The Application of Machine Learning-Based Prediction Models for Cardiometabolic Risk Among a Representative US Adult Population: A Cross-Sectional Study of NHANES 1999-2006

Jijie Xu

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science (Mathematics and Statistics) at

Concordia University

Montreal, Quebec, Canada

April 2023

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Jijie Xu

Entitled: The Application of Machine Learning-Based Prediction Models for Cardiometabolic Risk Among a Representative US Adult Population: A Cross-Sectional Study of NHANES 1999-2006

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics and Statistics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Arusharka Sen ________________________ Chair

Dr. Simone Brugiapaglia ___________________ Examiner

Dr. Lisa Kakinami ________________________ Thesis Supervisor(s)

Approved by Dr. Yogendra P. Chaubey

____

Chair of Department or Graduate Program Director

Dr. Pascale Sicotte

____

Dean of Faculty

ABSTRACT

The Application of Machine Learning-Based Prediction Models for Cardiometabolic Risk Among a Representative US Adult Population: A Cross-Sectional Study of NHANES 1999-2006

Jijie Xu

Introduction: Common measures of adiposity (such as body mass index) are only proxies. In contrast, dual-energy X-ray absorptiometry (DXA) is more precise to measure body composition. Therefore, this thesis utilized an unsupervised machine learning technique to group individuals based on similarities in their fat and muscle mass body-composition from DXA. Associations between the newly developed body-composition phenotypes with cardiometabolic risks were compared to phenotypes using a median split.

Methods: Data were collected from National Health and Nutrition Examination Survey (NHANES: 1999-2006 cycles, *n*=5,566; split into 70/30% training and test datasets), a representative U.S. population. The K-means cluster phenotypes based on partitioning observations from deciles of fat-mass and muscle-mass adjusted for age and sex were identified. Model fit was assessed using the silhouette and elbow method. Performance of logistic regression models to identify unfavorable cardiometabolic risks using either the K-means or the 50th percentile cut-off phenotypes was assessed with the area under the receiver operating characteristic (ROC-AUC). Analyses were performed separately for males and females and incorporated weighting and the complex sampling design.

Results: Optimal models were 2-means and 4-means k-clusters. ROC-AUCs from 2-means cluster models to identify cardiometabolic risk factors had the lowest predictive power (0.52 to 0.63). The ROC-AUCs from 50th percentile cut-off phenotypes and 4-means cluster phenotypes were higher (0.56-0.66, 0.57-0.67, respectively).

Discussion: Although the 4-means clustering was superior to the 50th percentile cut-off in predicting cardiometabolic risk, the ROC-AUCs were generally poor. Future work should investigate whether performing K-means clustering for each specific age improves their prediction.

ACKNOWLEDGEMENTS

Foremost, I would like to express my heartfelt thanks to my supervisor, Professor Dr. Lisa Kakinami, for her unwavering support, endless encouragement, and continuous guidance throughout my Master's studies and research. Her expertise, and valuable insights have been instrumental in shaping my research and academic development. I am truly grateful for her mentorship and inspiration.

I would also like to thank Professor Dr. Arusharka Sen for his kindness, and dedication to teaching during my Master's classes. His support, and encouragement have been invaluable in helping me navigate the challenges of graduate school and develop my research skills.

In addition, I would like to express my appreciation to the Department of Mathematics and Statistics of Concordia University for providing me this learning opportunity and I am grateful to Professor Dr. Galia Dafni for her support, and encouragement during my studies.

Finally, I would like to extend my thanks to my parents, Sen and Zhirong, for their continued support of my studies and their selfless love and devotion to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARI	Adjusted Rand Index
ASMI	Appendicular Skeletal Muscle Index
AUC	Area Under the ROC Curve
BMI	Body Mass Index
DBP	Diastolic Blood Pressure
DXA	Dual-energy X-ray Absorptiometry
FMI	Fat Mass Index
FPR	False Positive Rate
HA-HM	High Adiposity with High Muscle
HA-LM	High Adiposity with Low Muscle
HDL	High-density Lipoprotein
LA-HM	Low Adiposity with High Muscle
LA-LM	Low Adiposity with Low Muscle
LDL	Low-density Lipoprotein
LMS	Lambda-Mu-Sigma
MCAR	Missing Completely at Random
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
PSUs	Primary Sample Units
ROC	Receiver Operator Characteristic

SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
TPR	True Positive Rate
WCSS	Within-Cluster Sum of Squares

1. Introduction

1.1 Health Risks of Adiposity

Adiposity is defined as an excess accumulation of body fat (Prentice & Jebb, 2001) and is a significant public health concern because it is a risk factor for a number of chronic diseases, including diabetes, high blood pressure, cardiovascular disease, and some types of cancer (Freedman et al., 2007). In Canada, like many other countries, the prevalence of adiposity has been increasing over the years. According to data from Statistics Canada, in 2018, 26.8% of Canadians aged 18 and over were reported with obesity, while an additional 36.3% were reported with overweight (Connor Gorber et al., 2008; Government of Canada, 2019).

One of the most well-established health risks associated with adiposity is cardiovascular disease. Research has demonstrated that individuals with high levels of body fat are at an increased risk of developing cardiovascular diseases, including heart failure and coronary heart disease (Carbone et al., 2019; Hruby & Hu, 2015). In addition to increasing the risk of cardiovascular diseases, excess body fat deposits can also have an impact on other aspects of cardiovascular health, such as high blood pressure and inflammation (Goswami et al., 2020).

Another major health risk associated with adiposity is type 2 diabetes (Hruby & Hu, 2015; Tiwari & Balasundaram, 2022). Adiposity has been shown to impair the body's ability to effectively use insulin to regulate blood sugar levels, which can lead to the development of type 2 diabetes (Al-Goblan et al., 2014). This metabolic disorder can have serious health consequences, including an increased risk of cardiovascular disease and damage to the eyes, nerves and kidneys.

Adiposity has also been implicated in the development of several types of cancer, including breast, colon, endometrial, and kidney cancer (Friedenreich et al., 2021; Hruby & Hu, 2015; Tiwari & Balasundaram, 2022). Although the exact mechanisms by which adiposity increases the risk of these cancers are not fully understood, it is believed that hormonal imbalances and chronic inflammation may play a role (Ramos-Nino, 2013).

In addition to its effects on cardiovascular health and metabolism, adiposity can also lead to musculoskeletal problems, such as joint pain and osteoarthritis (King et al., 2013). The increased weight load on joints in individuals with excess body fat can cause damage and lead to the development of these conditions. Obstructive sleep apnea, a condition characterized by interrupted breathing during sleep, which is also associated with adiposity (Jehan et al., 2017). Excess body fat can put pressure on the airway, leading to partial or complete airway obstruction during sleep.

Thus, adiposity is associated with a range of health risks, including cardiovascular disease, type 2 diabetes, various types of cancer, musculoskeletal problems, and obstructive sleep apnea disease (Freedman et al., 2007; Jehan et al., 2017). Due to the global increase in obesity rates, decreasing obesity and therefore its associated health complications is a major public health concern (Tiwari & Balasundaram, 2022).

1.2 Adiposity Measurements

Accurately measuring adiposity is essential for tracking the prevalence of obesity and identifying individuals at risk for related health problems. This section provides an overview of the most commonly used methods for measuring adiposity.

Body Mass Index (BMI) is one of the most widely used anthropometric field measures used to determine adiposity (Fedewa et al., 2019). It is calculated as an individual's weight in kilograms divided by the square of their height in meters (kg/m^2). BMI is inexpensive and easy to perform to classify adults who are underweight or have normal weight (BMI < 25), as well as with overweight (BMI $\geq$ 25 and <30), or obesity (BMI $\geq$ 30) (*2018 Global Reference List of 100 Core Health Indicators (plus Health-Related SDGs)*, 2018).

Waist circumference is another common assessment of adiposity. It measures the circumference around the waist at its most constricted point. This measurement provides information about the amount of abdominal fat, which has been linked to an increased risk of health problems, such as cardiovascular diseases and type 2 diabetes (Cooper-DeHoff et al., 2010; Postorino et al., 2009; Tsujimoto & Kajio, 2017).

Skinfold thickness measurements are another method used to estimate adiposity. This method involves measuring the thickness of a fold of skin at various points on the body, typically the triceps, subscapular, and suprailiac regions, using a specialized tool called a caliper. The thickness of the skinfold is used to estimate body fat percentage based on regression equations (Ripka et al., 2017). This method is more accurate than BMI and waist circumference but can be subject to variability due to differences in the skill of the practitioner.

However, all of the measures previously described are proxies for assessing adiposity. In contrast, dual energy X-ray absorptiometry (DXA) is a highly accurate method used to measure the body fat and bone density. This method employs a low dose of ionizing radiation, which is absorbed differently by body fat, muscle, and bone. The DXA scanner captures the amount of radiation absorbed and utilizes this information to calculate body fat percentage and bone density (Laskey, 1996). This method provides a detailed breakdown of body composition, including total body fat, fat mass in specific regions, and lean mass.

1.3 Limitations of Measurements

While BMI is a simple and easy method to estimate body fat, it does have some limitations. For instance, it does not take into account differences in muscle mass, bone density, and distribution of fat, which can all affect an individual's weight status (Fedewa et al., 2019). Additionally, it is not suitable for individuals who have a lot of muscle mass, such as athletes as it may overestimate their body fat (Kraemer et al., 2005; Nevill et al., 2006). Thus, BMI presents as an inaccurate obesity classification method (Shah & Braverman, 2012).

Waist circumference only measures the circumference around the waist; as such it does not account for differences in muscle mass that can affect the measurement (Flint et al., 2010). In addition, ideal waist circumference measurements can vary among different ethnic and gender groups, making it difficult to establish a universal benchmark for all populations (Beydoun & Wang, 2009; Bosy-Westphal et al., 2010). Ultimately, waist circumference measures the total amount of fat in the abdominal area, but it cannot differentiate between subcutaneous fat (fat located beneath the skin) and visceral fat (fat located within the abdominal cavity), and it does not provide information about overall body fat or fat distribution (Bosy-Westphal et al., 2010).

Moreover, DXA also has some limitations. Firstly, this method is more expensive and is not as widely available. Secondly, DXA methods expose the individuals to ionizing radiation, which may be a concern for some individuals; children are more susceptible to the harmful effects of due to their longer life expectancy and greater latency period for the manifestation of delayed radiation effects. As a result, children may experience a higher cumulative radiation exposure over their lifetime. (Njeh et al., 1999). Thirdly, DXA measurements are influenced by the size and shape of bones (Blake et al., 2013). Lastly, DXA scans can be affected by various factors such as metal implants, movement during the scan and equipment technical problems, which can impact the accuracy of the results. Thus, the DXA is not clinically useful as a proxy for routine clinical adiposity management (Yumuk et al., 2015).

1.4 Lambda-Mu-Sigma (LMS) Methodology

Lambda-Mu-Sigma (LMS) methodology (Cole, 1990) is a statistical technique used to create sex- and age- specific references curves for growth and development indicators such as weight, height, BMI and others (Tambalis et al., 2015). It involves fitting three curves (Figure 1) to the data: the L-curve (lambda), which represents the Box-Cox power to transform the data to normality; the M-curve (mu), which represents the median of the transformed data; and the S-curve (sigma), which represents the coefficient of variation of the transformed data.

The variable of interest, (in this case weight) is denoted by $y \geq 0$. Suppose that y has median of μ_y and that y^λ is normally distributed. Based on the family of transformations proposed by Box and Cox (1964), it is then defined that the transformed variable:

$$x = \begin{cases} \dfrac{(y/\mu)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln\left(\dfrac{y}{\mu}\right), & \lambda = 0 \end{cases} \tag{1}$$

This transformation maps the median μ_y of y to $x = 0$ and is continuous at $\lambda = 0$. Since λ is a free parameter, the optimal value of λ is that which minimizes the standard deviation (SD) of x.

The SD of x and the coefficient of variation of y are denoted as σ. After x is calculated, the SD score (z-score) of x can be calculated through the standard normal distribution assumption.

$$z = \frac{x}{\sigma} = \begin{cases} \dfrac{(y/\mu)^\lambda - 1}{\lambda\sigma}, & \lambda \neq 0 \\ \dfrac{\ln\left(\dfrac{y}{\mu}\right)}{\sigma}, & \lambda = 0 \end{cases} \tag{2}$$

Assume now that the distribution of y varies with covariate t (in this study, t is age) and that μ, λ, and σ at t are $M(t)$, $L(t)$, and $S(t)$, respectively. To find the centile 100α of y at t is given by:

$$C_{100\alpha}(t) = \begin{cases} M(t)[1 + L(t)S(t)Z_\alpha]^{\frac{1}{L(t)}}, & L(t) \neq 0 \\ M(t)\exp[S(t)Z_\alpha], & L(t) = 0 \end{cases} \quad (3)$$

The curves $M(t)$, $L(t)$, and $S(t)$ are estimated by maximizing the penalized likelihood.

The LMS method allows for the creation of reference curves that can be used to determine whether an individual's growth is normal, delayed, or advanced compared to their peers of the same sex and age (Kelly et al., 2014). The LMS method is widely used in health research to create growth reference charts (e.g., height-for-age, or weight-for-age) and to monitor the health and development of children (Cole, 2021; Cole, 1990; Flegal & Cole, 2013; Cole, 2012).

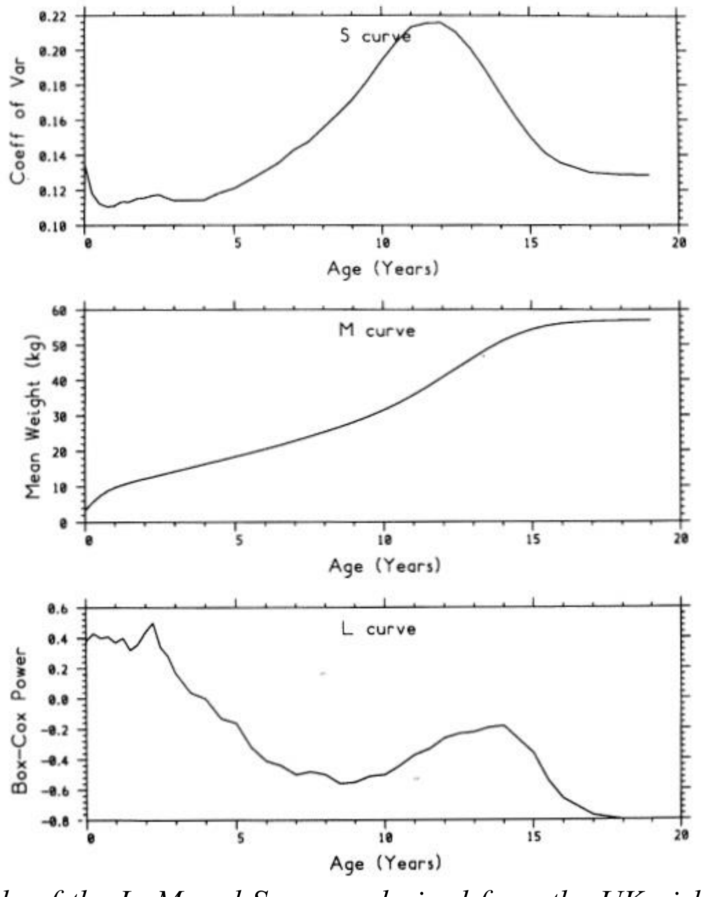


Figure 1: *An example of the L, M and S curves derived from the UK girls' weight standard of Tanner, Whitehouse & Takaishi (1966). (Cole, 1989)*

1.5 Prado et al.'s Phenotypes

In 2014, Prado et al. (2014) adapted the lambda-mu-sigma (LMS) methodology to develop sex- and age- specific references curves to classify dual energy X-ray absorptiometry (DXA) data into phenotypes of appendicular skeletal muscle index (ASMI; kg/m^2) and fat mass index (FMI; kg/m^2). Through the use of the LMS methodology, Prado et al. (2014) defined four distinct phenotypes based on whether an individual fell above (≥ 50th percentile) or below (< 50th percentile) combinations of sex- and age-specific references curves for ASMI and FMI. These phenotypes were labeled as: low adiposity with low muscle (LA-LM: <50th percentile FMI and <50th percentile ASMI), low adiposity with high muscle (LA-HM: <50th percentile FMI and ≥50th percentile ASMI), high adiposity with low muscle (HA-LM: ≥50th percentile FMI and <50th percentile ASMI), and lastly, high adiposity with high muscle (HA-HM: ≥50th percentile FMI and ≥50th percentile ASMI).

Previously, it was demonstrated that Prado's (2014) DXA phenotypes were not better than BMI in predicting cardiometabolic risk (Kakinami et al., 2021). While the results showed that Prado's (2014) DXA phenotypes did provide additional information on body composition, it did not improve the detection of cardiometabolic risks compared to using BMI alone. More specifically, while the ROC-AUCs obtained from BMI models for correctly identifying cardiometabolic risk were relatively weak (varied between 0.57 and 0.68), the ROC-AUCs from DXA phenotypes were the same or lower (ranging from 0.53 to 0.68), suggesting weaker predictive power than the BMI model.

As a result, this thesis aims to develop a better way to classify fat and muscle mass than using the Prado's four mutually exclusive phenotypes based on the median splits. In order to be able to say whether the newly developed phenotypes are superior to the existing Prado's 50th percentile cut-off phenotypes, the performance of the models in identifying cardiometabolic risk factors in a large representative sample will also be assessed.

The rest of this thesis will be organized in the following order: (I) method section expanding on Prado's phenotypes, the statistical approach for identifying phenotypes through clustering and the metrics to validate and assess their performance in predicting cardiometabolic risks; (II) results and (III) discussion and conclusion.

2. Methods

2.1 Data Sources

The data for this study were obtained from the National Health and Nutrition Examination Survey (NHANES). Its primary goal is to appraise and scrutinize the health and nutritional status of both adults and children residing in the United States. Since 1999, NHANES has collected data from ~10,000 people living in the United States every two years in biannual cycles. The samples were representative of the general population through a combination of sampling design (stratified, multistage probability sampling), oversampling, and weighting (Curtin et al., 2012; *National Health and Nutrition Examination Survey (NHANES) - Health, United States*, 2022; *NHANES - About the National Health and Nutrition Examination Survey*, 2022; *NHANES - History*, 2019).

The survey collects data on a wide range of health-related topics through interview and physical examinations, including information on health behaviours, chronic conditions, and use of health care services (King et al., 2021; *NHANES - About the National Health and Nutrition Examination Survey*, 2022) The survey results provide an objective assessment of health status and help people learn about the health of people in the United States (*NHIS - National Health Interview Survey*, 2023). Ethics approval was obtained from the National Center for Health Statistics Ethics Review Board. All participants provided informed consent.

The sample size for the 1999-2006 cycles of NHANES was 41,474. Of this, 18,850 participants were excluded from analysis since they were under the age of 18 years or above the age of 85 years. An additional 1,292 women were removed from analysis because a pregnancy test was positive at examination time, or they said they were pregnant and would thus be unable to have a DXA scan. Lastly, participants were excluded if they were missing data from other key measures of interest (such as BMI, cardiometabolic risk measures, sociodemographic characteristics, etc). The final sample size for analysis was 18,556 respondents, with DXA data in five imputed data sets (n=92,780).

2.2 Outcomes (Cardiometabolic Risk Factors)

The outcomes section of this thesis encompasses a comprehensive evaluation of several essential cardiometabolic risk factors. These factors play a vital role in understanding and assessing an individual's cardiovascular health and metabolic status. Specifically, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, as well as systolic and diastolic blood pressure are investigated.

2.2.1 Total Cholesterol

Total cholesterol contains both low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol. A high level of total cholesterol (>6.2 mmol/L) in the blood could increase a person's risk of developing cardiovascular disease and stroke (Grundy et al., 2019; *InformedHealth.Org*, 2006). The ideal level of total cholesterol could vary depending on an individual's age, sex and overall risk of heart disease.

2.2.2 Low-density Lipoprotein (LDL) Cholesterol

LDL cholesterol is a measure of the cholesterol carried by low-density lipoprotein in the blood. It is commonly referred to as "bad" cholesterol because high levels of LDL cholesterol are associated with an increased risk of heart disease and other cardiovascular problems (*InformedHealth.Org*, 2006). An optimal LDL cholesterol level is generally considered to be less than 2.6 mmol/L, while levels between 3.4 mmol/L and 4.1 mmol/L are considered borderline high. LDL cholesterol levels above 4.1 mmol/L are considered high (Grundy et al., 2019).

2.2.3 High-density Lipoprotein (HDL) Cholesterol

Higher levels of HDL cholesterol may help protect against heart disease by removing excess cholesterol from the blood and carrying it back to the liver for excretion (*InformedHealth.Org*, 2006; Sacks & Expert Group on HDL Cholesterol, 2002). An optimal HDL cholesterol level is generally considered to be 1.0 mmol/L or higher for men and 1.3 mmol/L or higher for women (Grundy et al., 2019).

2.2.4 Triglycerides

Triglycerides are a type of fat found in a person's blood. High triglycerides levels are also associated with an increased risk of cardiovascular disease (Miller et al., 2011). A high triglyceride level is generally considered to be greater than 2.3 mmol/L (Grundy et al., 2019). However, optimal levels may vary depending on individual's age, sex and other risk factors for heart disease.

2.2.5 Systolic and Diastolic Blood Pressure

Hypertension is the sustained elevation of systemic arterial blood pressure, most commonly defined as systolic blood pressure (SBP) ≥ 140 mm Hg or diastolic blood pressure (DBP) ≥ 90 mm Hg (Grundy et al., 2019). Both systolic and diastolic blood pressure are important for assessing individual's risk of developing heart diseases, stroke, and other health problem. The systolic and diastolic blood pressure are measured on the right arm with a sphygmomanometer after having the participant rest for five minutes in a seated position. After excluding the first measure, mean systolic and diastolic blood pressure were calculated based on three consecutive measurements (Muntner et al., 2019).

2.3 Body Composition

Body composition was determined using the whole-body dual-energy X-ray absorptiometry (DXA) with a Hologic QDR 4500A machine. These whole-body measurements were used to derive the ASMI and FMI (Heymsfield et al., 1990). The ASMI represents muscle mass, and its value provides information about an individual's muscle mass relative to their body size. The ASMI was calculated using the DXA lean mass results from the arm and legs divided by height (meters^2) (Hattori et al., 1997). The FMI reflects an individual's fat mass, and its value can be used to understand the distribution of fat in their body. FMI is calculated by dividing the total fat mass by height (meters^2) (Hattori et al., 1997). Both the ASMI and FMI values were classified into phenotypes, which allowed for a more detailed analysis of body composition as described further.

2.3.1 Prado et al.'s 50th Percentile Cut-off Classification

Prado et al. (2014) developed sex- and age- specific references curves for ASMI and FMI for the NHANES 1999-2004 data cycles. Based on these ASMI and FMI references curves, Prado (2014) recommended classifying participants into four mutually exclusive phenotype quadrants according to whether they were above or below the median ASMI or FMI for their age and sex. More specifically, the four phenotypes are:

1) low adiposity and low muscle (LA-LM; FMI deciles 0-49.99 and ASMI deciles 0-49.99)
2) low adiposity and high muscle (LA-HM; FMI deciles 0-49.99 and ASMI deciles 50-100);
3) high adiposity and low muscle (HA-LM; FMI deciles 50-100 and ASMI deciles 0-49.9);
4) high adiposity and high muscle (HA-HM; FMI deciles 50-100 and ASMI deciles 50-100).

2.4 Other Clusters

There are several limitations to using Prado et al. (2014) 50th percentile cut-off phenotypes. Firstly, using the 50th percentile cut-off to classify individuals based on ASMI or FMI for their age and sex may oversimplify the complexity of the data. Secondly, using the 50th percentile cut-off is a static approach and may not reflect changes in body composition over different ages in the life-course. Lastly, using the 50th percentile cut-off reduces the number of groups and may decrease the power to detect significant associations between different phenotypes and other variables of interest.

Thus, health based on concomitant muscle and fat mass measures may be better represented in phenotypes other than above/below the median. An alternative method to identifying these phenotypes is through clustering. It's common to try multiple clustering methods to determine which method best suits the particular dataset and problem. Different clustering methods are better suited for different types of data. For example, K-means is well suited for data that are compact and well separated from each other, while hierarchical clustering is better suited for data with a more complex structure. A brief overview is provided.

2.4.1 Hierarchical Clustering

Hierarchical clustering is a bottom-up approach to clustering (Sasirekha & Baby, 2013). Initially, the data is separated in *n* clusters of one point each (each point is a cluster). Then, two of the current clusters are merged based on similarity at each iteration. The algorithm stops when all points are regrouped in a single cluster. Hierarchical clustering merges (or fuses) clusters based on their similarity (Johnson, 1967; Sasirekha & Baby, 2013). In practice, there are different metrics to assess the similarity (or dissimilarity) between two clusters (Table 1).

Table 1: Different metrics for hierarchical clustering.

Name of metric	Description
Euclidean distance	Euclidean distance is used to measure the dissimilarity between two clusters
Complete linkage	Complete linkage is used to measure maximal intercluster point dissimilarity
Single linkage	Single linkage is used to measure minimal intercluster point dissimilarity
Average linkage	Average linkage is used to measure mean intercluster point dissimilarity
Centroid linkage	Dissimilarity between the centroid of both clusters

James et al., 2021

For a given dissimilarity measure, the hierarchical clustering algorithm goes as follows (James et al., 2021):

Iterate for i = n, …, 2:

1) Compute the $\frac{i\,(i-1)}{2}$ pairwise inter-cluster dissimilarities among the i remaining clusters.
2) Identify the pair of clusters that are the least dissimilar. Merge these two clusters.

The results of the process can then be visualized in a plot called a dendrogram, in order to choose an appropriate number of clusters. The following Figure 2 illustrates a dataset of nine points in $\mathbb{R}^2$ and the dendrogram (right panel) obtained by applying hierarchical clustering (complete linkage) to the raw data (left panel).

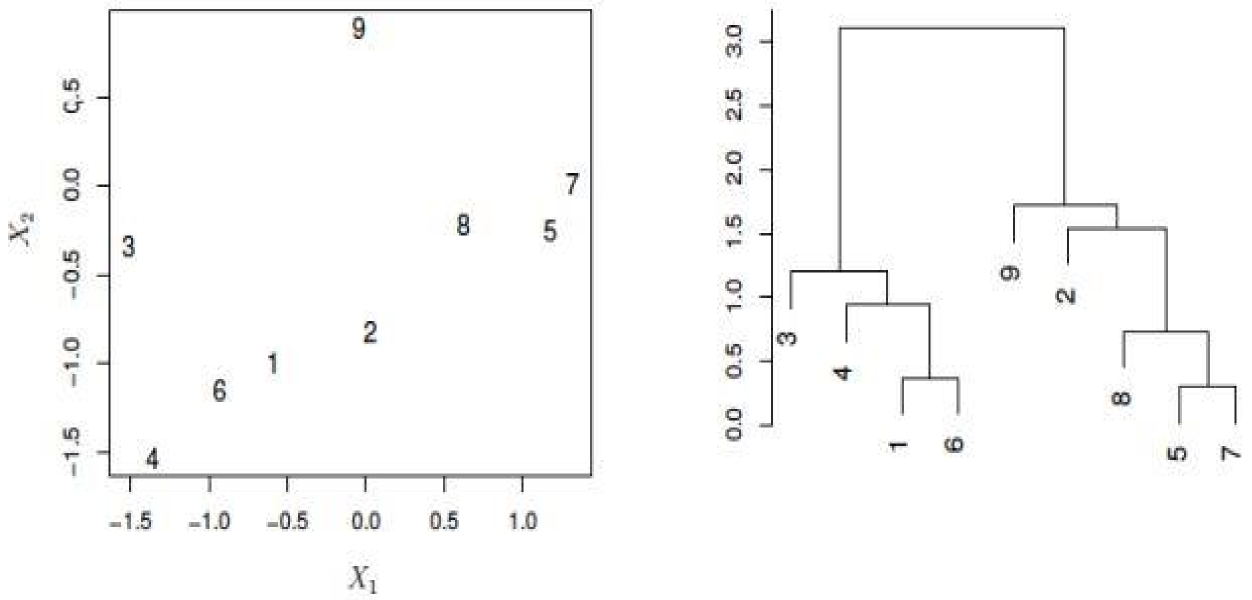


Figure 2: *An illustration of how to interpret a dendrogram with nine observations in two-dimensional space. (James et al., 2021)*

At the bottom of the dendrogram, all tree leaves are separated (all distinct clusters). Then, going progressively up, every time two clusters are merged, their respective branches fuse in the tree. Clusters were merged by order: cluster {5} with cluster {7}, cluster {1} with cluster {6}, and cluster 8 with cluster {5,7}, and so on.

Unlike other clustering methods, hierarchical clustering does not require the user to specify the number of clusters beforehand (Frigui & Krishnapuram, 1999). Instead, the user can determine the number of clusters by cutting the dendrogram at a certain height. However, hierarchical clustering has its limitations. Hierarchical clustering is a computationally expensive algorithm, especially for larger datasets (Murtagh & Contreras, 2012). This is because it requires the calculation of all pairwise distances between objectives, which can become time-consuming. Moreover, the results of hierarchical clustering can be sensitive to the choice of linkage method. Different linkage methods, such as single linkage, complete linkage, and average linkage, can produce different clustering results (Sasirekha & Baby, 2013).

2.4.2 K-means Clustering

K-means clustering is an iterative approach that divides a dataset into K distinct, non-overlapping clusters. To perform K-means clustering, the desired number of cluster K must be selected at the beginning, then the K-means algorithm will assign each observation to exactly one of the K clusters. This algorithm is laid out in the following steps (James et al., 2021):

1) First, simulate K random points in $\mathbb{R}^p$ which will act as initial centroids, then assign the data points to the closest centroid, each centroid corresponds to an initial cluster.

2) Iterate the following until the cluster assignments stop changing:

   a) For each of the K clusters, compute the cluster centroid. The *k*th cluster centroid is the vector of the *p* feature means for the observations in the *k*th cluster.

   b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance (4)).

$$\|x_i - x_{i'}\|^2 = \sum_{j=1}^{p} (x_{i,j} - x_{i,j'})^2 \tag{4}$$

$\|x_i - x_{i'}\|^2$ Is the Euclidean distance between p-dimensional points x_i and $x_{i'}$

Figure 3 shows the progress of the algorithm of K-means clustering algorithm for K=2.

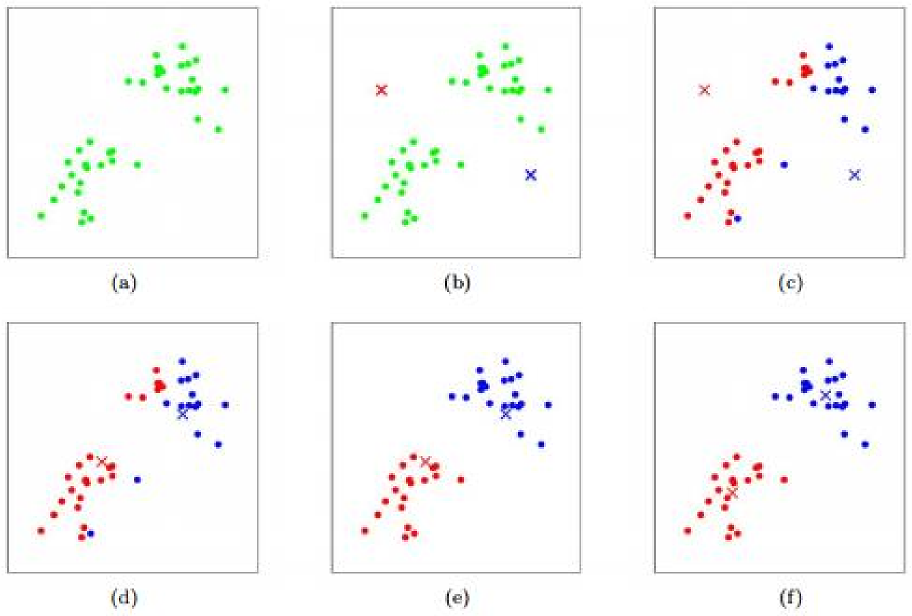


Figure 3: *An example of the K-means algorithm progression. Training data points are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training data points to the closest cluster centroid; then we move each cluster centroid to the mean of the points assigned to it. (Piech et al., 2013)*

The K-means algorithm is guaranteed to decrease the value of objective function (5) (James et al., 2021; Krishna & Narasimha Murty, 1999) at each step.

$$\mathbb{O}(C_1, \dots, C_k) = \left\{ \sum_{k=1}^{K} \frac{1}{card(C_k)} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|^2 \right\} \tag{5}$$

Card (C_k) Is the number of observations (cardinality) of the cluster C_k

Because the K-means algorithm finds a local rather than a global optimal, the results obtained will depend on the initial (random) cluster assignment of each observation (Krishna & Narasimha Murty, 1999; Nazeer & Sebastian, 2009). For this reason, it is important to run the algorithm multiple times from different random initial configurations. Then one selects the best solution, i.e., that for which the objective function (5) is smallest.

The principle behind a smaller objective function (5) indicates the data points within the same cluster are similar to each other, and data points in different clusters are dissimilar (Nazeer & Sebastian, 2009). Also, the clusters are well-separated from each other, and the boundaries between the clusters are clear.

K-means is a straightforward algorithm that can be easily understood and implemented. K-means is also computationally efficient, especially for large datasets, with the ability to process millions of data points in a matter of seconds (Nazeer & Sebastian, 2009). However, K-means requires the number of clusters to be specified beforehand, which can be difficult to tackle. In addition, the results of K-means can be sensitive to initial conditions, meaning that different starting points can produce different results (Nazeer & Sebastian, 2009).

Moreover, K-means is known for producing clear and well-defined clusters, which are advantageous for interpreting the results of the analysis. Additionally, K-means allows for controlling the number of clusters, enabling the specification of the desired number of clusters for the analysis with ease.

In comparison, hierarchical clustering is a more complex method that may not be as efficient for a large dataset (Murtagh & Contreras, 2012). Furthermore, hierarchical clustering can result in dendrogram structures that may not be as interpretable as the clear and well-studied clusters produced by K-means (Jain et al., 1999).

The NHANES dataset is a large dataset that encompasses a wide range of information related to health and nutrition. Due to the magnitude of the dataset and the necessity of conducting efficient analyses, the application of K-means analysis is deemed a suitable approach. K-means is a widely adopted and well-established clustering method renowned for its computational efficiency and scalability, particularly for large datasets (Nazeer & Sebastian, 2009). This attribute makes it an ideal option for analyzing the NHANES dataset. Therefore, based on the size of the NHANES dataset and the requirement for efficient and interpretable results, K-means clustering is deemed a more suitable option than hierarchical clustering.

2.5 Application of K-means to NHANES

The goal of applying K-means clustering to NHANES is to find clusters such that observations within each cluster were similar phenotypes (deciles ASMI and FMI) to each other and dissimilar to observations in others. All analyses were conducted with SAS 9.4 (SAS Institute, Cary, North Carolina) and R software (R Core Team, 2019) version 4.1.2.

2.5.1 Splitting the Sample

In order to begin the data analysis process, it is necessary to randomly split the sample (n=18,556) into a training dataset (70%, n=12,990) and a test dataset (30%, n=5,566) using the 'PROC SURVEYSELECT' statement in the SAS software. This procedure allows for the random sampling of a specified percentage of the data, using the 'SIZE' option to maintain a sex proportion consistent with the full dataset across all ages. The 'PROC SURVEYFREQ' statement was used to test whether the distribution of sex was different between the training and the test datasets. Also, the 'PROC SURVEYREG' was used for testing whether the mean age was different between the training and the test datasets. After the model was trained using the training set, then the statistical analysis was performed on the test dataset.

Given a set of n=12,990 training observations, each observation was grouped into a decile category (ranged from 0 to 9) for FMI or ASMI based on the Prado's (2014) reference curves which incorporated the subjects' sex and age. If a body-composition value was less than a decile cut-off but greater than or equal to the previous one, it was labeled according to the lower cut-off. (i.e. a FMI≥40th percentile cut-off and <50th percentile cut-off was grouped in the 40th percentile) (Prado et al., 2014), where each observation was a two-dimensional real vector, decile groups of ASMI and FMI.

2.5.2 Performing K-means Clustering

More specifically, the K-means clustering algorithm aims to partition a set of n=12,990 observations of each observation with decile groups of ASMI and FMI into K (≤ n) clusters to minimize the objective function (5). The objective function (5) was used for measuring the total within-cluster variation, summed over all K clusters, was as small as possible. The within-cluster variation for the *k*th cluster is the sum of all of the pairwise squared Euclidean distances between the observations in the *k*th cluster, divided by the total number of observations in the *k*th cluster.

2.5.3 Initial Centroids Selection

The K-means algorithm starts with a set of initial centroids, which are used as the center of each cluster. The algorithm then assigns each data point to the cluster whose centroid is closest to it. Once all the data points have been assigned to a cluster, the algorithm then iteratively updates centroids and the cluster assignments of the data points. The final clusters are determined by the positions of the initial centroids and the assignment of the data points to the nearest centroids (Krishna & Narasimha Murty, 1999; Nazeer & Sebastian, 2009). To demonstrate this, consider a simple example (Table 2) of nine observations (represented as points in a 2D space) that are to be grouped into three clusters using the K-means algorithm.

Table 2: An example of nine observations in a 2D space.

Observation	X	Y
1	2	5
2	4	7
3	5	4
4	3	8
5	1	9
6	7	6
7	8	9
8	7	8
9	2	4

If observations (2,5), (3,8) and (7,8) are chosen as the initial centroids, the 3-means algorithm will assign the observations 1, 3 and 9 to the same cluster as each other (Table 3).

Table 3: 3-means cluster results with initial centroids of (2,5), (3,8) and (7,8).

Observation	X	Y	Cluster centroid	Euclidean distance to cluster 1 centroid	Euclidean distance to cluster 2 centroid	Euclidean distance to cluster 3 centroid	Final Cluster
1	2	5	(3.0,4.3)	1.2	3.1	5.9	1
2	4	7	(2.7,8.0)	2.9	1.6	3.4	2
3	5	4	(3.0,4.3)	2.0	4.6	4.4	1
4	3	8	(2.7,8.0)	3.7	0.3	4.3	2
5	1	9	(2.7,8.0)	5.1	2.0	6.4	2
6	7	6	(7.3,7.7)	4.3	4.7	1.7	3
7	8	9	(7.3,7.7)	6.9	5.4	1.5	3
8	7	8	(7.3,7.7)	5.4	4.3	0.4	3
9	2	4	(3.0,4.3)	1.0	4.1	6.5	1

However, if observations (8,9), (7,8) and (2,4) are chosen as initial centroids, the algorithm will assign the observations differently (Table 4). For example, observation 1 and 9 belong in the same cluster as each other, but observation 3 belong to another cluster. This demonstrates how the choice of initial centroids can impact where the observations belong in the final cluster.

Table 4: 3-means cluster results with initial centroids of (8,9), (7,8) and (2,4).

Observation	X	Y	Cluster centroid	Euclidean distance to cluster 1 centroid	Euclidean distance to cluster 2 centroid	Euclidean distance to cluster 3 centroid	Final Cluster
1	2	5	(2.4,6.6)	6.5	4.0	1.6	3
2	4	7	(2.4,6.6)	3.8	2.8	1.6	3
3	5	4	(6.0,5.0)	5.1	1.4	3.7	2
4	3	8	(2.4,6.6)	4.5	4.2	1.5	3
5	1	9	(2.4,6.6)	6.5	6.4	2.8	3
6	7	6	(6.0,5.0)	2.6	1.4	4.6	2
7	8	9	(7.5,8.5)	0.7	4.5	6.1	1
8	7	8	(7.5,8.5)	0.7	3.2	4.8	1
9	2	4	(2.4,6.6)	7.1	4.1	2.6	3

If the initial centroids are chosen poorly, the algorithm may converge to a suboptimal solution because the initial centroids might not be representative of an actual cluster in the data. In this case, the algorithm may get stuck in a local minimum instead of finding the global optimal solution (Jain, 2010; Krishna & Narasimha Murty, 1999). This can lead to clusters that are not representative of the underlying structure of the data. Additionally, if the initial centroid is too far from the data points, the algorithm may not be able to find the true clusters, resulting in poor cluster assignments and suboptimal clusters (Krishna & Narasimha Murty, 1999). Therefore, the initial centroids selection can have a significant impact on the final cluster produced by the K-means algorithm, and it is important to choose them carefully.

2.6 Cluster Validation

The clustering algorithm assigns data points to clusters without prior knowledge of the underlying structure. Therefore, regardless of the specific clustering approach used, it is typically necessary to evaluate the final partition of data in most applications (Rezaee et al., 1998). The main objective of cluster validation is to identify the best clustering solution that meets the requirements of the problem at hand and generalizes well to new datasets. There are two main types of cluster validation measures:

1) Internal cluster validation measures, such as the silhouette score and elbow method (Kodinariya & Makwana, 2013). These validation measures use the internal information of the clustering process to assess the validity and stability of the clusters without reference to any external criterion. These measures are particularly useful for determining the optimal number of clusters and selecting the most appropriate clustering algorithm without the need for external data.

2) External cluster validation measures, such as the adjusted rand index (ARI) (Hubert & Arabie, 1985) is a technique that evaluates the quality of the clusters by comparing them to an externally known result, such as class labels known in advance. This approach is particularly useful when the true structure of the data is known or can be inferred, as it allows practitioners to compare the performance of different algorithms in order to choose the one that produces the best results in terms of similarity to the true clusters.

Internal cluster validation measures, including the silhouette score and the elbow method (Kodinariya & Makwana, 2013) are favored over external cluster validation measures, such as the adjusted rand index (ARI) (Hubert & Arabie, 1985) for several reasons:

Firstly, internal measures are derived from the intrinsic properties of the data and the clustering outcomes, whereas external measures require the availability of true class labels, which may not always be present in real-word data scenarios (Halkidi et al., 2001). This makes internal measures more practical and applicable. Secondly, internal measures are less susceptible to the subjective selection of the number of clusters, while external measures may be influenced by this choice. For instance, the silhouette score provides a metric of the similarity between each observation and its assigned cluster, enabling the assessment of the quality of the clustering solution without reference to the number of clusters (Rousseeuw, 1987). Finally, internal measures are more robust to noise and outliers in the data, and less susceptible to overfitting, making them a more dependable choice for evaluating the performance of a clustering algorithm.

In conclusion, internal cluster validation measures, such as the silhouette score and the elbow method (Kodinariya & Makwana, 2013) are preferred over external measures like the ARI (Hubert & Arabie, 1985) due to their derivation from intrinsic characteristics of the data, reduced sensitivity to subjective choices, and increased robustness against overfitting and noise. Thus, these internal cluster validation measures were selected for this thesis.

2.6.1 Silhouette Method

The silhouette method finds the optimal number of clusters by studying the separation distance between different numbers of clusters. The silhouette method computes silhouette coefficients of each point that measures how much a point is similar to its own cluster compared to other clusters (Rousseeuw, 1987). The value of the silhouette coefficients has a range of [-1, +1]. If the silhouette coefficient is close to +1, it suggests that the data points are significantly dissimilar from their adjacent clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters. On the other hand, a negative value suggests that the observation may have been wrongly assigned to its current cluster. An example of a plot of silhouette coefficients from the R "Zoo" dataset can be seen in Figure 4. Silhouette coefficients (x-axis) from two types of animals (y-axis) demonstrate the range of coefficients given the cluster assignment. At the bottom of the plot, silhouette identifies dolphin and porpoise with negative silhouette coefficients indicating that the dolphin and porpoise have been wrongly assigned to its current group of mammals.

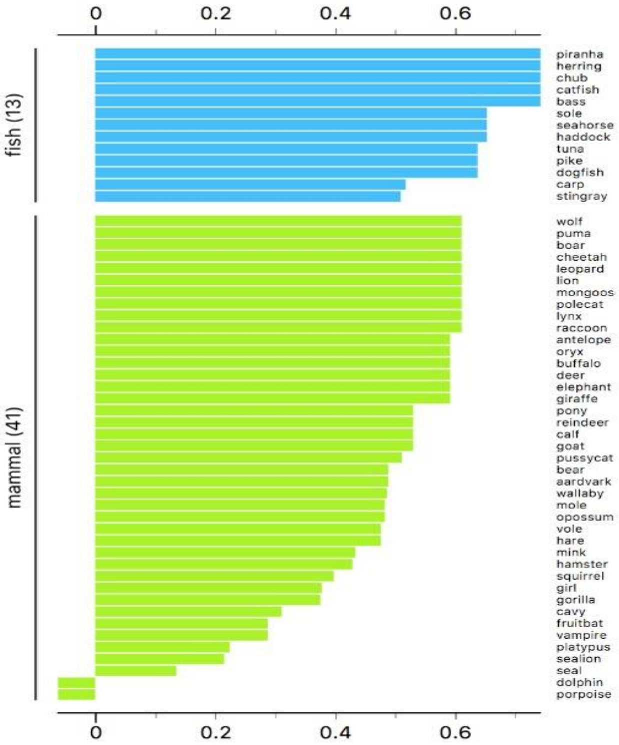


Figure 4: *An example of a plot of silhouette coefficients from the R "Zoo" dataset. Silhouette coefficients (x-axis) from two types of animals (y-axis) demonstrate the range of coefficients given the cluster assignment. (Silhouette: https://en.wikipedia.org/wiki/Silhouette_(clustering)).*

2.6.2 Computing the Silhouette Coefficient

Here are the steps to find the silhouette coefficients of ith point:

1. Compute a_i: The average Euclidean distance of that point with all other points in the same clusters.
2. Compute b_i: The average Euclidean distance of that point with all the points in the closest cluster to its cluster.
3. Compute silhouette coefficient S_i of ith point using the below formula.

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{6}$$

2.6.3 Elbow Method

The elbow method is an empirical method to find the optimal number of clusters for a dataset (Milligan & Cooper, 1985). The method is based on calculating the WCSS (within-cluster sum of square), which is the sum of the square distance between points in a cluster and the cluster centroid and represents it in a plot. The optimal number of clusters (K) can be identified based on where a steep fall of the WCSS is observed (Kodinariya & Makwana, 2013).

2.7 Other Design Features

Other design features section of this thesis elucidates the significance of multiple imputation techniques in handling missing data within the NHANES dataset, as well as the importance of weighting procedures to obtain unbiased and representative results from complex survey data.

2.7.1 Multiple Imputation Method

Multiple imputation (Rubin, 1987) is a statistical method used to handle incomplete data problems. In the context of NHANES, multiple imputation can be used to handle incomplete data on variables of interest. The specific method used would be dependent on the pattern of missing data and the variable types involved. For example, if the missing data are missing completely at random (MCAR), a simple imputation method such as mean imputation could be used (Enders, 2010). However, if the missing data are not MCAR, more sophisticated methods such as multiple imputation by chained equations (MICE) could be used (Buuren & Groothuis-Oudshoorn, 2011; Enders, 2010).

The MICE method is a powerful and flexible multiple imputation method. It works by creating multiple imputed datasets by iteratively imputing the missing data for each variable separately (Buuren & Groothuis-Oudshoorn, 2011). The imputed values for each variable are based on the observed values for that variable and the values of other related variables. This process is repeated multiple times (e.g. 5-10 times) to create multiple imputed datasets. These imputed datasets are then analyzed using standard statistical methods, and the results are combined to produce overall estimates and standard errors that account for the uncertainty due to the missing data (Enders, 2010). In NHANES, due to DXA data missingness, FMI and ASMI observations were provided in five multiply imputed datasets.

While multiple imputation is a valuable technique for handling missing data, it does come with potential challenges and limitations. One such problem is misspecification due to inadequate auxiliary variables. In multiple imputation, auxiliary variables are additional variables that are included in the imputation model to help predict the missing values. These variables should be correlated with both the missingness mechanism and the variables with missing data. Including relevant auxiliary variables improves the accuracy of imputations and reduces bias. However, if the set of auxiliary variables is inadequate or poorly chosen, it can lead to misspecification of the imputation model (Sullivan et al., 2015).

When the auxiliary variables are insufficient, the imputation model may not capture the underlying relationships between the missing data and the available variables accurately. This can result in biased imputations and potentially affect subsequent analyses and inferences based on the imputed data. Thus, it is essential to carefully consider the choice of auxiliary variables to ensure the robustness of the imputation procedures and the validity of subsequent analyses.

2.7.2 Weighting

NHANES is a complex survey that employs a multistage probability sampling design to represent the noninstitutionalized civilian population of the United States (Curtin et al., 2012; *NHANES - About the National Health and Nutrition Examination Survey*, 2022), the sample is not randomly selected from the entire population. Instead, participants are selected through a series of stages. For instance, selecting primary sample units (PSUs) and then selecting households within those PSUs (Montaquila et al., 2010). Additionally, some people within the selected households may not participate in the survey, which can lead to nonresponse bias. Thus, in NHANES, weights are constructed to consider the complex survey design (including oversampling), non-response to the survey, and post-stratification adjustment to produce unbiased estimates of population parameters (Johnson et al., 2013; *NHANES Tutorials - Weighting Module*, 2023).

The weighting process involves assigning a weight to each survey participant, which is based on their probability of selection and the characteristics of the population. The weight for each participant is calculated as the reciprocal of his/her probability of selection. By using these weights in the analysis, the sample is adjusted so it is more representative of the U.S. civilian noninstitutionalized population. This can be important for making valid inferences and generalization from the sample to population. Lastly, combining multiple survey years' weights were constructed in accordance with guidelines from NHANES.

2.8 Other Statistical Analysis Features

In this thesis, logistic regression was used to model the relationship between the probability K-means cluster phenotypes or 50th percentile cut-off phenotypes and having unfavorable cardiometabolic risks. In order to evaluate the performance of the logistic regression models, the area under the receiver operating characteristic (ROC-AUC) was used.

2.8.1 Logistic Regression

All logistic regression models were adjusted for age and sex. The general form of a logistic regression model can be expressed as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j \tag{7}$$

π Probability of participants' having unfavorable cardiometabolic risks

x_j Explanatory variables, $j = 1, \dots, p$

β_0 The intercept

β_j Regression coefficients associated with the x_j explanatory variables, $j = 1, \dots, p$

2.8.2 Logistic Regression Application

Logistic regression is commonly used when the outcome variable is binary or categorical, as it allows people to estimate the probability or odds of belonging to a particular category based on the values of the predictor variables.

In accordance with clinically relevant thresholds as previously described, logistic regression was performed with the cardiometabolic risk measures dichotomized as:

$$\text{Total Cholesterol} = \begin{cases} \text{Low}, \leq 6.2 \text{ mmol/L} \\ \text{High}, > 6.2 \text{ mmol/L} \end{cases}$$

$$\text{HDL Cholesterol} = \begin{cases} \text{High}, \begin{cases} \text{Men}, \geq 1.0 \text{ mmol/L} \\ \text{Women}, \geq 1.3 \text{ mmol/L} \end{cases} \\ \text{Low}, \begin{cases} \text{Men}, < 1.0 \text{ mmol/L} \\ \text{Women}, < 1.3 \text{ mmol/L} \end{cases} \end{cases}$$

$$\text{LDL Cholesterol} = \begin{cases} \text{Low}, \leq 4.1 \text{ mmol/L} \\ \text{High}, > 4.1 \text{ mmol/L} \end{cases}$$

$$\text{Triglycerides} = \begin{cases} \text{Low}, \leq 2.3 \text{ mmol/L} \\ \text{High}, > 2.3 \text{ mmol/L} \end{cases}$$

$$\text{Blood Pressure} = \begin{cases} \text{Low}, \begin{cases} \text{SBP} < 140 \text{ mm Hg} \\ \text{OR} \\ \text{DBP} < 90\text{mm Hg} \end{cases} \\ \text{High}, \begin{cases} \text{SBP} \geq 140 \text{ mm Hg} \\ \text{OR} \\ \text{DBP} \geq 90 \text{ mm Hg} \end{cases} \end{cases}$$

As the phenotypes (using either the 50th percentile cut-offs, or the K-means clustering) are categorical, these were modeled through dummy variables:

Class	Value	Codes		
The 50th percentile cut-off defined phenotypes:	**LA_HM**: low adiposity with high muscle (Reference category)	0	0	0
	HA_HM: high adiposity with high muscle	1	0	0
	HA_LM: high adiposity with low muscle	0	1	0
	LA_LM: low adiposity with low muscle	0	0	1

Thus, the LA-HM was the reference category, and the logistic regression model with the 50th percentile cut-off defined phenotypes was the following:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * (HA\_HM) + \beta_2 * (HA\_LM) + \beta_3 * (LA\_LM) \qquad (8)$$

2.8.3 ROC Curve and AUC

The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different classification thresholds. The term “thresholds” refers to the specific values used to determine the classification of data into different categories or classes. In the context of the ROC curve, thresholds represent the points at which a binary classification decision is made. These thresholds are used to determine whether a given observation should be classified as positive or negative based on the output of a classifier or predictive model.

By adjusting the classification threshold, it is possible to influence the trade-off between the TPR and FPR. A lower threshold may lead to a higher TPR, meaning more true positives are detected, but it can also result in a higher FPR, leading to an increased number of false positives. Conversely, a higher threshold may decrease the TPR but also reduce the FPR. The ROC curve illustrates the performance of a classification model across a range of thresholds, showcasing how the TPR and FPR change simultaneously. It provides a visual representation of the trade-off between correctly identifying positive cases and incorrectly classifying negative cases as positive, allowing for an evaluation of the model’s discriminatory power.

The AUC represents the area under this ROC curve (Figure 5) and measures the model’s ability to distinguish between positive and negative classes. A higher ROC-AUC indicates better performance, with an ideal value of 1.0 indicating a perfect classifier (Lohr, 2012) to identify unfavorable levels of cardiometabolic risk based on the phenotype. The calculation of the ROC-AUC provides a robust evaluation of the logistic regression model, as it is independent of the choice of threshold and provides a good indicator of the model’s overall performance.

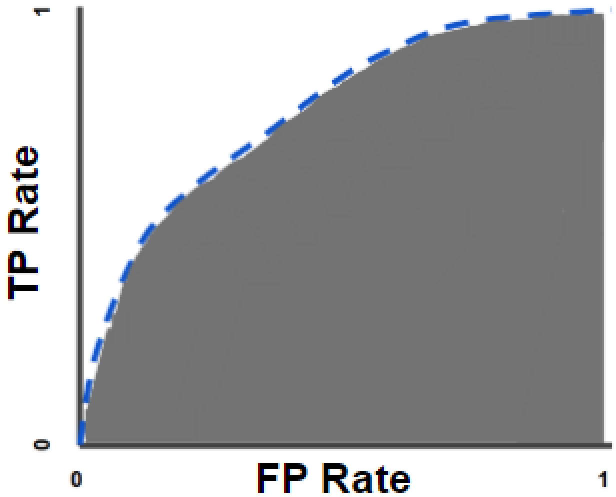


Figure 5: *AUC (Area under the ROC Curve). AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). (Classification: ROC Curve and AUC, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc)*

2.8.4 Framework of the Study

The study utilized a framework (Figure 6) that began with a total of 18,556 observations extracted from the NHANES dataset. The observations were subsequently divided into two distinct datasets based on gender (male and female). To create the male dataset, a random sampling technique was applied to assign observations to both a training dataset and a test dataset. Within the male training dataset, the value of FMI and ASMI were obtained for each observation. These measures were then incorporated into Prado's (2014) sex- and age- specific references curves, resulting in the calculation of decile_FMI and decile_ASMI values for each observation, ranging from 0 to 9.

Following this, the K-means algorithm was employed to group the observations based on their decile_FMI and decile_ASMI values. The algorithm was executed for various values of K, ranging from 1 to 10. To determine the optimal number of clusters, the silhouette method and the elbow method were utilized. Once the optimal number of clusters was confirmed, the decile_FMI and decile_ASMI values for each cluster's centroids were calculated. These centroid values were instrumental in determining the phenotypes associated with each cluster. For instance, if a cluster's centroid had a low decile_FMI and low decile_ASMI values, it was labeled as LA-LM, and all observations within that cluster exhibited the same phenotype.

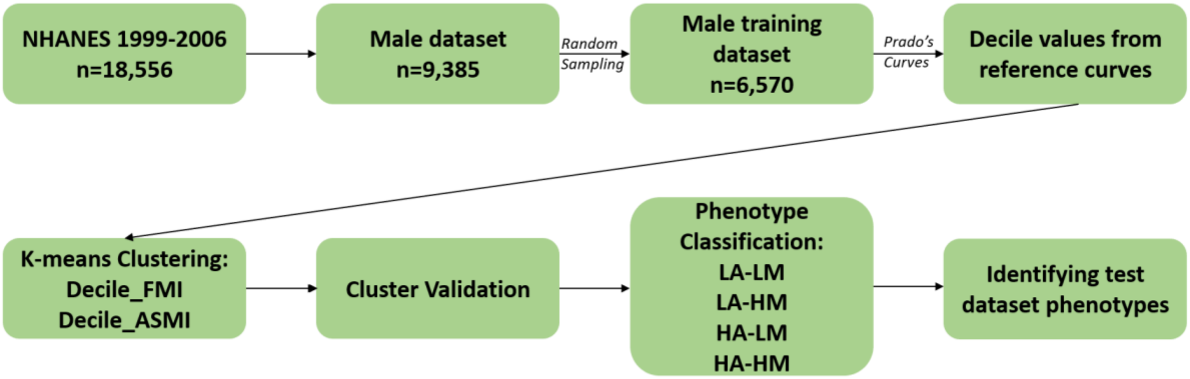


Figure 6: *Summary of research methodology.*

Similar procedures were then conducted on the female dataset, which consisted of 9,171 observations. The same steps of random sampling, calculation decile_FMI and decile_ASMI values, application of the K-means algorithm, determination of optimal clusters, calculation of cluster centroid values, and assignment of phenotypes based on the centroid values were followed.

In the subsequent stage, the decile_FMI and decile_ASMI values for the observations in the test dataset were obtained. The Euclidean distance between each observation and the centroid values of the clusters was calculated. The observation was then assigned a closest phenotype based on the cluster with the lowest Euclidean distance.

In summary, this study employed a framework that involved the division of observations by gender, utilization of FMI, ASMI and references curves to determine decile values, application of the K-means algorithm for clustering, determination of phenotypes based on cluster centroids, and assignment of closest phenotypes to observations in the test dataset using Euclidean distance calculations. The same procedures were carried out separately on the male and female datasets.

2.8.5 Statistical Analysis

Data from cycles 1999-2006 were performed separately for imputed male and female datasets. As previously described, there were many sampling designs and data features in this dataset that needed to be incorporated into the analysis for proper inference. Specific to the use of ROC-AUC curves in a dataset with a complex sampling design, SAS macros were necessary to calculate the standard errors of ROC-AUC using bootstrapping (n=1,000) (Izrael, 2002), alongside the calculation of a single, summary AUC given that the predictor was categorical and from a dataset with a complex sampling design with multiple imputation. The standard errors derived from the bootstrap samples, along with the ROC-AUCs obtained from each of the multiple imputation datasets, were combined with PROC MIANALYZE to perform standard statistical inferences through multiple imputation (Rubin & Schenker, 1986; Schomaker & Heumann, 2018). Weighted descriptive statistics of the sample characteristics were calculated. PROC SURVEYLOGISTIC was used for the regression models and ROC curves were also conducted as further described.

3. Results

Each of the data files for 1999-2000, 2001-2002, 2003-2004, and 2005-2006 contains five sets of measured and imputed values. The preferred statistical approach is to analyze each of the five datasets separately using K-means clustering and then pooling the estimates and standard errors in accordance with the literature (Enders, 2010; Kakinami et al., 2022).

In this study, data were separated into training sets for males and females. The 'PROC SURVEYFREQ' statement showed that the distribution of male and female participants were not significantly different between training and test datasets (*p* = 0.72). Moreover, the 'PROC SURVEYREG'' statement showed that the mean age was not significantly different between training and test datasets (*p* = 0.79). More on this is described later. As cluster validation results were consistent between men and women, results from the men's datasets will be uniformly presented.

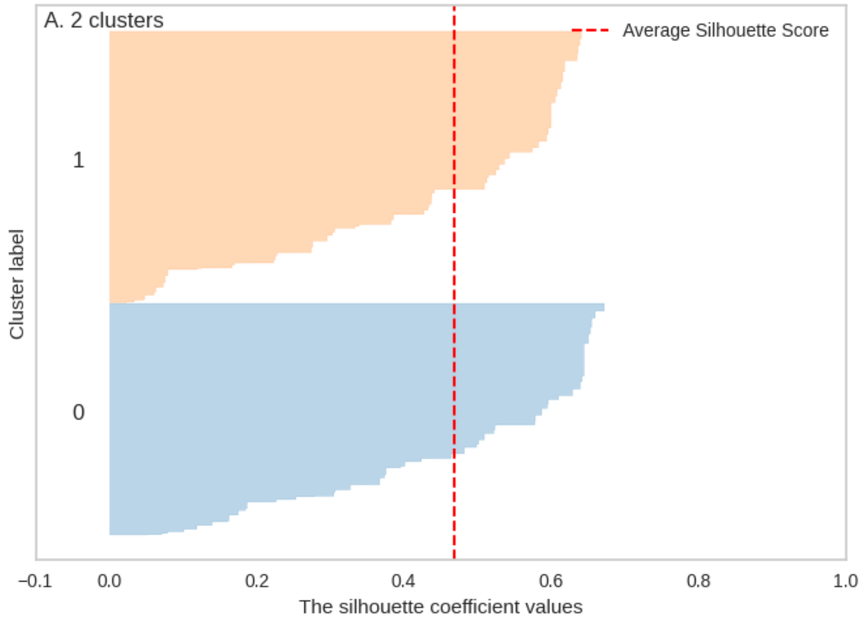
3.1 Cluster Validation: Silhouette Method

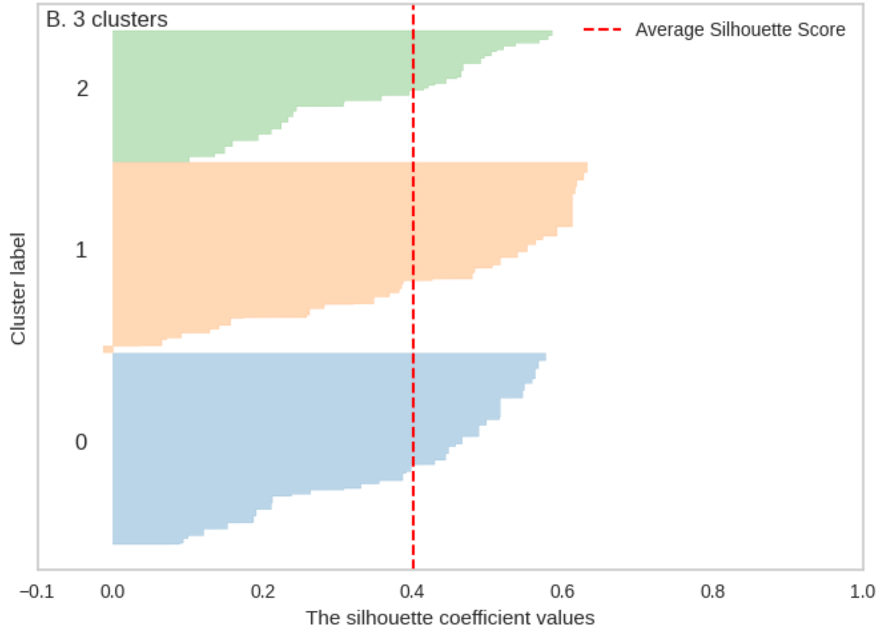
After completing the silhouette coefficient for each data point, it can be averaged out for all the samples to calculate the silhouette score (Table 5).

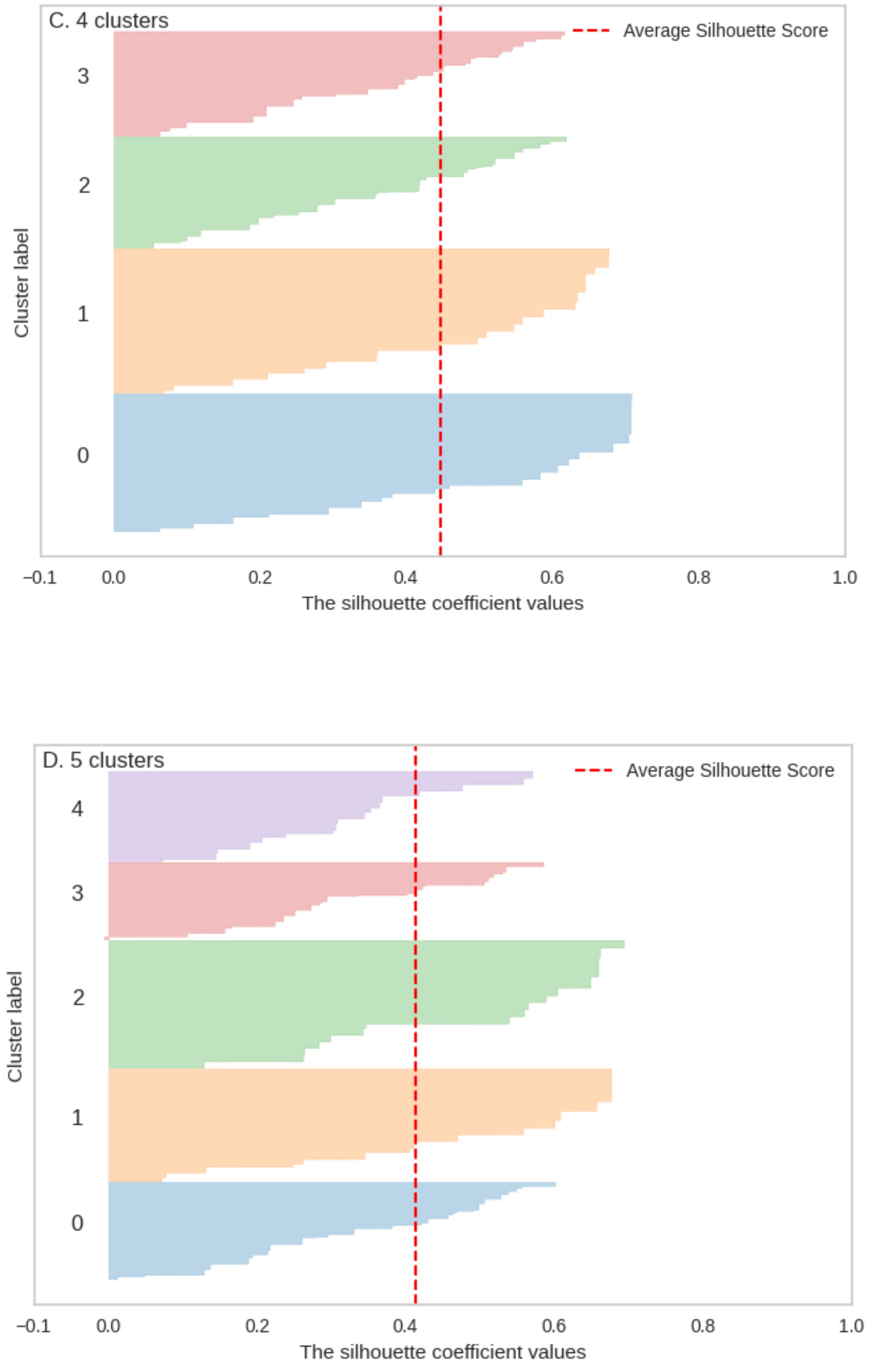
Table 5: The results of silhouette scores from the men's training dataset.

Number of clusters	Average silhouette score
2	0.468
3	0.402
4	0.446
5	0.416
6	0.415

The silhouette plots (Figure 7) shows that the number of clusters of 2, 3, 4, 5, and 6 are good selections for the given training data due to the clusters with above average silhouette score. However, the silhouette plots show that the number of clusters of 5 and 6 are suboptimal due to wide fluctuations in the size of the silhouette plots. Furthermore, the cluster size can be visualized from the thickness of the silhouette plot. The silhouette plots for 2, 3 and 4 clusters, are all more or less of similar thickness and hence are of similar sizes, as can be considered as best 'K'. In contrast, the silhouette plots for clusters 5 and 6 are more varying in size. However, the silhouette analysis is less clear in selecting between 2, 3 and 4. Nevertheless, as 2 and 4 clusters have a higher average silhouette score (0.468 and 0.446, respectively) compared to 3 clusters (0.402), either 2 or 4 clusters should be the ideal number and were explored further.


A. 2 clusters
Average Silhouette Score
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


B. 3 clusters
Average Silhouette Score
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


C. 4 clusters
Average Silhouette Score
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
D. 5 clusters
Average Silhouette Score
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values

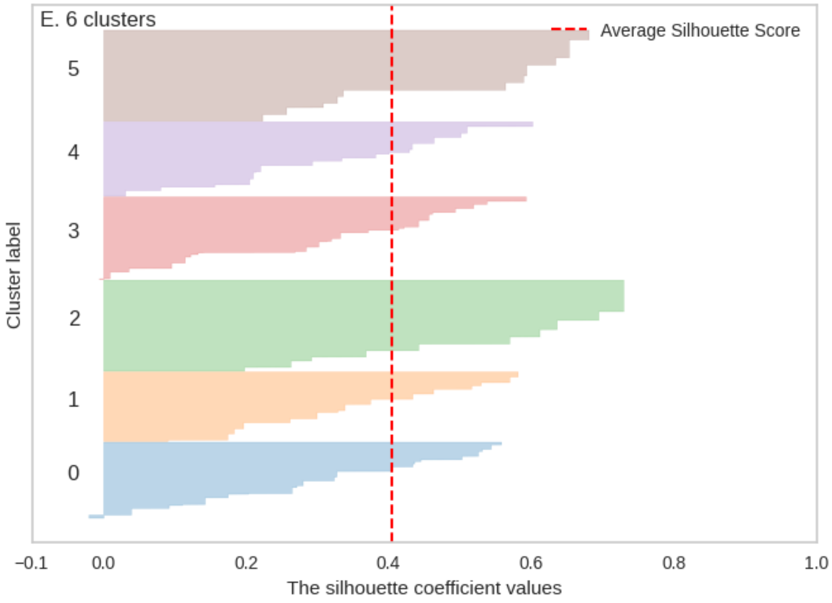


Figure 7: *Silhouette analysis for each cluster in K-means clustering on the male training dataset. The silhouette measures how similar an object is to the other objects in its own cluster versus those in some other cluster. Values for S_j range from 1 to -1, with values close to 1 indicating that the item is well clustered (is similar to the other objects in its group) and values near -1 indicating it is poorly cluster.*

3.2 Cluster Validation: Elbow Method

The elbow method can also be used to choose an optimal number of clusters. Consistent with the silhouette plots, the elbow plots (Figure 8) show a sharp fall of WCSS at k=2 and 4. It follows that the optimal number of clusters should be either 2 or 4.

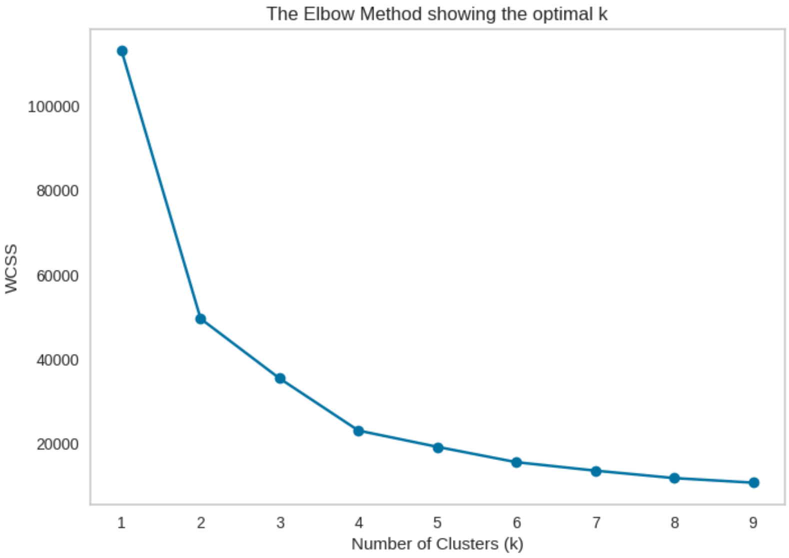


Figure 8: *Elbow method to find optimal cluster number K.*

As mentioned previously, due to the multiple imputation, analyses were conducted for each dataset separately. For instance, within each training set, 2-means and 4-means cluster analysis were conducted on each imputed dataset individually and compared the phenotype of each participant in our five imputed sets. In the tables below, the percentage of concordance in phenotype in all five datasets is provided. From these results, the results were largely consistent across all imputed datasets for both male and female groups. For example, in Table 6, there is 9.7% variation of phenotypes across all five imputed datasets; the variation is lower among females (7.5%, Table 7). Thus, over 90% of phenotypes were classified consistently across all five datasets. Therefore, in order to simplify some of the analyses, the phenotypes classification from the first imputed dataset were used for the rest of the thesis.

Table 6: 4-means cluster analysis results for five imputed male training datasets.

Count	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	634	9.65	634	9.65
5	5,936	90.35	6,570	100.00

Table 7: 4-means cluster analysis results for five imputed female training datasets.

Count	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	478	7.45	478	7.45
5	5,942	92.55	6,420	100.00

For both 2-means and 4-means, once the cluster centroids for male and female in the training datasets were identified, they were used to calculate the Euclidean distance for each observation in the test dataset and to choose the closest phenotype for each observation as described previously (Section 2.5.3). The 'nstart' parameter in the R 'kmeans' function was used to control the number of times the algorithm is run with different initial centroids configurations. These configurations involved randomly selecting the initial centroids for each run. For both male and female training datasets, 25 and 50 different initial centroids configurations were tested; results showed that both 25 and 50 initial configurations consistently produced the same smallest sum of squared distances as the final result. This indicated that the clustering algorithm converged to a stable solution for these initial configurations (Gentleman & Carey, 2008).

3.3 Clustering: 2-Means

There were 12,990 observations in the training dataset, with 6,570 males (51%) and 6,420 females (49%). As the silhouette and elbow method demonstrated that two clusters, and four clusters were optimal, both were explored further. The following Table 8 shows the mean value of each cluster for the male and female training dataset. For both the male and female training datasets, although the mean values suggested LA-LM (cluster 1) and HA-HM (cluster 2) phenotypes, the plot of these data illustrate that labeling them as such was an oversimplification (Figure 9).

Table 8: 2-means cluster analysis results.

Characteristics	Male training set	Female training set
LA-LM Means (Decile_ASMI, Decile_FMI), *N*	(2.57, 2.42), 3,545	(2.41, 2.51), 3,291
HA-HM Means (Decile_ASMI, Decile_FMI), *N*	(7.01, 6.79), 3,025	(7.17, 7.09), 3,129

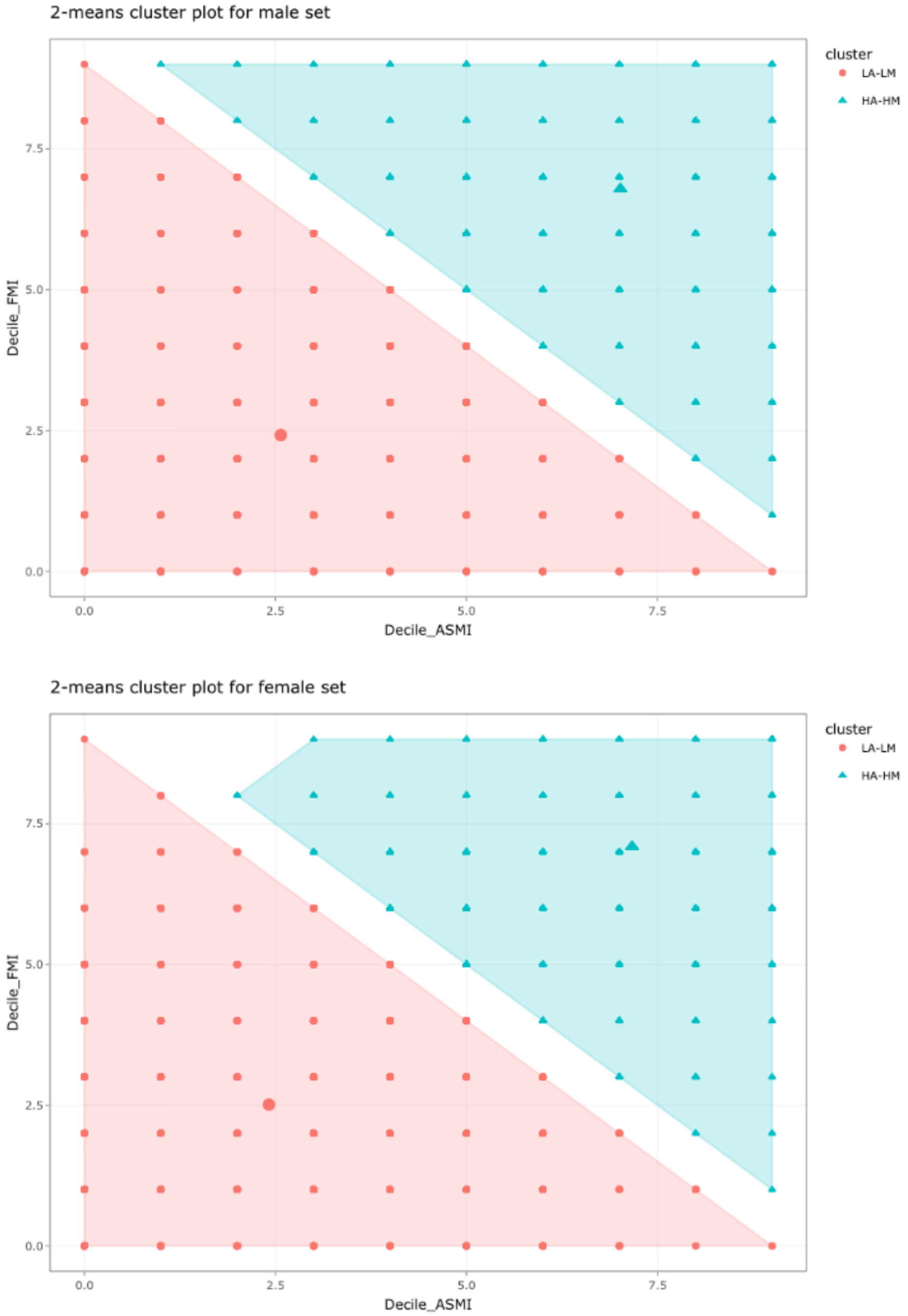


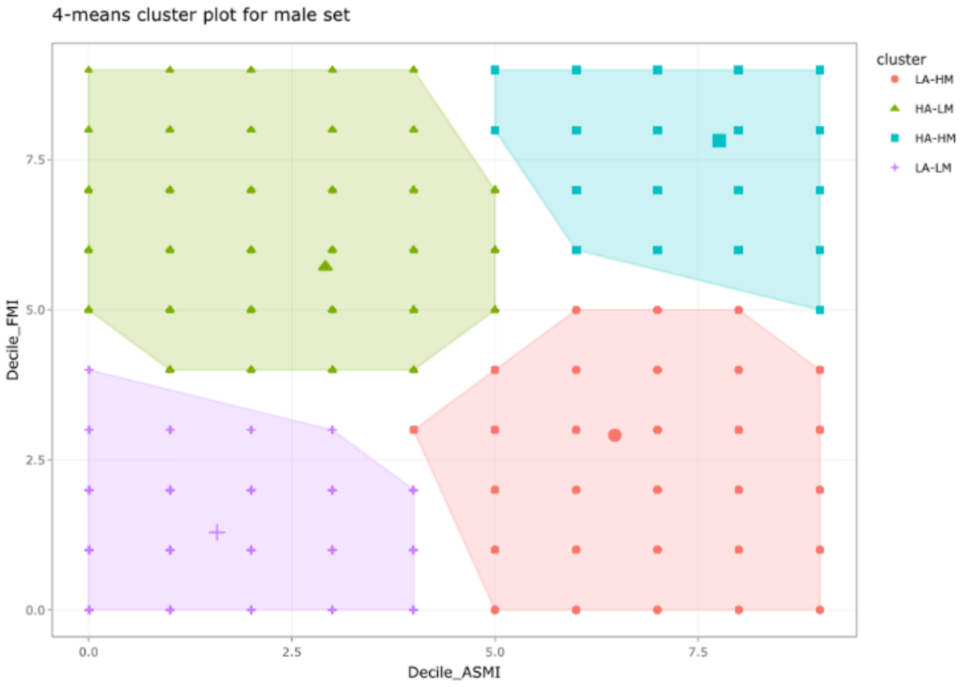
Figure 9: *2-means cluster plot for male and female datasets with two variables decile_FMI and decile_ASMI. The plots show two clusters with their corresponding centroids and phenotypes.*

3.4 Clustering: 4-Means

For the 4-means cluster, the mean values of each cluster were also measured. Table 9 provides the results of the mean value of each cluster for the male and female training datasets. For the male training set, Cluster 1 contains centroids with a high decile ASMI and a low decile FMI (Figure 10). Based on these, the data points in this cluster are tightly grouped together, indicating a low adiposity and high muscle (LA-HM) phenotype within the cluster. Cluster 2 contains centroids with a low decile ASMI and low decile FMI, indicating a low adiposity and low muscle (LA-LM) phenotype within the cluster. Cluster 3 contains centroids with high decile ASMI and high decile FMI, indicating a high adiposity and high muscle (HA-HM) phenotype within the cluster. Cluster 4 contains centroids with low decile ASMI and high decile FMI, indicating a high adiposity and low muscle (HA-LM) phenotype within the cluster. For both the male and female training datasets, similar ASMI and FMI cut-points were detected as useful for identifying distinctive phenotypes.

Table 9: 4-means cluster analysis results.

Cluster	Characteristic	Male training set	Female training set
1	LA-HM Means (Decile_ASMI, Decile_FMI), *N*	(6.48, 2.91), 1,379	(5.75, 2.84), 1,103
2	LA-LM Means (Decile_ASMI, Decile_FMI), *N*	(1.56, 1.39), 1,976	(1.32, 1.62), 1,821
3	HA-HM Means (Decile_ASMI, Decile_FMI), *N*	(7.76, 7.82), 1,819	(7.86, 7.75), 2,235
4	HA-LM Means (Decile_ASMI, Decile_FMI), *N*	(3.01, 5.81), 1,396	(3.22, 5.60), 1,261



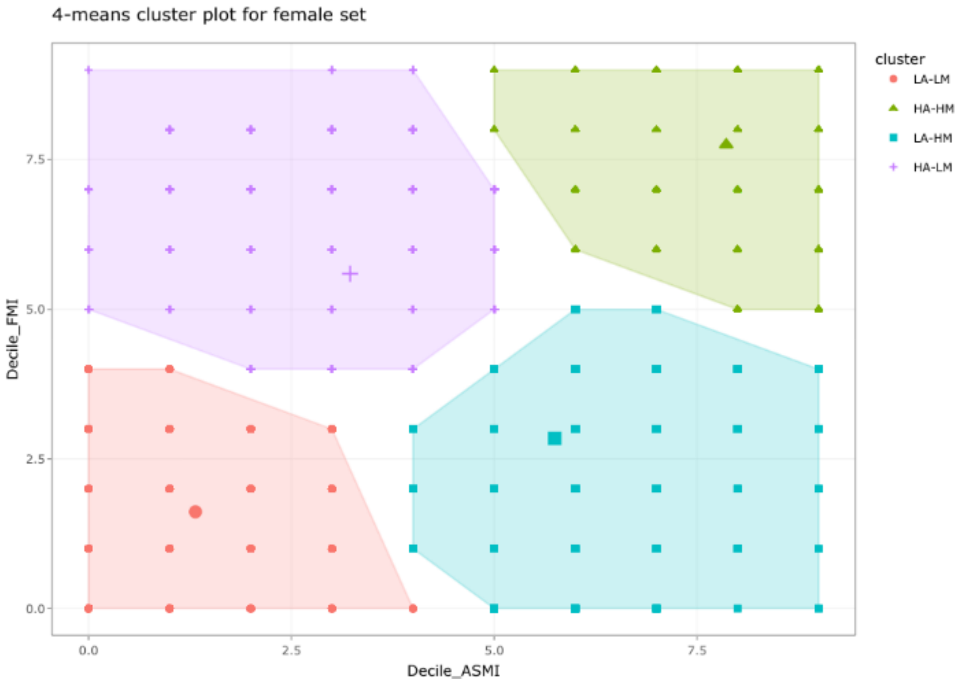


Figure 10: *4-means cluster plot for male and female datasets with two variables decile_FMI and decile_ASMI. The plots show four clusters with their corresponding centroids and phenotypes.*

As the NHANES dataset is a complex heterogeneous dataset, it is likely to contain multiple subgroups or phenotypes (Patel et al., 2016). A 4-means clustering approach may be better suited for capturing this heterogeneity than a 2-means clustering approach. In particular, the use of 4-means clustering approach allows for more clusters to be formed, which can capture more nuanced differences in the data. The result is a more accurate and detailed representation of the data, which can lead to more interpretable and meaningful results. Specifically, a 4-means clustering approach can identify specific phenotypes of individuals who share similar health characteristics (Nazeer & Sebastian, 2009). The identification of these phenotypes can inform the development of targeted public health interventions. In contrast, a 2-means clustering approach may oversimplify the data and miss important subgroups or phenotypes, which can result in less accurate and less meaningful results. Therefore, the use of 4-means clustering approach was deemed as preferable in this thesis, as it can lead to a more comprehensive and accurate understanding of the data.

3.5 Sample Characteristics

The general characteristics of the participants are presented in Table 10. There were 18,556 participants included in this study. The data included 12,990 (70.0%) participants in the training dataset and 5,566 (30.0%) participants in the test dataset. From Table 10, the training set and test set significantly differed in percentage of hypertension (3.49 vs 2.68, p = 0.04). These were no other significant differences in the rest of characteristics between the training and test dataset.

Table 10: Weighted means and frequencies for demographic and health characteristics of the total study population, NHANES 1999 to 2006 (*n* = 18,556).

Characteristic[a]	**All Participants** *n* **= 18,556**	**Training set participants (*n* = 12,990)**	**Test set participants (*n* = 5,566)**	***p* value**[b]
Age (y)	44.41 (0.25)	44.38 (0.26)	44.47(0.35)	0.79
BMI	28.10 (0.10)	28.07 (0.10)	28.18 (0.13)	0.33
FMI	9.92 (0.07)	9.90 (0.07)	9.97 (0.09)	0.32
ASMI	7.66 (0.02)	7.65 (0.02)	7.67 (0.03)	0.58
DXA: 4-means clusters (%)				
HA-HM	30.71	30.43	31.36	0.51
HA-LM	20.89	20.80	21.11	
LA-HM	17.33	17.61	16.69	
LA-LM	31.06	31.16	30.84	
DXA: 50th percentile (%)				
HA-HM	37.68	37.47	38.18	0.43
HA-LM	13.23	13.12	13.48	
LA-HM	12.02	12.33	11.30	
LA-LM	37.07	37.08	37.04	
Cardiometabolic measures				
Total cholesterol (mmol/L)	5.18 (0.01)	5.19 (0.02)	5.18 (0.02)	0.52
HDL cholesterol (mmol/L)	1.35 (0.01)	1.36 (0.01)	1.35 (0.01)	0.16
LDL cholesterol (mmol/L)	3.08 (0.02)	3.08 (0.02)	3.08 (0.03)	0.77
Triglycerides (mmol/L)	1.64 (0.02)	1.65 (0.03)	1.62 (0.03)	0.39
SBP (mm Hg)	122.04 (0.28)	122.19 (0.31)	121.68 (0.36)	0.17
DBP (mm Hg)	71.57 (0.20)	71.63 (0.24)	71.42 (0.22)	0.45
Abnormal cardiometabolic measures (%)[c]				
High total cholesterol	16.18	16.19	16.15	0.96
Low HDL cholesterol	31.32	31.19	31.64	0.68
High LDL cholesterol	12.98	12.83	13.33	0.66
High triglycerides	16.66	16.80	16.33	0.67
High SBP or High DBP	3.25	3.49	2.68	0.04

Abbreviations: DBP, diastolic blood pressure; DXA, dual-energy x-ray absorptiometry; FMI, fat mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NHANES, National Health and Nutrition Examination Survey; SBP, systolic blood pressure; ASMI, appendicular lean mass index. HA-HM: high adiposity, high muscle; HA-LM: high adiposity, low muscle; LA-HM: low adiposity, high muscle; LA-LM: low adiposity, low muscle.

[a] Weighted mean (SE) unless otherwise noted.

[b] *P* comparing training set participants with test set participants.

[c] Total cholesterol>6.2 mmol/L; HDL cholesterol<1 mmol/L [men], <1.3 mmol/L [women]; LDL cholesterol>4.1 mmol/L; Triglycerides>2.3 mmol/L; SBP≥140 mm Hg or DBP≥90 mm Hg.

Demographic and health characteristic comparisons between men and women are presented in Table 11. Mean BMI and total cholesterol levels were not significantly different between male and female participants, but significant differences were observed in nearly all other demographic and health characteristics. Approximately 52% had high adiposity using our 4-means cluster phenotypes. Similarly, 52% of participants were classified as having high adiposity using the 50th percentile cut-off phenotypes. The distribution of 4-means cluster phenotypes and 50th percentile cut-off phenotypes were statistically different between male and female participants ($p < 0.01$).

For male participants, 4-means clustering and 50th percentile cut-off showed a concordance rate of 2,521 out of 2,815 individuals, indicating a strong agreement of 90%. This implies that the two methods consistently classified the phenotypes of male participants in a highly similar manner. Similarly, among female participants, 4-means clustering and 50th percentile cut-off demonstrated a concordance rate of 2,342 out of 2,751 individuals, resulting in a concordance percentage of 85%. This finding suggests a substantial level of agreement between the two methods in accurately defining the phenotypes of female participants. Thus, the high concordance rates for both genders indicate a reliable consistency between 4-means clustering and 50th percentile cut-off in categorizing phenotypes.

Table 11: Weighted means and frequencies for demographic and health characteristics of the test dataset population, NHANES 1999 to 2006 (*n* = 5,566).

Characteristic[a]	**All Participants *n* = 5,566**	**Male participants (*n* = 2,815)**	**Female participants (*n* = 2,751)**	***p* value**[b]
Age (y)	44.47 (0.35)	43.44 (0.45)	45.49 (0.42)	<0.01
BMI	28.18 (0.13)	28.12 (0.17)	28.23 (0.19)	0.66
FMI	9.97 (0.09)	8.24 (0.10)	11.70 (0.13)	<0.01
ASMI	7.67 (0.03)	8.60 (0.04)	6.74 (0.04)	<0.01
DXA: 4-means clusters (%)				
HA-HM	31.36	30.14	32.57	<0.01
HA-LM	21.11	23.78	18.46	
LA-HM	16.69	16.76	16.62	
LA-LM	30.84	29.32	32.35	
DXA: 50th percentile (%)				
HA-HM	38.18	36.52	39.82	<0.01
HA-LM	13.48	16.84	10.14	
LA-HM	11.30	12.80	9.82	
LA-LM	37.04	33.84	40.22	
Cardiometabolic measures				
Total cholesterol (mmol/L)	5.18 (0.02)	5.16 (0.03)	5.19 (0.03)	0.62
HDL cholesterol (mmol/L)	1.35 (0.01)	1.22 (0.01)	1.47 (0.01)	<0.01
LDL cholesterol (mmol/L)	3.08 (0.03)	3.14 (0.04)	3.02 (0.04)	0.03
Triglycerides (mmol/L)	1.62 (0.03)	1.76 (0.05)	1.47 (0.03)	<0.01
SBP (mm Hg)	121.68 (0.36)	122.37 (0.43)	120.99 (0.53)	0.03
DBP (mm Hg)	71.42 (0.22)	72.49 (0.31)	70.34 (0.30)	<0.01
Abnormal cardiometabolic measures (%)[c]				
High total cholesterol	16.15	16.04	16.26	0.90
Low HDL cholesterol	31.64	25.30	37.96	<0.01
High LDL cholesterol	13.33	14.98	11.67	0.06
High triglycerides	16.33	20.15	12.33	<0.01
High SBP or High DBP	2.68	2.87	2.48	0.50

Abbreviations: DBP, diastolic blood pressure; DXA, dual-energy x-ray absorptiometry; FMI, fat mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NHANES, National Health and Nutrition Examination Survey; SBP, systolic blood pressure; ASMI, appendicular lean mass index. HA-HM: high adiposity, high muscle; HA-LM: high adiposity, low muscle; LA-HM: low adiposity, high muscle; LA-LM: low adiposity, low muscle.

[a] Weighted mean (SE) unless otherwise noted.

[b] *P* comparing male participants with female participants.

[c] Total cholesterol>6.2 mmol/L; HDL cholesterol<1 mmol/L [men], <1.3 mmol/L [women]; LDL cholesterol>4.1 mmol/L; Triglycerides>2.3 mmol/L; SBP≥140 mm Hg or DBP≥90 mm Hg.

3.6 Assessing Model Performance

The ROC-AUCs from 2-means cluster models used to correctly identify cardiometabolic risk factors ranged from 0.52 to 0.63 (Table 12), indicating generally weak predictive power. The ROC-AUCs from 50th percentile cut-off phenotypes and 4-means cluster phenotypes were higher (0.56-0.66, 0.57-0.67, respectively).

Overall comparison of model performance based on ROC-AUCs suggested that for this study 4-means clustering was superior to 50th percentile cut-off in predicting cardiometabolic risk. For example, for total cholesterol, the 4-means cluster phenotype showed a higher ROC-AUCs value of 0.60 compared to the 50th percentile cut-off phenotype value of 0.58, indicating that the 4-means cluster phenotype is a better predictor. However, most of these comparisons were not statistically significant.

Table 12: AUC from ROC curves comparing 2-means cluster, 4-means cluster and 50th percentile DXA phenotype, NHANES 1999 to 2006 (*n* = 5,566).

	AUC (SE) from ROC curves[a]				
Unfavorable cardiometabolic risk factor	**Model 1: 50th percentile**	**Model 2: 2-means clusters**	***p***[b]	**Model 3: 4-means clusters**	***p***[c]
High total cholesterol[d]					
Male participants (*n* = 2,815)	0.562 (0.02)	0.519 (0.01)	0.02	0.572 (0.02)	0.44
Female participants (*n* = 2,751)	0.577 (0.02)	0.533 (0.01)	<0.01	0.603 (0.02)	0.09
Low HDL cholesterol[e]					
Male participants	0.623 (0.02)	0.611 (0.02)	0.23	0.608 (0.02)	0.08
Female participants	0.655 (0.01)	0.633 (0.01)	<0.01	0.667 (0.01)	0.06
High LDL cholesterol[f]					
Male participants	0.583 (0.02)	0.529 (0.02)	0.03	0.596 (0.02)	0.51
Female participants	0.578 (0.03)	0.562 (0.03)	0.37	0.592 (0.03)	0.58
High triglycerides[g]					
Male participants	0.598 (0.02)	0.579 (0.02)	0.12	0.595 (0.02)	0.78
Female participants	0.587 (0.02)	0.539 (0.02)	0.06	0.599 (0.02)	0.61
High SBP or High DBP[h]					
Male participants[i]	0.617 (0.03)	0.562 (0.03)	<0.01	0.606 (0.03)	0.67
Female participants[i]	0.606 (0.05)	0.551 (0.03)	0.13	0.589 (0.04)	0.60

Abbreviations: AUC, area under the curve; DBP, diastolic blood pressure; DXA, dual-energy x-ray absorptiometry; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NHANES, National Health and Nutrition Examination Survey; ROC, receiver operating characteristic; SBP, systolic blood pressure.

[a] ROC-AUC in correctly identifying unfavorable cardiometabolic risk factors. ROC-AUC values ≤ 0.50, 0.51 to 0.70, 0.71 to 0.80, 0.81 to 0.90, and 0.91 to 1.00 indicate zero, weak, acceptable, good, and exceptional predictive powers, respectively.

[b] *P* comparing Model 1 with Model 2.

[c] *P* comparing Model 1 with Model 3.

[d] Total cholesterol>6.2 mmol/L.

[e] HDL cholesterol<1 mmol/L [men], <1.3 mmol/L [women].

[f] LDL cholesterol>4.1 mmol/L.

[g] Triglycerides>2.3 mmol/L.

[h] SBP≥140 mm Hg or DBP≥90 mm Hg.

[i] Male participants (*n* = 2,815); female participants (*n* = 2,751).

4. Discussion and Conclusions

It has been widely acknowledged that adiposity is a significant public health concern because it is a risk factor for a number of chronic diseases, including diabetes, high blood pressure, cardiovascular disease, and some types of cancer (Freedman et al., 2007). Accurately measuring adiposity is essential for tracking the prevalence of obesity and identifying individuals at risk of related health problems. For this reason, this thesis attempted to assess whether phenotypes identified through K-means clustering would outperform phenotypes based on being above/below the 50th percentile for identifying cardiometabolic risks.

Overall comparison of model performance based on ROC-AUCs suggested that in this study, 4-means clustering was superior to the 50th percentile cut-off in predicting cardiometabolic risk, and 2-means clustering was inferior. For instance, the ROC-AUCs from 2-means cluster models used to correctly identify cardiometabolic risk factors in this representative sample of the general US population ranged from 0.52 to 0.63, indicating generally weak predictive power. As the NHANES dataset is a complex heterogeneous dataset, the use of 2-means clustering approach likely was unable to capture more nuanced differences in the data. A 2-means clustering approach may oversimplify the data and miss important subgroups or phenotypes, which can result in less accurate and less meaningful results.

In contrast, the ROC-AUCs from the 50th percentile cut-off phenotypes for correctly identifying cardiometabolic risk were relative stronger (varied between 0.56 and 0.66). However, the ROC-AUCs from 4-means cluster phenotypes were higher (ranging from 0.57 and 0.67), suggesting stronger predictive power than the 50th percentile phenotypes and 2-means cluster phenotypes. The statistical inferiority of 50th percentile cut-off phenotypes' predictive power may reflect the limitations of the approach used to identify them. Although Prado et al.'s (2014) 50th percentile cut-off phenotypes adapted the well-validated lambda-mu-sigma methodology (Cole, 1990; Cole, 2012) to develop sex- and age- specific references curves to classify DXA data into phenotypes of fat and muscle mass, the phenotypes were defined based on whether an individual fell above (≥ 50th percentile) or below (< 50th percentile) permutations of adiposity and muscle mass. This binary categorization might have oversimplified the intricate interplay between adiposity and muscle.

In contrast, K-means clustering is an intuitive algorithm that can be easily implemented, it is also computationally efficient, rendering it suitable for large datasets. Despite these advantages, a significant challenge in applying K-means is specifying the number of clusters beforehand, which can be difficult to tackle. If the number of clusters is too large, the clusters may be too fine-grained, making it difficult to interpret the results or leading to overfitting. There are several methods to determine the optimal number of clusters in K-means, one approach is to use the elbow method, which involves plotting the WCSS against the number of clusters and identifying the "elbow" point where the rate of decrease in WCSS start to level off. Another approach is the silhouette method, which calculates the silhouette coefficient for each observation and then averages them over all observations. The silhouette coefficient measures how similar an observation is to its own cluster compared to other clusters, and a higher average silhouette coefficient indicates better clustering. This thesis used both methods to identify the optimal number of clusters in a more robust and comprehensive way.

Furthermore, the outcomes of K-means may be influenced by initial conditions, leading to differing results depending on the selection of starting points (Nazeer & Sebastian, 2009). If the initial centroids are poorly chosen, the algorithm may converge to a suboptimal solution because the initial centroids might not be representative of an actual cluster in the data. This can lead to an algorithm getting "stuck" in local minimum instead of finding the global optimal solution (Jain, 2010; Krishna & Narasimha Murty, 1999). To address this issue, this thesis followed the general recommendations in the literature by running K-means multiple times with different initializations and choosing the clustering that produced the smallest sum of squared distances (Gentleman & Carey, 2008). This indicated that the clustering algorithm converged to a stable solution for these initial configurations. As the initial selection of cluster centers can have a significant impact on the final clustering results, this was an important step to choose them carefully.

Although the elbow method is commonly employed to determine the optimal number of clusters, it relies on manual identification of the elbow points on the visualization curve. However, a challenge with the elbow method is that analysts may struggle to distinguish the elbow point accurately when the plotted curve is smooth. When a clear elbow is observed in the line chart, the corresponding point likely represents the estimated optimal cluster number. Conversely, if no clear elbow is discernible, the elbow method may not provide reliable results.

Shi et al. (2021) proposed a novel quantitative discriminant method for identifying the potential optimal cluster number in clustering algorithms. They utilized the interaction angle between adjacent elbow points as a criterion to determine a discriminant elbow point. The proposed method relies on estimating the range of the cluster number. For each estimated number of clusters, the entire dataset needs to be trained, resulting in increased computational costs. Thus, in order to successfully implement this method, a high-performance computing environment is essential and was not feasible for this thesis.

Although 4-means cluster defined phenotypes had higher predictive power in identifying cardiometabolic risk compared with 50th percentile cut-off phenotypes and 2-means cluster phenotypes, it was not statistically significantly higher than the 50th percentile cut-offs. This is likely because of the high concordance between the 50th percentile cut-off phenotypes and the 4-means cluster defined phenotypes. To better understand how the 4-means clusters can improve disease identification and prevention, further research on the observations that were discordant between the two methodologies is needed. In addition, AUCs were all still generally poor. As this study evaluated the overall performance of all different phenotypes concomitantly in identifying cardiometabolic risk, it's possible that the performance is being lowered because one or more phenotypes are not informative in predicting cardiometabolic risk. For instance, it is possible that perhaps only one or more specific phenotypes are most closely linked with the cardiovascular risk measures explored in this study, and the others are not informative. It is also possible that phenotypes may be better at identifying some other health event or outcome rather than lipids. In addition to exploring these relationships, other outcomes beyond just current cardiometabolic risks should be assessed in future studies.

This study is not without limitations. Firstly, K-means clustering was applied to all ages in sex-stratified samples. It may not reflect the changes in body composition over different ages in the life-course. This may be improved by performing K-means clustering across different ages for males and females. Despite how large NHANES is, one would need an even larger dataset of all ages to be able to accomplish this and the results of K-means clustering can be more difficult to interpret. Secondly, the data for this study were from individuals who fit within the parameters of the DXA machine (e.g., height limited to no taller than 197 cm, and no wider than 65 cm, with weight limitations from 114-159 kg) (Brownbill & Ilich, 2005). Thus, the findings of this research are only generalizable to the general population that fits the parameters of a DXA machine. However, as this work might be most useful for people with severe obesity, further study on identifying phenotypes when DXA is contraindicated is needed. Lastly, due to the cross-sectional design of this study, causality cannot be inferred (Hill, 1965). For example, it is impossible to determine whether the exposure caused the outcome or whether the outcomes caused the exposure. Future research should incorporate longitudinal designs to determine whether K-means cluster phenotypes or 50th percentile cut-off phenotypes significantly improve prediction power of cardiometabolic risk and health outcomes.

Bibliography

2018 Global reference list of 100 core health indicators (plus health-related SDGs). (2018, April 16). https://apps.who.int/iris/handle/10665/259951

Al-Goblan, A. S., Al-Alfi, M. A., & Khan, M. Z. (2014). Mechanism linking diabetes mellitus and obesity. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *7*, 587–591. https://doi.org/10.2147/DMSO.S67400

Beydoun, M. A., & Wang, Y. (2009). Gender–ethnic Disparity in BMI and Waist Circumference Distribution Shifts in US Adults. *Obesity (Silver Spring, Md.)*, *17*(1), 169–176. https://doi.org/10.1038/oby.2008.492

Blake, G., E. Adams, J., & Bishop, N. (2013). DXA in Adults and Children. In *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism* (pp. 249–263). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118453926.ch30

Bosy-Westphal, A., Booke, C.-A., Blöcker, T., Kossel, E., Goele, K., Later, W., Hitze, B., Heller, M., Glüer, C.-C., & Müller, M. J. (2010). Measurement Site for Waist Circumference Affects Its Accuracy As an Index of Visceral and Abdominal Subcutaneous Fat in a Caucasian Population. *The Journal of Nutrition*, *140*(5), 954–961. https://doi.org/10.3945/jn.109.118737

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Brownbill, R. A., & Ilich, J. Z. (2005). Measuring body composition in overweight individuals by dual energy x-ray absorptiometry. *BMC Medical Imaging*, *5*, 1. https://doi.org/10.1186/1471-2342-5-1

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Carbone, S., Canada, J. M., Billingsley, H. E., Siddiqui, M. S., Elagizi, A., & Lavie, C. J. (2019). Obesity paradox in cardiovascular disease: Where do we stand? *Vascular Health and Risk Management*, *15*, 89–100. https://doi.org/10.2147/VHRM.S168946

Classification: ROC Curve and AUC | Machine Learning. Google for Developers. Retrieved May 30, 2023, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults—The Evidence Report. National Institutes of Health. (1998). *Obesity Research*, *6 Suppl 2*, 51S-209S.

Cole, T. (2021). Sample size and sample composition for constructing growth reference centiles. *Statistical Methods in Medical Research*, *30*(2), 488–507. https://doi.org/10.1177/0962280220958438

Cole, T. J. (1990). The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutrition*, *44*(1), 45–60.

Cole, T. J. (2012). The development of growth references and growth charts. *Annals of Human Biology*, *39*(5), 382–394. https://doi.org/10.3109/03014460.2012.694475

Connor Gorber, S., Shields, M., Tremblay, M. S., & McDowell, I. (2008). The feasibility of establishing correction factors to adjust self-reported estimates of obesity. *Health Reports*, *19*(3), 71–82.

Cooper-DeHoff, R. M., Wen, S., Beitelshees, A. L., Zineh, I., Gums, J. G., Turner, S. T., Gong, Y., Hall, K., Parekh, V., Chapman, A. B., Boerwinkle, E., & Johnson, J. A. (2010). Impact of abdominal obesity on incidence of adverse metabolic effects associated with antihypertensive medications. *Hypertension (Dallas, Tex.: 1979)*, *55*(1), 61–68. https://doi.org/10.1161/hypertensionaha.109.139592

Curtin, L. R., Mohadjer, L. K., Dohrmann, S. M., Montaquila, J. M., Kruszan-Moran, D., Mirel, L. B., Carroll, M. D., Hirsch, R., Schober, S., & Johnson, C. L. (2012). The National Health and Nutrition Examination Survey: Sample Design, 1999-2006. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, *155*, 1–39.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fedewa, M. V., Nickerson, B. S., & Esco, M. R. (2019). Associations of body adiposity index, waist circumference, and body mass index in young adults. *Clinical Nutrition*, *38*(2), 715–720. https://doi.org/10.1016/j.clnu.2018.03.014

Flegal, K. M., & Cole, T. J. (2013). Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts. *National Health Statistics Reports*, *63*, 1–3.

Flint, A. J., Rexrode, K. M., Hu, F. B., Glynn, R. J., Caspard, H., Manson, J. E., Willett, W. C., & Rimm, E. B. (2010). Body mass index, waist circumference, and risk of coronary heart disease: A prospective study among men and women. *Obesity Research & Clinical Practice*, *4*(3), e171–e181. https://doi.org/10.1016/j.orcp.2010.01.001

Freedman, D. S., Mei, Z., Srinivasan, S. R., Berenson, G. S., & Dietz, W. H. (2007). Cardiovascular risk factors and excess adiposity among overweight children and adolescents: The Bogalusa Heart Study. *The Journal of Pediatrics*, *150*(1), 12-17.e2. https://doi.org/10.1016/j.jpeds.2006.08.042

Friedenreich, C. M., Ryder-Burbidge, C., & McNeil, J. (2021). Physical activity, obesity and sedentary behavior in cancer etiology: Epidemiologic evidence and biologic mechanisms. *Molecular Oncology*, *15*(3), 790–800. https://doi.org/10.1002/1878-0261.12772

Frigui, H., & Krishnapuram, R. (1999). A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(5), 450–465. https://doi.org/10.1109/34.765656

Gentleman, R., & Carey, V. J. (2008). Unsupervised Machine Learning. In F. Hahne, W. Huber, R. Gentleman, & S. Falcon (Eds.), *Bioconductor Case Studies* (pp. 137–157). Springer. https://doi.org/10.1007/978-0-387-77240-0_10

Goswami, B., Reang, T., Sarkar, S., Sengupta, S., & Bhattacharjee, B. (2020). Role of body visceral fat in hypertension and dyslipidemia among the diabetic and nondiabetic ethnic population of Tripura—A comparative study. *Journal of Family Medicine and Primary Care*, *9*(6), 2885–2890. https://doi.org/10.4103/jfmpc.jfmpc_187_20

Government of Canada, S. C. (2019, June 25). *Overweight and obese adults, 2018*. https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00005-eng.htm#n1

Grundy, S. M., Stone, N. J., Bailey, A. L., Beam, C., Birtcher, K. K., Blumenthal, R. S., Braun, L. T., de Ferranti, S., Faiella-Tommasino, J., Forman, D. E., Goldberg, R., Heidenreich, P. A., Hlatky, M. A., Jones, D. W., Lloyd-Jones, D., Lopez-Pajares, N., Ndumele, C. E., Orringer, C. E., Peralta, C. A., … Yeboah, J. (2019). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical

Practice Guidelines. *Journal of the American College of Cardiology*, *73*(24), 3168–3209. https://doi.org/10.1016/j.jacc.2018.11.002

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, *17*(2), 107–145. https://doi.org/10.1023/A:1012801612483

Hattori, K., Tatsumi, N., & Tanaka, S. (1997). Assessment of body composition by using a new chart method. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, *9*(5), 573–578. https://doi.org/10.1002/(SICI)1520-6300(1997)9:5<573::AID-AJHB5>3.0.CO;2-V

Heymsfield, S. B., Smith, R., Aulet, M., Bensen, B., Lichtman, S., Wang, J., & Pierson, R. N. (1990). Appendicular skeletal muscle mass: Measurement by dual-photon absorptiometry. *The American Journal of Clinical Nutrition*, *52*(2), 214–218. https://doi.org/10.1093/ajcn/52.2.214

High cholesterol: Overview. (2017). In *InformedHealth.org*. Institute for Quality and Efficiency in Health Care (IQWiG). Retrieved April 10, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK279318/

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300.

Hruby, A., & Hu, F. B. (2015). The Epidemiology of Obesity: A Big Picture. *PharmacoEconomics*, *33*(7), 673–689. https://doi.org/10.1007/s40273-014-0243-x

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Izrael, D. (2002). *Use of the ROC Curve and the Bootstrap in Comparing Weighted Logistic Regression Models*.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Unsupervised Learning. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 497–552). Springer US. https://doi.org/10.1007/978-1-0716-1418-1_12

Janssen, I., Katzmarzyk, P. T., & Ross, R. (2004). Waist circumference and not body mass index explains obesity-related health risk. *The American Journal of Clinical Nutrition*, *79*(3), 379–384. https://doi.org/10.1093/ajcn/79.3.379

Jehan, S., Zizi, F., Pandi-Perumal, S. R., Wall, S., Auguste, E., Myers, A. K., Jean-Louis, G., & McFarlane, S. I. (2017). Obstructive Sleep Apnea and Obesity: Implications for Public Health. *Sleep Medicine and Disorders: International Journal*, *1*(4), 00019.

Johnson, C. L., Paulose-Ram, R., Ogden, C. L., Carroll, M. D., Kruszon-Moran, D., Dohrmann, S. M., & Curtin, L. R. (2013). National health and nutrition examination survey: Analytic guidelines, 1999-2010. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, *161*, 1–24.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. https://doi.org/10.1007/BF02289588

Kakinami, L., Danieles, P. K., Ajibade, K., Santosa, S., & Murphy, J. (2021). Adiposity and muscle mass phenotyping is not superior to BMI in detecting cardiometabolic risk in a cross-sectional study. *Obesity*, *29*(8), 1279–1284. https://doi.org/10.1002/oby.23197

Kakinami, L., Plummer, S., Cohen, T. R., Santosa, S., & Murphy, J. (2022). Body-composition phenotypes and their associations with cardiometabolic risks and health behaviours in a representative general US sample. *Preventive Medicine*, *164*, 107282. https://doi.org/10.1016/j.ypmed.2022.107282

Kelly, A., Winer, K. K., Kalkwarf, H., Oberfield, S. E., Lappe, J., Gilsanz, V., & Zemel, B. S. (2014). Age-Based Reference Ranges for Annual Height Velocity in US Children. *The Journal of Clinical Endocrinology and Metabolism*, *99*(6), 2104–2112. https://doi.org/10.1210/jc.2013-4455

King, J. H., Hall, M. A. K., Goodman, R. A., & Posner, S. F. (2021). Life in Data Sets: Locating and Accessing Data on the Health of Americans Across the Life Span. *Journal of Public Health Management and Practice: JPHMP*, *27*(3), E126–E142. https://doi.org/10.1097/PHH.0000000000001079

King, L. K., March, L., & Anandacoomarasamy, A. (2013). Obesity & osteoarthritis. *The Indian Journal of Medical Research*, *138*(2), 185–193.

Kodinariya, T., & Makwana, P. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*, 90–95.

Kraemer, W. J., Torine, J. C., Silvestre, R., French, D. N., Ratamess, N. A., Spiering, B. A., Hatfield, D. L., Vingren, J. L., & Volek, J. S. (2005). Body size and composition of

National Football League players. *Journal of Strength and Conditioning Research*, *19*(3), 485–489. https://doi.org/10.1519/18175.1

Krishna, K., & Narasimha Murty, M. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 433–439. https://doi.org/10.1109/3477.764879

Laskey, M. A. (1996). Dual-energy X-ray absorptiometry and body composition. *Nutrition*, *12*(1), 45–51. https://doi.org/10.1016/0899-9007(95)00017-8

Lean, M. E., Han, T. S., & Morrison, C. E. (1995). Waist circumference as a measure for indicating need for weight management. *BMJ (Clinical Research Ed.)*, *311*(6998), 158–161. https://doi.org/10.1136/bmj.311.6998.158

Lohr, S. (2012). *343-2012: Using SAS® for the Design, Analysis, and Visualization of Complex Surveys*.

Measurement Site for Waist Circumference Affects Its Accuracy As an Index of Visceral and Abdominal Subcutaneous Fat in a Caucasian Population | The Journal of Nutrition | Oxford Academic. (2010). Retrieved March 2, 2023, from https://academic.oup.com/jn/article/140/5/954/4689066

Miller, M., Stone, N. J., Ballantyne, C., Bittner, V., Criqui, M. H., Ginsberg, H. N., Goldberg, A. C., Howard, W. J., Jacobson, M. S., Kris-Etherton, P. M., Lennie, T. A., Levi, M., Mazzone, T., & Pennathur, S. (2011). Triglycerides and Cardiovascular Disease. *Circulation*, *123*(20), 2292–2333. https://doi.org/10.1161/CIR.0b013e3182160726

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245

Montaquila, J. M., Brick, J. M., & Curtin, L. R. (2010). Statistical and Practical Issues in the Design of a National Probability Sample of Births for the Vanguard Study of the National Children’s Study. *Statistics in Medicine*, *29*(13), 1368–1376. https://doi.org/10.1002/sim.3891

Muntner, P., Einhorn, P. T., Cushman, W. C., Whelton, P. K., Bello, N. A., Drawz, P. E., Green, B. B., Jones, D. W., Juraschek, S. P., Margolis, K. L., Miller, E. R., Marie Navar, A., Ostchega, Y., Rakotz, M. K., Rosner, B., Schwartz, J. E., Shimbo, D., Stergiou, G. S., Townsend, R. R., … Appel, L. J. (2019). Blood Pressure Assessment in Adults in Clinical Practice and Clinic-Based Research: JACC Scientific Expert Panel. *Journal of the American College of Cardiology*, *73*(3), 317–335. https://doi.org/10.1016/j.jacc.2018.10.069

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

National Health and Nutrition Examination Survey (NHANES)—Health, United States. (2022, August 8). https://www.cdc.gov/nchs/hus/sources-definitions/nhanes.htm

Nazeer, K. A. A., & Sebastian, M. P. (2009). *Improving the Accuracy and Efficiency of the k-means Clustering Algorithm*.

Nevill, A. M., Stewart, A. D., Olds, T., & Holder, R. (2006). Relationship between adiposity and body size reveals limitations of BMI. *American Journal of Physical Anthropology*, *129*(1), 151–156. https://doi.org/10.1002/ajpa.20262

NHANES - About the National Health and Nutrition Examination Survey. (2022, December 21). Retrieved April 12, 2023, from https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

NHANES - History. (2019, May 8). Retrieved April 12, 2023, from https://www.cdc.gov/nchs/nhanes/history.htm

NHIS - National Health Interview Survey. (2023, February 9). Retrieved April 12, 2023, from https://www.cdc.gov/nchs/nhis/index.htm

Njeh, C. F., Fuerst, T., Hans, D., Blake, G. M., & Genant, H. K. (1999). Radiation exposure in bone mineral density assessment. *Applied Radiation and Isotopes*, *50*(1), 215–236. https://doi.org/10.1016/S0969-8043(98)00026-8

Patel, C. J., Pho, N., McDuffie, M., Easton-Marks, J., Kothari, C., Kohane, I. S., & Avillach, P. (2016). A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.96

Piech, C. (2013). *K-means algorithm*. Retrieved May 30, 2023, from https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Postorino, M., Marino, C., Tripepi, G., Zoccali, C., & CREDIT (Calabria Registry of Dialysis and Transplantation) Working Group. (2009). Abdominal obesity and all-cause and cardiovascular mortality in end-stage renal disease. *Journal of the American College of Cardiology*, *53*(15), 1265–1272. https://doi.org/10.1016/j.jacc.2008.12.040

Prado, C. M. M., Siervo, M., Mire, E., Heymsfield, S. B., Stephan, B. C. M., Broyles, S., Smith, S. R., Wells, J. C. K., & Katzmarzyk, P. T. (2014). A population-based approach to define body-composition phenotypes. *The American Journal of Clinical Nutrition*, *99*(6), 1369–1377. https://doi.org/10.3945/ajcn.113.078576

Prentice, A. M., & Jebb, S. A. (2001). Beyond body mass index. *Obesity Reviews*, *2*(3), 141–147. https://doi.org/10.1046/j.1467-789x.2001.00031.x

Public Health Considerations Regarding Obesity—StatPearls—NCBI Bookshelf. (2023). Retrieved March 1, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK572122/

Ramos-Nino, M. E. (2013). The Role of Chronic Inflammation in Obesity-Associated Cancers. *ISRN Oncology*, *2013*, 697521. https://doi.org/10.1155/2013/697521

Rezaee, M., Lelieveldt, B. P. F., & Reiber, J. H. C. (1998). A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, *19*(3), 237–246. https://doi.org/10.1016/S0167-8655(97)00168-2

Ripka, W. L., Ulbricht, L., & Gewehr, P. M. (2017). Body composition and prediction equations using skinfold thickness for body fat percentage in Southern Brazilian adolescents. *PLoS ONE*, *12*(9), e0184854. https://doi.org/10.1371/journal.pone.0184854

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rubin, D. B. (1987). Statistical Background. In *Multiple Imputation for Nonresponse in Surveys* (pp. 27–74). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316696.indauth

Rubin, D. B., & Schenker, N. (1986). Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, *81*(394), 366–374. https://doi.org/10.2307/2289225

Sacks, F. M. & Expert Group on HDL Cholesterol. (2002). The role of high-density lipoprotein (HDL) cholesterol in the prevention and treatment of coronary heart disease: Expert group recommendations. *The American Journal of Cardiology*, *90*(2), 139–143. https://doi.org/10.1016/s0002-9149(02)02436-0

Sasirekha, K., & Baby, P. (2013). *Agglomerative Hierarchical Clustering Algorithm- A Review*. *3*(3).

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine*, *37*(14), 2252–2266. https://doi.org/10.1002/sim.7654

Shah, N. R., & Braverman, E. R. (2012). Measuring Adiposity in Patients: The Utility of Body Mass Index (BMI), Percent Body Fat, and Leptin. *PLoS ONE*, *7*(4), e33308. https://doi.org/10.1371/journal.pone.0033308

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 31. https://doi.org/10.1186/s13638-021-01910-w

Silhouette (clustering). In Wikipedia, the free encyclopedia. Retrieved May 30, 2023, from https://en.wikipedia.org/w/index.php?title=Silhouette_(clustering)&oldid=1155324336

Sullivan, T. R., Salter, A. B., Ryan, P., & Lee, K. J. (2015). Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data. *American Journal of Epidemiology*, *182*(6), 528–534. https://doi.org/10.1093/aje/kwv100

Tambalis, K. D., Panagiotakos, D. B., Arnaoutis, G., Psarra, G., Maraki, M., Mourtakos, S., Grigorakis, D., & Sidossis, L. S. (2015). Establishing cross-sectional curves for height, weight, body mass index and waist circumference for 4- to 18-year-old Greek children, using the Lambda Mu and Sigma (LMS) statistical method. *Hippokratia*, *19*(3), 239–248.

Tiwari, A., & Balasundaram, P. (2022). Public Health Considerations Regarding Obesity. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK572122/

Tsujimoto, T., & Kajio, H. (2017). Abdominal Obesity Is Associated With an Increased Risk of All-Cause Mortality in Patients With HFpEF. *Journal of the American College of Cardiology*, *70*(22), 2739–2749. https://doi.org/10.1016/j.jacc.2017.09.1111

Yumuk, V., Tsigos, C., Fried, M., Schindler, K., Busetto, L., Micic, D., Toplak, H., & Obesity Management Task Force of the European Association for the Study of Obesity. (2015). European Guidelines for Obesity Management in Adults. *Obesity Facts*, *8*(6), 402–424. https://doi.org/10.1159/000442721