1 mmol/l then the likelihood that they will have MS decreases by the log odds of 0.7150.

- If the participant's level of glucose decreases by 0.1 mmol/l then the likelihood that they will have MS decreases by the log odds of 0.1002.
- If the participant's level of high-density lipoprotein cholesterol increases by 0.1 mmol/l then the likelihood that they will have MS increases by the log odds of 3.1663.
- If the participant's level of systolic blood pressure decreases by 1 mmHg then the likelihood that they will have MS decreases by the log odds of 0.0158.
- If the participant's level of diastolic blood pressure decreases by 1 mmHg then the likelihood that they will have MS decreases by the log odds of 0.0494.

Table 4.17: Fitting logistic regression model (European standard)

Coefficients	Estimate	SE	P-value
Intercept	3.9368	0.8575	< 0.001
Age	-0.0050	0.0069	0.4668
Trig	-0.8046	0.1554	< 0.001
Gluc	-0.2306	0.0487	< 0.001
Hdl	4.6529	0.5289	< 0.001
Sbp	-0.0209	0.0063	< 0.001
Dbp	0.0438	0.0112	< 0.001

Table 4.17 presents the y-intercept and the coefficients of variables used to fit binary logistic regression model for the European standard. Then, the fitted logistic regression model can be expressed as:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

$$= \beta_0 + \beta_1 Age + \beta_2 Trig + \beta_3 Gluc + \beta_4 Hdl + \beta_5 Sbp + \beta_6 Dbp$$

$$= 3.9368 - 0.0050Age - 0.8046Trig - 0.2306Gluc + 4.6529Hdl - 0.0209Sbp + 0.0438Dbp$$

Table 4.18: Logistic regression model predictions

	20% dataset		2005 dataset		2012 dataset	
	WC≥88	WC≥80	WC≥88	WC≥80	WC≥88	WC≥80
Sensitivity	0.6667	0.8125	0.7258	0.8871	0.6000	0.8261
Specificity	0.8929	0.7500	0.9761	0.9043	0.9409	0.8424
PPV	0.6400	0.4815	0.8182	0.5789	0.7419	0.5975
NPV	0.9036	0.9333	0.9600	0.9818	0.8925	0.9448
Accuracy	0.8426	0.7639	0.9438	0.9021	0.8656	0.8388

It can be seen in Table 4.18 that sensitivity and the negative predictive value are higher for the European standard than the newly derived cut-off, but specificity and the positive predictive value are higher for the newly derived cut-off point than the European standard. However, accuracy, the overall model prediction is higher for the newly derived cut-off point than the European standard. Therefore we can conclude that the newly derived cut-off point best discriminate participants with and without MS than the European standard.

4.6.2 Linear discriminant analysis model

This section presents the results of a linear discriminant analysis model that classifies participants into two groups, with and without metabolic syndrome. This gives a linear combination of the original variables that provide the best possible separation between the two groups.

Table 4.19: Prior probabilites

	Newly derived	European standard
MS	0.2968	0.4088
NoMS	0.7032	0.5912

Table 4.19 presents the proportion of each group using the newly derived cut-off point and the European standard. According to the newly derived cut-off point, 29.68% participants have MS and 70.32% participants do not have MS. Using the European standard 40.88% participants have MS and 59.12% participants do not have MS.

Table 4.20: Group means

	Newly derived		European standard	
	MS	NoMS	MS	NoMS
Age	6.0019	3.9149	4.0010	3.8990
Trig	0.2416	-0.0358	0.2131	-0.0686
Gluc	1.7697	1.5708	1.7597	1.5400
Hdl	0.1279	0.2000	0.1191	0.2198
Sbp	4.9889	4.8592	4.9714	4.8466
Dbp	4.5302	4.4146	4.5122	4.4051

The mean values for each predictor variable of each group are presented in Table 4.20.

Table 4.21: Coefficients of linear dicriminants

	Newly derived	European standard
Age	0.1201	-0.2157
Trig	-0.7226	-0.6602
Gluc	-1.0561	-1.2396
Hdl	2.6437	3.6108
Sbp	-2.2939	-2.3643
Dbp	-3.2102	-2.5493

The coefficients of the linear dicriminants in Table 4.21 display the linear combination of the predictor variables that form the decision rule of the linear

discriminant analysis model. Thus the discriminant function of the linear combination using the newly derived cut-off point is given by:

$$0.1201Age - 0.7226Trig - 1.0561Gluc + 2.6437Hdl - 2.2939Sbp - 3.2102Dbp$$

and the discriminant function of the linear combination using the European standard is given by:

$$-0.2157Age - 0.6602Trig - 1.2396Gluc + 3.6108Hdl - 2.3643Sbp - 2.5493Dbp$$

Table 4.22: Linear discriminant analysis model predictions

	20% dataset		2005 dataset		2012 dataset	
	WC≥88	WC≥80	WC≥88	WC≥80	WC≥88	WC≥80
Sensitivity	0.6458	0.8333	0.6774	0.8548	0.5565	0.8261
Specificity	0.9167	0.7500	0.9833	0.9091	0.9409	0.8448
PPV	0.6889	0.4878	0.8571	0.5824	0.7273	0.6013
NPV	0.9006	0.9403	0.9536	0.9769	0.8822	0.9449
Accuracy	0.8565	0.7685	0.9438	0.9021	0.8560	0.8407

It can be seen in Table 4.22 that sensitivity and the negative predictive value are higher for the European standard than the newly derived cut-off, but specificity and the positive predictive value are higher for the newly derived cut-off point than the European standard. However, accuracy, the overall model prediction is higher for the newly derived cut-off point than the European standard. Therefore we can conclude that the newly derived cut-off point best discriminate participants with and without MS than the European standard.

4.6.3 Random forest model

This section presents the results of the random forest model of a classification type since we are classifying participants into two groups, with and without MS. The optimised number of random variables to be tried at each split was found to be 4 using the tuneRF function. The default number of trees is 500 and the results were based on the compilation of all 500 decision trees.

Table 4.23: Random forest model results

	Newly derived		European standard	
	MS	NoMS	MS	NoMS
Class error	0.3658	0.1658	0.0960	0.1777
Error rate	22.52%		14.43%	

Table 4.23 presents the results of the random forest model. The model error rate for the newly derived cut-off point is much quiet higher than the model error rate for the European standard. This implies that the model accurary is higher when using the European standard than the newly derived cut-off point. However, when testing the model using new datasets, the overall prediction is quite higher for the newly derived cut-off point than the European standard. The model prediction results are presented in Table 4.24.

Table 4.24: Random forest model predictions

	20% dataset		2005 dataset		2012 dataset	
	WC≥88	WC≥80	WC≥88	WC≥80	WC≥88	WC≥80
Sensitivity	0.7708	0.9583	0.6936	0.8710	0.6348	0.8783
Specificity	0.8095	0.6190	0.9546	0.8876	0.9064	0.7463
PPV	0.5362	0.4182	0.6936	0.5347	0.6577	0.4951
NPV	0.9252	0.9811	0.9546	0.9789	0.8976	0.9558
Accuracy	0.8009	0.6944	0.9208	0.8854	0.8464	0.7754

It can be seen in Table 4.24 that sensitivity and the negative predictive value

are higher for the European standard than the newly derived cut-off, but specificity and the positive predictive value are higher for the newly derived cut-off point than the European standard. However accuracy the overall model prediction is higher for the newly derived cut-off point than the European standard. Therefore we can conclude that the newly derived cut-off point best discriminate participants with and without MS than the European standard.

4.6.4 Model selection

This section focuses on comparing the three different fitted models and selecting the model that best discrimanate participants with MS.

Table 4.25: Prediction measures (2005 dataset)

	LR	LDA	RF
Accuracy	0.9438	0.9438	0.9208
Kappa	0.7373	0.7254	0.6481
F1 score	0.7692	0.7567	0.6936

Table 4.26: Prediction measures (2012 dataset)

	LR	LDA	RF
Accuracy	0.8656	0.8560	0.8464
Kappa	0.5807	0.5431	0.5480
F1 score	0.6634	0.6305	0.6460

Table 4.25 and 4.26 present the prediction measures for the three models, LR, LDA and RF using the 2005 and 2012 datasets. The model assessment measures, accuracy which is the overall classification, kappa which is the level of agreement between the actual and the predicted and F1 score which is the pre-

dictive power of the classification procedure are all higher in LR than in LDA and RF models.

4.6.5 Sensitivity analysis

Sensitivity analysis assesses the impact on model outputs of independently changing a single model input throughout the course of its whole range of likely values (Razavi et al., 2021). Sensitivity analysis in the context of a classification model usually includes analysing how changes in the model input data affect the model's performance. This section presents the model's performance when the waist circumference cut-off point varies.

Table 4.27: Model performance

	Different thresholds						
	WC75	WC80	WC85	WC88	WC90	WC95	WC100
Sensitivity	0.9516	0.8871	0.7903	0.7258	0.6129	0.3710	0.2419
Specificity	0.8421	0.9043	0.9498	0.9761	0.9857	0.99520	1.0000
Accuracy	0.8562	0.9021	0.9292	0.9438	0.9375	0.9146	0.9021
Kappa	0.5540	0.6452	0.7015	0.7373	0.6830	0.4909	0.3573

It can be seen in Table 4.27 that as the waist circumference decreases from the derived cut-off point, sensitivity increases and as the waist circumference increases from the derived cut-off point, sensitivity decreases. Specificity does the opposite of sensitivity. However, accuracy and kappa decrease on both sides of the derived cut-off point, making the model to perform best at our derived cut-off point.

4.7 Summary of the chapter

In this chapter we were determining the optimal cut-off point that detects metabolic disorders in ga-Dikgale area, Limpopo province. The data was tested for normality. Direct evaluation was performed. The three models were fitted using the 1997 dataset. The evaluation was done using the test data from the split of 80:20 ratio, where 80 was the training data and 20 the testing data. Further evaluation was done using the 2005 and 2012 datasets. The data were analysed using R software package. The findings are as follows:

1. The prevalence of MS was found to be low when compared with International Diabetes Federation prescribed definition of MS.

2. The distribution of the data is positive, sharp-peaked and heavy-tailed as shown by skewness and kurtosis values.

3. The Shapiro-wilk test was used to test for normality of the data. The findings revealed that some of the variable were not normally distributed.

4. The optimal cut-off point was obtained using ROC curve analysis and it was found to be 88cm with an area under the ROC curve 0.7 with implies that this cut-off point can better discriminate participants with and without MS.

5. Direct evaluation was done to check which cut-off point between the European standard and the newly derived best discriminate participants with and without MS for this study group. The findings revealed that the newly derived cut-off point discriminate participants with and without MS than the European standard.

6. The three classification models: logistic regression,- linear discriminant analysis and random forest were fitted to check which cut-off point between the European standard and the newly derived best discriminate

participants with and without MS. The findings also revealed that the newly derived cut-off point discriminate participants with and without MS than the European standard.

7. These models were evaluated to check which one performs better. The findings revealed that logistics regression performs better.

8. Sensitivity analysis was done to assess the model's performance at different thresholds. The findings revealed that the model performs better at the newly derived cut-off point.

Chapter 5

Conclusion

5.1 Conclusion

It has been argued that the European based WC cut-off point, as agreed by the International Diabetes Federations may not be good enough to detect MS among the African population. Based on this premise, this study set out to determine if there is a WC cut-off point that will improve on the IDF pronoucements in discriminating between African women with MS from those without. The first step was to identify women suffering from MS using the ATP III definition that was set as a starting point to modify and update to the IDF definition of MS. In this definition, one suffers from MS if they presented at least three of the following components: Trig $\geq$1.7 mmol/l, Hdl $<$1.29 mmol/l, Gluc $\geq$ 5.6 mmol/l and Sbp $\geq$ 130 mmHg or Dbp $\geq$ 85 mmHg. In this case there were 209 (19%) participants suffering from MS.

The second step involved deriving a sample-based cut-off point using the receiver operating characteristics curve analysis, an approach that was used by

Crowther and Norris (2012). A cut-off point of 88 cm was obtained. The number of participants with waist circumference of 88 cm were identified and were found to be 483 (45%). Of the 483, those who had MS were 142 (29%). Lastly, we used machine learning methods to determine a cut-off point that best discriminates the classification of having metabolic syndrome and non-metabolic syndrome between the new cut-off point of 88 cm and the European based cut-off point of 80 cm. Machine learning methods include Logistic regression (LR), linear discriminant analysis (LDA) and random forest (RF). It was shown that the newly derived cut-off point of 88 cm was better than the European based cut-off point of 80 cm at discriminating MS based on prediction accuracy, specificity, and positive predictive value. Using 5-fold validation, the LR, LDA, and LR gave a prediction accuracy of 0.8656, 0.8560, and 0.8464, respectively using the 2012 dataset. The LR method was also found to be the best discriminator of MS with the highest prediction accuracy of 0.8656. Sensitivity analysis showed that the cut-off point is robust with prediction accuracy reducing as one moves either side of the derived (88 cm) cut-off point.

According to the studies conducted in the other parts of the Sub-Saharan Africa, the following cut-off points were obtained for women: in Angola 80.5 cm by Magalhães et al. (2014); in Ethiopia 78 cm by Sinaga et al. (2018); and in Botswana 82.3 cm by Tladi et al. (2020) among others. These cut-off points are less than the derived cut-off point of 88 cm and closer to the European cut-off point of 80 cm as compared to 88 cm. However, for studies conducted in South Africa, the following cut-off points were obtained for women: in KwaZulu-Natal 92 cm by Motala et al. (2011); in Johannesburg 91.5 cm by Crowther and Norris (2012); and in the Eastern Cape 89.45 cm by Owolabi et al. (2018). These cut-off points are above, but closer to the derived cut-off point of 88 cm as compared to the European cut-off point of 80 cm.

Most studies used only receiver operating characteristics curve analysis to determine the optimal cut-off point. In this study we further applied sensitivity analysis to determine if the derived cut-off point occurred at the true inflection points. Additionally, the was no study found discriminating MS and non-MS participants using machine learning methods for three independent samples from the same population taken 7 years apart.

It is important to establish population-specific cut-off points to accurately detect metabolic syndrome in diverse populations. Correct subject classification is crucial since misdiagnosing a patient can result in consequences. For example, a false negative classification may lead to expanding or spreading disease as a result of not taking medicine and a false positive classification may lead to a reaction to a prescription drug. The findings from this study will assist obtaining health outcomes quick and cost-effective by using a tape measure to determine obesity from cut-off point of 88 cm, instead of doing biochemical tests. The findings will also be helpful in developing more effective public health interventions.

5.2 Recommendations of the study

The findings of this study will guide health departments in raising awareness tailored to specific population groups to mitigate obesity challenges.

5.3 Limitations of the study

The study's cross-sectional design precludes causal inferences, and the sample was limited to one rural area.

5.4 Future studies

Further research across Sub-Saharan Africa is needed to establish a common WC cut-off point, given the regional variations identified. Additionally, machine learning models such as gradient boosting machines should be explored for their sequential error-correcting capabilities.

References

ABUBAKARI, A.-R., LAUDER, W., AGYEMANG, C., JONES, M., KIRK, A., AND BHOPAL, R. (2008). Prevalence and time trends in obesity among adult West African populations: a meta-analysis. *Obesity Reviews*, **9** (4), 297–311.

AGUILAR-SALINAS, C. A., ROJAS, R., GÓMEZ-PÉREZ, F. J., VALLES, V., RÍOS-TORRES, J. M., FRANCO, A., OLAIZ, G., RULL, J. A., AND SEPULVEDA, J. (2003). Analysis of the agreement between the World Health Organization criteria and the National Cholesterol Education Program-III definition of the metabolic syndrome: results from a population-based survey. *Diabetes Care*, **26** (5), 1635–1635.

AGYEMANG, C., BOATEMAA, S., FREMPONG, G. A., AND DE GRAFT AIKINS, A. (2015). Obesity in Sub-Saharan Africa. Springer International Publishing Switzerland 2015.

AJAYI, I. O., ADEBAMOWO, C., ADAMI, H.-O., DALAL, S., DIAMOND, M. B., BAJUNIRWE, F., GUWATUDDE, D., NJELEKELA, M., NANKYA-MUTYOBA, J., CHIWANGA, F. S., ET AL. (2016). Urban–rural and geographic differences in overweight and obesity in four Sub-Saharan African adult populations: a multi-country cross-sectional study. *BMC Public Health*, **16** (1), 1–13.

ALBERTI, K. G., ECKEL, R. H., GRUNDY, S. M., ZIMMET, P. Z., CLEEMAN, J. I., DONATO, K. A., FRUCHART, J.-C., JAMES, W. P. T., LORIA, C. M., AND SMITH JR, S. C. (2009). Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epi-

demiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, **120** (16), 1640–1645.

ALBERTI, K. G. M. M., ZIMMET, P., AND SHAW, J. (2006). Metabolic syndrome—a new world-wide definition. a consensus statement from the International Diabetes Federation. *Diabetes Medicine*, **23** (5), 469–480.

AMIT, Y. AND GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, **9** (7), 1545–1588.

ANEKWE, C. V., JARRELL, A. R., TOWNSEND, M. J., GAUDIER, G. I., HISERODT, J. M., AND STANFORD, F. C. (2020). Socioeconomics of obesity. *Current Obesity Reports*, **9**, 272–279.

AZIZI, F., SALEHI, P., ETEMADI, A., AND ZAHEDI-ASL, S. (2003). Prevalence of metabolic syndrome in an urban population: Tehran lipid and glucose study. *Diabetes Research and Clinical Practice*, **61** (1), 29–37.

BALKAU, B. AND CHARLES, M. (1999). The European Group for the study of Insulin Resistance (EGIR) comment on the provisional report from the WHO consultation. *Diabetes Medicine*, **16** (5), 442–443.

BALKAU, B., VERNAY, M., MHAMDI, L., NOVAK, M., ARONDEL, D., TICHET, J., ESCHWEGE, E., GROUP, D. S., ET AL. (2003). The incidence and persistence of the National Cholesterol Education Program (NCEP) metabolic syndrome: The French D.E.S.I.R study. *Diabetes and Metabolism*, **29** (5), 526–532.

BALL, K., MISHRA, G., AND CRAWFORD, D. (2002). Which aspects of socioeconomic status are related to obesity among men and women? *International Journal of Obesity*, **26** (4), 559–565.

BARROSO, T. A., MARINS, L. B., ALVES, R., GONÇALVES, A. C. S., BARROSO, S. G., AND ROCHA, G. D. S. (2017). Association of central obesity with the incidence of cardiovascular diseases and risk factors. *International Journal of Cardiovascular Sciences*, **30**, 416–424.

BLÜHER, M. (2019). Obesity: global epidemiology and pathogenesis. *Nature Reviews Endocrinology*, **15** (5), 288–298.

BRAY, G., KIM, K., WILDING, J., AND FEDERATION, W. O. (2017). Obesity: a chronic relapsing progressive disease process. a position statement of the World Obesity Federation. *Obesity Reviews*, **18** (7), 715–723.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

CAMERON, A. J., MAGLIANO, D. J., ZIMMET, P. Z., WELBORN, T., AND SHAW, J. E. (2007). The metabolic syndrome in Australia: prevalence using four definitions. *Diabetes Research and Clinical Practice*, **77** (3), 471–478.

CAMERON, A. J., SHAW, J. E., AND ZIMMET, P. Z. (2004). The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and Metabolism Clinics*, **33** (2), 351–375.

CARR, D. B., UTZSCHNEIDER, K. M., HULL, R. L., KODAMA, K., RETZLAFF, B. M., BRUNZELL, J. D., SHOFER, J. B., FISH, B. E., KNOPP, R. H., AND KAHN, S. E. (2004). Intra-abdominal fat is a major determinant of the National Cholesterol Education Program Adult Treatment Panel III criteria for the metabolic syndrome. *Diabetes*, **53** (8), 2087–2094.

CASSELLS, H. B. AND HAFFNER, S. M. (2006). The metabolic syndrome: risk factors and management. *Journal of Cardiovascular Nursing*, **21** (4), 306–313.

CHIEN, K.-L., LEE, B.-C., HSU, H.-C., LIN, H.-J., CHEN, M.-F., AND LEE, Y.-T. (2008). Prevalence, agreement and classification of various metabolic

syndrome criteria among ethnic Chinese: a report on the hospital-based health diagnosis of the adult population. *Atherosclerosis*, **196** (2), 764–771.

CHOOI, Y. C., DING, C., AND MAGKOS, F. (2019). The epidemiology of obesity. *Metabolism*, **92**, 6–10.

COLLABORATORS, G. . O. (2017). Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine*, **377** (1), 13–27.

COOPER, A. J., GUPTA, S. R., MOUSTAFA, A. F., AND CHAO, A. M. (2021). Sex/gender differences in obesity prevalence, comorbidities, and treatment. *Current Obesity Reports*, 1–9.

CORNIER, M.-A., DABELEA, D., HERNANDEZ, T. L., LINDSTROM, R. C., STEIG, A. J., STOB, N. R., VAN PELT, R. E., WANG, H., AND ECKEL, R. H. (2008). The metabolic syndrome. *Endocrine Reviews*, **29** (7), 777–822.

CROWTHER, N. J. AND NORRIS, S. A. (2012). The current waist circumference cut point used for the diagnosis of metabolic syndrome in Sub-Saharan African women is not appropriate. *PloS One*, **7** (11), e48883.

DAKE, F. A., TAWIAH, E. O., AND BADASU, D. M. (2011). Sociodemographic correlates of obesity among Ghanaian women. *Public Health Nutrition*, **14** (7), 1285–1291.

DE HOOP, M. M., MOODLEY, M. V., AND MORRAN, M. D. (2017). Review of the national strategy for the prevention and control of obesity in South Africa (2015-2020). *Health Promotion*, **17**, 12.

DEEPA, R., SHANTHIRANI, C., PREMALATHA, G., SASTRY, N., AND MOHAN, V. (2002). Prevalence of insulin resistance syndrome in a selected South Indian population: the Chennai urban population study-7 [cups-7]. *The Indian Journal of Medical Research*, **115**, 118–27.

DELISLE, H., NTANDOU-BOUZITOU, G., AGUEH, V., SODJINOU, R., AND FAYOMI, B. (2012). Urbanisation, nutrition transition and cardiometabolic risk: the Benin study. *British Journal of Nutrition*, **107** (10), 1534–1544.

DIETTERICH, T. G. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, **32**, 1–22.

DOBBS, R. AND MANYIKA, J. (2015). The obesity crisis. *The Cairo Review of Global Affairs*, **5**, 44–57.

ECKEL, R. H., GRUNDY, S. M., AND ZIMMET, P. Z. (2005). The metabolic syndrome. *The Lancet*, **365** (9468), 1415–1428.

EKORU, K., MURPHY, G., YOUNG, E., DELISLE, H., JEROME, C., ASSAH, F., LONGO-MBENZA, B., NZAMBI, J., ON'KIN, J., BUNTIX, F., ET AL. (2018). Deriving an optimal threshold of waist circumference for detecting cardiometabolic risk in Sub-Saharan Africa. *International Journal of Obesity*, **42** (3), 487–494.

ERASMUS, R. T., SOITA, D. J., HASSAN, M. S., BLANCO-BLANCO, E., VERGOTINE, Z., KENGNE, A. P., AND MATSHA, T. E. (2012). High prevalence of diabetes mellitus and metabolic syndrome in a South African coloured population: Baseline data of a study in Bellville, Cape Town. *South African Medical Journal*, **102** (11), 841–844.

FAIJER-WESTERINK, H. J., KENGNE, A. P., MEEKS, K. A., AND AGYEMANG, C. (2020). Prevalence of metabolic syndrome in Sub-Saharan Africa: A systematic review and meta-analysis. *Nutrition, Metabolism and Cardiovascular Diseases*, **30** (4), 547–565.

FEZEU, L., BALKAU, B., KENGNE, A.-P., SOBNGWI, E., AND MBANYA, J.-C. (2007). Metabolic syndrome in a Sub-Saharan African setting: central obesity may be the key determinant. *Atherosclerosis*, **193** (1), 70–76.

GBADAMOSI, M. A. AND TLOU, B. (2020). Modifiable risk factors associated with non-communicable diseases among adult outpatients in Manzini, Swaziland: a cross-sectional study. *BMC Public Health*, **20**, 1–12.

GEBREYES, Y. F., GOSHU, D. Y., GELETEW, T. K., ARGEFA, T. G., ZEMEDU, T. G., LEMU, K. A., WAKA, F. C., MENGESHA, A. B., DEGEFU, F. S., DEGHEBO, A. D., ET AL. (2018). Prevalence of high bloodpressure, hyperglycemia, dyslipidemia, metabolic syndrome and their determinants in Ethiopia: Evidences from the National NCDs STEPS survey, 2015. *PloS One*, **13** (5), e0194819.

GOH, L. G., DHALIWAL, S. S., WELBORN, T. A., LEE, A. H., AND DELLA, P. R. (2014). Anthropometric measurements of general and central obesity and the prediction of cardiovascular disease risk in women: a cross-sectional study. *BMJ Open*, **4** (2), e004138.

GRUNDY, S. M. (2004a). Obesity, metabolic syndrome, and cardiovascular disease. *The Journal of Clinical Endocrinology and Metabolism*, **89** (6), 2595–2600.

GRUNDY, S. M. (2004b). What is the contribution of obesity to the metabolic syndrome? *Endocrinology and Metabolism Clinics*, **33** (2), 267–282.

GRUNDY, S. M. (2008). Metabolic syndrome pandemic. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **28** (4), 629–636.

GUPTA, A., GUPTA, R., SARNA, M., RASTOGI, S., GUPTA, V., AND KOTHARI, K. (2003). Prevalence of diabetes, impaired fasting glucose and insulin resistance syndrome in an urban Indian population. *Diabetes Research and Clinical Practice*, **61** (1), 69–76.

GYAKOBO, M., AMOAH, A. G., MARTEY-MARBELL, D.-A., AND SNOW, R. C. (2012). Prevalence of the metabolic syndrome in a rural population in Ghana. *BMC Endocrine Disorders*, **12**, 1–9.

HAN, T. S. AND LEAN, M. E. (2015). Metabolic syndrome. *Medicine*, **43** (2), 80–87.

HANLEY, J. A. AND MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143** (1), 29–36.

HEALTH, N. AND AFRICA, D. (2015). Strategy for the prevention and control of obesity in South Africa 2015–2020.

HEART, N., LUNG, INSTITUTE, B., OF DIABETES, N. I., AND (US), K. D. (1998). *Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: the evidence report*. 98.

HO, T. K. (1995). Random decision forests. *In Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1. IEEE, pp. 278–282.

HOSMER, D. W. AND LEMESHOW, S. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

HUANG, P. L. (2009). A comprehensive definition for metabolic syndrome. *Disease Models and Mechanisms*, **2** (5-6), 231–237.

JEBB, S. (2004). Obesity: causes and consequences. *Women's Health Medicine*, **1** (1), 38–41.

KANTER, R. AND CABALLERO, B. (2012). Global gender disparities in obesity: a review. *Advances in Nutrition*, **3** (4), 491–498.

KHAN, L. K. AND BOWMAN, B. (1999). Obesity: a major global public health problem. *Annual Review of Nutrition*, **19** (1), xiii–xvii.

KOSTER, A., LEITZMANN, M. F., SCHATZKIN, A., MOUW, T., ADAMS, K. F., VAN EIJK, J. T. M., HOLLENBECK, A. R., AND HARRIS, T. B. (2008). Waist

circumference and mortality. *American Journal of Epidemiology*, **167** (12), 1465–1475.

KRAMER, C. K., ZINMAN, B., AND RETNAKARAN, R. (2013). Are metabolically healthy overweight and obesity benign conditions? a systematic review and meta-analysis. *Annals of Internal Medicine*, **159** (11), 758–769.

KRUGER, A., WISSING, M. P., TOWERS, G. W., AND DOAK, C. M. (2012). Sex differences independent of other psycho-sociodemographic factors as a predictor of body mass index in black South African adults. *Journal of health, population, and nutrition*, **30** (1), 56.

KRUGER, H. S., PUOANE, T., SENEKAL, M., AND VAN DER MERWE, M.-T. (2005). Obesity in South Africa: challenges for government and health professionals. *Public Health Nutrition*, **8** (5), 491–500.

KUMAR, R. AND INDRAYAN, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, **48**, 277–287.

KYLIN, E. (1923). Studies of the hypertension-hyperglycemia-hyperuricemia syndrome. *Zentralbl Inn Med*, **44**, 105–27.

KYROU, I., RANDEVA, H. S., TSIGOS, C., KALTSAS, G., AND WEICKERT, M. O. (2018). Clinical problems caused by obesity. *Endotext [Internet]*.

LAWES, C. M., VANDER HOORN, S., RODGERS, A., ET AL. (2008). Global burden of blood-pressure-related disease, 2001. *The Lancet*, **371** (9623), 1513–1518.

LEMESHOW, S., STURDIVANT, R. X., AND HOSMER JR, D. W. (2013). *Applied logistic regression*. John Wiley & Sons.

LEO, B. (2001). Random forests. *Machine Learning*, **45** (1), 5–32.

LEVENSON, J. W., SKERRETT, P. J., AND GAZIANO, J. M. (2002). Reducing the global burden of cardiovascular disease: the role of risk factors. *Preventive Cardiology*, **5** (4), 188–199.

LOPEZ, A. D., MATHERS, C. D., EZZATI, M., JAMISON, D. T., AND MURRAY, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, **367** (9524), 1747–1757.

LUSTED, L. B. (1971). Signal detectability and medical decision-making: Signal detectability studies help radiologists evaluate equipment systems and performance of assistants. *Science*, **171** (3977), 1217–1219.

MAGALHÃES, P., CAPINGANA, D. P., AND MILL, J. G. (2014). Prevalence of the metabolic syndrome and determination of optimal cut-off values of waist circumference in university employees from angola: cardiovascular topic. *Cardiovascular Journal of Africa*, **25** (1), 27–33.

MALHOTRA, R., HOYO, C., ØSTBYE, T., HUGHES, G., SCHWARTZ, D., TSOLEKILE, L., ZULU, J., AND PUOANE, T. (2008). Determinants of obesity in an urban township of south africa. *South African Journal of Clinical Nutrition*, **21** (4), 315–320.

MANAFE, M., CHELULE, P. K., AND MADIBA, S. (2022). The perception of overweight and obesity among South African adults: Implications for intervention strategies. *International Journal of Environmental Research and Public Health*, **19** (19), 12335.

MASIBO, P., BULUKU, E., MENYA, D., AND MALIT, V. (2013). Prevalence and determinants of under-and over-nutrition among adult kenyan women; evidence from the Kenya Demographic and Health Survey 2008–09. *East Afr J Public Health*, **10**, 611–622.

MATTA, J., CARETTE, C., CZERNICHOW, S., ET AL. (2018). French and worldwide epidemiology of obesity. *Presse Medicale (Paris, France: 1983)*, **47** (5), 434–438.

MAY, J. (2000). Poverty and inequality in South Africa: Meeting the challenge. *(No Title)*.

MEHRZAD, R. (2020). The global impact of obesity. *In Obesity*. Elsevier, pp. 55–72.

MEIGS, J. B., WILSON, P. W., NATHAN, D. M., D'AGOSTINO SR, R. B., WILLIAMS, K., AND HAFFNER, S. M. (2003). Prevalence and characteristics of the metabolic syndrome in the San Antonio Heart and Framingham Offspring Studies. *Diabetes*, **52** (8), 2160–2167.

MICKLESFIELD, L. K., LAMBERT, E. V., HUME, D. J., CHANTLER, S., PIENAAR, P. R., DICKIE, K., GOEDECKE, J. H., AND PUOANE, T. (2013). Sociocultural, environmental and behavioural determinants of obesity in black South African women: review articles. *Cardiovascular Journal of Africa*, **24** (9), 369–375.

MOTADI, S., VELDSMAN, T., MOHLALA, M., AND MABAPA, N. (2018). Overweight and obesity among adults aged 18–45 years residing in and around Giyani town in Mopani district of Limpopo province, South Africa. *J Nutr Heal Sci*, **5** (1), 102.

MOTALA, A. A., ESTERHUIZEN, T., PIRIE, F. J., AND OMAR, M. A. (2011). The prevalence of metabolic syndrome and determination of the optimal waist circumference cutoff points in a rural South African community. *Diabetes Care*, **34** (4), 1032–1037.

MUHIHI, A. J., NJELEKELA, M. A., MPEMBENI, R., MWIRU, R. S., MLIGILICHE, N., AND MTABAJI, J. (2012). Obesity, overweight, and perceptions

about body weight among middle-aged adults in Dar es Salaam, Tanzania. *International Scholarly Research Notices*, **2012**.

NCEP (2001). Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *Jama*, **285** (19), 2486–2497.

NELDER, J. A. AND WEDDERBURN, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135** (3), 370–384.

NESS-ABRAMOF, R. AND APOVIAN, C. M. (2008). Waist circumference measurement in clinical practice. *Nutrition in Clinical Practice*, **23** (4), 397–404.

NEUPANE, S., KC, P., AND DOKU, D. T. (2015). Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African countries. *BMC Public Health*, **16**, 1–9.

NICHOLS, G. A., HILLIER, T. A., AND BROWN, J. B. (2008). Normal fasting plasma glucose and risk of type 2 diabetes diagnosis. *The American Journal of Medicine*, **121** (6), 519–524.

OBUCHOWSKI, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology*, **229** (1), 3–8.

OFORI, E. K. AND ANGMORTERH, S. K. (2019). Relationship between physical activity, body mass index (BMI) and lipid profile of students in Ghana. *The Pan African Medical Journal*, **33**.

OGDEN, C. L., CARROLL, M. D., FRYAR, C. D., AND FLEGAL, K. M. (2015). Prevalence of obesity among adults and youth: United States, 2011-2014.

OKOP, K. J., MUKUMBANG, F. C., MATHOLE, T., LEVITT, N., AND PUOANE, T. (2016). Perceptions of body size, obesity threat and the willingness to lose

weight among black South African adults: a qualitative study. *BMC Public Health*, **16** (1), 1–13.

OKWUOSA, I. S., LEWSEY, S. C., ADESIYUN, T., BLUMENTHAL, R. S., AND YANCY, C. W. (2016). Worldwide disparities in cardiovascular disease: challenges and solutions. *International Journal of Cardiology*, **202**, 433–440.

ORGANIZATION, W. H. (2000). Obesity: preventing and managing the global epidemic: report of a WHO consultation.

OWOLABI, E. (2017). Prevalence and associated factors of obesity among South African adults: A cross-sectional study. *Online Journal of Health and Allied Sciences*, **16** (2).

OWOLABI, E. O., TER GOON, D., ADENIYI, O. V., AND AJAYI, A. I. (2018). Optimal waist circumference cut-off points for predicting metabolic syndrome among low-income black South African adults. *BMC Research Notes*, **11** (1), 1–5.

PAUSOVA, Z. (2006). From big fat cells to high blood pressure: a pathway to obesity-associated hypertension. *Current Opinion in Nephrology and Hypertension*, **15** (2), 173–178.

PEREKO, K. K., SETORGLO, J., OWUSU, W. B., TIWEH, J. M., AND ACHAMPONG, E. K. (2013). Overnutrition and associated factors among adults aged 20 years and above in fishing communities in the urban Cape Coast Metropolis, Ghana. *Public Health Nutrition*, **16** (4), 591–595.

PI-SUNYER, F. X. (2000). Obesity: criteria and classification. *Proceedings of the Nutrition Society*, **59** (4), 505–509.

POPKIN, B. M. (1994). The nutrition transition in low-income countries: an emerging crisis. *Nutrition Reviews*, **52** (9), 285–298.

PRENTICE, A. M. (2006). The emerging epidemic of obesity in developing countries. *International Journal of Epidemiology*, **35** (1), 93–99.

PRINSLOO, J., MALAN, L., DE RIDDER, J., POTGIETER, J., AND STEYN, H. (2011). Determining the waist circumference cut off which best predicts the metabolic syndrome components in urban Africans: the SABPA study. *Experimental and Clinical Endocrinology and Diabetes*, **119** (10), 599–603.

PUOANE, T., STEYN, K., BRADSHAW, D., LAUBSCHER, R., FOURIE, J., LAMBERT, V., AND MBANANGA, N. (2002). Obesity in South Africa: the South African demographic and health survey. *Obesity Research*, **10** (10), 1038–1048.

RAZAVI, S., JAKEMAN, A., SALTELLI, A., PRIEUR, C., IOOSS, B., BORGONOVO, E., PLISCHKE, E., PIANO, S. L., IWANAGA, T., BECKER, W., ET AL. (2021). The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, **137**, 104954.

REAVEN, G. M. (1988). Role of insulin resistance in human disease: Banting Lecture 1988. *Diabetes*, **37**, 1595–1607.

REGION, W. W. P. (2000). International Association for the Study of Obesity (IASO), International Obesity Task Force (IOTF). *The Asia-Pacific perspective: Redefining obesity and its treatment. Sydney, Australia: Health Communications Australia Pty Limit*.

ROTH, G. A., ABATE, D., ABATE, K. H., ABAY, S. M., ABBAFATI, C., ABBASI, N., ABBASTABAR, H., ABD-ALLAH, F., ABDELA, J., ABDELALIM, A., ET AL. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, **392** (10159), 1736–1788.

SAMUEL, K., DAVID, B., STEVEN, B., RUDOLPH, L., CATHY, N., AND RICHARD, K. (2007). Waist circumference and cardiometabolic risk. *Diabetes Care*, **30**, 1647–1652.

SEIDELL, J. C. AND HALBERSTADT, J. (2015). The global burden of obesity and the challenges of prevention. *Annals of Nutrition and Metabolism*, **66** (Suppl. 2), 7–12.

SEKGALA, M., MONYEKI, K., MOGALE, A., MCHIZA, Z., PARKER, W., CHOMA, S., AND MAKGOPA, H. (2018a). The risk of metabolic syndrome as a result of lifestyle among Ellisras rural young adults. *Journal of Human Hypertension*, **32** (8-9), 572–584.

SEKGALA, M. D., MCHIZA, Z. J., PARKER, W.-A., AND MONYEKI, K. D. (2018b). Dietary fiber intake and metabolic syndrome risk factors among young South African adults. *Nutrients*, **10** (4), 504.

SENEKAL, M., STEYN, N. P., AND NEL, J. H. (2003). Factors associated with overweight/obesity in economically active South African populations. *Ethnicity and Disease*, **13** (1), 109–116.

SHAYO, G. A. AND MUGUSI, F. M. (2011). Prevalence of obesity and associated risk factors among adults in Kinondoni Municipal district, Dar es Salaa tanzania. *BMC Public Health*, **11** (1), 1–7.

SIMFUKWE, P., VAN WYK, B., AND SWART, C. (2017). Perceptions, attitudes and challenges about obesity and adopting a healthy lifestyle among health workers in Pietermaritzburg, KwaZulu-Natal province. *African Journal of Primary Health Care and Family Medicine*, **9** (1), 1–9.

SINAGA, M., WORKU, M., YEMANE, T., TEGENE, E., WAKAYO, T., GIRMA, T., LINDSTROM, D., AND BELACHEW, T. (2018). Optimal cut-off for obesity and markers of metabolic syndrome for Ethiopian adults. *Nutrition Journal*, **17** (1), 1–12.

STEYN, K., BOURNE, L., JOOSTE, P., FOURIE, J., ROSSOUW, K., AND LOMBARD, C. (1998). Anthropometric profile of a black population of the Cape Peninsula in South Africa. *East African Medical Journal*, **75** (1), 35–40.

SUMNER, A. E., ZHOU, J., DOUMATEY, A., IMOISILI, O. E., AMOAH, A., ACHEAMPONG, J., OLI, J., JOHNSON, T., ADEBAMOWO, C., AND ROTIMI, C. N. (2010). Low HDL-cholesterol with normal triglyceride levels is the most common lipid pattern in West Africans and African Americans with metabolic syndrome: implications for cardiovascular disease prevention. *CVD Prevention and Control*, **5** (3), 75–80.

SWETS, J. A. (1979). Roc analysis applied to the evaluation of medical imaging techniques. *Investigative radiology*, **14** (2), 109–121.

SWETS, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, **99** (1), 100.

TAL, S., LITOVCHIK, I., KLAR, M. M., MARESKY, H. S., GRYSMAN, N., WISER, I., VITKON-BARKAY, I., MARCUS, G., TZUMAN, O., PEREG, D., ET AL. (2019). The association between neck adiposity and long-term outcome. *Plos One*, **14** (4), e0215538.

TLADI, D., MOKGATLHE, L., SHAIBU, S., NELL, T., MITCHELL, R., MOKGOTHU, C., GABONTHONE, T., AND HUBONA, O. (2020). Determination of optimal cut-off values for waist circumferences used for the diagnosis of the metabolic syndrome among Batswana adults (ELS 32). *Cardiovascular Journal of Africa*, **31** (6), 314–318.

TUGENDHAFT, A., MANYEMA, M., VEERMAN, L. J., CHOLA, L., LABADARIOS, D., AND HOFMAN, K. J. (2016). Cost of inaction on sugar-sweetened beverage consumption: implications for obesity in South Africa. *Public Health Nutrition*, **19** (13), 2296–2304.

VAGUE, J. (1947). Sexual differentiation; factor determining forms of obesity. *La Presse Médicale*, **55** (30), 339.

VAN DER LINDEN, E. L., MEEKS, K., BEUNE, E., DE GRAFT AIKINS, A., ADDO, J., OWUSU-DABO, E., MOCKENHAUPT, F. P., BAHENDEKA, S., DANQUAH, I., SCHULZE, M. B., ET AL. (2019). The prevalence of metabolic syndrome among Ghanaian migrants and their homeland counterparts: the Research on Obesity and Type 2 Diabetes among African Migrants (rodam) study. *European Journal of Public Health*, **29** (5), 906–913.

VAN DER MERWE, M.-T. AND PEPPER, M. (2006). Obesity in South Africa. *Obesity Reviews*, **7** (4), 315–322.

VAN VOLLENSTEE, F. A. AND VAN DER MERWE, M.-T. (2021). Obesity and its implications for COVID-19 pandemic in South Africa. *Southern African Journal of Infectious Diseases*, **36** (1).

VAN ZYL, S., VAN DER MERWE, L. J., WALSH, C. M., GROENEWALD, A. J., AND VAN ROOYEN, F. C. (2012). Risk-factor profiles for chronic diseases of lifestyle and metabolic syndrome in an urban and rural setting in South Africa. *African Journal of Primary Health Care and Family Medicine*, **4** (1), 1–10.

VAZQUEZ, G., DUVAL, S., JACOBS JR, D. R., AND SILVENTOINEN, K. (2007). Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiologic Reviews*, **29** (1), 115–128.

VORSTER, H., VENTER, C., KRUGER, H., KRUGER, A., MALAN, N., WISSING, M., DE RIDDER, J., VELDMAN, F. J., STEYN, H., MARGETTS, B., ET AL. (2000). The impact of urbanization on physical, physiological and mental health of Africans in the North West province of South Africa: the THUSA study. *South African Journal of Science*, **96**.

WARDLE, J., WALLER, J., AND JARVIS, M. J. (2002). Sex differences in the association of socioeconomic status with obesity. *American Journal of Public Health*, **92** (8), 1299–1304.

WELLS, J. C., MARPHATIA, A. A., COLE, T. J., AND MCCOY, D. (2012). Associations of economic and gender inequality with global obesity prevalence: understanding the female excess. *Social Science and Medicine*, **75** (3), 482–490.

WHO (1999). Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus. Technical report, World health organization.

ZHOU, X.-H., MCCLISH, D. K., AND OBUCHOWSKI, N. A. (2009). *Statistical methods in diagnostic medicine*. John Wiley & Sons.

ZIMMET, P., ALBERTI, K., AND SHAW, J. (2001). Global and societal implications of the diabetes epidemic. *Nature*, **414** (6865), 782–787.

ZIRABA, A. K., FOTSO, J. C., AND OCHAKO, R. (2009). Overweight and obesity in urban Africa: a problem of the rich or the poor? *BioMed Central Public Health*, **9** (1), 1–9.