# Assessing the Impact of Social Determinants of Health on Multiple Binary Outcomes Using Multivariate Statistical Methods

**Musyoki, Faith Mutheu**

**Strathmore Institute of Mathematical Sciences**

**Strathmore University**

**Nairobi, Kenya**

**Submitted in partial fulfillment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University**

**June 2025**

# Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the proposal itself.



Name: **Musyoki Faith Mutheu**

Signature: .....

Date: March 28, 2025

## Approval

The thesis of Musyoki Faith Mutheu was reviewed and approved by the following:

**Dr. Evans Otieno Omondi**

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

March 28, 2025

**Dr. Collins Odhiambo**

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

28th March, 2025

**Dr. Godfrey Madigu**

Dean,

Institute of Mathematical Sciences, Strathmore University.

**Dr. Bernard Shibwabo**

Director,

Office of Graduate Studies, Strathmore University.

# Abstract

Social Determinants of Health (SDOH) are the non-medical factors for example; housing, employment, education, food security, etc., influencing health conditions/outcomes including chronic respiratory diseases and cardiovascular diseases. It is important to understand the social factors to reduce SDOH gaps across different social groups and improve public health outcomes.

The objectives examined the impact of SDOH, modeling the interaction between SDOH and health outcomes, and assessing covariate distribution. This research employed multivariate statistical methods; Multivariate Logistic Regression (MLR) and Canonical Correlation Analysis (CCA). Simulations of different covariate shapes of the Wishart distribution was run to understand how they affect model fit. The data utilized is a multidimensional survey conducted in Costa Rica's provinces of San José, Alajuela, Cartago, and Heredia between February 2019 and December 2022.

The findings showed that significant associations between SDOH and health outcomes. Middle-income and unemployed individuals showed higher odds of developing diabetes, mental health disorders, obesity, and CRD. Smoking and physical exercise are strongly associated with increased odds of obesity, CVD, and diabetes. CCA revealed a strong canonical correlation between SDOH and health outcomes. Smoking, informal housing, income, and employment influenced health outcomes, CVD, CKD, and CRD. Covariate assessment showed robustness of logistic regression models under different covariance structure of Wishart-distribution, with higher degrees of freedom improving model performance.

It highlighted insight into application of multivariate statistical methods in public health policymakers. Health-care practitioners should focus on lifestyle factors and socioeconomic factors to help reduce health disparities. There is also need for targeted health interventions and policy addressing to promote healthier lifestyle. Further research should explore longitudinal data to understand SDOH impacts on population health.

# Contents

# List of abbreviations

| | | | |
|---|---|---|---|
| SDOH | Social Determinants Of Health | SDH | Social Determinants of Health |
| MLR | Multivariat Logistic Regression | CCA | Canonical Correlation Analysis |
| CSDH | Commission on Social Determinants of Health | WHO | World Health Organization |
| CVD | CardioVascular Diseases | CKD | Chronic Kidney Disease |

# Acknowledgement

First and above all, I am grateful to the Almighty God for His provision, grace, and for granting me good health throughout the study period.

I am grateful to my parents, John Musyoki and Dr. Josephine Musyoki, my sister, Melody Mawia and my friends, Eng.Kithia Collins Maxwell and Diana Waithira Maingi for their unwavering support and cheering me on throughout the study period.

I am indebted to my supervisors, Dr. Evans Otieno Omondi and Dr. Collins Odhiambo for their availability, kind guidance, and support without which this thesis would not have been a success.

I am grateful to my classmates for their support and encouragement.

# Dedication

*This thesis is dedicated to God Almighty for giving me wisdom and good health.*

# Chapter 1

# Introduction

## 1.1 Background of the study

Social Determinants of Health (SDOH) are the conditions in which people are born, grow, live, work, and age that affect their health outcomes. They are crucial in shaping someone's health standard, functioning, and quality of lifestyle (Council, 2023). Social determinants of health that can influence health equity in positive and negative ways can include income, protection, education, job insecurity, working conditions, food insecurity, housing, basic amenities, social inclusion, and access to affordable health services of standard quality. Studies indicate social determinants are more essential than health care in influencing people's health (World Health Organization, 2024). Interventions to lessen health disparities can enhance population health and encourage health equality (Chelak.K, 2023).

Khetpal et al. (2021) describes health equity as a situation where everyone has equal opportunity to reach their full health potential and no one is deprived of grasping this potential because of their social standing or a different socially fixed set of situations. SDOH such as poverty, unequal access to health care, lack of education and stigma play a notable role in contributing to disparities in health outcomes (C.D.C, 2024a). Some social determinants have long-term effects on health. For instance, a person with less education can have confined knowledge on how to make the most of available resources, affecting their ability to use them sufficiently (Chelak.K, 2023).

There are multiple metrics, indices, and rankings available for the assessment of health care in relation to SDOH (Krause et al., 2021). SDOH data has many health outcomes with a wide range of occurrences (Mahendran et al., 2022). Therefore, to analyze SDOH impact in relation to outcomes we may use multivariate analysis. Multivariate analysis

can explain interactions among variables present in a data set (Ramadan and Abdel-Fatah, 2020). Author (2023) in addition, it can help understand the relationships between variables with health outcomes. Researchers can study gaps in social inequalities, conduct surveys, and analyze survey data. ABConvert (2024) multivariate analysis enables the discovery of complex relationships and patterns in a data set. Different strategies can be utilized to acquire collectible insights and make informed decisions.

## 1.2 Problem statement

Understanding how SDOH impacts health outcomes is crucial for developing effective public health interventions. By analyzing multiple health outcome along with various social determinants, we can gain a comprehensive understanding of health, resulting to improved health outcomes and reduced disparities across populations. Given the complex nature of SDOH and its diverse influence on health, it is important to ensure that interventions are effective and equitable (Bhavnani et al., 2023).

Previous studies have explored the impact of SDOH on individual health outcomes. In contrast, very little is known about the simultaneous impact of SDOH on multiple binary health outcomes, dismissing the complex relationship between various SDOH and their combined impact on health disparities. This study focuses on addressing this knowledge gap by analyzing how different social factors influence multiple health outcomes simultaneously. Despite CCA and MLR being commonly used in statistical analysis, their application in examining the impact of SDOH on multiple binary health outcomes is limited, indicating a methodological gap. This research extends multivariate techniques to this context to improve the modeling of these interactions.

Understanding how different Wishart distribution shapes affect model outcomes is crucial for enhancing the validity and reliability of statistical analysis in public health research. This study examined complex relationships and interactions between covariates and response variables, contributing to theoretical understanding by analyzing how social factors interact and contribute to health disparities.

This aim of this study was these gaps by employing CCA and MLR to analyze the impact of Wishart-distributed covariates on multiple binary health outcomes. This research aimed to give significant insights into the correlation between SDOH and health outcomes, leading to more informed public health interventions and policies. In addition, to contribute to a better understanding of health inequalities and the importance of multivariate statistical methods in this context by investigating the simultaneous impact of different social determinants on multiple binary health outcomes.

## 1.3 Research objectives

### 1.3.1 General objective

To systematically analyze the impact of various SDOH on multiple binary outcomes using multivariate statistical methods.

### 1.3.2 Specific objectives

1. To examine the impact of SDOH on multiple binary health outcomes.
2. To model the interaction between SDOH and health outcomes using MLR and CCA.
3. To assess the distribution of covariates affecting health outcomes and their influence on model fit.

## 1.4 Justification of the study

The World Health Organization's Commission on Social Determinants of Health (CSDH) has stressed that raising people's health and lowering disparities depend on improving SDOH (World Health Organization, 2024). The number of research evaluating the outcome of public policies on health equity is growing. However, there is still a greater need for attention to this

area of policy assessment because of the growing number of inequities (Hall and Jacobson, 2018).

Parsons et al. (2024) Commented that more study to understand how social and environmental variables impact health outcomes. This research aimed to improve public health interventions and policies to reduce health disparities. Public health practitioners can understand SDOH that influence health and test which strategies are more efficient and effective. Additionally, this research fills a gap in the analysis of multiple outcomes to obtain a better understanding of the impact of SDOH.

## 1.5 Significance of the study

The study provides significant insight on how different SDOH impact multiple health outcomes. Understanding there relationship helps in developing targeted strategies to address health disparities, resulting in more efficient and equitable health care services. By identifying the significant SDOH that impact health outcomes, policy makers can develop and implement policies to address these factors thus improving general public health and reducing health disparities.

The study applied MLR and CCA to model SDOH and health outcomes. Therefore, the application and validation of these methods contribute to methodological literature providing a framework for future research in related fields. Most studies have focused on the impact of SDOH on specific health outcomes. Therefore, by analyzing simultaneous effects of SDOH and multiple binary health outcomes, we have a comprehensive understanding of health disparities and fill existing research gap.

# Chapter 2

# Literature Review

## 2.1 Introduction to social determinants of health and health outcomes

International health agendas have shifted between two main agenda; 1) focusing on public health action and technology based health care and 2) health as a social trend that requires elaborate forms of policy action in different sectors (Alderwick et al., 2021). Historical records emphasize the susceptibility of health policy approaches integrating social determinants; indicating these approaches can encounter opposition(from stakeholders,institutional entities etc). The resistance can pose a significant challenge to effective implementation of policies aimed to address SDOH. World Health Organization established The Commission on Social Determinants of Health (CSDH) in March 2005 to support countries and global health partners in addressing the social factors leading to ill health and health inequities (Osmick and Wilson, 2020). Given the rising focus on SDOH, health professionals worldwide are seeking effective ways to improve the health outcomes across all the social levels (Gómez et al., 2021).

It has been evident that enhancing medical care is not sufficient to improve overall health or reduce health disparities; there is need to address the environmental and socioeconomic factors that influence people's live (Whitman et al., 2022). Reducing SDOH gaps encourages health equity therefore eliminating unfair health disparities among population groups based on social, economic, geographical or demographic factors (C.D.C, 2024b; Organization et al., 2010). Healthy People 2030 et al. (2021) identified 5 primary domains of SDOH; health care access and quality, education access and quality, social and community context,

economic stability, and neighborhood and built environment which is line with (Organization et al., 2010). The Dahlgren-Whitehead model has been used for more than three decades to explain the main determinants of health (Dahlgren and Whitehead, 2021). In addition, Hunter et al. (2011) noted that the problem is not the social determinants themselves, but rather the opportunities available to individuals due to relative differences in the distribution of these determinants.

The relationship between social determinants and population health and health disparities involves complex mediation factors (Hahn, 2021). A number of this factors tend to be among individuals in underprivileged conditions in both developed and developing countries (Beech et al., 2021). Most of the factors addressed are the downstream determinants (education, neighborhood condition, risk behaviors and access to health care) of health. However, it's important to understand and curb the underlying upstream factor (governance and policy) that play a role in shaping and impacting people's health (Ray et al., 2023).

In his study, Krause et al. (2021) notes different SDOH affects populations to different extent. The acknowledgment of social determinants as underlying cause of health conditions continues to rise. They have an impact on both chronic and infectious diseases (Dhlamini, 2023). Chronic diseases are conditions that persist a year or longer and require continue medical attention and have long-lasting effects that limit activities of daily living (diabetes, cancer, heart disease, etc) (Nowrozy, 2023). A research aimed to investigate the ecological association between SDOH factors and population health outcomes by Vo et al. (2023) revealed that all 9 selected SDOH variables had a statistically significant impact on population health outcomes.

Wan et al. (2022) found the relationships between cluster scores( social background, social insecurities, insurance/employment) and uncontrolled diabetes and/or hypertension were positively associated. In Germany, Ziegler et al. (2024) notes a higher level of education increased the chances of patients getting information about peer support throughout cancer care treatment. During the analysis of risk factors predicting recurrence of chronic subdural hematoma Yogi et al. (2018) concluded that chronic alcohol use(a SDH risk behavior)and

intraoperative brain enlargement(a binary health outcome) were both substantially associated. There is a strong evidence linking SDOH to various health outcomes.

## 2.2 Multivariate statistical methods

Multivariate statistics is a set of statistical methods developed to handle data with multiple variables. An analysis of more than two variables or measures can be categorized as a multivariate statistical analysis. Multivariate analysis is used when; 1) we have multiple variables and wan to investigate each variables, 2)we have a set of variables that we want to analyze as a set and 3)we are not interested in the raw variables themselves as we are i the us of all or a subset of them (Camacho et al., 2024).

Multiple binary outcomes occur often in healthcare research. Due to its simplicity one might underestimate the complexity of modeling or analyzing binary data. Assume a study with multiple binary outcomes of interest; modeling each outcome one needs to consider if the outcomes are associated (Lupparelli and Mattei, 2020). The multivariate techniques were considered for analysis in this research are Multivariate Logistic Regression and Canonical Correlation Analysis.

### 2.2.1 Multivariate logistic regression

Multivariate logistic regression involves multiple outcomes and predictors Ebrahimi Kalan et al. (2021), the logistic regression can be expressed as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \alpha_i$$

Where $\pi_{ij}$ represents the probability of outcome $j$ in cluster $i$ and $\alpha_i$ is a random effect that accounts for within-cluster correlation.

According to Fernandes et al. (2021) MLR provides a comprehensive understanding of association between multiple outcomes and their predictors. In addition, it takes into account

the correlation between multiple binary outcomes; each outcome has a unique logistic equation but they are analyzed simultaneously within the same model. The model can estimate the joint probability for the outcomes given their interdependence.

Victorino and Gauthier (2009) applied MLR to study the relationship between children health outcome and SDOH. It was noted that children likely to suffer from these health condition were from low-income families. Furthermore, after adjusting for some variables there still existed a gradient relationship between household income and a child's likelihood to getting asthma, migraines or headaches or ear infections.

From Gu et al. (2021) low social and economic capital have a significant relationship with poor health status. It contributes to evidence that both economic difficulties and social capital impact variety of health outcomes and when combined they intensify the burden of poor health. As stated by Mandalia et al. (2023) delays in access to health care and increase in severity of diseases can be attributed to SDH (gender, occupation, tobacco use, insurance, lower education level, racial or ethnic minority status, low-income, place of residence,unemployment, and preoperative narcotic use etc). During rotator cuff repair, these factors lead to poor health care and patient reported outcomes such as unsuccessful repair, high risk of revision surgery, greater risk of postoperative complications and reduced capacity to resume work.

Holbert et al. (2022) carried out a study to identify SDOH for patients getting a lumbar spine surgery and evaluate their significance to the postoperative outcomes such as length of stay, discharge disposition and re-admissions. They observed that across the SDOH looked at, financial resources and strong support systems was significantly associated with better post-operative outcomes. Patients with household income were likely to be discharged to a skilled nursing facility, while patients with financial difficulties were likely to have a longer length of stay. Individuals with a life partner/married had reduced length of stay. From their conclusions, patients with low financial resources, less in-home support and low social relations (do not attend church) are at high risk of sub-optimal outcomes.

A study between SDOH with blood screening for hypertension, hypercholesterolemia and hyperglycemia, by Nguyen et al. (2021) indicated household income was significantly

correlated with the probability of getting a blood screen. The state of employment had a notable association with blood glucose testing; employed individuals were less likely to check their blood glucose in comparison to unemployed individuals. Individuals who were married or living with a partner are more likely to take a blood screen. Education level had significant relationship with getting blood tests. Furthermore, marginalized people seek treatment only when they are in immediate need.

### 2.2.2 Canonical correlation analysis

Everitt and Hothorn (2011) Canonical Correlation Analysis (CCA) analyzes the relationship between response and predictor variables. It finds the most highly correlated linear combination of the predictor variables and the response variable. Each set partner combination is expected to reveal something unique, however, all the combination are expected to be mutually uncorrelated of each other.

According to Bell et al. (2021), CCA is able to perform data reduction by determining highly correlated variable. This study indicated a strong correlation between men view on gender and masculinity, alcohol consumption, testing and treatment behavior with HIV-related anxieties and testing methods. Furthermore Bíró et al. (2021) applied CCA to determine variables to include in the model. The CCA developed two dimension; mental and physical health. In adolescence, mental health was strongly correlated with drugs(excluding smoking), social support and sports while physical health was highly correlated with age, drugs(apart from marijuana) and dance.

Adza et al. (2023) explored the association between air pollution and traffic noise. The study focused on the relationship between hypertension cases and rates and exposure to air pollution and traffic noise. Based on CCA by canonical loading reading, joint air pollution and traffic noise across different frequency components were significantly correlated with the total reported hypertension cases and rates.

## 2.3 Covariate distribution

Covariates are predictor variables that possibly influence the response variable. There are two types of covariate variables: Independent covariate variables(this are the variables of direct interest in the study) and Confounding variables(covariates that are not of interest in the study). They are used to explain variation in the outcome of interest. The importance of covariates are;when included in models they help to control for confounding factors, improve model accuracy and reduce bias in estimating the effect of the primary variable of interest (Tippins, 2023).

Covariate distribution is often assumed to follow a multivariate normal distribution. From multivariate normal random variables, the sampling distribution of sample covariance matrix will follow a **The Wishart distribution**. The Wishart distribution is a multivariate generalization of univariate $\chi^2$ distribution. The Wishart properties; independence and additive: If $A_1$and$A_2$ are independently distributed as $W_{m1}(A_1 \mid \Sigma)$ and $W_{m2}(A_2 \mid \Sigma)$ respectively then the sum is distributed as $A_1 + A_2 \sim W_{m1+m2}(A_1 + A_2 \mid \Sigma)$. Transformation: If $A \sim W_m(A \mid \Sigma)$ then for any constant matrix $C$ therefore $CAC'$is distributed as $W_m(CAC' \mid CAC')$ which can be expressed as $W_m(CAC')$.

The Wishart properties allows one to make inference about the population covariance matrix (Johnson and Wichern, 2007; Wilkinson, 2024). CCA relies on multivariate normal and Wishart distribution to draw conclusion abut the correlation between original variables and canonical variables. It also helps estimate and test hypothesis related to covariance matrices (Bykhovskaya and Gorin, 2023).

# Chapter 3

# Methodology

## 3.1 Introduction

This chapter introduces methods for analyzing SDOH's impact on multiple binary health outcomes. The analysis included data processing, explanatory data analysis, MLR, CCA, assessment of covariates, and model validation. MLR and CCA were used to model the relationship between the explanatory and response variables while accounting for the covariance structure in the binary response. Covariate assessment examined how the distributions of the covariates, which take different shapes of Wishart distribution, may affect model fit and result interpretation. Model validation using cross-validation assessed model performance and accuracy.

## 3.2 Research design

### 3.2.1 Study

The analysis utilized data from a multidimensional survey conducted in Costa Rica's provinces of San José, Alajuela, Cartago, and Heredia between February 2019 and December 2022. To ensure representativeness, the samples were selected using stratified sampling approach, and the survey contained questions regarding areas in health, socioeconomic, and environmental variables.

### 3.2.2 Response variables

These are the variables of interest that are being explained or predicted. The response variables in this study represent the binary health outcomes, i.e., the presence or absence of a specific chronic condition. The chronic conditions to be included are: Cardiovascular Diseases (CVD), diabetes, Chronic Respiratory Diseases (CRD), obesity, mental health disorders, and Chronic Kidney Disease (CKD). Each health condition was coded as "1" (presence/yes) or "0" (absence/no). Understanding the simultaneous effects of SDOH on these multiple health conditions helped reveal how different social factors influence health outcomes, thus offering a more holistic understanding of health disparities.

### 3.2.3 Exploratory variables

These are factors that influence or explain variations in the response variable. They represent the SDOH and these factors included: age, income, education level, employment, informal housing, any form of food insecurity, does regular physical exercise, smoking, exposure to air pollution, and exposure to violent crime. The explanatory variables are important for understanding how socioeconomic and lifestyle factors affect health outcomes. SDOH contributes significantly to health disparities, and by analyzing these determinants along with health outcomes, we determined which factors are most influential in the development of chronic health conditions.

## 3.3 Data analysis

### 3.3.1 Data processing and explanatory data analysis

Data processing entailed data cleaning: to handle missing values, outliers, or inconsistencies in the data set. The explanatory data analysis included several steps. Firstly, descriptive statistics were used to summarize the variable distribution. Visualizations techniques included box plots, histogram, and bar graph were applied to examine the distribution of the covariates

and their relationship with health outcomes. Additionally, multicollinearity among the predictor variables was checked to ensure independence of explanatory variables. Lastly, covariate distribution was assessed to ensure covariates follow the normality assumption in multivariate statistics.

### 3.3.2 Joint cumulative distribution function

To understand the dependencies among multiple health outcomes. When dealing with a random vector $\mathbf{X}$ consisting of $p$ components, namely $X_1, X_2, \ldots, X_p$, we define the joint cumulative distribution function (CDF) as:

$$F(\mathbf{a}) = F(a_1, a_2, \ldots, a_p) = P(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_p \leq a_p) \tag{3.1}$$

This function shows the probability that all components of the random vector $\mathbf{X}$ take values less than or equal to the corresponding elements of the vector $\mathbf{a}$.

The probability that $\mathbf{X}$ lies within a hyper-rectangle, rather than just within an interval:

$$F(\mathbf{b}) - F(\mathbf{a}) = P(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \ldots, a_p < X_p \leq b_p) \tag{3.2}$$

This equation quantifies the probability of $\mathbf{X}$ falling within the range defined by the vectors $\mathbf{a}$ and $\mathbf{b}$.

The joint probability density function (PDF) is derived by differentiating the joint CDF:

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_p) = \frac{\partial^p F(a_1, \ldots, a_p)}{\partial a_1 \cdots \partial a_p} \tag{3.3}$$

From the joint PDF, we can obtain the CDF through integration:

$$F(\mathbf{a}) = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_p} p(x_1, x_2, \ldots, x_p)\, dx_p \ldots dx_2 dx_1 \tag{3.4}$$

The integration order is not necessary since from the joint PDF, we can recover the marginal PDF of any subset of variables, those numbered $1 \ldots q$:

$$p(x_1, x_2, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_p)\, dx_{q+1} dx_{q+2} \ldots dx_p \tag{3.5}$$

The limits of integration are the conditional PDF for some variables given by using variables 1 through $q$ to condition those numbered $q+1$ through $p$ is given by:

$$p(x_{q+1}, x_{q+2}, \ldots, x_p | X_1 = x_1, \ldots, X_q = x_q) = \frac{p(x_1, x_2, \ldots, x_p)}{p(x_1, x_2, \ldots, x_q)} \tag{3.6}$$

These two steps can be iterated as:

$$p(x_3|x_1) = \int p(x_3, x_2|x_1)\, dx_2 \tag{3.7}$$

This helps to account for the probability of simultaneous occurrences of multiple binary health outcomes in the data.

### 3.3.3 Multivariate gaussian distribution

The explanatory variables were assumed to follow a multivariate Gaussian distribution. This allows for effective estimation of association among explanatory variables and their influence on health outcomes.

The multivariate gaussian distribution is a generalization of the ordinary gaussian distribution to vectors. Scalar gaussians are denoted by a mean $\mu$ and variance $\sigma^2$, denoted:

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{3.8}$$

Multivariate gaussians has parameter, mean vector $\mu$ and a covariance matrix $\Sigma$, written as:

$$X \sim \mathcal{N}(\mu, \Sigma). \tag{3.9}$$

The $\mu$ show the means of the different components of $X$. The $(i,j)$-th component of $\Sigma$ represent the covariance between $X_i$ and $X_j$, meaning that the diagonal elements of $\Sigma$ provide the variances of the individual components.

The multivariate gaussian PDF generalizes to:

$$p(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right). \tag{3.10}$$

The parameters of a gaussian distribution transform predictably under linear transformations. If

$$X \sim \mathcal{N}(\mu, \sigma^2), \tag{3.11}$$

then applying an affine transformation gives:

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2). \tag{3.12}$$

Similarly, for the multivariate Gaussian:

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{3.13}$$

leading to transformed variable

$$aX + b \sim \mathcal{N}(a\mu + b, a\Sigma a^T). \tag{3.14}$$

### 3.3.4 Linear algebra and the covariance matrix

The covariance matrix plays a central role in multivariate analysis. From linear algebra the general pattern for arbitrary multivariate gaussians in an arbitrary number of dimensions. The covariance matrix $\Sigma$ is symmetric and positive-definite, from matrix algebra that it can be written in terms of its eigenvalues and eigenvectors:

$$\Sigma = v^T d v \tag{3.15}$$

Let $d$ denote the diagonal matrix with the eigenvalues of $\Sigma$, and let $v$ is the matrix whose columns to the eigenvectors of $\Sigma$. The eigenvalues are arranged in decreasing order, with the eigenvectors in $v$ likewise. However, any consistent arrangement is acceptable. Since the eigenvectors have 1 unit length, and they are mutually orthogonal, it follows that $v^T v = I$, implying $v^{-1} = v^T$. Therefore, $v$ is an orthogonal matrix. In the context of multivariate gaussian density function, the term $\Sigma^{-1}$, can be expressed as:

$$(v^T d v)^{-1} = v^{-1} d^{-1} (v^T)^{-1} = v^T d^{-1} v \tag{3.16}$$

Orthogonal matrices are associated with rotational transformation. When vector $v$ is multiplied by orthogonal matrix, it rotates aligning them with the eigenvectors of covariance matrix $\Sigma$. The principle axes of the probability-contour ellipse corresponds to these eigenvectors. The length of these axes are proportional to the square roots of the associated eigenvalues. This relationship can be established by considering the following derivation. Let $f_0$ denote the fixed probability density level, define the contour of interest. Then we obtain:

$$f_0 = (2\pi \det \Sigma)^{-p/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right) \tag{3.17}$$

Defining a constant $c$, we get:

$$c = (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) \tag{3.18}$$

Substituting the spectral decomposition:

$$c = (\mathbf{x}-\mu)^T v^T d^{-1} v(\mathbf{x}-\mu) \tag{3.19}$$
$$= (\mathbf{x}-\mu)^T v^T d^{-1/2} d^{-1/2} v(\mathbf{x}-\mu) \tag{3.20}$$
$$= \left(d^{-1/2} v(\mathbf{x}-\mu)\right)^T \left(d^{-1/2} v(\mathbf{x}-\mu)\right) \tag{3.21}$$

which simplifies to:

$$\|d^{-1/2} v(\mathbf{x}-\mu)\|^2 \tag{3.22}$$

Where, $c$ represents the combination of $f_0$ and other constant terms, while $d^{-1/2}$ is the diagonal matrix whose elements are the reciprocals of the square roots of the eigenvalues of $\Sigma$. The term $v(\mathbf{x}-\mu)$ takes the displacement of $\mathbf{x}$ from the mean, $\mu$, and replaces the components of that vector with its projection onto the eigenvectors. Multiplication by $d^{-1/2}$ then standardizes the projections, ensuring the radii are proportional to the square roots of the eigenvalues. This matrix captures the variance and the linear relationship between variables, creating basis for correlation analysis.

### 3.3.5 Multivariate logistic regression

MLR was used to model the relation between the binary health outcome (response variables) and SDOH (explanatory variables). Although MLR has been extensively utilized, its application in this context was to model the combined effect of SDOH on multiple health conditions simultaneously. Each health outcome with its probability of developing is modeled as a function of the SDOH factors taking into account the correlation between the binary outcomes. This sheds insight into the individual as well as the collective effect of socioeconomic and lifestyle factors on different health conditions.

Tabachnick (2007) states the multivariate logistics regression model will be fitted as follows:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \tag{3.23}$$

where; $P(Y=1)$ is the probability of the outcome,, $\beta_0$ is the intercept term, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for each predictor variable, and $X_1, X_2, \ldots, X_k$ are the predictor variables.

Moreover, the study extended MLR by examining how Wishart-distributed covariates influence model performance and reliability. The Wishart distribution described the covariance matrix of the covariates, which accounts for the complex interrelationship between the SDOH variables. A feature often overlooked in a typical logistic regression model.

### 3.3.6 Canonical correlation analysis

CCA examined the linear relationship between the two sets of variables. That is, the data will be divided into $X$ and $Y$ as explanatory and response variables, respectively. This enabled us to understand SDOH's effect on health outcomes. Through the examination of canonical covariate pairs, we determined which linear combination of SDOH is highly associated with specific health outcomes. Everitt and Hothorn (2011) fitted the canonical variable as follows;

$$U = a'X \quad \text{and} \quad V = b'Y \tag{3.24}$$

where; $U$ and $V$ are the canonical variables, $a$ and $b$ are the coefficient variables that define the linear combination, $X$ are explanatory variables $(X_1, X_2 \ldots, X_p)$, and $Y$ are response variables $(Y_1, Y_2 \ldots, Y_q)$.

Then the canonical correlation $\rho$ was maximized as follows;

$$\rho = \frac{Cov(U,V)}{\sqrt{Var(U)Var(V)}} = \frac{a'\Sigma_{XY}b}{\sqrt{a'\Sigma_{XX}ab'\Sigma_{YY}b}}$$

where; $\Sigma_{XX}$ covariance matrix of explanatory variables, $\Sigma_{YY}$ covariance matrix of response variables, and $\Sigma_{XY}$ covariance matrix of response and explanatory variables.

### 3.3.7 Covariate distribution

This study evaluated the impact of different Wishart distribution shapes in the covariates and how they affect model performance and interpretation. Moreover, a simulation was ran to evaluate model sensitivity to changes in the distribution in the covariate distribution. Johnson and Wichern (2007) states the probability density function (PDF) of a Wishart-distributed random matrix $X$ is as follows:

$$X \sim W_p(n, \Sigma) \text{ for } n \geq p$$

$$f(X) = \frac{|X|^{\frac{(n-p-2)}{2}} exp - \frac{1}{2} tr(\Sigma^{-1} X)}{2^{\frac{np}{2}} |\Sigma|^{\frac{n}{2}} \Gamma_p(\frac{n}{2})} \tag{3.25}$$

Where; $p$ Number of variables being modeled (dimensionality of the matrix), $n$ Degrees of freedom, representing the sample size used to estimate covariance, and $\Sigma$ Scale matrix represents the true covariance structure. $|X|$ is the determinant of X and $\Gamma_p$ is multivariate generalization of gamma function defined as:

$$\Gamma_p(\frac{n}{2}) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma(\frac{n-i+1}{2})$$

Condition; $X$ must be positive definite to represent valid covariance matrices and $n \geq p$ to ensure density exists, which is critical for valid simulations.

### 3.3.8 Conditional distributions and least squares estimation

The study further refined predictive modeling by leveraging conditional distributions and least squares estimation. Conditional Distributions in the bivariate case. Suppose that **X** is

bivariate, so $p = 2$, with mean vector

$$\mu = (\mu_1, \mu_2)^T$$

and variance-covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

The conditional distribution of $X_2$ given $X_1$ is gaussian, and can be defined as:

$$X_2 \mid X_1 = x_1 \sim \mathcal{N}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right). \tag{3.26}$$

The optimal slope for linearly regressing $X_2$ on $X_1$ is given by:

$$\frac{\text{Cov}[X_2, X_1]}{\text{Var}[X_1]} = \Sigma_{21}\Sigma_{11}^{-1}.$$

In bivariate gaussian, the best linear regression and the optimal regression coincide, eliminating the necessity of considering nonlinear regressions. Additionally, the conditional variance for each value of $x_1$, indicating that the regression of $X_2$ on $X_1$ is homoskedastic, with gaussian noise that is independently distributed. This aligns with the assumptions of standard regression models.

Conditional Distributions in the Multivariate Case

More generally, if $X_1, X_2, \ldots, X_p$ follow a multivariate Gaussian distribution, then conditioning on $X_1, \ldots, X_q$ ensures that the remaining variables $X_{q+1}, \ldots, X_p$ also follow a Gaussian distribution.

If we partition the mean vector and covariance matrix as:

$$\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix},$$

where $A$ corresponds to the conditioning variables and $B$ to the conditioned variables, then:

$$\mathbf{X}_B \mid \mathbf{X}_A = \mathbf{x}_A \sim \mathcal{N}\left(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(\mathbf{x}_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\right). \tag{3.27}$$

Note that $\Sigma_{BA} = \Sigma_{AB}^T$. This result effectively expresses a linear regression of $\mathbf{X}_B$ on $\mathbf{X}_A$.

This approach ensured estimates coefficients minimized residual errors while maintaining statistical interpretation.

### 3.3.9 Model comparison and validation

To compare model robustness, statistical criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used:

$$AIC = 2k - 2\ln(L) \tag{3.28}$$

where $k$ is the number of parameters and $L$ is the likelihood function. Similarly, BIC was computed as:

$$BIC = k\ln(n) - 2\ln(L) \tag{3.29}$$

. These metrics ensured an optimal trade-off between model complexity and explanatory power.

Models can be compared based on log-likelihood. When a strict out-of-sample comparison is not possible, cross-validation can be used.

A likelihood ratio test can be used. This has two forms, depending on the relationship between the models. Suppose that there is a large or wide model with parameter $\Theta$, and a narrow or small model with parameter $\theta$, which we get by fixing some of the components of $\Theta$. Thus, the dimension of $\Theta$ is $q$ and that of $\theta$ is $r < q$. Since every distribution from the narrow model can also be obtained from the wide model, the likelihood of the wide model must always be larger. Thus,

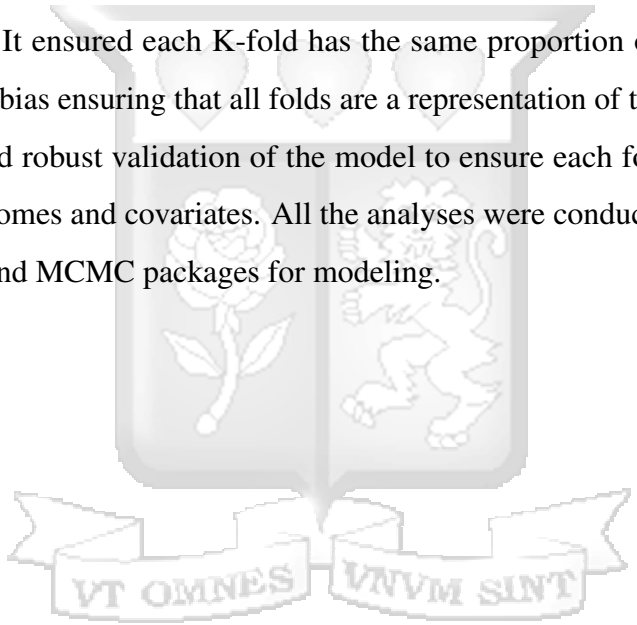
$$\ell(\hat{\Theta}) - \ell(\hat{\theta}) \geq 0. \tag{3.30}$$

Here, the null hypothesis assumes that the data comes from the narrower, smaller model. Under this null hypothesis, as $n \to \infty$,

$$2[\ell(\hat{\Theta}) - \ell(\hat{\theta})] \sim \chi^2_{q-r}, \tag{3.31}$$

provided that the restriction imposed by the small model does not place it on the boundary of the parameter space of $\Theta$.

Model validation was done using cross-validation techniques. Dealing with binary health outcomes there may be data imbalance, since one outcome may occur more frequently than the other. Due to this, the best cross-validation technique to be used was **Stratified K-fold Cross-validation.** It ensured each K-fold has the same proportion of both outcomes. In addition, it reduces bias ensuring that all folds are a representation of the entire original data set. It also provided robust validation of the model to ensure each fold copies the overall distribution of outcomes and covariates. All the analyses were conducted using R software, the MASS, CCA, and MCMC packages for modeling.



# Chapter 4

# Results and Interpretation

## 4.1 Introduction

This chapter shows the results of data analysis. It entails exploration of the dataset to understand variable distributions and relationships. Key assumptions were then examined to ensure robustness of the models. The results of the MLR are presented by identifying the significant SDOH of various health conditions. CCA is applied to examine the association between SDOH and health outcomes. Lastly, the impact of different covariance structures on model performance is evaluated using Wishart distributed covariates, and model validation is applied to assess model accuracy.

## 4.2 Descriptive statistics

Figure 4.1 shows that, the distribution indicates multiple age groups with higher frequencies. The most frequently observed ages are around 30 and 40 years suggesting that these age groups are the most common. The age ranges from approximately 20 to 70 years, showing wide spread of observation. The age variables shows a diverse age representation, indicating different subgroups within the population.

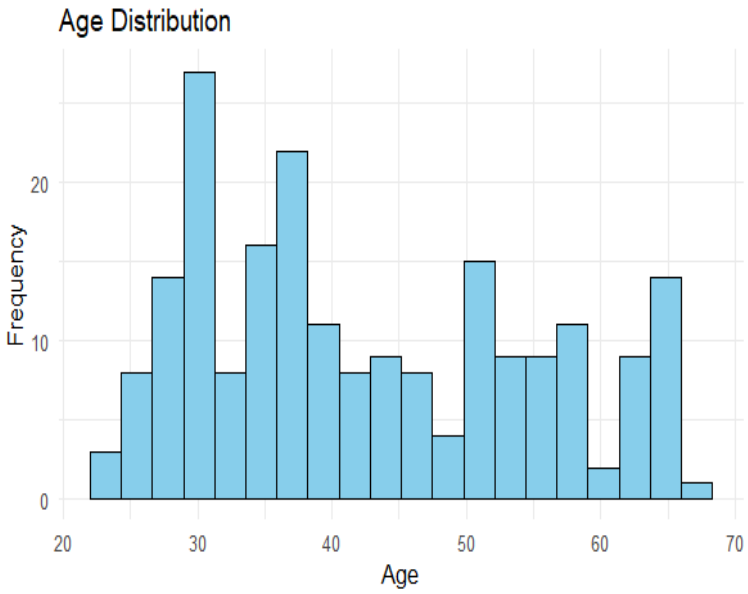



Figure 4.1: Histogram of Age

From Figure 4.2, low income individuals have the highest cases of diabetes indicating high prevalence. Middle income persons have a nearly equal portion of "Yes" and "No". High income individuals have the lowest prevalence where almost all individuals being free of diabetes.

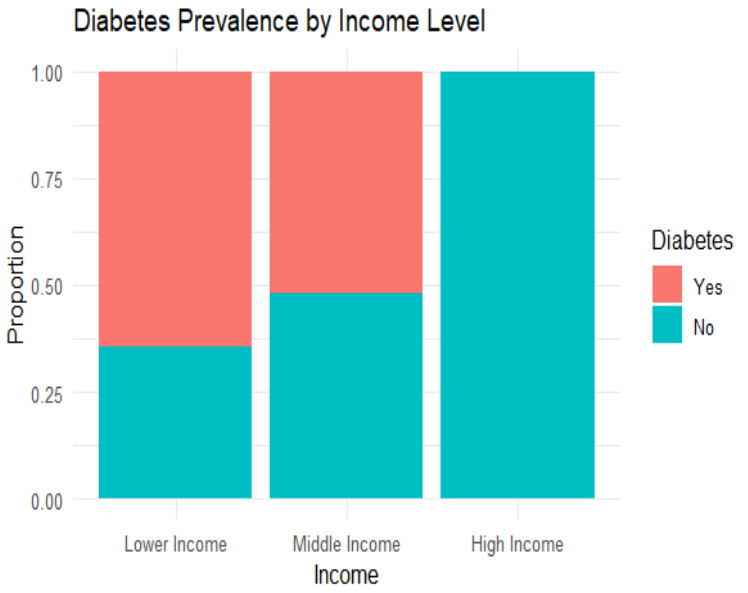



Figure 4.2: Income level among diabetes individuals

From Figure 4.3 among individuals with obesity, there is a higher proportion of those with cardiovascular disease compared to non-obese individuals.

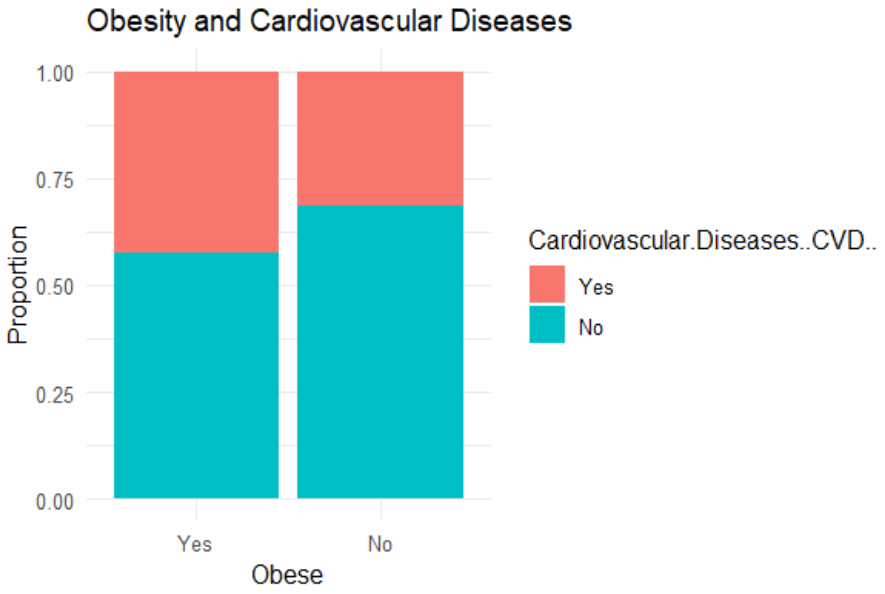



Figure 4.3: Cardiovascular diseases prevalence among the obese

The Figure 4.4 shows the smokers are associated with with higher mental health disorder compared to non-smokers. The difference is however not as significant.

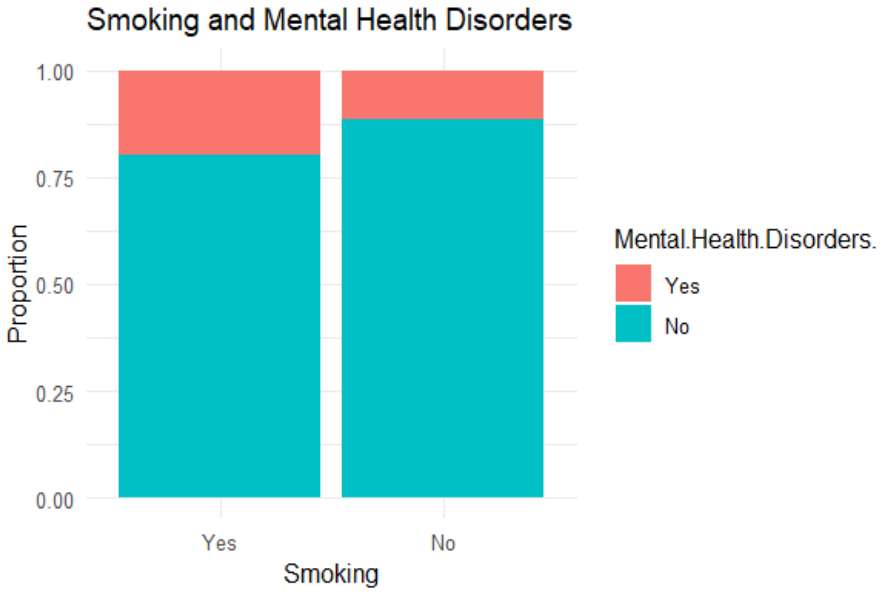



Figure 4.4: Mental health disorder prevalence in relation to smoking

## 4.3 Multivariate logistic regression

### 4.3.1 Multicollinearity assessment

Multicollinearity was checked using Variance Inflation Factor(VIF) for explanatory variables in a separate logistic regression models for each health outcome. Chronic respiratory diseases had extremely high VIF for education level (60.998) and employment(3691.4577). Obesity, mental health disorders, CVD, diabetes, CRD, and CKD had multicollinearity pattern is observed in informal housing with VIF 9.26, 9.41, 10.177, 9.17, 9.86, and 9.42 respectively. Despite the high VIF values, the explanatory variables were retained in the initial models because they are theoretically important health outcome predictors. However , regularization technique was employed to address multicollinearity and improve model stability.

### 4.3.2 Linearity in the logit and outlier detection

Linearity in the logit was tested for the continuous predictor variable, Age, using Box-Tidwell test. The interaction term Age_log was significant at $p = 0.046$. This suggests age violates the linearity assumption of the logit. This suggests a transformation or polynomial terms might be necessary. However, given the complexity that will be introduced, regularization approach was considered.

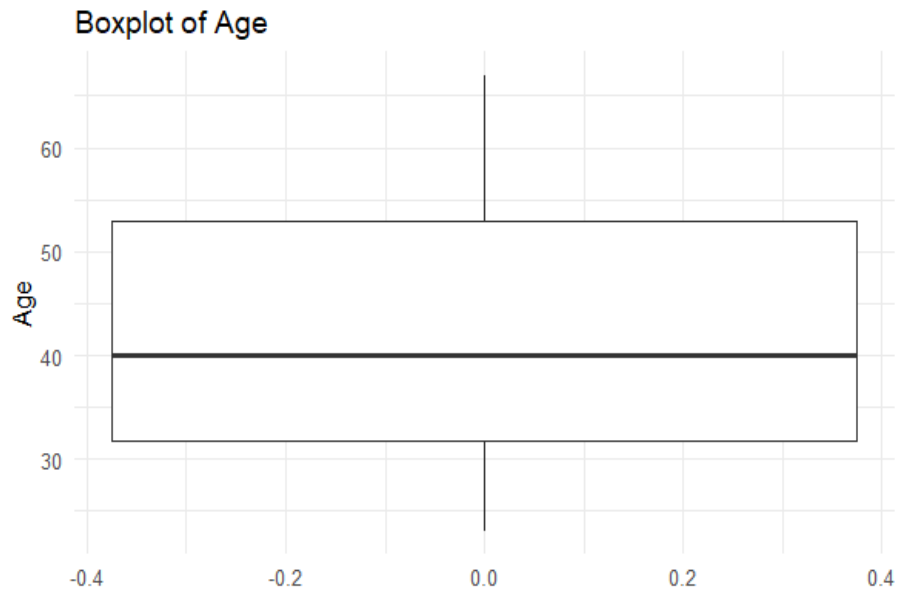



Figure 4.5: Plot of age outliers

From Figure 4.5 there were no outliers detected in Age.

Cook's distance was computed to detect influential observations. The observations were classified as influential if their Cook's distance exceeded the $4/n$ where $n$ is the total number of observations in the data set. The following observations were detected as influential: $11, 12, 13, 15, 20, 35, 44, 56, 82, 92, 98, 126, 133, 163$. The influential observations were retained for analysis since they were not due to data entry errors or measurement issues. Their impact on the model results were mitigated through robust regularization modeling technique.

### 4.3.3 Model regularization

To address the limitation of the multivariate logistic regression, LASSO (Least Absolute Shrinkage and Selection Operator) regularization was employed. LASSO introduces penalty term to the loss function, which shrinks less important coefficients to zero, thus reducing over-fitting. In addition, due to the multicollinearity consistent in the informal housing, and high in education level and employment LASSO will handle multicollinearity by its ability to select relevant features by setting highly correlated variables to zero.

LASSO regression was applied to each binary health outcome. Categorical predictors were converted to dummy variables and a design matrix $X$ was created. The optimal lambda values for each response variable were identified through 10-fold cross-validation and selected based on; lambda_min where the $\lambda$ that minimizes cross-validation error and lambda_1se which is the most regularized $\lambda$ with one standard error of the minimum.

Table 4.1: Lambda values

| **Outcome** | $\lambda_{min}$ | $\lambda_{1se}$ |
|---|---|---|
| Cardiovascular diseases | 0.00418 | 0.03896 |
| Diabetes | 0.00559 | 0.02982 |
| Chronic respiratory diseases | 0.00280 | 0.04154 |
| Obesity | 0.00531 | 0.02351 |
| Mental health disorder | 0.00556 | 0.05181 |
| Chronic respiratory diseases | 0.00187 | 0.04024 |

Table 4.1 the low $\lambda_{min}$ values suggest that many variables were needed to explain variability in the outcome. Higher $\lambda_{1se}$ show more regularization was done meaning some predictors shrunk to zero.

### 4.3.4 Cross-validation

Cross-validation from the LASSO models showed improved model performance. A stratified 10-fold cross-validation resulted the following:

Table 4.2: Stratified 10-fold cross-validation

| **Outcome** | **ROC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| Cardiovascular diseases | 0.862 | 0.946 | 0.713 |
| Diabetes | 0.873 | 0.696 | 0.815 |
| Chronic respiratory diseases | 0.923 | 0.992 | 0.636 |
| Obesity | 0.8631 | 0.720 | 0.846 |
| Mental health disorder | 0.784 | 0.092 | 0.994 |
| Chronic respiratory diseases | 0.801 | 0.250 | 0.875 |

CRD has the highest ROC indicating excellent model performance. CVD, diabetes, and obesity have ROC suggesting good classification model ability. MHD has the lowest sensitivity, despite high specificity, suggesting models poor ability to detect positive cases. CKD as low sensitivity but good specificity, indicating the model struggles to identify positive cases.

Based on Table 4.3 individuals who do not exercise regularly have a 3.80 times higher odds of having cardiovascular diseases compared to those who exercise regularly.

Table 4.3: Multivariate logistic regression for cardiovascular diseases

| **Outcome** | **Covariate** | **Parameter estimate** | **Standard error** | **p-value** |
|---|---|---|---|---|
| Cardiovascular diseases | Intercept | -2.035 | 1.035 | 0.0494 |
| | No regular excercise | 1.335 | 6.395e-01 | 0.0369 |

Table 4.4 individuals with middle income have 5.85 times higher odds of having diabetes compared to those of low income. The unemployed individuals have a 6.80 times high odds of having diabetes compared to the employed. Individuals who do not exercise regularly have 4.331 times higher odds of having diabetes compared to those who exercise regularly. Non-smokers have 87.5% lower odds of having diabetes compared to smokers.

Table 4.4: Multivariate logistic regression for diabetes

| **Outcome** | **Covariate** | **Parameter estimate** | **Standard error** | **p-value** |
|---|---|---|---|---|
| Diabetes | Intercept | -2.556 | 1.056 | 0.015510 |
| | Middle income | 1.767 | 4.792e-01 | 0.000227 |
| | Unemployment | 1.916 | 5.366e-01 | 0.000356 |
| | No regular excercise | 1.460 | 5.487e-01 | 0.007789 |
| | No smoking | -2.083 | 5.311e-01 | 8.77e-05 |

From Table 4.5, middle income individual have 2.82 times high odds to mental health disorders compared to those with low income.

Table 4.5: Multivariate logistic regression for mental health disorders

| **Outcome** | **Covariate** | **Parameter estimate** | **Standard error** | **p-value** |
|---|---|---|---|---|
| Mental health disorders | Intercept | 2.70911 | 1.09969 | 0.0138 |
| | Middle Income | 1.03182 | 0.48558 | 0.0336 |

Table 4.6 the unemployed have 15.89 times high odds of CRD compared to employed individuals.

Table 4.6: Multivariate logistic regression for chronic respiratory diseases

| **Outcome** | **Covariate** | **Parameter estimate** | **Standard error** | **p-value** |
|---|---|---|---|---|
| Chronic respiratory diseases | Intercept | -3.409 | 1.446 | 0.01841 |
| | Unemployment | 2.766 | 1.054 | 0.00868 |

From Table 4.7 middle income individuals have 9.88 times higher odds of obesity compared to low income individuals. Unemployed have 7.60 times higher odds of obesity compared to the unemployed. People who do not exercise regularly have 3.54 times higher odds to becoming obese compared to those exercise. Non-smokers have 85% lower odds of obesity compare to smokers.

Table 4.7: Multivariate logistic regression for obesity

| **Outcome** | **Covariate** | **Parameter estimate** | **Standard error** | **p-value** |
|---|---|---|---|---|
| Obesity | Intercept | -2.631 | 1.085 | 0.015364 |
| | Middle income | 2.291 | 5.143e-01 | 8.43e-06 |
| | Unemployment | 2.028 | 5.652e-01 | 0.000332 |
| | No regular excercise | 1.263 | 5.046e-01 | 0.012318 |
| | No smoking | -1.906 | 4.743e-01 | 5.84e-05 |

It was noted that the chronic kidney disease model had no significant predictors.

## 4.4 Canonical correlation analysis

CCA was employed to explore the relationship between two sets of variables. In this analysis we examined SDOH as the explanatory variables and health outcomes as the response variables. The canonical correlation $\rho_k$, measures the strength of the linear relationship between the canonical covariates of the explanatory and response variables. The results show six canonical correlations:

$$\hat{\rho} = (0.88376030, 0.78486237, 0.67567489, 0.30198022, 0.19194451, 0.0937527)6$$

The first three canonical correlations $\rho_1 = 0.8838, \rho_2 = 0.7849$, and $\rho_3 = 0.6757$ show a strong relationship between the pairs of canonical variables. The subsequent pairs show a weak relationship, as the magnitude decreases, suggesting the components do not contribute significantly to explaining the relationship between response and explanatory variables, as shown by Figure 4.6.

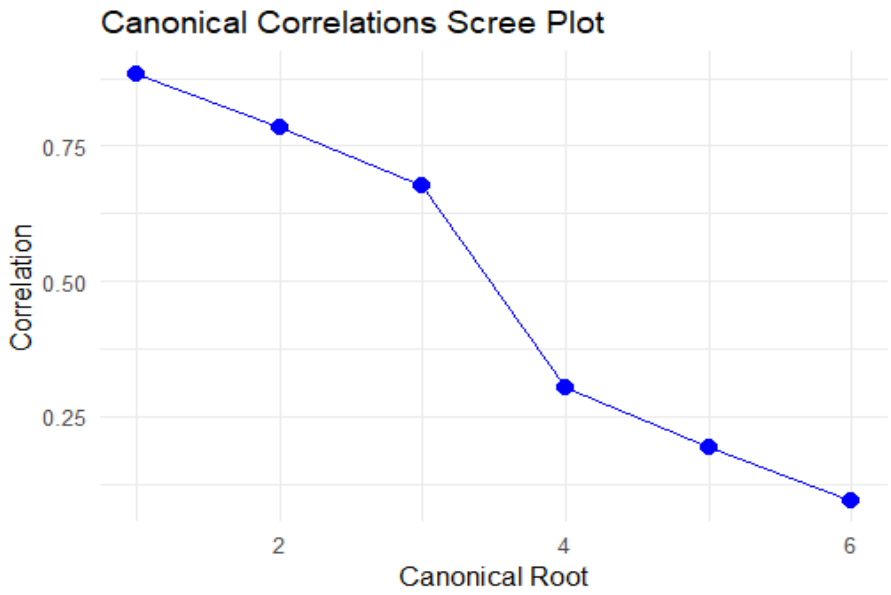



Figure 4.6: A scree plot of canonical correlations

Coefficients show the weights assigned to each variable in the canonical variates for both explanatory and response variables. For each column it corresponds to a canonical covariate,

and for each row to a variable defined in Equation 3.24. The first explanatory canonical varaite; smoking(0.067) and informal housing (0.079) have relatively he highest positive contributions. The second explanatory canonical variate; informal housing (0.158) and employment (0.046) contribute positively. Smoking (-0.112) contribute negatively. The third explanatory canonical variate; income contributes the highest positively while informal hosing (-0.1428) has the highest negative contribution.

Response canonical variates is computed as defined in Equation 3.24. The first response canonical variate; CVD(0.075) and CKD (0.098) contribute positively. The second response canonical variate; CRD (0.21) contributes strongly positively and CVD (-0.162) contributes strongly negatively. The third response canonical variate; CRD (-0.1621) contributes negatively the highest. Diabetes and obese (0.1092) contribute relatively the highest positively.

This measures the total variance in the explanatory and response variables captured. Explanatory variable variance explained, $5.348624e-05, 2.648624e-06, 3.754055e-06$, the values show canonical variates explain a small fraction of variance suggesting while canonical correlations are strong, explanatory variables do not have a strong linear relationship with the response variable. Response variable variance explained, $0.0131379, 0.0008587, 0.003828$, the variance is also low.

## 4.5 Covariate assessment

Covariate distribution was simulated using the Wishart distribution, to assess the impact of different covariance structure on model performance. The simulation began by computing the empirical covariance matrix of the explanatory variables and ensuring its positive definiteness. Multiple covariance matrices with varying degrees of freedom (ranging from 100 to 500 in increments of 100) were generated using the Wishart distribution. For each simulated covariance matrix, new explanatory variables were generated and fitted logistic regression model to the original response variables. The variation of model performance was assessed using AIC, BIC and log-likelihood.

AIC and BIC values were computed for each model under different covariance structures. The lower the AIC and BIC, the better the model fit. The log-likelihood ranges from -86.0 to -139.5. Models with lower AIC and BIC values have a higher log-likelihood values. The results, as indicated in Table 4.8, indicated the performance of the logistic regression model are relatively stable across different degrees of freedom. Models with higher degrees of freedom, that is DF = 500, show better performance, with lower AIC and BIC and higher log-likelihood values.

Table 4.8: A summary of AIC, BIC and log-likelihood values for each degree of freedom

| **Degrees of freedom** | **Variable** | **AIC** | **BIC** | **Log-likelihood** |
|---|---|---|---|---|
| 100 | 1 | 275.2732 | 312.1402 | -126.7136 |
| | 2 | 299.3179 | 336.0308 | -138.659 |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| 200 | 1 | 282.1344 | 318.8473 | -130.0672 |
| | 2 | 300.9734 | 337.6863 | -139.4867 |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| 300 | 1 | 286.3236 | 323.0365 | -132.1618 |
| | 2 | 292.2975 | 329.0105 | -135.1488 |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| 400 | 1 | 283.5209 | 320.2338 | -130.7605 |
| | 2 | 290.3394 | 327.0523 | -134.1697 |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| 500 | 1 | 274.9474 | 311.6603 | -126.4737 |
| | 2 | 288.8967 | 325.6096 | -133.4483 |
| | ⋮ | ⋮ | ⋮ | ⋮ |

# Chapter 5

# Discussions, Conclusions and Recommendations

## 5.1 Introduction

This chapter provides a summary of the findings. The study employed multivariate logistic regression and canonical correlation analysis to analyze the impact of SDOH on multiple binary health outcomes, while assessing the covariates. The implication of the findings, conclusions and recommendation are also highlighted in this study. Furthermore, the limitation for this study are stated to provide a unbiased view of the research outcomes.

## 5.2 Discussions

Comparing income level with diabetes, high income individuals have the lowest prevalence to diabetes. This is because most high income individuals have better access to treatment and can adopt healthy lifestyle. Park et al. (2023), explored income and diabetes risk using cox proportional hazard model, and identifies the similar pattern. The higher income someone gets the lower the risk of diabetes.

Income and employment, middle income (compared to high income) and unemployed individuals have higher odds to develop health conditions like diabetes, mental health disorders, obesity and chronic respiratory diseases. This is consistent with studies that show the impact of socioeconomic instability on health van de Ven et al. (2023), aimed to investigate employment and its impact to having chronic diseases using restricted mean

survival time. In addition, this study reveled that middle income had higher odds of getting diabetes compared to low-income individuals. However, Horestani and M.MehdiOwrang (2024) identified that people with low income have a high prevalence of diabetes using machine learning for prediction.

Based on this study, middle income individuals have a higher risk of diabetes despite previous literature showing otherwise. This may be because, middle income individuals have access to processed foods with high calories. In addition, they may experience job related stress and long office hours, leading to poor lifestyle decisions. Compared to the low income individuals who have o do more physical activities therefore reducing risks to diabetes.

Lifestyle factors, smoking and lack of physical exercise were associated with increased odds to obesity, cardiovascular diseases, and diabetes. Szydlowski et al. (2020), emphasized on individuals with low socioeconomic backgrounds have a higher likelihood of developing chronic health conditions, showing the role of SDOH in health disparities. Mumtaz et al. (2023), analyzed cross sectional survey to analyze the relation between exercise and CVD reduction and revealed those who exercised showed a 76% lower risk of CVD.

The model in this study did not find CKD significant to any SDOH. However, Al Kibria and Hasan (2022), analyses CKD prevalence in relation to income, cited that middle and low income individuals had a higher prevalence to CKD compared to high income individuals. In addition, Antunes et al. (2023) highlighted that lifestyle factors such as physical activities help manage CKD with the aging generation. CKD however not significant may develop due to underlying conditions such as diabetes and obesity rather than influencing SDOH directly.

SDOH such as age, education level, informal housing, any form of food insecurity, exposure to air pollution, and exposure to violent crime were not found to be significant to any health outcomes. This variables could have been insignificant because of non-linear association among them. Ozieh et al. (2021), proved that food insecurity was statistically insignificant in influencing CKD and diabetes. In addition, Dominguez-Dominguez et al. (2022) highlighted that food and housing insecurity were associated with mental health disorders like depression. Low income and aged individuals were at increased risk of food insecurity with chronic conditions (Jih et al., 2018). Hamilton et al. (2024) posed a question if the burden of health

outcomes led to food insecurity or it is the burden food insecurities that has led to chronic health outcomes.

Bailey et al. (2020) identified that air pollution can be inhaled or ingested. Increased exposure to any form of air pollution increases the risk of obesity and diabetes since air pollution can cause inflammation. Through cross sectional study, Aguiar et al. (2024), proved educational level had a negative step-wise association with adults with diabetes. This means that increased level of education led to decreased risks of diabetes, since more knowledge helped people have better lifestyle to manage diabetes.

High risk neighborhoods especially those with high violent or crime rate, were associated with obesity. This is because, as noted in our study, physical exercise,can reduce the odds of obesity. Living areas with violent crimes can reduce walk ability and exercise hence increased risk to obesity (Cunningham-Myrie et al., 2021).

In this study, diabetes and obesity have shared SDOH predictors. This highlights that they stem from overlapping inequalities, suggesting integrated prevention strategies. This shows that some health outcomes may also be associated simultaneously. Antunes et al. (2023), stated that CKD is also influenced by diabetes. Individuals with obesity were associated with high prevalence to cardiovascular diseases. Asif (2024), highlighted the impact of obesity to increasing CVD and the need for early detection of obesity to reduce advance chronic diseases.

LASSO regularization was able to address high multicollinearity without removing important variables. MLR showed a good predictive performance for CRD, diabetes, and CVD, with ROC above 0.85. Extreme values in covariates reduce predictive accuracy of logistic models making logistic highly sensitive (Idris et al., 2024). This study, Figure 4.5, proved that there were no outliers in the covariates therefore the results in this study are reliable.

CCA showed a strong correlation between SDOH and health outcomes. It identified a correlation between cardiovascular diseases, chronic respiratory and kidney diseases, and diabetes. It identifies smoking, informal housing, income, and employment with multiple health conditions. However, Loperfido (2021) points out that CCA may not capture the

nonlinear relation between the SDOH and health outcomes and this might contribute to the low variance explained by the CCA model.

The simulation of covariates following Wishart distribution, showed logistic regression models were robust to changes in the covariance structure. The higher the degree of freedom, the better the model performance. Makled and Cheng (2024), analyzed how degrees of freedom and covariance matrix affect Wishart distribution and concluded that larger degrees of freedom led to a stable distribution

## 5.3 Conclusions

The study identifies the significant SDOH, lifestyle and socioeconomic factors, providing insight to the health inventions. The findings show that there is need to address these factors in order to reduce health disparities.The use of MLR and CCA proved the effectiveness of multivariate statistical methods to help understand the interaction between the SDOH and health outcomes. This findings will contribute to the existing literature on SDOH and health outcomes and disparities by offering methodological framework for future research. The study emphasizes on targeted interventions in order to improve the overall population health outcome.

## 5.4 Recommendations

Policy makers should focus on addressing the socioeconomic inequalities. They should improve employment opportunities, and create affordable housing. This will help reduce prevalence of chronic health conditions. Health initiatives should emphasize on promoting physical activities and reducing smoking rates. The campaigns should encourage lifestyle modification to reduce the risk of obesity and diabetes. Targeted intervention could be developed for the low-income and unemployed individuals, for those at higher risk of

developing chronic condition. In addition, providing subsidize gym memberships for low income individuals instead of disease specific approach

Studies should explore the simultaneous impact of SDOH on multiple health outcomes, using large and more diverse datasets. One can consider longitudinal data would provide robust understanding the relationship between the variables. The role of upstream determinants such as governance and policy in health disparities. Alternative statistical methods might be considered especially those that capture complex interaction beyond CCA and MLR.

## 5.5 Limitations of the Study

While the study provided valuable insights, there were limitations which included; Low variance explained: CCA showed strong canonical correlation bt the variance explained by the canonical variates was low. This indicated that the model explanatory power was limited. Data limitation: cross-sectional data limits establishment of relationships between SDOH and health outcomes. In addition, the dataset sample size was relatively small at 208.

# Bibliography


ABConvert (2024). The power of multivariate analysis exploring techniques and benefits. *ABConvert*.

Adza, W. K., Hursthouse, A. S., Miller, J., and Boakye, D. (2023). Exploring the joint association of road traffic noise and air quality with hypertension using qgis. *International Journal of Environmental Research and Public Health*, 20(3):2238.

Aguiar, C., Hurwitz, E. L., Wu, Y. Y., and Yamanaka, A. B. (2024). Examining diabetes status by the social determinants of health among adults in hawai 'i. *Hawai'i Journal of Health & Social Welfare*, 83(8):216.

Al Kibria, G. M. and Hasan, M. Z. (2022). Income disparities in prevalence and trends of chronic kidney disease among us adults, 2003–18. *Journal of Public Health*, pages 1–9.

Alderwick, H., Hutchings, A., Briggs, A., and Mays, N. (2021). The impacts of collaboration between local health care and non-health care organizations and factors shaping how they work: a systematic review of reviews. *BMC Public Health*, 21:1–16.

Antunes, C., Antunes, D., Ponciano, A., Abrantes, R., and Miragaia, A. (2023). The role of physical exercise in chronic kidney disease.

Asif, M. (2024). Obesity: A profoundly under recognized chronic disease; and its impacts on cardiovascular disease. *South Dakota Medicine*, 77(8).

Author, G. (2023). Multivariate analysis. *code institute*.

Bailey, M. J., Naik, N. N., Wild, L. E., Patterson, W. B., and Alderete, T. L. (2020). Exposure to air pollutants and the gut microbiota: a potential link between exposure, obesity, and type 2 diabetes. *Gut Microbes*, 11(5):1188–1202.

Beech, B. M., Ford, C., Thorpe Jr, R. J., Bruce, M. A., and Norris, K. C. (2021). Poverty, racism, and the public health crisis in america. *Frontiers in public health*, 9:699049.

Bell, J., Sharma, S., Malone, S., Levy, M., Reast, J., Ciecieląg, J., Gogolina, S., Ansons, T., Fourie, S., Braz, R., et al. (2021). Targeting interventions for hiv testing and treatment uptake: An attitudinal and behavioural segmentation of men aged 20–34 in kwazulu-natal and mpumalanga, south africa. *PloS one*, 16(3):e0247483.

Bhavnani, S. K., Zhang, W., Bao, D., Raji, M., Ajewole, V., Hunter, R., Kuo, Y.-F., Schmidt, S., Pappadis, M. R., Smith, E., et al. (2023). Subtyping social determinants of health in all of us: Network analysis and visualization approach. *Medrxiv*.

Bíró, É., Kovács, S., Veres-Balajti, I., Ádány, R., and Kósa, K. (2021). Modelling health in university students: Are young women more complicated than men? *International Journal of Environmental Research and Public Health*, 18(14):7310.

Bykhovskaya, A. and Gorin, V. (2023). High-dimensional canonical correlation analysis. *arXiv preprint arXiv:2306.16393*.

Camacho, J., Wasielewska, K., Bro, R., and Kotz, D. (2024). Interpretable feature learning in multivariate big data analysis for network monitoring. *IEEE Transactions on Network and Service Management*.

C.D.C (2024a). Health disparities in hiv, viral hepatitis,stds and tuberclosis.

C.D.C (2024b). Social determinants of health (sdoh).

Chelak.K, . C. (2023). The role of social determinants of health in promoting health equality: A narrative review. *Cureus*, Cureus vol. 15,1 e33425.

Council, D. P. (2023). The us playbook to address social determinants of health. *The White House: Office of Science and Technology Policy*.

Cunningham-Myrie, C., Theall, K. P., Younger-Coleman, N., Greene, L.-G., Lyew-Ayee, P., and Wilks, R. (2021). Associations of neighborhood physical and crime environments with obesity-related outcomes in jamaica. *PLoS One*, 16(4):e0249619.

Dahlgren, G. and Whitehead, M. (2021). The dahlgren-whitehead model of health determinants: 30 years on and still chasing rainbows. *Public health*, 199:20–24.

Dhlamini, B. (2023). Redefining social determinants of health: The social realities of health. *The Student Midwife*.

Dominguez-Dominguez, L., Campbell, L., Barbini, B., Fox, J., Nikiphorou, E., Goff, L., Lempp, H., Tariq, S., Hamzah, L., and Post, F. A. (2022). Associations between social determinants of health and co-and multi-morbidity in people of black ethnicities with hiv. *Aids*, pages 10–1097.

Ebrahimi Kalan, M., Jebai, R., Zarafshan, E., and Bursac, Z. (2021). Distinction between two statistical terms: multivariable and multivariate logistic regression. *Nicotine and Tobacco Research*, 23(8):1446–1447.

Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer-Verlag New York, New York, NY.

Fernandes, A. A. T., Figueiredo Filho, D. B., Rocha, E. C. d., and Nascimento, W. d. S. (2021). Read this paper if you want to learn logistic regression. *Revista de Sociologia e Política*, 28:006.

Gómez, C. A., Kleinman, D. V., Pronk, N., Gordon, G. L. W., Ochiai, E., Blakey, C., Johnson, A., and Brewer, K. H. (2021). Addressing health equity and social determinants of health through healthy people 2030. *Journal of public health management and practice*, 27(Supplement 6):S249–S257.

Gu, L., Cheng, Y., Phillips, D. R., Rosenberg, M., Yang, L., Wang, L., and Li, H. (2021). Does social capital interact with economic hardships in influencing older adults' health? a study from china. *International Journal for Equity in Health*, 20:1–12.

Hahn, R. A. (2021). What is a social determinant of health? back to basics. *Journal of public health research*, 10(4):jphr–2021.

Hall, R. L. and Jacobson, P. D. (2018). Examining whether the health-in-all-policies approach promotes health equity. *Health Affairs*, 37(3):364–370.

Hamilton, A., Beneke, A. A., Meisel, E., Zhang, C., Gao, H., and Portillo-Romero, J. (2024). Associations between social determinants of health and outcomes of chronic medical conditions. *Cureus*, 16(8).

Healthy People 2030, U. D. o. H., Human Services, O. o. D. P., and Promotion, H. (2021). Social determinants of health.

Holbert, S. E., Andersen, K., Stone, D., Pipkin, K., Turcotte, J., and Patton, C. (2022). Social determinants of health influence early outcomes following lumbar spine surgery. *Ochsner Journal*, 22(4):299–306.

Horestani, F. J. and M.MehdiOwrang, O. (2024). Predicting diabetes with machine learning analysis of income and health factors. *ArXiv*, abs/2404.13260.

Hunter, B. D., Neiger, B., and West, J. (2011). The importance of addressing social determinants of health at the local level: the case for social capital. *Health & social care in the community*, 19(5):522–530.

Idris, H. I., Mohammed, A., Salisu, U. F., Balansana, K. I., Abdulazeez, D., and Danrimi, N. H. (2024). Evaluating the performances of robust logistic regression models in the presence of outliers.

Jih, J., Stijacic-Cenzer, I., Seligman, H. K., Boscardin, W. J., Nguyen, T. T., and Ritchie, C. S. (2018). Chronic disease burden predicts food insecurity among older adults. *Public health nutrition*, 21(9):1737–1742.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6th edition.

Khetpal, S., Lopez, J., Redett, R. J., and Steinbacher, D. M. (2021). Health equity and healthcare disparities in plastic surgery: what we can do. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 74(12):3251–3259.

Krause, T. M., Schaefer, C., and Highfield, L. (2021). The association of social determinants of health with health outcomes. *Am J Manag Care*, 27(3):e89–e96.

Loperfido, N. M. R. (2021). Canonical correlations and nonlinear dependencies. *Symmetry*, 13(7):1308.

Lupparelli, M. and Mattei, A. (2020). Joint and marginal causal effects for binary non-independent outcomes. *Journal of Multivariate Analysis*, 178:104609.

Mahendran, M., Lizotte, D., and Bauer, G. R. (2022). Quantitative methods for descriptive intersectional analysis with binary health outcomes. *SSM-population health*, 17:101032.

Makled, R. A. and Cheng, W. (2024). Exploring multivariate statistics: Unveiling the power of eigenvalues in wishart distribution analysis. *Contemporary Mathematics*, pages 4054–4063.

Mandalia, K., Ames, A., Parzick, J. C., Ives, K., Ross, G., and Shah, S. (2023). Social determinants of health influence clinical outcomes of patients undergoing rotator cuff repair: a systematic review. *Journal of Shoulder and Elbow Surgery*, 32(2):419–434.

Mumtaz, M. T., Khan, S., Ikram, U., Khan, M. M. F., Khan, M. A., and Salman, A. (2023). Investigating the dynamic relationship between exercise and cardiovascular disease risk reduction. *Pakistan Journal of Medical & Health Sciences*, 17(04):642–642.

Nguyen, T., Barefield, A., and Nguyen, G.-T. (2021). Social determinants of health associated with the use of screenings for hypertension, hypercholesterolemia, and hyperglycemia among american adults. *Medical Sciences*, 9(1):19.

Nowrozy, R. (2023). Machine learning model for chronic disease prediction. *Journal of Biomedical Research Environmental Sciences*.

Organization, W. H. et al. (2010). A conceptual framework for action on the social determinants of health.

Osmick, M. J. and Wilson, M. (2020). Social determinants of health—relevant history, a call to action, an organization's transformational story, and what can employers do?

Ozieh, M. N., Garacci, E., Walker, R. J., Palatnik, A., and Egede, L. E. (2021). The cumulative impact of social determinants of health factors on mortality in adults with diabetes and chronic kidney disease. *BMC nephrology*, 22:1–10.

Park, J. C., Nam, G. E., Yu, J., McWhorter, K. L., Liu, J., Lee, H. S., Lee, S.-S., and Han, K. (2023). Association of sustained low or high income and income changes with risk of incident type 2 diabetes among individuals aged 30 to 64 years. *JAMA network open*, 6(8):e2330024–e2330024.

Parsons, K., Mulugeta, M. G., Bailey, G., Gillespie, S., Johnson, L. M., Myers, H. E., Reisner, A., and Blackwell, L. S. (2024). Association between social determinants of health and pediatric traumatic brain injury outcomes. *Frontiers in neurology*, 15:1339255.

Ramadan, Ahmed, A. K. A. T. A. E.-S. and Abdel-Fatah, N. A. (2020). A multivariate data analysis approach for investigating daily statistics of countries affected with covid-19 pandemic. *Heliyon*, 6, no. 11.

Ray, R., Lantz, P. M., and Williams, D. (2023). Upstream policy changes to improve population health and health equity: a priority agenda. *The Milbank Quarterly*, 101(Suppl 1):20.

Szydlowski, S., Szydlowski, S., Luliak, M., and Luliak, M. (2020). Prevention of disease-related mortality from chronic non-communicable diseases. 11:28–33.

Tabachnick, B. G. (2007). Using multivariate statistics. *Alyn and Bacon*.

Tippins, N. (2023). What is covariate in statistics? Blog post. Accessed: 2024-06-02.

van de Ven, D., Robroek, S. J., Burdorf, A., and Schuring, M. (2023). Inequalities in the impact of having a chronic disease on entering permanent paid employment: a registry-based 10-year follow-up study. *J Epidemiol Community Health*, 77(7):474–480.

Victorino, C. C. and Gauthier, A. H. (2009). The social determinants of child health: variations across health outcomes–a population-based cross-sectional analysis. *BMC pediatrics*, 9:1–12.

Vo, A., Tao, Y., Li, Y., Albarrak, A., et al. (2023). The association between social determinants of health and population health outcomes: Ecological analysis. *JMIR Public Health and Surveillance*, 9(1):e44070.

Wan, W., Li, V., Chin, M. H., Faldmo, D. N., Hoefling, E., Proser, M., and Weir, R. C. (2022). Development of prapare social determinants of health clusters and correlation with diabetes and hypertension outcomes. *The Journal of the American Board of Family Medicine*, 35(4):668–679.

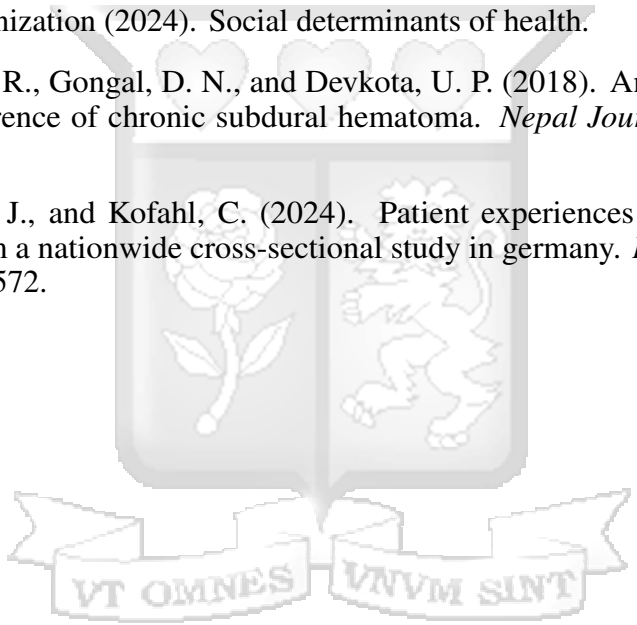
Whitman, A., De Lew, N., Chappel, A., Aysola, V., Zuckerman, R., and Sommers, B. D. (2022). Addressing social determinants of health: Examples of successful evidence-based strategies and current federal efforts. *Off Heal Policy*, 1:1–30.

Wilkinson, R. D. (2024). 7.2 the wishart distribution | multivariate statistics. https://rich-d-wilkinson.github.io/MATH3030/7.2-the-wishart-distribution.html. Accessed: July 2, 2024.

World Health Organization (2024). Social determinants of health.

Yogi, N., Nepal, P. R., Gongal, D. N., and Devkota, U. P. (2018). Analysis of risk factors predicting recurrence of chronic subdural hematoma. *Nepal Journal of Neuroscience*, 15(3):32–38.

Ziegler, E., Klein, J., and Kofahl, C. (2024). Patient experiences and needs in cancer care–results from a nationwide cross-sectional study in germany. *BMC Health Services Research*, 24(1):572.



VT OMNES
VNVM SINT

# Appendix A

# Similarity report

Faith-Musyoki.pdf

ORIGINALITY REPORT

| 13% | 12% | 7% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.stat.cmu.edu<br>Internet Source | 4% |
| 2 | Submitted to Strathmore University<br>Student Paper | 2% |
| 3 | documents.mx<br>Internet Source | 2% |
| 4 | slidelegend.com<br>Internet Source | 1% |
| 5 | publichealth.jmir.org<br>Internet Source | <1% |
| 6 | cdr.lib.unc.edu<br>Internet Source | <1% |
| 7 | Nathan W. Link, Meghan A. Novisky, Chantal Fahmy. "Handbook on Contemporary Issues in Health, Crime, and Punishment", Routledge, 2024<br>Publication | <1% |
| 8 | Submitted to National University<br>Student Paper | <1% |
| 9 | stat.cmu.edu<br>Internet Source | <1% |
| 10 | Samuel E. Holbert, Kristina Andersen, Deborah Stone, Karen Pipkin, Justin Turcotte, Chad Patton. "Social Determinants of Health | <1% |

| | | |
|---|---|---|
| | Influence Early Outcomes Following Lumbar Spine Surgery", Ochsner Journal, 2022<br>Publication | |
| 11 | www.mdpi.com<br>Internet Source | <1% |
| 12 | Priyanka Subramani, Kalpanapriya Dhakshnamoorthy. "Spatio-Temporal Analysis and Prediction by Logistic Regression of Respiratory Diseases in India", Contemporary Mathematics, 2025<br>Publication | <1% |
| 13 | www.ncbi.nlm.nih.gov<br>Internet Source | <1% |
| 14 | Submitted to Middle Tennessee State University<br>Student Paper | <1% |
| 15 | colors-newyork.com<br>Internet Source | <1% |
| 16 | Submitted to University of Mississippi Medical Center (Jackson, MS)<br>Student Paper | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 25 words |
| Exclude bibliography | On | | |

# Appendix B

# Ethical clearance confirmation

**11th February 2025**

Ms Musyoki Faith,
faith.musyoki@strathmore.edu

Dear Ms Musyoki,

**RE: Analyzing the Impact of Social Determinants of Health on Multiple Binary Outcome using Multivariate Statistical Method**

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2644/25.** The approval period is from **11th February 2025 to 10th February 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

**Mr Ambrose Rachier,**
**Chairperson; SU-ISERC**

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

# Appendix C

# R code

https://github.com/Faith897/Thesis-analysis-code