University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Katrin Kulberg

Relationships Between Genetic and Lifestyle Factors and Weight in the Estonian Biobank

Mathematical Statistics

Bachelor's Thesis (9 ECTS)

Supervisors: PhD Jon Anders Eriksson
PhD Kristi Kuljus

Tartu 2024

RELATIONSHIPS BETWEEN GENETIC AND LIFESTYLE FACTORS AND WEIGHT IN THE ESTONIAN BIOBANK

Bachelor's thesis

Katrin Kulberg

Abstract

The aim of this bachelor's thesis is to explore relationships between weight and a range of lifestyle and genetic factors. In this study, the genetic factors under consideration are single nucleotide polymorphisms. The analysis is performed on data originating from the Estonian Biobank and focuses on a cohort of slightly over 40 000 gene donors. In addition to identifying associations, emphasis is also placed on quantifying them by finding the effect sizes of relevant factors on weight. To better assess age and gender-specific patterns, the analysis is conducted separately across eight groups, by considering both females and males in the following four age categories: individuals under 30, 30-45, 46-65, and those aged 66 or older. The thesis gives an overview of the necessary theoretical background to enhance understanding of the analysis, introduces the statistical methodology used, then describes the data, and concludes by presenting and discussing the results.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Body weight, obesity, genetic factors, lifestyle factors, linear regression.

SEOSED KAALU JA GENEETILISTE TEGURITE JA ELUSTIILITEGURITE VAHEL TÜ EESTI GEENIVARAMU ANDMETEL

Bakalaureusetöö

Katrin Kulberg

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on uurida seoseid kaalu ja mitmete erinevate elustiilitegurite ja geneetiliste tegurite vahel. Antud töös käsitletakse geneetiliste teguritena üksiku nukleotiidi polümorfisme. Analüüs põhineb TÜ Eesti Geenivaramu andmetel, kusjuures vaatluse alla võetakse veidi enam kui 40 000 geenidoonorit. Lisaks seoste tuvastamisele, pööratakse tähelepanu ka erinevate faktorite mõju arvulisele kirjeldamisele. Hindamaks paremini vanusest ja soost tulenevaid eripärasid, viiakse analüüs eraldi läbi kaheksas grupis, nii naiste kui meeste lõikes vaadeldakse nelja vanuserühma: kuni 30-aastased, 30–45-aastased, 46–65-aastased ning vähemalt 66-aastased. Töös antakse ülevaade vajalikust teoreetilisest taustast analüüsi paremaks mõistmiseks, tutvustatakse kasutatud statistilist metoodikat, kirjeldatakse uuritavat andmestikku ning seejärel arutletakse tulemuste üle.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Kehakaal, rasvumine, geneetilised tegurid, elustiilitegurid, lineaarne regressioon.

Contents

Introduction

Overweight and obesity have become increasingly prevalent worldwide over the past couple of decades. While overweight is simply defined as excessive fat accumulation, obesity is recognised as a chronic complex disease. However, both of them can lead to numerous health problems, as well as reduce the quality of life, and thus pose a serious concern to society. Overweight and obesity are most commonly characterised by the body mass index (BMI): adult individuals with a BMI of 25 or higher are regarded as overweight, and those with a BMI of at least 30 as obese. (World Health Organization, 2024)

Obesity, which is considered the more severe of the two conditions, stems from the interplay of environmental and innate biological factors. While the role of the environment when it comes to the development of obesity can not be underestimated by any means, an individual's susceptibility to obesity also relies heavily on his or her genetic determinants. (Loos and Yeo, 2022)

The aim of this thesis is to explore associations between environmental factors, specifically lifestyle determinants, and weight, as well as genetic factors and weight. In addition to identifying associations, the purpose of the thesis is also to try to quantify them and to provide a better understanding of their effect. The analysis in the thesis is based on data obtained from the Estonian Biobank and focuses on a cohort of slightly over 40 000 gene donors.

The thesis consists of five parts. Section 1 gives an overview of obesity from a genetic perspective, discusses methods for identifying genes linked to phenotypes, delves into gene-environment interactions as well as different lifestyle factors, and lastly provides insight into overweight and obesity trends in Estonia. The second section is concerned with the statistical methodology used in this thesis. Section 3 offers an overview of the data, while also placing emphasis on discussing the genetic data used in this thesis.

The last two parts are empirical. Section 4 provides a descriptive overview of the key variables used in the analysis. Finally, the fifth section is dedicated to the presentation and discussion of the results.

The author is grateful to her supervisors Kristi Kuljus and Jon Anders Eriksson, for their valuable guidance and dedicated time.

1 Background

This section aims to give a better understanding of obesity from a genetic perspective and discuss the most prevalent contemporary method for identifying gene-obesity associations. The section also looks at key environmental factors, as well as the interaction between genetic and environmental elements. In the last part, emphasis is placed on the overweight and obesity trends in Estonia, as the forthcoming analysis will focus specifically on the part of the Estonian population.

1.1 Obesity

According to the classical framework, there are two types of obesity: monogenic and polygenic. Monogenic obesity results from a single-gene mutation and follows, in terms of heritability, a Mendelian pattern. It is considered to be rare and usually occurs early in the life. (Loos and Yeo, 2022)

Polygenic obesity, on the other hand, comes from an interaction between the so-called obesogenic environment, an environment in favour of weight gain, and numerous genetic variants, where each variant on its own has a small effect. As a result, the heritability of polygenic obesity is far more complex and no longer adheres to simple Mendelian inheritance. (Loos and Yeo, 2022)

Due to their characteristics, the two forms of obesity can be also regarded as rare and common obesity, respectively. Despite them being initially seen as separate entities, advancements in genetic studies over the past decades have led to the line between them becoming much more blurred. (Loos and Yeo, 2022)

1.2 Genome-Wide Association Studies

The methods to identify genes linked to obesity have seen many changes over the years since the start in the 1990s. Progressing from the hypothesis-driven approach,

which in principle makes use of the underlying biology by selecting a set of candidate genes that are thought to be associated with body-weight regulation, to the hypothesis-generating approach, which makes no such assumptions in advance, the breakthrough was made by the introduction of genome-wide association studies (GWAS) in 2005. GWAS, as a hypothesis-generating approach, makes use of the entire genome, unlike some of the previous methods that focused on specific genes or regions of interest, without prior assumptions. As a result, it is much more effective at identifying associations between genetic variants and phenotypes, such as obesity, than its predecessors. (Loos and Yeo, 2022)

The most common genetic variants used in GWAS are single nucleotide polymorphisms, abbreviated as SNPs (Tam et al., 2019). In essence, an SNP represents a difference between two DNA sequences at a single nucleotide level, where one of the nucleotides, adenine (A), thymine (T), cytosine (C), or guanine (G), in an individual's genome differs from the reference sequence, with the variation typically occurring in at least 1% of the population (National Cancer Institute, 2012). It is worth mentioning that the reference sequence does not originate from a single individual, but rather represents a composite from many that should accurately reflect the human genome, thus acting as a framework that can be used to identify deviations (National Human Genome Research Institute, 2024).

While GWAS have demonstrated to be highly successful in discovering numerous genetic variants associated with phenotypes, they also entail one significant disadvantage. Namely, due to their hypothesis-free approach, explaining the causality aspect has proven to be difficult, thus in turn hindering the interpretation of results. (Tam et al., 2019)

One prime illustration in the context of obesity is the FTO locus, the genomic region surrounding the FTO gene, which stands as the most prominent obesity-associated locus up to date that has been discovered with GWAS. Despite having been thoroughly studied for over a decade, a total understanding of biological mech-

anisms that would explain the causal relationship between the locus and obesity poses a challenge to this day. (Loos and Yeo, 2022)

The role of GWAS, however, should not be underestimated by any means, as they have truly revolutionised the field of genetic studies. For obesity alone, in less than 20 years, over 50 GWAS have been published, uncovering more than 1100 loci linked with various obesity-related traits, which represents a valuable contribution to the genetics behind obesity. (Loos and Yeo, 2022)

Furthermore, sometimes it is specifically the hypothesis-free nature of GWAS that paves the way to new biological insights behind the diseases or traits, as such associations could not have been detected otherwise. On top of that, it is believed that many of the current challenges, including the causality aspect, could be solved or at least extenuated in effect in the future thanks to improved methodology. (Tam et al., 2019)

1.3 Gene-Environment Interactions

Genetic factors alone are not able to fully account for the heritability of complex traits, but it is also the environment that plays a crucial role, and must be, therefore, taken into account. Such an approach of considering both genetic and environmental elements together is known as a gene-environment interaction. (Simon et al., 2016)

In principle, a gene-environment interaction is a relationship between a genetic variant and an environmental factor. In particular, it aims to explain how the environmental determinants impact their genetic counterparts. For instance, exposure to certain environmental factors can increase the risk of developing a disease for individuals with a genetic predisposition to it. It is important to highlight that within this framework, environmental factors can be any external source of stress or risk. Common examples include socioeconomic indicators, as well as lifestyle habits such as diet or smoking. (Simon et al., 2016)

Gene-environment interactions are dynamic phenomena that change across the lifespan, rather than remain constant. This is best illustrated by examining the characteristics of the individual components and considering how they depend on time. It is well-known that the genetic component usually takes action early in life, whereas the environmental factors require longer exposure time to them in order to take effect. Furthermore, while the genetic structure of a human remains stable over time, the environment, in particular, has the power to vary considerably. For these reasons, it is therefore crucial to regard time as a significant element in studying gene-environment interactions. (Simon et al., 2016)

The second important factor to consider in the light of gene-environment interactions is sex, as genetic variants can sometimes play different roles depending on the gender. For instance, by acting as a protective variant in one and harmful for the other gender. On top of gene-age interactions, with the additional term sex, it is also possible to look at the sex-age interactions, helping to uncover the associations that might otherwise be overlooked. In practice, this is achieved by stratifying data into groups based on both gender and age. (Simon et al., 2016)

Additional challenges encountered in studying gene-environment interactions include the impact of medication and underlying diseases, both of which can alter the effects of genetic variants. Similarly, it is essential to take into account the physiological factors, as the stress level, specifically, tends to play a role in determining the extent to which a genetic variant affects the phenotype of interest. Although both physical and mental stress are considered, the latter has a stronger impact and can thus lead to more serious consequences. (Simon et al., 2016)

Some of the most extensively studied lifestyle factors are physical activity and diet. On the one hand, in today's obesogenic environment, where the balance between energy intake and expenditure is often disrupted by reduced exercise and consumption of high-calorie food, both are recognised as significant contributors to weight gain from a purely physical perspective. On the other hand, the aforementioned

two factors are also believed to modulate the degree to which genetic factors influence the development of obesity, highlighting once again the importance of gene-environment interactions. In particular, adopting a physically active lifestyle and following a certain diet can weaken the effects of genetic susceptibility to obesity. (Albuquerque et al., 2017)

Analysing the role of socioeconomic factors in the context of obesity all comes down to the inequalities in society, primarily driven by the available resources, especially financial. In developed countries, the relationship between higher socioeconomic status and higher rates of obesity is inverse, meaning obesity is more common amongst underprivileged individuals. In developing countries, however, the trend is the other way around. (Adams, 2020)

Financial constraints are also relevant in terms of diet and exercise, as it is rather the quality of food and intensity of physical activity that is believed to be the main determinant instead of the mere quantity. Due to the fact that both high-quality nutritious food and regular workouts, including the necessary equipment, represent additional expenditures, practising a healthier lifestyle in this regard is inevitably tied to financial opportunities. (Adams, 2020)

But it is also beliefs and outlooks on life that can influence decisions related to healthy habits. Especially relevant is the role of education, as individuals with higher levels of education are more likely to follow health-related advice, as well as implement it in their own lives. This has been best illustrated in studies focused on childhood obesity, which highlight the importance of parental education level being one of the key determinants in a child's initial exposure to a healthy lifestyle. (Albuquerque et al., 2017)

1.4 Trends in Estonia

The problems related to excess body weight have not left Estonia untouched. In Estonia, approximately half of the adult population is either overweight or obese.

An age-period-cohort study, based on data from biennial Health Behaviour Surveys from years 1996 to 2018 focusing on individuals aged 16 to 64, revealed that between survey years 1996/1998 and 2016/2018, average BMI increased by 1.4 points among men and 0.64 points among women. It is worth noting that the analysis also took into account the education level and ethnic background, as both have been shown to be relevant factors in similar studies previously. (Reile et al., 2020)

The same study also highlighted the fact that there is a significant relationship between age and BMI, with some key differences between genders. For men, an increase in average BMI was observed until late 40s, after which it slowed down and even showed a decline around the age of 60, thus resembling a curvilinear relationship. However, for women in the observed age group of 16 to 64 years, there was an almost positive linear relationship between mean BMI and age. (Reile et al., 2020)

Obesity in Estonia is also relevant in the context of premature death, as the country tends to experience higher rates of mortality from preventable and treatable causes than the European Union average. For instance, in 2020, cardiovascular diseases, for which obesity is a considerable risk factor, were the most prevalent treatable cause of death. (OECD and European Observatory on Health Systems and Policies, 2023)

As growing rates of obese individuals in Estonia pose a serious concern, there have been several attempts to alleviate the problem at the national level. One of the most notable examples is a green paper on nutrition and physical activity, which was started back in 2014. Although it has yet to be implemented, there are plans to do so by the end of 2024. (OECD and European Observatory on Health Systems and Policies, 2023). The green paper on nutrition and physical activity is a governmental document consisting of proposals, aiming to tackle the problem of sedentary lifestyle and unhealthy diet (Ministry of Social Affairs, 2023).

2 Multiple Linear Regression

The following section is written based on Chatterjee and Hadi (2006), and gives an overview of the statistical methodology used in this thesis.

Linear regression model is a statistical model used in case of a linear relationship between the response variable Y and explanatory variables $X_1, \ldots, X_p$. Mathematically, the model can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon, \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are called regression coefficients and ε is referred to as a random disturbance or error term.

By defining the following matrices based on a data set of n observations,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \ldots & x_{1p} \\ x_{20} & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

model (1) can be written in matrix notation as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Here, σ^2 represents the variance of the error term, and $\boldsymbol{I}$ is an identity matrix. The estimates of the coefficients $\boldsymbol{\beta}$ are found by considering the sum of squares of errors:

$$S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{2}$$

In order to determine the estimates $\hat{\boldsymbol{\beta}}$, the sum of squares (2) is minimised with respect to $\boldsymbol{\beta}$, yielding the following system of equations, also known as normal

equations:

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}. \tag{3}$$

If $\mathbf{X}^T\mathbf{X}$ is invertible, the least squares estimate of $\boldsymbol{\beta}$ can be expressed from (4) as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The fundamental quantities for linear regression are the following three types of sum of squares. Firstly, the sum of squares due to regression, regarded as SSR:

$$SSR = (\hat{\boldsymbol{Y}} - \overline{\mathbf{Y}})^T(\hat{\boldsymbol{Y}} - \overline{\mathbf{Y}}),$$

where $\hat{\boldsymbol{Y}}$ denotes the vector of predicted values by the model, i.e. $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and $\overline{\mathbf{Y}}$ is a vector of n elements, where each element is calculated as $\frac{1}{n}\sum_{i=1}^{n} y_i$, i.e. the sample mean of the response variable measurements. Secondly, the sum of squared residuals, also known as SSE, is given by:

$$SSE = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Lastly, the total sum of squared errors, abbreviated as SST, which is also the sum of SSR and SSE, is written as

$$SST = SSR + SSE = (\mathbf{Y} - \overline{\mathbf{Y}})^T(\mathbf{Y} - \overline{\mathbf{Y}}).$$

In order to test hypotheses about regression coefficients, variances of them are needed, for which an estimator of σ^2 has to be defined:

$$\hat{\sigma}^2 = \frac{SSE}{n - (p+1)}.$$

Now it is possible to express the standard error of $\hat{\beta}_j$ $(j = 0, \ldots, p)$ as follows:

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{c_{jj}},$$

where c_{jj} is the jth diagonal element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. It is evident that as sample size increases, the values of c_{jj} decrease, while $\hat{\sigma}^2$ remains about the same. As a result, the standard error $\hat{\sigma}_{\hat{\beta}_j}$ of an estimate $\hat{\beta}_j$ is smaller in a larger sample.

For multiple linear regression, there are two types of considerations regarding significance. The first of them is concerned with the overall significance of the model, where the null hypothesis states

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0,$$

i.e. that all regression coefficients, except for the intercept term, are equal to zero. Therefore, it implies that the model is considered statistically significant if at least one of the regression coefficients of predictor variables differs from zero. This can be tested by first defining the F-statistic:

$$F = \frac{SSR/p}{SSE/(n-p-1)}. \tag{4}$$

The F-statistic follows an F distribution under the null hypothesis, $F \sim F_{p,n-p-1}$. The null hypothesis is rejected if $F > F_{p,n-p-1;\alpha}$, where α is a chosen significance threshold.

Similarly, it is possible to test significance of a single parameter β_j $(j = 1, \ldots, p)$, where under the null hypothesis it is equal to zero. This can be tested by using the t-statistic defined as follows:

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}. \tag{5}$$

Under the null hypothesis, the t-statistic follows a Student's t-distribution $t \sim t_{n-p-1}$. The null hypothesis is rejected if $|t| > t_{n-p-1;\alpha}$.

It is worth noting that with an increased sample size, the likelihood of rejecting the null hypothesis when testing for significance also increases. This tendency is particularly relevant for the t-statistic (5), in which increase is caused by the lower $\hat{\sigma}_{\hat{\beta}_j}$, as discussed before. As a result, sample size can strongly influence what statistically significant associations would the model yield. For the F-statistic (4), increase in value is driven by the numerator.

Since the role of sample size for the t-statistic will be an important aspect in the analysis later, an illustrative example is presented, based on the data used in this thesis. Let us consider a variable called *potatoes* and two samples, where one is a subset of the other, consisting of 2119 and 9288 observations, respectively. The effect sizes for this variable are -0.32 in the smaller sample and -0.34 in the bigger sample. The standard errors in the smaller and larger sample are 0.20 and 0.09, respectively. This leads us to the corresponding t-values, which are -1.60 and -3.78. Thus, it can be concluded that the effect of potatoes is significant in the larger sample, whereas in the smaller sample, it is no longer the case. This represents an outcome, which is consistent with the theory discussed above.

To assess the goodness of fit of a linear regression model, the coefficient of determination R^2 is used, which illustrates the proportion of the total variance in the response variable Y that is explained by the predictors $X_1, \ldots, X_p$. Within the established framework, R^2 can be written as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Interpretation of a linear regression model is relatively straightforward. If a coefficient, for example β_1, is positive, then an increase in the corresponding explanatory variable X_1 consequently increases Y. On the other hand, if β_1 is negative, an increase in X_1 leads to a decrease in the value of Y. In particular, change by one unit in X_1 changes Y by β_1 units.

The most important assumptions of a linear regression model are concerned with

the error term. Specifically, it is assumed that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ $(i = 1, \ldots, n)$, which means that the errors are assumed to be normally distributed with a mean of 0 and constant variance, also regarded as homoscedasticity.

3 Overview of Data

Data used in this thesis originates from the Estonian Biobank, a population-based biobank of Estonia comprising information about more than 200 000 individuals. Upon enrollment in the Estonian Biobank, all participants, in addition to providing a DNA sample, also complete a questionnaire, which thematically focuses on four subjects: personal information, genealogical data, educational and occupational history, and lifestyle. (Institute of Genomics, 2021)

The analysis performed in this thesis is only based on the data from the participants who were recruited during the initial phase, which took place between the years 2002 and 2010. It is also worth mentioning that only the records of gene donors whose data consists of a single set of answers, as opposed to filling in the questionnaire multiple times during the recontacting phase, are considered. Due to the aim of this thesis, mainly the variables concerned with physical characteristics and lifestyle are used. Additionally, a selection of 10 key obesity-related genetic variants are studied.

In the following, an overview of the dataset is provided. It is divided into two parts by first considering variables related to physical characteristics and lifestyle determinants, followed by those pertaining to genetic variants. On top of it, the preparation of the dataset for the forthcoming analysis is discussed.

3.1 Physical Characteristics and Lifestyle

The first part of the dataset is concerned with general information about a gene donor, as well as characteristics describing his or her lifestyle. In terms of physical attributes, the variables of interest are weight, the response variable in the upcoming analysis, measured in kilograms, height measured in centimeters and the age of a gene donor at the time of recruitment. Moving on, the highest level of completed education is considered, which is categorised into three groups as follows:

- higher education: university education including professional higher education studies;
- secondary education: general upper secondary and professional secondary education;
- basic education or lower: levels up to and including basic education, as well as individuals without any educational qualifications.

Transitioning to the main section of lifestyle, the first part consists of 12 spare time activities which are all measured in hours per week. More specifically, the considered activities are as follows:

- slow walking;
- moderate walking;
- speed walking;
- cooking;
- shopping;
- cleaning;
- laundry;
- childcare;
- elderly care;
- gardening;
- household repair;
- physical exercise different from walking.

The second portion of the variables related to lifestyle focuses on the dietary habits of gene donors. The majority of them are concerned with the frequency of food and drink consumption. In particular, the studied products are the following:

- potatoes;
- rice and pasta;
- porridge, muesli and flakes;
- milk products;
- fish;
- meat;
- meat products (sausage, frankfurters);
- fresh vegetables;
- boiled vegetables;
- fresh fruits and berries;
- compotes and jams;
- sweets;
- soft drinks;
- eggs.

Additionally, the dietary block includes variables of daily coffee and tea consumption, as well as both bread and white bread intake, measured in cups and slices per day, respectively. It is worth mentioning that initially, the frequency of consumption was considered in intervals: 1-2 days, 3-5 days, 6-7 days, or 0 for individuals who do not consume a given product. However, for modelling purposes, they are

converted into numeric values by taking the midpoint of each interval, except for the zero class, where it is not needed.

Other variables of interest regarding lifestyle pertain to unhealthy habits. In particular, they are smoking and alcohol consumption, both of which are considered in 3 categories, indicating whether a participant is a current or a former smoker or an alcohol consumer, or has never smoked or consumed alcohol.

3.2 Genetic Variants

The second part of the dataset used in this thesis comprises genetic variants, which were chosen according to the study Locke et al. (2015) as the 10 strongest BMI-associated SNPs. Each variant is characterised by a chromosome, marking its position in the genome, reference allele and alternative allele, giving it a code name where the described information is separated by colons. The reference allele is based on the reference sequence, which is familiar from Section 1, while the alternative allele represents the nucleotide found in the individual's DNA. The data about the genetic variants is given in dosage, meaning the number of copies an individual carries, of the alternative allele. The variants of interest in this thesis are given in Table 1, accompanied by their nearest gene based on Locke et al. (2015).

Table 1: Selection of key obesity-associated genetic variants.

Chromosome:position:reference allele:alternative allele	**Nearest gene**
11:27684517:A:G	BDNF
1:72751185:T:C	NEGR1
1:177889480:A:G	SEC16B
12:50247468:G:A	BCDIN3D
16:53803574:T:A	FTO
18:57829135:T:C	MC4R
2:632348:A:G	TMEM18
2:25150296:A:G	ADCY3
4:45182527:A:G	GNPDA2
6:50845490:A:G	TFAP2B

In the following part of the subsection, a brief summary of each gene is provided, giving an overview of its role within the human body. On top of that, the most important associations with obesity are discussed.

The first gene, BDNF, is a protein-encoding gene, encoding a protein that is involved in the growth, survival, and maintenance of nerve cells. Furthermore, the gene is believed to be associated with the regulation of the stress response as well as mood disorders. It is worth mentioning that BDNF has also been linked to the binge-eating disorder Bulimia Nervosa. (GeneCards, 2024c)

Moving on, NEGR1, similarly to BDNF, is also a protein-encoding gene. NEGR1 is involved in several processes, for instance, it plays an important role in feeding and movement activities. One of the most notable diseases associated with NEGR1 is Niemann-Pick Disease (NPD). (GeneCards, 2024g). NPD is a rare inherited metabolic disorder characterised by abnormal accumulation of fats in various parts of the body (MalaCards, 2024b).

The third gene, SEC16B, is concerned with certain parts of the cell called transitional endoplasmic reticulum sites, which are important for moving proteins from one part of the cell to another. SEC16B is also a protein-encoding gene. (GeneCards, 2024h)

Continuing with the next gene, we have BCDIN3D. Conceptually speaking, BCDIN3D belongs to the family of protein-encoding genes. It is worth pointing out that the gene is over-expressed in breast cancer cells, meaning it produces more of its protein than usual. As a result, it may contribute to the more aggressive behaviour of cancer cells and make the prognosis of cancer more difficult. (GeneCards, 2024b)

The next gene, FTO, is a protein-encoding gene strongly linked to BMI and obesity risk. In terms of functions, the gene is involved in various processes, such as the regulation of fat mass. FTO also plays a role in energy homeostasis, in other words, it is maintaining the balance between energy intake and expenditure. (GeneCards, 2024d). Despite its frequent association with obesity, FTO is also linked with a

disorder characterised by distinct facial features and shorter growth called Growth Retardation, Developmental Delay, And Facial Dysmorphism (MalaCards, 2024a).

Similarly to FTO, MC4R is also involved in energy homeostasis. Being a protein-encoding gene, in terms of functions, it plays a crucial role in the physical growth and development of human body tissues and organs, a process known as somatic growth. It is noteworthy that mutations in MC4R are one of the most prevalent causes of monogenic obesity. (GeneCards, 2024f)

The next gene of interest, TMEM18, is also a protein-encoding gene. When it comes to its role, TMEM18 acts as a transcription repressor, meaning that it helps to control the activity of other genes. The gene is also concerned with cell migration. Mutations in TMEM18 have been linked to problems with leptin, in particular its deficiency or dysfunction. (GeneCards, 2024j)

Leptin is a hormone secreted by fat cells and acts as an indicator for the brain on the body's energy reserves, where lower levels correspond to energy deficit and vice versa. It is the activation force for the melanocortin pathway, which can be thought of as a series of signals in the brain resulting in the regulation of food intake and food preference. What is striking, however, is that a great proportion of single-gene mutations giving rise to monogenic obesity are part of this pathway. The aforementioned gene MC4R is also specifically associated with this pathway. (Loos and Yeo, 2022)

Returning to the genes, ADCY3 is involved in various metabolic processes, such as the regulation of body fat and insulin levels. Furthermore, ADCY3 is significant in the context of perception of odorants, as well as crucial for normal male fertility. Conceptually, ADCY3 is also a protein-encoding gene. It is worth mentioning that abnormalities in this gene, similar to a few previous ones, can lead to conditions associated with increased body weight. (GeneCards, 2024a)

The penultimate gene of interest is a protein-encoding gene GNPDA2, which plays a role in metabolic and cellular processes. However, the gene is also relevant in

the light of excess weight. Namely, the variations of GNPDA2 have been found to affect BMI, as well as the risk of obesity. (GeneCards, 2024e)

Concluding the exploration of the genes, there is TFAP2B. TFAP2B is a protein-encoding gene, encoding a protein which belongs to the AP-2 family of transcription factors. AP-2 proteins play a crucial role in regulating gene expression by binding themselves to specific DNA sequences. Moreover, the TFAP2B gene holds significance in normal development, particularly in processes such as the proper development of facial features and limbs. (GeneCards, 2024i)

3.3 Further Considerations Regarding the Dataset

It is worth mentioning that the dataset contained a few extreme observations within the section of 12 leisure activities, where the indicated hours were too high to be realistic. For that reason, the cap of 30 hours was set for all activities, except for child and elderly care, both of which were assigned a limit of 84 hours per week instead. Observations exceeding these limits were excluded from further analysis. On top of that, rows with missing values were also discarded.

In this thesis, data analysis is conducted across 8 groups that are formed based on the gender and age of the gene donors. The groups are outlined in Table 2, along with the number of observations in each category. It is apparent that the group sizes differ quite drastically, representing a potential problem for linear regression, as larger sample sizes tend to yield more statistically significant associations as discussed in Section 2. This issue will be appropriately addressed in Section 5.

Table 2: Distribution of gene donors by gender and age groups.

Age group (years)	**Female**	**Male**
<30	5585	4069
30-45	8167	3812
46-65	9288	4430
66+	4255	2119

4 Descriptive Overview

The aim of this section is to provide an overview of the key variables used in modelling. This section also endeavours to build an intuition for the upcoming analysis by showcasing the relationships between different explanatory variables and weight.

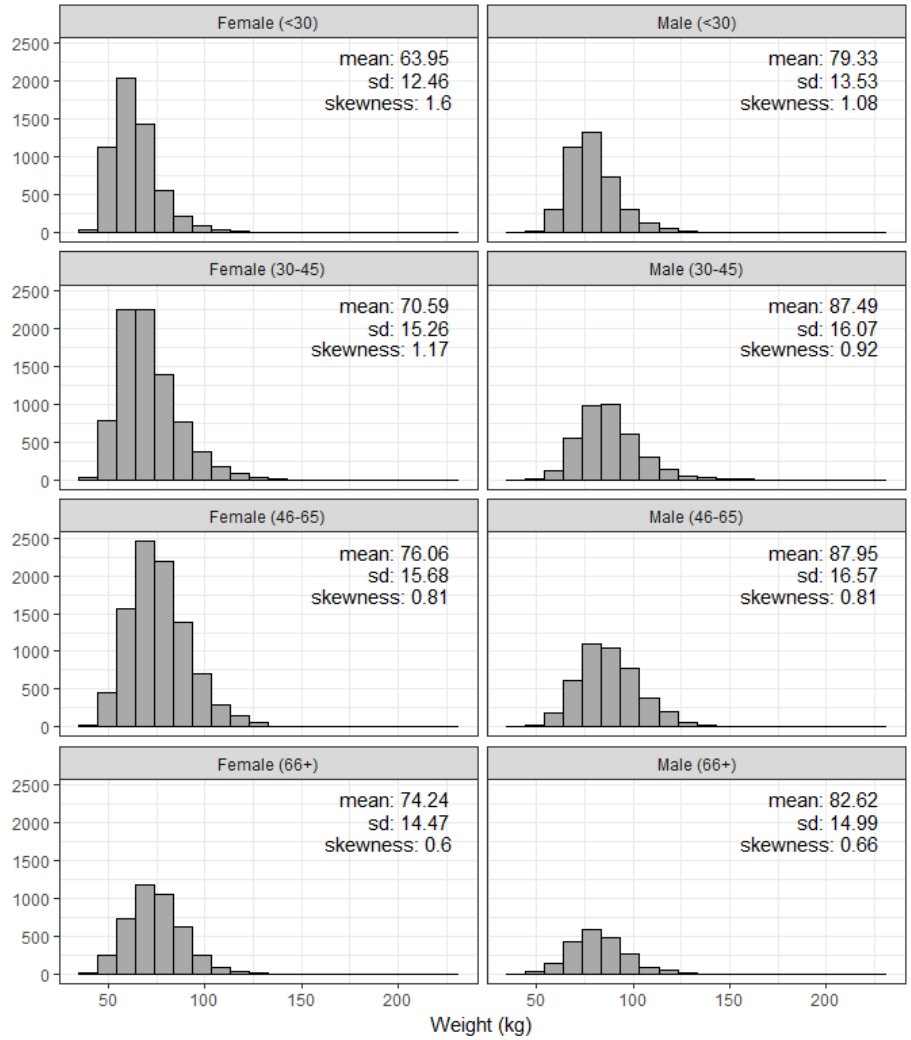

Figure 1: Distribution of weight by gender and age groups.

The distribution of weight, among the eight groups, along with its mean, standard deviation and skewness, is illustrated in Figure 1. It is evident that weight is

not exactly normally distributed and tends to have a heavier right tail, especially among younger women. However, as indicated by both the skewness coefficient and the shapes of the distributions, skewness appears to diminish with increasing age. Overall, the weight trends across age groups are quite similar for women and men. However, for the former, an increase is observed until the age category of 46-65, where the highest average weight is attained, whereas for men, average weights in the two middle age brackets are pretty much the same. Among the individuals aged 66 or older, a subtle decline takes place for both sexes. When it comes to variance, it is relatively stable, with standard deviation ranging from 12-17 kilograms.

As suggested by the heavier right tails seen in Figure 1, excessive body weight is a problem among gene donors of the Estonian Biobank. Table 3 gives an overview of the percentages of overweight and obese individuals, as defined by the WHO criteria. As a reminder, individuals with a BMI of at least 25 are considered overweight, and those with a BMI of at least 30 are obese. For a better overview, percentages for the conditions have been calculated separately, i.e. overweight does not include obesity. The table demonstrates that the condition of being overweight is more prevalent among men, whose corresponding proportions are always higher. On the other hand, when it comes to obesity, it is also the age factor that has to be taken into account. In particular, among younger individuals, the difference between women and men is quite small. However, in the second half of the age spectrum, the gender disparity becomes more pronounced, with women showing significantly higher rates of obesity. It is noteworthy that the difference between sexes in the last age bracket is approximately 13%.

Table 3: Prevalence of overweight and obesity in percentage by gender and age groups.

Condition	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Overweight	15.4	25.9	34.9	38.0	25.2	41.9	41.9	43.6
Obese	6.2	18.2	35.0	37.5	6.91	20.8	30.9	24.2

In the following, a closer look at the primary categorical variables in the dataset is provided. Starting with education, Table 4 gives an overview of its distribution across the eight groups of interest. Firstly, it is apparent that high school education represents the dominating level, except for males in the last age bracket, with proportions in the groups typically ranging between 50-65%. The table also suggests that among younger individuals, higher education is more prevalent for women, whose corresponding percentages are, on average, roughly 10% higher compared to men. However, among gene donors aged 66 or older, men surpass women in the proportion of those with higher education instead.

Table 4: Proportions of education levels in percentage by age and gender groups.

Education level	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Up to basic school	21	6	11	38	27	11	20	39
High school	55	63	63	45	60	66	56	38
Higher education	24	31	26	17	13	23	24	23

As education is often seen as an important socioeconomic factor shaping beliefs and behaviors related to healthy lifestyles, it would be interesting to also examine potential differences in average weights. For comparison, both now and henceforth, the mean weight in every age-gender group is presented once more (in the same order as in tables):

$$63.95,\ \ 70.59,\ \ 76.06,\ \ 74.24,\ \ 79.33,\ \ 87.49,\ \ 87.95,\ \ 82.62.$$

Average weights across the three levels of education have been summarised in Table 5. It is interesting that for women, an increase in educational level is typically associated with a slightly lower average weight. Conversely, for men, this trend is reversed, with the highest average weight observed among individuals with higher education, although differences remain marginal.

Table 5: Average weights across education levels by age and gender groups.

Education level	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Up to basic school	63.6	72.4	79.0	75.2	76.6	83.1	85.8	81.6
High school	64.3	71.2	76.3	74.1	80.1	87.8	88.2	82.9
Higher education	63.5	69.0	74.2	72.4	81.5	88.7	89.1	83.9

The next two variables of interest are associated with unhealthy lifestyle choices and will be analysed similarly to education. The first one of them is smoking status, for which the proportions of each category across the eight groups are outlined in Table 6. If among women, the most prevalent group is non-smokers, then for men, in all age categories except for the last one, the current smokers take the lead. It is also worth mentioning that the proportion of former smokers among older individuals is significantly higher for men, with the difference between sexes in the age group of 66 and older reaching 34%.

Table 6: Proportions of smoking status in percentage by age and gender groups.

Smoking status	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Non-smoker	62	60	66	89	43	37	33	42
Current	31	29	23	4	50	46	41	17
Former	7	11	11	7	7	17	26	41

Table 7: Average weights across smoking status by age and gender groups.

Smoking status	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Non-smoker	63.4	70.1	76.6	74.4	79.6	87.7	89.1	83.4
Current	64.9	70.9	73.3	68.8	78.7	86.0	84.4	78.3
Former	64.4	72.6	78.5	75.6	82.0	91.0	92.0	83.6

The average weights for the three categories of smoking classification are presented in Table 7. The table suggests that, typically, except for younger females, the

average weight is lower among current smokers compared to non-smokers. It can be also observed that this phenomenon seems to be most pronounced among the two older age groups. However, when it comes to the comparison between non-smokers and former smokers, the table shows that the latter weigh, on average, slightly more.

Secondly, from the section of unhealthy habits, alcohol consumption is considered. Table 8 gives an overview of its distribution in percentages across the eight age-gender groups. In contrast to smoking, current consumers consistently represent the dominant group in alcohol consumption. However, it has to be noted that among women, there is a higher proportion of non-consumers compared to men, especially in the oldest age bracket, where the difference is 24%. When it comes to former consumers, the trends are quite similar for both sexes, with a slight uptick in proportion occurring with increasing age.

Table 8: Proportions of alcohol consumption status in percentage by age and gender groups.

Drinking status	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Non-consumer	11	12	15	31	7	5	4	7
Current	85	83	79	60	91	91	87	79
Former	4	5	6	9	2	4	9	14

Table 9: Average weights across alcohol consumption status by age and gender groups.

Drinking status	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Never	62.8	70.9	77.3	73.7	77.6	84.2	86.7	82.4
Current	64.0	70.5	75.7	74.4	79.5	87.8	88.2	82.9
Former	65.5	72.0	77.5	74.8	78.6	83.6	86.3	81.2

The average weights based on alcohol consumption status across the eight groups have been summarised in Table 9. Although differences in average weights here are

even more marginal compared to previous variables, and thus not much of interest, it appears that for women, former consumers have the highest average weight, whereas for men, current consumers are the leading group.

The dataset included many other lifestyle-related variables, providing a detailed overview of all them would be too lengthy. Instead, a selection of the most important dietary and leisure activities are presented, while keeping their relevance in the following statistical analysis in mind.

Within the food categories discussed in this section, the first one is fish. The distribution of fish consumption in days per week among the eight groups is presented in Table 10. The key insight, visible from the table, is that the majority of individuals consume fish at a frequency of 1-2 days per week. The table also suggests that when it comes to very frequent consumption, i.e. 6-7 days per week, it is more prevalent among older age groups. However, total avoidance of fish is much more pronounced among younger age brackets, especially for individuals aged under 30, whose corresponding proportion is approximately 30% among both sexes.

Table 10: Proportions of fish consumption in days per week in percentage by age and gender groups.

Days	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
0 days	30	17	10	13	27	14	9	9
1-2 days	61	71	71	65	64	72	69	66
3-5 days	8	11	16	18	8	12	18	20
6-7 days	1	1	3	4	1	2	4	5

Table 11: Average weights across fish consumption in days per week by age and gender groups.

Days	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
0 days	63.7	69.9	74.8	73.6	77.6	85.2	85.3	81.4
1-2 days	64.1	70.7	75.8	74.2	79.8	87.6	87.8	81.9
3-5 days	63.7	71.0	77.5	74.9	81.6	89.4	89.9	85.3
6-7 days	64.5	70.5	78.0	74.5	80.3	89.0	88.8	84.1

Average weights across the four rates of fish consumption in days per week have been summarised in Table 11. The table illustrates one rather peculiar tendency, which is better observed for men. Namely, within the first three frequency categories, there is typically a slight positive relationship between the number of days and average weight. However, in the group of the highest frequency, i.e. 6-7 days per week, there is a slight decrease in average weight compared to the preceding category. Nevertheless, the individuals in the non-consumer group still have the lowest average weight.

In addition to fish, sweets are studied in greater detail, with the intention to also highlight the potential differences between salty and sugary products. Table 12 gives an overview of the consumption of sweets in days per week across the observed groups. As it turns out, the most prevalent frequency to consume sweets is typically 1-2 days. However, the occurrence of 3-5 days is also rather significant, especially among the younger individuals. The table also highlights a tendency that, with increasing age, the proportion of those not consuming sweets increases, reaching almost 25% among individuals aged 66 or older. Conversely, consumption of sweets at a rate of 6-7 days per week decreases with age. However, this is in comparison to the previous trend, much more subtle.

Table 12: Proportions of sweets consumption in days per week in percentage by age and gender groups.

Days	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
0 days	4	7	14	23	7	10	17	24
1-2 days	37	40	44	46	38	43	44	45
3-5 days	38	35	28	19	37	31	26	19
6-7 days	21	18	14	12	18	16	13	12

Moving on to the average weights across different rates of sweets consumption, an overview is provided in Table 13. While one might expect that consuming sweets more frequently leads to higher body weight, the table reveals the opposite trend,

with individuals who consume sweets almost every day of the week having the lowest average weight instead. However, it is unlikely that the relationship between sweets and weight is causal. For instance, it is possible that individuals who consume sweets more frequently also engage in more physical activity or maintain a more balanced diet overall.

Table 13: Average weights across sweets consumption in days per week by age and gender groups.

Days	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
0 days	65.3	72.7	79.4	76.8	80.8	91.1	90.5	85.9
1-2 days	65.3	72.2	76.9	74.6	80.3	88.8	87.9	82.4
3-5 days	63.7	69.6	74.8	72.5	78.9	86.1	87.3	80.5
6-7 days	61.8	68.2	72.9	70.7	77.5	84.5	86.2	80.3

Compotes and jams is another category of food that is generally classified as sweet products. However, in the questionnaire of the Estonian Biobank, they are regarded as separate entities. The consumption of compotes and jams in days per week across the eight groups has been summarised in Table 14. Compared to sweets, there are far more individuals who belong to the group of non-consumers. The table also illustrates the fact that consuming compotes and jams at least three days a week is quite uncommon, which is also a notable difference in terms of consumption patterns observed for sweets.

Table 14: Proportions of compotes and jams consumption in days per week in percentage by age and gender groups.

Days	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
0 days	49	42	41	35	42	39	38	32
1-2 days	41	45	45	45	46	48	47	45
3-5 days	8	10	11	14	10	10	11	16
6-7 days	2	3	3	6	2	3	4	7

When it comes to the average weight among the different rates of consumption of

compotes and jams in a week, the patterns observed were similar to what was seen for sweets, especially for women, where a higher frequency of consumption of the products is characterised by lower average weight. Thus, a more explicit overview of average weights for compotes and jams will not be provided.

In the following, insights into the most significant leisure activities are presented. In particular, the observed activities are speed walking and physical exercise different from walking, whose distributions are presented in Figure 2 and Figure 3, respectively.

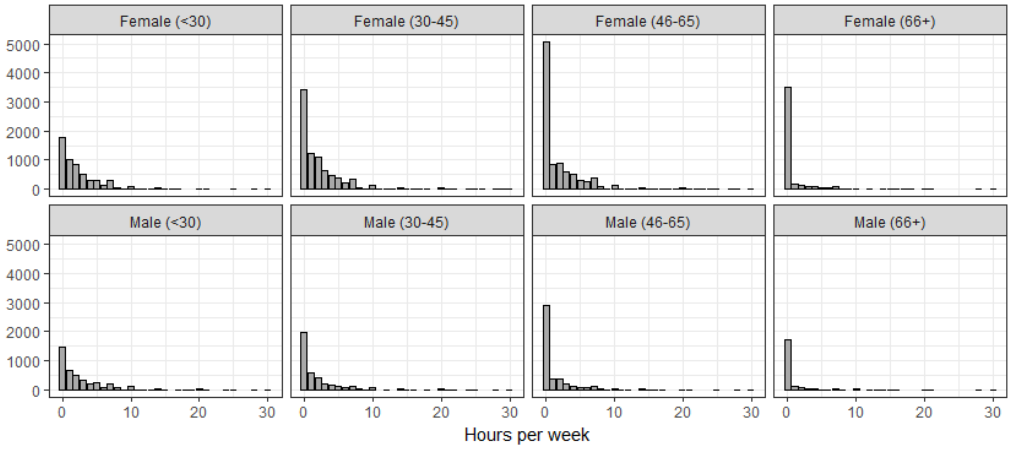

Figure 2: Distribution of speed walking in hours per week by gender and age groups.

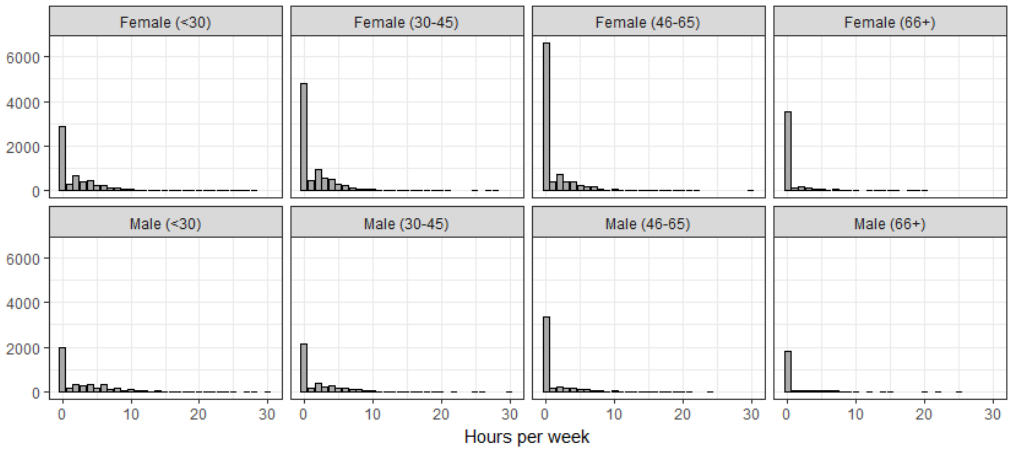

Figure 3: Distribution of physical exercise in hours per week by gender and age groups.

As depicted by the figures, the predominant group in both activities consists of those who do not engage in either of them. In fact, this phenomenon is also present in all the other activities, which will not be explicitly described here. What is also visible from the plots is that for speed walking, the frequencies decrease steadily, whereas for physical exercise, there is first a slight uptick at around 2-3 hours, and only then a decline. Lastly, it is worth pointing out that the majority of individuals' activity patterns fall within the first 10 hours. While exceptions exist, they are fairly rare.

The relationship between weight and the two leisure activities, speed walking and physical activity, is described through correlations in Table 15. Firstly, it is apparent that the associations are quite weak, which is also logical, due to weight being a very complex trait. The table also highlights that there is generally a marginal negative relationship between physical activity and weight, with the exceptions for physical exercise observed among women aged under 30 years and men across all age groups, where the corresponding correlations are positive but basically zero, especially for males.

Table 15: Correlations between weight and speed walking and physical exercise by age and gender groups.

Leisure activity	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Speed walking	-0.065	-0.061	-0.111	-0.074	-0.051	-0.099	-0.085	-0.029
Physical exercise	0.012	-0.054	-0.082	-0.067	0.005	0.007	0.003	0.008

The remaining part of this section is devoted to the genetic variants, for which the distributions, once again in percentages, are presented in Table 16 and Table 17 for females and males, respectively. Most notably, the tables illustrate that the distributions of the number of copies of the alternative allele are very similar between women and men, as well as across the age groups. This is quite reasonable, as one would expect the genetic structure of a human to remain consistent. However, there is some variation in the distribution of the proportions of having 0, 1, or

2 alternative alleles between the genetic variants. For variants 11:27684517:A:G, 1:177889480:A:G, 18:57829135:T:C and 6:50845490:A:G, the predominant group consists of individuals having a count of zero for the alternative allele, indicating its absence. For the remaining variants, excluding 2:632348:A:G, the most common group comprises individuals with one copy of the alternative allele.

Table 16: Proportions of the number of copies of the alternative allele across genetic variants in percentage among females by age groups.

Variant	**<30**			**30-45**			**46-65**			**66+**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
11:27684517:A:G	70	28	2	68	29	3	69	28	3	69	28	3
1:72751185:T:C	13	45	42	13	46	41	13	46	41	14	46	40
1:177889480:A:G	71	27	2	73	25	2	72	26	2	72	26	2
12:50247468:G:A	29	50	21	29	49	22	29	49	22	29	49	22
16:53803574:T:A	31	48	21	30	49	21	31	49	20	30	49	21
18:57829135:T:C	67	29	4	67	30	3	66	30	4	67	30	3
2:632348:A:G	3	29	68	3	29	68	3	28	69	3	29	68
2:25150296:A:G	32	49	19	32	49	19	32	49	19	32	48	20
4:45182527:A:G	33	49	18	33	48	19	33	49	18	34	48	18
6:50845490:A:G	56	38	6	56	37	7	57	37	6	56	38	6

Table 17: Proportions of the number of copies of the alternative allele across genetic variants in percentage among males by age groups.

Variant	**<30**			**30-45**			**46-65**			**66+**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
11:27684517:A:G	70	27	3	69	28	3	69	28	3	69	28	3
1:72751185:T:C	14	46	40	13	47	40	13	47	40	14	46	40
1:177889480:A:G	72	26	2	72	26	2	72	26	2	72	25	3
12:50247468:G:A	29	50	21	28	51	21	28	49	23	27	50	23
16:53803574:T:A	31	49	20	29	50	21	28	50	22	29	49	22
18:57829135:T:C	67	30	3	67	30	3	66	31	3	66	31	3
2:632348:A:G	3	29	68	3	29	68	2	31	67	3	29	68
2:25150296:A:G	31	49	20	32	50	18	32	49	19	34	47	19
4:45182527:A:G	32	49	19	35	46	19	33	49	18	33	49	18
6:50845490:A:G	55	39	6	57	36	7	56	38	6	55	39	6

Table 18: Average weights across the number of copies of the alternative allele in genetic variants among females by age groups.

Variant	**<30**			**30-45**			**46-65**			**66+**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
11:27684517:A:G	64.1	63.7	62.8	70.9	70.1	69.7	76.4	75.5	74.3	74.4	73.7	75.0
1:72751185:T:C	64.0	63.9	64.0	70.2	70.3	71.1	75.2	76.2	76.2	74.1	74.4	74.1
1:177889480:A:G	63.5	65.2	64.5	70.3	71.6	69.2	75.7	76.7	79.7	74.0	75.0	74.4
12:50247468:G:A	63.6	64.0	64.4	70.6	70.3	71.2	75.4	76.2	76.7	73.4	74.2	75.4
16:53803574:T:A	63.1	63.5	66.3	69.4	70.4	72.7	74.2	76.3	78.2	73.6	74.1	75.5
18:57829135:T:C	63.7	64.4	64.9	70.0	71.5	73.7	75.5	76.8	79.8	74.0	74.6	74.7
2:632348:A:G	63.3	62.8	64.4	67.7	69.8	71.0	75.0	75.3	76.4	73.1	73.7	74.5
2:25150296:A:G	63.3	64.0	65.0	70.3	70.7	70.8	75.9	76.0	76.5	74.1	74.4	74.1
4:45182527:A:G	63.8	64.1	63.9	70.3	70.6	71.0	75.3	76.1	77.3	74.2	74.6	73.2
6:50845490:A:G	63.7	64.0	66.0	70.1	71.3	71.2	75.7	76.4	77.2	74.0	74.6	74.3

Table 19: Average weights across the number of copies of the alternative allele in genetic variants among males by age groups.

Variant	**<30**			**30-45**			**46-65**			**66+**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
11:27684517:A:G	79.4	79.1	79.6	87.7	87.2	84.6	88.0	87.8	88.7	82.9	82.0	81.4
1:72751185:T:C	79.0	79.5	79.3	88.0	87.3	87.5	87.9	87.3	88.7	83.6	82.5	82.5
1:177889480:A:G	78.9	80.3	82.1	87.2	88.0	90.5	87.3	89.3	92.7	82.7	82.4	83.0
12:50247468:G:A	78.8	79.1	80.5	87.0	87.5	87.9	86.7	88.6	88.1	81.3	82.8	83.7
16:53803574:T:A	78.6	79.0	81.4	85.7	87.6	89.8	86.5	87.9	89.9	81.7	82.5	84.1
18:57829135:T:C	78.8	80.2	82.0	86.8	88.8	88.9	87.6	88.3	90.7	82.5	82.9	82.4
2:632348:A:G	77.8	78.1	80.0	86.4	86.7	87.9	87.9	87.4	88.2	81.8	81.9	83.0
2:25150296:A:G	79.4	79.5	78.8	87.6	87.7	86.8	87.8	87.9	88.3	82.8	82.6	82.3
4:45182527:A:G	78.9	79.4	80.0	86.9	87.7	88.0	87.9	87.9	88.1	82.3	82.7	82.9
6:50845490:A:G	79.4	79.3	79.0	87.4	87.4	88.2	87.5	88.5	88.4	82.4	82.9	82.6

Average weights across the number of copies of the alternative allele among different age groups for females and males have been summarised in Table 18 and Table 19, respectively. It appears that the individuals with at least one copy of the alternative allele generally weigh, on average, slightly more than those with zero copies, although the differences are marginal.

Additionally, it is worth mentioning that sometimes the relationship between the

number of copies of the alternative allele and weight is linear, whereas in other instances, this is not the case, illustrating the importance of considering these variables as categorical. For example, a slight positive linear relationship is observed for variant 16:53803574:T:A (FTO)[1] among both women and men. On the other hand, when considering variant 2:632348:A:G (TMEM18) for females aged under 30, the lowest average weight occurs among individuals having one copy of the alternative allele, while the highest is found among those having two copies.

[1]Now and henceforth, when discussing the effects of genetic variants on weight, the nearest gene to them will be indicated in parentheses.

5 Regression Analysis for Finding Weight Determinants

The aim of this section is to first give an overview of the steps taken to address the concern with drastically different group sizes that was introduced in Section 3. The central focus of the section is the presentation and discussion of the results of the analysis.

In order to make the eight groups comparable with each other, several steps were taken. Firstly, a random sampling technique was implemented. Specifically, the number of observations of the smallest group, men aged 66 or older, which comprised 2119 observations, served as the reference size. In other groups, random samples of 2119 observations were drawn 5 times. Random sampling was repeated several times (5 times) for each age-gender group to better understand the impact of the randomness component on estimated models. In particular, the concern is how it affects which variables will have significant effects in the models later on. If a certain variable really is related to body weight, we would like it to be stable in a sense within the samples.

Following the random sampling procedure, linear regression models for weight were constructed for each sample. For categorical variables the base classes were as follows: basic school or lower for education, non-consumers for both smoking and alcohol consumption, and absence of the alternative allele for genetic variants, i.e. the count of zero.

For fitting models, a backward stepwise regression approach based on p-values was employed. In practice, this was achieved with the help of package `olsrr` in R (Hebbali, A., 2024). In rare cases, a few manual adjustments were needed to further remove non-significant variables from the models. It is also worth pointing out that the significance level used for backward stepwise regression approach in this thesis is 0.07, which is less strict, thus facilitating the identification of associations

between weight and explanatory variables introduced in Section 3. This choice also aligns more closely with the aims of the thesis. The counts of statistically significant variables across the eight gender and age groups, hereafter referred to as intermediate results, are presented in Table 21 in Appendix.

It is worth mentioning that after examining the residuals of the models, where violations of the normality assumption were identified, logarithmic transformation was considered. However, since the improvements were marginal and the models are only used for detecting associations, as opposed to prediction, the idea of transformation was discarded to enhance the interpretability of the results.

Another noteworthy observation made at this stage was the fact that the models had very low R^2, thus explaining only a small portion of the variance in weight. More specifically, R^2 remained between 0.15 and 0.30. However, it was interesting that models fitted on samples of men generally had a slightly higher R^2. In particular, R^2 values less than 0.20 were only observed among samples of women, whereas for men, R^2 typically ranged from 0.25 to 0.30.

The next step involved selecting variables that were statistically significant in at least four models out of five within one group among those where samples were drawn, or significant among men aged 66 or older. For the categorical variables within five samples, in particular, this criterion required that the variable, taking into account all its levels, demonstrated statistical significance in at least four instances. In other words, the variable was kept in the final model at least four times. For the selected variables, the average effect size was calculated by finding the mean of the corresponding beta coefficients, except for the group of men over 66 years old, where effects were directly taken from the model.

It is important to note that the average effects based on the five samples were calculated only if a variable had a significant effect at least four times within a particular group. This criterion ensures greater stability, as a result making findings more reliable. Moreover, it must also be mentioned that for the selected categorical

variables, averages were calculated for all their levels, regardless of whether a particular level had a statistically significant effect or not. The central result of this thesis, comprising the average effect sizes on weight of various lifestyle and genetic factors, is presented in Table 20. In the following sections, both intermediate results and average effect sizes are discussed.

Table 20: Average effect sizes for statistically significant variables by gender and age groups. Effect sizes denoted by * are calculated based on 4 samples. For males in the age group of 66 and older, the effect sizes are directly taken from the model.

Variable	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
General information								
Age	0.44	0.51	0.22	-0.49	0.79	0.36	-0.13*	-0.49
Height	0.68	0.72	0.79	0.77	0.86	0.98	0.99	0.86
Education								
-Secondary education	-	-1.51	-1.96	-3.10	-	-	-	-
-Higher education	-	-3.80	-4.38	-4.79	-	-	-	-
Diet								
Potatoes	-	-	-	-0.32*	-	-0.42*	-	-0.37
Porridge, muesli, flakes	-	-	-	-0.38	-	-	-	-
Fish	-	-	0.71*	-	0.59	0.83	0.62	0.63
Meat	-	-	-	-	0.36*	-	0.49*	0.38
Meat products (sausage, frankfurters)	-	-	0.50	0.44*	-	0.45	-	-
Fresh fruits, berries	-	-	-	-	-	0.48*	0.37*	-
Compotes, jams	-	-0.58	-0.79	-0.37	-	-0.59	-0.79	-0.45
Sweets	-0.63	-0.66	-0.81	-0.61	-0.75	-1.07	-0.65	-0.64
Soft drinks	-	0.39*	-	-	-	-	-	-
Coffee	-	-	-0.66	-	-0.47*	-0.51*	-	-
Tea	-0.60*	-	-	-	-	-	-	-
Bread	0.59	0.68	-	-	-	-	-	-
White bread	-	-	-	-0.47*	-	-	-	-
Habits								
Smoking status								
-Current	-	-	-3.04	-6.57	-	-1.24	-4.84	-6.39
-Former	-	-	1.23	1.13	-	2.38	2.03	0.09
Spare time activities								
Slow walking	0.12*	0.20*	-	-	-	-	-	0.12
Moderate Walking	-	-	-	-	-	-	-	-0.10
Speed walking	-0.26	-0.28	-0.38	-0.52	-	-0.31	-0.36	-0.35

Continued on next page.

Variable	Female				Male			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Shopping	-	-	-	-0.31	0.33	-	-	-
Laundry	-	-	-	-	-	-1.20	-	-
Household repair	-	-	-	-	-	-	-	0.14
Physical exercise	-	-0.44	-0.60	-0.33	-	-	-0.35	-
Genetic variants								
11:27684517:A:G (1)	-	-	-1.61*	-	-	-1.46*	-	-
11:27684517:A:G (2)	-	-	-2.27*	-	-	-3.38*	-	-
1:177889480:A:G (1)	1.68	-	-	-	-	-	1.42*	-
1:177889480:A:G (2)	2.24	-	-	-	-	-	4.68*	-
12:50247468:G:A (1)	-	-	-	1.10*	0.51	-	-	1.18
12:50247468:G:A (2)	-	-	-	2.30*	2.01	-	-	2.18
16:53803574:T:A (1)	-0.01	0.75	2.44	-	0.10	1.32	1.00	0.67
16:53803574:T:A (2)	2.82	3.05	4.22	-	2.86	3.86	2.98	2.23
18:57829135:T:C (1)	-	2.04	1.79*	-	1.11*	-	-	-
18:57829135:T:C (2)	-	2.48	4.74*	-	3.54*	-	-	-
2:632348:A:G (1)	-	-	-	-	1.59	-	-	-
2:632348:A:G (2)	-	-	-	-	3.02	-	-	-
2:25150296:A:G (1)	0.93*	-	-	-	-	-	-	-
2:25150296:A:G (2)	2.23*	-	-	-	-	-	-	-
6:50845490:A:G (1)	0.62*	-	-	-	-	-	-	-
6:50845490:A:G (2)	2.81*	-	-	-	-	-	-	-

5.1 Significant Variables in Different Subsamples

In the first part of the discussion, emphasis is directed towards the intermediate results, which present the frequencies for statistically significant variables across the eight gender and age groups (see Table 21 in Appendix). Overall, the effects of age and height were most consistent, being most frequently statistically significant across the 8 age-gender groups. This phenomenon is logical, especially since the latter is one of the key determinants of weight. When it comes to diet, the most prevalent food items influencing weight were fish, particularly among men, along with compotes and jams, as well as sweets. It is worth mentioning that higher-calorie foods and drinks generally had more often significant effects on weight. This is also a logical outcome from a physiological perspective, as a higher amount of calories can disrupt the balance between energy intake and expenditure, potentially

leading to surplus and subsequent weight gain.

Continuing on the note of habits and preferences, smoking status also stood out, as it showed a significant effect on weight at least once in nearly all groups. On the other hand, the impact of alcohol consumption on weight, according to this analysis, is negligible, being statistically significant only a handful of times.

The category of leisure activities in terms of having significant effects on weight was dominated by physical exercise and speed walking, though the first was more frequently significant among females. Since both of them can be physically demanding and thus play a role in the usage of energy, this result can also be considered reasonable in the context of weight and its management.

When it comes to the genetic variants, the one most frequently associated with weight was 16:53803574:T:A (FTO). Considering that FTO is one of the prominent genes associated with obesity, as discussed in Section 3, this finding is also highly plausible. The effects of other variants were not as consistent, though it is worth mentioning 18:57829135:T:C (MC4R), which was identified as having a significant impact on weight in most groups at least once.

The analysis also revealed several gender-specific relationships between explanatory variables and weight, some of which also depend on age. One of the most interesting tendencies observed pertains to education. Specifically, education had a significant effect on weight among women, whereas for men its impact on weight was rarely identified. It is worth noting that for the former, significant differences occurred more frequently between the higher education and the lowest level, i.e. levels up to and including basic school.

Regarding the dietary variables, there were also a few intriguing differences between the sexes. It is interesting that for women, the effect of bread on weight was often significant, particularly among the younger age brackets, but for men, it was identified only once among individuals aged under 30 years old. Furthermore, white bread was more frequently significantly associated with weight among women as

well, although often not significant in at least 4 instances within an age group, thus excluding it from further analysis for the most part. On the other hand, among men, the impact of coffee was more pronounced compared to females. Interestingly, this pattern was also consistently observed among younger males. Moreover, it is worth pointing out that for men, certain staple foods, such as meat and fresh fruits and berries, showed more frequent significant effects on weight in comparison to females.

Although the effect of smoking status was identified for both females and males, it was the latter, among whom the impact of the former smokers' class was slightly more pronounced. It is plausible that this can be partly explained by the higher proportion of former smokers among men.

When considering leisure activities, the significant effects of gardening were more prevalent among women, although the activity never reached the threshold of being significant in at least four subsample models in any age group, whereas for men it was somewhat surprisingly simple chores, such as cleaning and laundry, that showed significant associations with weight more often. However, the latter ones were also significant in less than four instances quite often, making them not so intriguing for the second stage of the analysis. One other rather peculiar finding relates to shopping. In particular, the impact of shopping on weight was highly significant among women aged 66 or older, while for men, it was significant most consistently among the youngest age bracket. However, this outcome is quite difficult to interpret and may suggest underlying behavioural patterns that shopping may embody.

Lastly, genetic variants also have to be considered in the context of gender-specific associations. Although here significant effects were not as strongly dependent on the gender, it was still intriguing that variants 1:72751185:T:C (NEGR1) and 2:25150296:A:G (ADCY3), were identified among women, albeit often less frequently than four times, but never among the opposite sex.

On top of discussing significant effects, it is also worth mentioning the variables, whose effects either remained entirely non-significant or were significant too few times to be included in the second stage. Regarding the former, there was only one such case with boiled vegetables, which was always absent from the final models. The other variables, which rarely had significant effect on weight are the following: rice and pasta, milk products, fresh vegetables, eggs, alcohol consumption status, cooking, cleaning, gardening, childcare, elderly care, as well as two genetic variants 1:72751185:T:C (NEGR1) and 4:45182527:A:G (GNPDA2).

5.2 Average Effect Sizes

In the second part of the discussion, the focus is shifted to the average effects (see Table 20). Given the diversity in the nature of the variables, the discussion on average effects will be done similarly as before, by considering them in more general groups. Starting with the most consistent variables, height and age, the analysis showed that the former has a positive relationship with weight, which is a highly expected result. However, it was interesting that a one-unit change in height has a greater effect among males compared to females. When it comes to age, it was found that it initially maintains a positive relationship with weight, but after reaching certain age groups, 46-65 and over 66 for males and females, respectively, the direction reverses.

Since education exhibited significant effects on weight pretty much only among women, average effects for men will not be considered. Based on the results, the relationship between education level and weight is negative, meaning a higher level of education is associated with a lower average weight, consistent with what was already discussed in Section 4. In light of the effect sizes, it was observed that as age increases, the impact of education on weight becomes stronger, as the corresponding coefficients increased in magnitude.

Direct comparisons in the section of variables related to diet are much more chal-

lenging to make, as the significant effects were not as consistent across the groups. However, it is possible to categorise foods and drinks into two groups, based on how they effect weight. Firstly, there are those that showed a positive relationship with weight, such as fish, meat and meat products, fresh fruits and berries, soft drinks and bread. When it comes to the other group of variables that were uniformly negatively associated with weight, there are porridge, muesli and flakes, potatoes, compotes and jams, sweets, coffee and tea, as well as white bread. What is also worth mentioning is that among the category of foods and drinks, the strongest effects on weight in magnitude were most often observed for sweets, occasionally for fish, compotes and jams, or bread.

It is difficult to find a reasonable explanation for why some of the unhealthier foods or drinks tend to have decreasing effects on weight. One potential cause for this phenomenon is that overall healthier individuals, who generally weigh less, may consume them more frequently, as a result drawing out these associations and making them statistically significant. On the other hand, more frequent consumption of staple foods, such as fish or meat, which had an increasing effect on weight, may be more prevalent among overweight or obese individuals.

Transitioning to the section of habits, the only variable discussed at this stage is smoking status, the effects of which on weight are more or less known from Section 4. Specifically, being an active smoker tends to have a decreasing effect on weight in comparison to non-smokers, while for former smokers, an opposite trend was observed. However, it is still worth pointing out that the impact of being an active smoker increases in strength with age. On the other hand, the effect of being a former smoker on weight, especially among men, seems to diminish in magnitude with age, being no longer statistically significant in the oldest age bracket.

The next section of interest is leisure activities, where, similarly to variables related to diet, two opposing groups can be formed. In particular, slow walking and household repair, albeit not significant in most groups, showed a positive relation-

ship with weight, whereas laundry, moderate and speed walking, as well as physical exercise, were consistently negatively associated with it. Once again, it is difficult to interpret why, slow walking, for example, has a weight-increasing effect, while a more vigorous activity results in an opposite outcome. However, it is possible that the intensity of an activity may play a role, as the former tends to be more relaxed, making it more suitable for overweight or obese individuals. As a consequence, a statistically significant association may have emerged.

It is also interesting that the strongest average effect size among leisure activities was observed for laundry among males aged 30-45 years, which is, similarly to many other phenomena in this analysis, difficult to interpret. The remaining activity, shopping, showed a positive relationship with weight for men and a negative for the opposite gender, although only in the youngest and oldest age brackets, respectively, as previously mentioned. Moreover, the absolute effect sizes were quite similar between these groups.

Concluding the discussion with the genetic variants, the average effect sizes for a given variant, were generally uniform in terms of sign, although often varying in magnitude. As mentioned previously, the effect of 16:53803574:T:A (FTO) was the most consistent across the groups. However, it was not the variant with the strongest average effect on weight in absolute terms. Instead, it was observed for variant 18:57829135:T:C (MC4R), with two copies of the alternative allele, among women aged 46-65 years, closely followed by 1:177889480:A:G (SEC16B), also with two copies, among men in the same age group. It is also worth pointing out, that the effect of having two alleles, is generally stronger compared to having only one.

When it comes to signs, all variants apart from 11:27684517:A:G (BDNF), tend to have a positive relationship with weight. Although, 16:53803574:T:A (FTO) with one copy of the alternative allele among women in the youngest age group was an exception in this regard, its effect was never significant, being on average essentially zero.

Lastly, it is interesting that the highest number of genetic variants having a stable significant effect on weight (i.e., being significant in 4 or all 5 subsample models) was observed among individuals in the youngest age bracket. Although once again difficult to explain, it may stem from the fact that in the earlier stages of life, an individual has not yet been exposed to an obesogenic environment for that long, and thus, the impact of genetic factors may be more pronounced.

Conclusions

The aim of this bachelor's thesis was to investigate relationships between weight and a range of lifestyle and genetic factors. In this study, the genetic factors under consideration were single nucleotide polymorphisms. The analysis was performed on data originating from the Estonian Biobank and focused on a cohort of slightly over 40 000 gene donors. On top of identifying associations, emphasis was also placed on quantifying them by finding the average effects of relevant factors on weight. The analysis was conducted separately across eight groups, by considering both females and males in four age categories: individuals under 30, 30-45, 46-65, and those aged 66 or older.

Based on the analysis, the most consistently significant effects on weight from dietary habits were observed for consumption of fish (albeit more so among males), compotes and jams, as well as sweets. It is worth mentioning that the effect of the latter was generally the strongest among foods and drinks. However, explaining the directions of relationships was not always straightforward, as illustrated by sweets, where higher consumption had a decreasing effect on weight, highlighting the need for cautious interpretation.

Continuing on the note of lifestyle factors, the analysis also revealed that smoking status has a significant effect on weight, whereas the impact of alcohol consumption is negligible. When it comes to spare time activities, the most frequent significant effects on weight were observed for physical exercise and speed walking. Although here, the directions were logical, as both showed a negative relationship with weight.

Among the ten studied genetic variants, the one most frequently associated with weight was 16:53803574:T:A, the variant in proximity to the FTO gene. It was also observed that the effect of having two copies of the alternative allele was generally stronger compared to having only one.

The analysis also uncovered a few gender and age-specific patterns. One of the

most notable findings in this regard is related to education, for which a significant effect on weight was often identified among women, but rarely for the opposite sex. For females, a higher level of education was associated with lower average weight. On the other hand, an age-specific phenomenon was observed for genetic variants. Specifically, the highest number of variants having a stable significant effect on weight was identified among individuals in the youngest age category, i.e. those aged under 30 years, distinguishing them from the subsequent age groups.

The author would also like to highlight certain limitations present in conducting such studies, in particular exploring the associations between lifestyle factors and weight or BMI, using data from the Estonian Biobank. Due to their focus on the frequency of food consumption, rather than the quantity for most products, assessing the effects of food items is challenging. For example, if one individual consumes unhealthy products, such as sweets, daily but in small quantities, while another consumes large amounts but only two days per week, it is not evident from the data whose habits are actually more concerning. As a result, this can lead to surprising outcomes, as demonstrated in this study as well. But it is also leisure activities that have to be treated with caution since the causal effects were not under consideration in the analysis. For this reason, some of the significant associations may be more strongly influenced by the underlying behavioural patterns associated with them, rather than solely by the activities themselves.

In addition, the author would like to suggest a potential further development for conducting similar studies. The idea involves analysing effects of different lifestyle and genetic factors on weight among three conditions, normal weight, overweight and obese, separately. By already taking into account an individual's condition, the analysis could identify more relevant associations.

References

Adams, J. (2020). "Addressing socioeconomic inequalities in obesity: Democratising access to resources for achieving and maintaining a healthy weight". In: *PLOS Medicine* 17.7, e1003243. URL: https://doi.org/10.1371/journal.pmed.1003243.

Albuquerque, D., C. Nóbrega, L. Manco and C. Padez (2017). "The contribution of genetics and environment to obesity". In: *British Medical Bulletin* 123.1, pp. 159–173. URL: https://doi.org/10.1093/bmb/ldx022.

Chatterjee, S. and A. S. Hadi (2006). *Regression Analysis by Example*. 4th ed. Wiley.

GeneCards (2024a). *ADCY3 Gene - Adenylate Cyclase 3*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=ADCY3&keywords=ADCY3 (visited on 20/04/2024).

GeneCards (2024b). *BCDIN3D Gene - BCDIN3 Domain Containing RNA Methyltransferase*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=BCDIN3D&keywords=BCDIN3D (visited on 20/04/2024).

GeneCards (2024c). *BDNF Gene - Brain Derived Neurotrophic Factor*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=BDNF (visited on 20/04/2024).

GeneCards (2024d). *FTO Gene - FTO Alpha-Ketoglutarate Dependent Dioxygenase*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=FTO&keywords=FTO (visited on 20/04/2024).

GeneCards (2024e). *GNPDA2 Gene - Glucosamine-6-Phosphate Deaminase 2*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=GNPDA2&keywords=GNPDA2 (visited on 20/04/2024).

GeneCards (2024f). *MC4R Gene - Melanocortin 4 Receptor*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MC4R&keywords=rs6567160 (visited on 20/04/2024).

GeneCards (2024g). *NEGR1 Gene - Neuronal Growth Regulator 1*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=NEGR1&keywords=NEGR1 (visited on 20/04/2024).

GeneCards (2024h). *SEC16B Gene - SEC16 Homolog B, Endoplasmic Reticulum Export Factor*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SEC16B&keywords=SEC16B (visited on 20/04/2024).

GeneCards (2024i). *TFAP2B Gene - Transcription Factor AP-2 Beta*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=TFAP2B&keywords=rs2207139 (visited on 20/04/2024).

GeneCards (2024j). *TMEM18 Gene - Transmembrane Protein 18*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=TMEM18&keywords=TMEM18 (visited on 20/04/2024).

Hebbali, A. (2024). *Package 'olsrr'*. URL: https://cran.r-project.org/web/packages/olsrr/olsrr.pdf (visited on 20/04/2024).

Institute of Genomics (2021). *Estonian Biobank*. URL: https://genomics.ut.ee/en/content/estonian-biobank (visited on 23/03/2024).

Locke, A. and Kahali, B. and Berndt, S. et al. (2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.2, pp. 197–206. URL: https://doi.org/10.1038/nature14177.

Loos, R.J.F. and G.S.H. Yeo (2022). "The genetics of obesity: from discovery to biology". In: *Nature Reviews Genetics* 23.2, pp. 120–133. URL: https://doi.org/10.1038/s41576-021-00414-z.

MalaCards (2024a). *Growth Retardation, Developmental Delay, and Facial Dysmorphism (GDFD)*. URL: https://www.malacards.org/card/

growth_retardation_developmental_delay_and_facial_dysmorphism (visited on 20/04/2024).

MalaCards (2024b). *Niemann-Pick Disease (NPD)*. URL: https://www.malacards.org/card/niemann_pick_disease (visited on 20/04/2024).

Ministry of Social Affairs (2023). *Toitumise ja liikumise roheline raamat*. URL: https://www.sm.ee/toitumise-ja-liikumise-roheline-raamat (visited on 01/05/2024).

National Cancer Institute (2012). *Snp*. URL: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/snp (visited on 03/05/2024).

National Human Genome Research Institute (2024). *Human genome reference sequence*. URL: https://www.genome.gov/genetics-glossary/Human-Genome-Reference-Sequence (visited on 03/05/2024).

OECD and European Observatory on Health Systems and Policies (2023). *Estonia: Country Health Profile 2023*. OECD Publishing. URL: https://doi.org/10.1787/bc733713-en.

Reile, R., A. Baburin, T. Veideman and M. Leinsalu (2020). "Long-term trends in the body mass index and obesity risk in Estonia: an age–period–cohort approach". In: *International Journal of Public Health* 65.6, pp. 859–869. URL: https://doi-org.ezproxy.utlib.ut.ee/10.1007/s00038-020-01447-7.

Simon, P. H. G., M.-P. Sylvestre, J. Tremblay and P. Hamet (2016). "Key Considerations and Methods in the Study of Gene–Environment Interactions". In: *American Journal of Hypertension* 29.8, pp. 891–899. URL: https://doi.org/10.1093/ajh/hpw021.

Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré and D. Meyre (2019). "Benefits and limitations of genome-wide association studies". In: *Nature*

Reviews Genetics 20.8, pp. 467–484. URL: https://doi.org/10.1038/s41576-019-0127-1.

World Health Organization (2024). *Obesity and overweight*. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Worldwide%20adult%20obesity%20has%20more,16%25%20were%20living%20with%20obesity. (visited on 04/05/2024).

Appendix. Intermediate Results

Table 21: Counts of statistically significant variables by gender and age groups. Significance of a variable in the 66+ age group for males is indicated by "yes". Non-significant variables are denoted by dash. Counts for categorical variables are reported separately for each level.

Variable	**Female**				**Male**			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
General information								
Age	5	5	5	5	5	5	4	yes
Height	5	5	5	5	5	5	5	yes
Education								
-Secondary education	1	1	3	5	2	-	-	-
-Higher education	2	5	5	5	-	-	-	-
Diet								
Potatoes	-	1	1	4	-	4	1	yes
Rice, pasta	-	3	-	-	1	-	-	-
Porridge, muesli, flakes	-	1	2	5	-	2	-	-
Milk products	2	-	-	-	-	-	1	-
Fish	-	2	4	3	5	5	5	yes
Meat	1	-	-	3	4	-	4	yes
Meat products (sausage, frankfurters)	1	-	5	4	-	5	-	-
Fresh vegetables	1	2	1	-	1	1	-	-
Boiled vegetables	-	-	-	-	-	-	-	-
Fresh fruits, berries	-	-	-	1	1	4	4	-
Compotes, jams	-	5	5	5	-	5	5	yes
Sweets	5	5	5	5	5	5	5	yes
Soft drinks	2	4	2	-	3	2	3	-
Eggs	-	-	1	1	-	3	2	-
Coffee	1	-	5	-	4	4	2	-
Tea	4	-	-	-	3	-	1	-
Bread	5	5	2	1	1	-	-	-
White bread	1	2	1	4	-	-	2	-
Habits								
Smoking status								
-Current	2	-	5	5	-	2	5	yes
-Former	-	2	2	-	-	5	4	-
Alcohol consumption status								
-Current	-	-	-	3	-	1	-	-
-Former	1	-	-	-	-	-	-	-

Continued on next page.

Variable	Female				Male			
	<30	**30-45**	**46-65**	**66+**	**<30**	**30-45**	**46-65**	**66+**
Spare time activities								
Slow walking	4	4	1	-	-	3	2	yes
Moderate walking	-	1	1	-	1	1	-	yes
Speed walking	5	5	5	5	3	5	5	yes
Cooking	-	-	2	-	2	-	2	-
Shopping	-	-	1	5	5	2	2	-
Cleaning	1	-	1	1	3	-	3	-
Laundry	-	-	1	-	-	5	2	-
Childcare	-	-	2	2	1	-	-	-
Elderly care	-	-	1	-	-	1	-	-
Gardening	3	3	-	2	-	1	-	-
Household repair	2	2	1	-	-	-	-	yes
Physical exercise	-	5	5	5	-	-	5	-
Genetic variants								
11:27684517:A:G (1)	-	1	4	-	-	3	-	-
11:27684517:A:G (2)	-	-	2	-	-	2	-	-
1:72751185:T:C (1)	2	1	-	-	-	-	-	-
1:72751185:T:C (2)	2	2	-	-	-	-	-	-
1:177889480:A:G (1)	4	1	1	-	3	-	2	-
1:177889480:A:G (2)	2	-	-	-	2	-	3	-
12:50247468:G:A (1)	-	-	-	2	1	-	2	-
12:50247468:G:A (2)	-	-	1	4	5	-	1	yes
16:53803574:T:A (1)	-	1	5	-	-	2	1	-
16:53803574:T:A (2)	5	5	5	2	5	5	5	yes
18:57829135:T:C (1)	-	5	3	1	3	1	-	-
18:57829135:T:C (2)	-	2	3	-	3	-	2	-
2:632348:A:G (1)	-	1	1	-	-	-	-	-
2:632348:A:G (2)	1	1	1	1	5	1	-	-
2:25150296:A:G (1)	1	1	2	-	-	-	-	-
2:25150296:A:G (2)	4	-	-	-	-	-	-	-
4:45182527:A:G (1)	-	-	-	-	-	1	-	-
4:45182527:A:G (2)	-	-	2	-	-	2	-	-
6:50845490:A:G (1)	1	2	-	1	-	-	1	-
6:50845490:A:G (2)	4	1	-	-	-	-	-	-

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Katrin Kulberg,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis, Relationships Between Genetic and Lifestyle Factors and Weight in the Estonian Biobank, supervised by Jon Anders Eriksson and Kristi Kuljus.

2. I grant the University of Tartu a permit to make the work specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Katrin Kulberg
14/05/2024