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

Liis Hiie

# Human Metabolic Pathways as Predictors for Hypertension Based on Estonian Biobank Data

Mathematics and Statistics Curriculum

Mathematical Statistics

Master's Thesis (30 ECTS)

Supervisors: Jaanika Kronberg, PhD
Prof. Krista Fischer, PhD

Tartu 2023

# INIMESE AINEVAHETUSRAJAD KUI KÕRGVERERÕHKTÕVE RISKITEGURID EESTI GEENIVARAMU ANDMETE PÕHJAL

Magistritöö

Liis Hiie

**Lühikokkuvõte**

Magistritöö eesmärk on leida statistiliselt olulisi inimese ainevahetusradu kõrgvererõhktõve ennustamisel Eesti Geenivaramu andmete põhjal. Kõrgvererõhktõbi on kõige olulisem ennetatav riskifaktor nii südame-veresoonkonna haiguste, üldhaigestumuse kui ka suremuse puhul. Kõrgvererõhktõve ja metaboliitide vaheliste seoste uurimine aitab leida olulisi ainevahetusradu ja biomarkereid haiguse varajaseks diagnoosiks. Ulatuslikust massispektromeetria andmestikust pärit metaboliidid kaardistatakse ainevahetusradadele. Igal ainevahetusrajal sooritatakse peakomponentanalüüs ning valitakse esimesed kolm peakomponenti. Kõrge korrelatsiooniga komponendid eemaldatakse korrelatsioonanalüüsi tulemusena. Kõrgvererõhktõve ennustamiseks olulised komponendid leitakse Coxi võrdeliste riskide mudelite abil paremalt tsenseeritud ja vasakult tõkestatud andmetel, kasutades vanust ajaskaalana. Töö tulemusena leiti, et nii kehamassiindeks kui süsiniku ainevahetusrada aitavad kõrgvererõhktõve tekkimist ennustada enne diagnoosi. Tulemused valideeriti varasemate uuringute põhjal.

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

**Märksõnad:** Peakomponentanalüüs, elukestusanalüüs, kõrgvererõhktõbi, metaboloomika.

**HUMAN METABOLIC PATHWAYS AS PREDICTORS FOR HYPERTENSION BASED ON ESTONIAN BIOBANK DATA**

Master's Thesis

Liis Hiie

**Abstract**

The aim of this master's thesis is to identify relevant human metabolic pathways for incident hypertension prediction in the Estonian Biobank. Hypertension is the most important preventable risk factor for cardiovascular disease, as well as mortality and all-cause morbidity. Studying the relationship between hypertension and metabolites can reveal important pathways and biomarkers for early diagnosis. Metabolites from an extensive mass spectrometry metabolomics dataset are mapped to human metabolic pathways. Principal component analysis is performed on each pathway and the first three principal components are chosen. Highly correlated components are removed as a result of the correlation analysis. Relevant pathway components for the prediction of incident hypertension are found using Cox proportional hazards models on right censored and left truncated data with age as time scale. As a result, it was found that both BMI and carbon metabolism pathway can help predict incident hypertension before the diagnosis. The results were validated against the previous research.

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

**Key Words:** Principal component analysis, survival analysis, hypertension, metabolomics.

# Contents

# Introduction

Hypertension is a chronic medical condition affecting millions of people worldwide. It is also the most important preventable risk factor for cardiovascular disease, mortality and all-cause morbidity. (Oparil et al., 2018) Hypertension has been researched for years using genetics, physiology and immunology. In addition to this, there has been an increased interest in using metabolomics recently. Studying connections between hypertension and metabolites can provide valuable insight into the underlying mechanisms of the disease. Furthermore, it can identify the biomarkers relevant in the development of hypertension for earlier diagnosis. (Chakraborty et al., 2020)

The aim of this Master's thesis is to find the relevant human metabolomic pathways for predicting incident hypertension in the Estonian Biobank. From the extensive mass spectrometry metabolomics dataset of 1505 variables, 236 metabolites are mapped to 35 human metabolomic pathways. Principal component analysis is performed on the pathways and first three principal components are chosen from each. Correlation analysis is performed to remove highly correlated components. The samples belong to 999 pseudonymised participants that have filled questionnaires for background data and are also linked periodically to electronic health records. Incident hypertension is modelled using Cox proportional hazards models on right censored and left truncated data with age as time scale. First, a model is fitted with sex, BMI, smoking status, education class, region type, and time of day as predictors. Irrelevant variables are removed using backward stepwise elimination to obtain the baseline model. Then, relevant pathway components for predicting incident hypertension are found using the forward stepwise selection. The results are validated against the prior research. The data manipulation and statistical analysis is conducted using R statistics software.

This thesis covers all of these topics in the following structure. The first chapter gives an overview of hypertension and metabolomics. The second chapter describes the data. The third chapter addresses the methodology, including principal component analysis, survival analysis, and the multiple testing problem. The fourth chapter presents the compilation of

the pathway components and survival modelling. The fifth chapter discusses the findings, validates them against previous research, and lists the strengths and limitations of this thesis.

# 1 Hypertension and Metabolomics

Systemic arterial hypertension (commonly known and referred to as hypertension) is persistently high blood pressure in the systemic arteries. Blood pressure means the ratio of systolic blood pressure (pressure on the arterial walls when the heart contracts) and diastolic blood pressure (pressure on the arterial walls when the heart relaxes). The majority of people with hypertension have essential (also known as primary) hypertension with ICD-10 (International Classification of Diseases 10th Revision) code I10. This is defined as persistently elevated systolic blood pressure over 140 mmHg and persistently elevated diastolic pressure over 90 mmHg. (Oparil et al., 2018) There are also other hypertensive diseases: hypertensive heart disease with ICD-10 code I11, hypertensive renal disease with ICD-10 code I12, hypertensive heart and renal disease with ICD-10 code I13 (World Health Organization, 2023).

Hypertension is the most important preventable risk factor for cardiovascular diseases, as well as for mortality and all-cause morbidity. Less than half of people with hypertension are aware of their condition, many are aware but not treated. This is unfortunate as both lifestyle changes (dietary modifications and increased physical activity) and pharmacological therapy are effective in preventing hypertension. (Oparil et al., 2018)

Hypertension has been researched for years using various approaches, for example genetics, physiology and immunology. In the recent years, there has been growing interest in using metabolomics to better understand hypertension and the biomarkers relevant in its development. (Chakraborty et al., 2020)

Metabolomics is the study of biological systems and the prediction of their behaviour through the profiling of metabolites. Two main technologies used for analysing metabolite profiles are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). (Dona et al., 2016) Metabolites are intermediates and products of metabolism, which aggregates all the biochemical reactions in the organism (Thirumurugan et al., 2018).

Buergel et al. (2022) studied metabolomic profiles derived from NMR spectroscopy and their predictive power for 24 common conditions. They found that metabolomic profiles are associated with incident event rates in all of the conditions, except breast cancer. Hypertension was not included in the study but several cardiovascular diseases, such as coronary heart disease and heart failure, showed significant results.

# 2 Data

The data used in this thesis was obtained from the Estonian Biobank (data release R26, ethics approval 234T-12 "Omics for Health"). Multiple datasets were manipulated to put together necessary information using the R statistics software (The R Foundation, 2023). The samples were collected from the Estonian Biobank cohort (Leitsalu et al., 2015). The data includes mass spectrometry metabolomics data generated by Metabolon (Metabolon, Inc, 2023), combined background data from questionnaires filled by participants and electronic health records linked periodically to Estonian Biobank participants. Further details about the data are explained in this chapter.

## 2.1 Metabolomics Data

The first and most important dataset used was the batch-normalised metabolomics data from mass spectrometry, generated by Metabolon and provided by prof. Tõnu Esko. The data consists of 999 observations. The plasma samples used are from population-based participants of Estonian Biobank cohort with different loss of function mutations. The plasma samples were originally selected for a different project (Yu et al., 2023).

In the dataset of 999 observations, there are some that can't be used further in the analysis. One sample is provided without the sample code and two samples without the sample date, these samples were left out of the analysis. There are also seven individuals who have given the sample twice. For the analysis, the sample with the earliest known date was chosen. Therefore, a total of 989 samples can be used for further analysis.

The dataset includes 1505 variable metabolites. Each of these has a chemical ID, an unique identifier. There is chemical annotation available including respective biochemical classes, chemical names and KEGG codes. KEGG code is an alphanumeric compound identifier from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa Laboratories, 2023b). The proportion of different biochemical classes in the initial dataset is shown in Table 1.

Table 1: Number of metabolites in the dataset by biochemical classes.

| Biochemical class | Number of original metabolites | Number of metabolites with less than 20% missingness |
|---|---|---|
| Amino Acid | 223 | 196 |
| Carbohydrate | 26 | 23 |
| Cofactors and Vitamins | 36 | 29 |
| Energy | 9 | 9 |
| Lipid | 484 | 402 |
| Nucleotide | 42 | 37 |
| Partially Characterized Molecules | 28 | 20 |
| Peptide | 35 | 25 |
| Xenobiotics | 294 | 109 |
| Not known | 328 | 205 |
| Total | 1505 | 1055 |

A metabolite can have missing values in the dataset for three main reasons: it's not present in the individual, it's below the detection threshold or for other technical reasons (Reinhold et al., 2019). In this thesis, the main reason for missing data is that the value is below the detection threshold. Xenobiotics, metabolites not part of human metabolic pathways, can also have missing values because they are not present in the individual. Therefore the values are missing not at random. If a metabolite has too many missing values, it's considered best to leave it out of the analysis. In this case, if a metabolite had more than 20% missingness, it was removed from the analysis. The threshold was chosen as optimal because it leaves a relatively clean dataset where imputation and normalisation are expected to work well. There were 450 metabolites that had more than 20% of missing values and were removed from the analysis. This leaves 1055 metabolites. The proportion of different biochemical classes in the remaining metabolites is again shown in Table 1.

Remaining metabolites in the dataset were imputed using half of the minimum value in

the given variable. There are different methods available for imputation but as the missing values are expected to be missing not as random, the chosen method is appropriate (Reinhold et al., 2019). The imputed data was then log transformed using the natural log. It's a standard and recommended procedure while dealing with metabolomics data as it's typically right-skewed (ibid.).

## 2.2 Hypertension Data

The cases of hypertension in this paper are defined according to advice from Anu Reigo, the senior specialist of medical genomics in the Estonian Biobank. Database queries based on definitions were performed by Jaanika Kronberg, resulting in the following classes: cases 1, cases 2, cases 3, suspicious 1, suspicious 2, suspicious 3, exclude 1, exclude 2, exclude 3.

Exclude 1 includes people with no DNA sample, exclude 2 includes people with no information from electronic health records linking but some answers from the questionnaire, exclude 3 includes people with no information for any of the diseases. These people should be excluded from the analysis as they cannot be used with certainty with regards to hypertension. In this case, there were no overlapping participants in the metabolomics dataset and exclude datasets.

Suspicious 1 includes people that have received only one diagnosis of hypertension (ICD-10 codes I10, I11, I12 or I13) but no other signal of the disease. Suspicious 2 includes people with diagnoses of hypertensive disorders in pregnancy, childbirth and the puerperium (ICD-10 codes O10, O11, O13, O14, O15, O16), elevated blood-pressure reading without diagnosis of hypertension (ICD-10 code R03.0), background retinopathy and retinal vascular changes (ICD-10 code H35.0), hypertensive encephalopathy (ICD-10 code I67.4) and secondary hypertension (ICD-10 code I15). Suspicious 3 includes people that have objective measurements of systolic blood pressure over 140 mmHg or diastolic blood pressure over 90 mmHg but no other proof of hypertension. Suspicious 4 includes people with no information about hypertension from linked electronic health records but with

the disease mentioned in their questionnaire. These participants do not meet the criteria for cases of hypertension and at the same time are suspicious enough to not be considered clean from the disease. In this case, there were 77 people with the suspicion of hypertension. Therefore they were neither included in cases nor controls and were excluded from further analysis. A dataset of 912 participants remains.

Cases 1 includes people who have received two diagnoses of hypertension (I10, I11, I12 or I13) within a 6 months period. Cases 2 includes people who have gotten a diagnosis of hypertension (I10, I11, I12 or I13) and a prescription of medicine against hypertension (ATC codes C02*, C03*, C04*, C07*, C08* or C09*). Cases 3 includes people who have gotten a diagnosis of hypertension (I10, I11, I12 or I13) and have died with a reason of death being hypertension (I10, I11, I12 or I13). These people are defined as participants with hypertension at the end of the follow-up period on 31.12.2021.

These cases are divided into two categories: incident cases of hypertension and prevalent cases of hypertension. Prevalent case of hypertension means that the first diagnosis of hypertension occurred before the sample date. Incident case of hypertension means that the first diagnosis of hypertension occurred after the sample date. In this dataset, there are 443 people with a confirmed diagnosis of hypertension at the end of the follow-up period, 342 are prevalent and 101 incident cases of hypertension. The other 469 participants are defined as controls for hypertension because they are clean from the disease at the end of the follow-up period on 31.12.2021.

## 2.3 Questionnaire Data

Each participant in the Estonian Biobank cohort fills a questionnaire when first joining and occasionally also later in time when participating in follow-up studies. One important part of data in this thesis is the background data from questionnaires. It was selected and combined from different datasets and recoded according to necessity. The final variables are listed below:

- Sex is a categorical variable with two possible values: male and female. Sex is included because men are generally known to have higher blood pressure than women (Sandberg and Ji, 2012).

- Age is a numeric variable that reflects the age of the participant at the given time. Hypertension is more common among older age groups (Buford, 2016).

- Body mass index (BMI) is a numeric variable that reflects the BMI of the participant at the closest possible time to the sample. BMI is calculated by a formula which divides the person's weight in kilograms with the square of their height in metres. BMI is included because it is known as an independent risk factor for hypertension (Landi et al., 2018).

- Smoking status is a categorical variable with three possible values: current, former and never. This reflects the most recent available smoking status of the participant. It is included as smoking is a cardiovascular risk factor and often linked to hypertension (Virdis et al., 2010).

- Education class is a categorical variable with three possible values: low, intermediate and high. Low education group includes people both without primary education and with primary education, also with basic education and with vocational education based on basic education. Intermediate education group includes people with upper secondary education and vocational education based on upper secondary education. High education group includes people with Bachelor's degrees, Master's degrees and Doctorate degrees or equal diplomas. This reflects the most recent available education status reported by the participant. Education is included as it can guide healthy behaviours throughout an entire lifetime and therefore improve life expectancy. It has also been found that people with primary education and below have higher risks of hypertension and worse blood pressure control. (Sun et al., 2022)

- Region type is a categorical variable with four possible values: rural area, town,

city and unknown. This reflects the most recent available area of residence of the participant. The group "unknown" is created for 85 participants with missing values in region type. Region type is included because it might describe the environmental effects.

- Time of day is a categorical variable with three possible values: before 10, 10-15 and after 15. This reflects at what time in the day was the sample of the participant taken. Time of day is included because the metabolic profiles vary in different parts on the day (Sato et al., 2018).

Missing values occured in three variables. There were 24 participants with missing values in smoking status, 20 participants with missing values in BMI and one participant with missing value in time of day. It is noted here that there is also an overlap among these participants. All the participants with missing values in at least one of these variables were excluded from the analysis. As a result, 27 people were removed and the questionnaire data of 885 participants remains, including 101 incident cases of hypertension, 334 prevalent cases of hypertension and 450 control cases.

The descriptive statistics of the variables is listed in Table 2. There's a larger proportion of males in the prevalent hypertension group (44.9%) compared to incident hypertension group (34.7%) and control group (31.3%). There are also large differences in the average age at sample between the groups. The prevalent group has an average age of 61 years, incident group 48 years and control group 37 years. The highest average BMI of 29.9 is in the prevalent group and 28.2 in the incident group. People with a BMI greater than or equal to 25 are defined as being overweight (World Health Organization, 2021). The average BMI in the control group is 24.3. The proportion of people that have never smoked is quite equal in each of the groups (between 55-57%). The proportion of current smokers is larger in the incident group (28.7%) compared to prevalent group (14.1%) and control group (23.6%). Therefore the largest proportion of former smokers is in the prevalent group. The largest proportion of people with low education is in the prevalent group (14.1%) and largest proportion of people with high education in the control group

Table 2: The descriptive statistics of variables.

| Variable | Prevalent hypertension | Incident hypertension | Control group |
|---|---|---|---|
| Sex, male (%) | 44.9 | 34.7 | 31.3 |
| Sex, female (%) | 55.1 | 65.3 | 68.7 |
| Age at sample (mean±sd) | 61±13 | 48±13 | 37±13 |
| BMI (mean±sd) | 29.9±5.7 | 28.2±5.2 | 24.3±4.7 |
| Smoking, current (%) | 14.1 | 28.7 | 23.6 |
| Smoking, former (%) | 29.6 | 14.9 | 21.3 |
| Smoking, never (%) | 56.3 | 56.4 | 55.1 |
| Education, low (%) | 14.1 | 6.9 | 4.7 |
| Education, intermediate (%) | 56.8 | 56.4 | 45.3 |
| Education, high (%) | 29.1 | 36.6 | 50.0 |
| Region type, rural area (%) | 32.9 | 39.6 | 32.0 |
| Region type, town (%) | 9.3 | 8.9 | 12.0 |
| Region type, city (%) | 43.4 | 44.6 | 49.6 |
| Region type, unknown (%) | 14.4 | 6.9 | 6.4 |
| Time of day, before 10 (%) | 22.5 | 26.7 | 25.6 |
| Time of day, 10-15 (%) | 54.5 | 49.5 | 54.4 |
| Time of day, after 15 (%) | 23.1 | 23.8 | 20.0 |

(50%). The largest proportion of people living in rural areas is in the incident group (39.6%) and the largest proportion of people living in towns and cities in the control group (12% and 49.6% respectively).

# 3 Methodology

In this chapter, the methodology used in the analysis is introduced. The theoretical framework for principal component analysis (PCA) is discussed in chapter 3.1, survival analysis in chapter 3.2 and multiple testing problem in chapter 3.3.

## 3.1 Principal Component Analysis

This chapter includes the theoretical overview of PCA: the derivation of principal components is introduced in chapter 3.1.1 and variable scaling in chapter 3.1.2. The chapter is based on (Izenman, 2008) unless stated otherwise.

### 3.1.1 Derivation of Principal Components

Random vector $\mathbf{X} = (X_1, ..., X_p)^{\mathrm{T}}$ is considered with mean $\boldsymbol{\mu} = (EX_1, ..., EX_p)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}]$.

The aim of PCA is to replace $p$ correlated input variables with a set of $t$ linear projections $\xi_1, ..., \xi_t$ $(t \leq p)$ of the input variables that are uncorrelated and ordered in terms of variance. These linear projections,

$$\xi_j = \mathbf{b}_j^{\mathrm{T}}\mathbf{X} = b_{j1}X_1 + ... + b_{jp}X_p, \quad j = 1, ...t,$$

are called the first $t$ principal components of $\mathbf{X}$.

From the spectral decomposition theorem it is known that

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}, \quad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_p,$$

where the columns of $\mathbf{U}$ are the eigenvectors of $\boldsymbol{\Sigma}$ and the diagonal elements of the diagonal matrix $\boldsymbol{\Lambda}$ are the eigenvalues $\lambda_1, ..., \lambda_p$ of $\boldsymbol{\Sigma}$. Therefore the total variation of the

original input variables can be calculated as the sum of the eigenvalues of $\mathbf{\Sigma}$:

$$\sum_{j=1}^{p} \text{var}(X_j) = \text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j.$$

The coefficient vectors, $\mathbf{b}_j = (b_{1j}, ..., b_{pj}), \quad j = 1, ..., t$, are chosen so that:

- the first $t$ linear projections $\xi_j$, $j = 1, ..., t$ are ranked in decreasing order of their variances $\text{var}(\xi_1) \geq \text{var}(\xi_2) \geq ... \geq \text{var}(\xi_t)$,
- $\xi_j$ is uncorrelated with all $\xi_k, \quad k < j$.

Therefore the coefficient vectors are chosen in a sequential manner so that the variances of the derived variables ($\text{var}(\xi_j) = \mathbf{b}_j^{\text{T}} \mathbf{\Sigma} \mathbf{b}_j$) are arranged in descending order with two restrictions:

- $\mathbf{b}_j^{\text{T}} \mathbf{b}_j = 1, \quad j = 1, ..., t,$
- $\text{cov}(\xi_i, \xi_j) = \mathbf{b}_i^{\text{T}} \mathbf{\Sigma} \mathbf{b}_j = 0.$

The first principal component, $\xi_1$, is found when $p$ coefficients of $\mathbf{b}_1$ are chosen for the linear projection $\xi_1$ so that it maximizes the variance of this projection. The selection of $\xi_j$ is unique for all $j = 1, ..., t$ because of the normalisation constraint $\mathbf{b}_j^{\text{T}} \mathbf{b}_j = 1$.

The function

$$f(\mathbf{b}_1) = \mathbf{b}_1^{\text{T}} \mathbf{\Sigma} \mathbf{b}_1 + \lambda_1 (1 - \mathbf{b}_1^{\text{T}} \mathbf{b}_1)$$

is formed, where $\lambda_1$ is a Lagrangian multiplier. It's then differentiated with respect to $\mathbf{b}_1$. When the result is set equal to zero, the following equation is achieved:

$$\frac{\partial f(\mathbf{b}_1)}{\partial \mathbf{b}_1} = 2(\mathbf{\Sigma} - \lambda_1 \mathbf{I}_p)\mathbf{b}_1 = \mathbf{0}.$$

If $\mathbf{b}_1 \neq \mathbf{0}$ then $\lambda_1$ must satisfy equation

$$|\mathbf{\Sigma} - \lambda_1 \mathbf{I}_p| = 0.$$

Therefore, as $\text{var}(\xi_1) = \mathbf{b}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{b}_1 = \lambda_1$, $\lambda_1$ must be the largest eigenvalue of $\boldsymbol{\Sigma}$ and $\mathbf{b}_1$ the the eigenvector related to that eigenvalue.

The second principal component, $\xi_2$, is found when coefficients $\mathbf{b}_2$ are chosen for the linear projection $\xi_2$ so that it maximizes the variance of $\xi_2$ among all linear projections that are uncorrelated with $\xi_1$. The variance of $\xi_2$ is $\text{var}(\xi_2) = \mathbf{b}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{b}_2$ and has to be maximised subject to constraints $\mathbf{b}_2^{\mathrm{T}}\mathbf{b}_2 = 1$ and $\mathbf{b}_1^{\mathrm{T}}\mathbf{b}_2 = 0$.

The function

$$f(\mathbf{b}_2) = \mathbf{b}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{b}_2 + \lambda_2(1 - \mathbf{b}_2^{\mathrm{T}}\mathbf{b}_2) + \mu\mathbf{b}_1^{\mathrm{T}}\mathbf{b}_2$$

is formed, where $\lambda_2$ and $\mu$ are Lagrangian multipliers. It's then differentiated with respect to $\mathbf{b}_2$ and the result is set equal to zero. This equation is then solved and coefficients $\mathbf{b}_2$ found.

In sequential manner, all the remaining sets of coefficients for principal component are found. The coefficients are given by the ordered sequence of eigenvectors $\{\mathbf{v}_j\}$ where each eigenvector is associated with the $j$th largest eigenvalue $\lambda_j$ of $\boldsymbol{\Sigma}$.

### 3.1.2 Variable Scaling

Principal components are sensitive to scaling, therefore the units of measurement of different variables matter. The variables with the largest variances overwhelm the first principal components with other variables contributing minimally. Therefore it's advised to standardise the variables by centering and scaling:

$$\mathbf{Z} = (\text{diag}\{\boldsymbol{\Sigma}\})^{-1/2}(\mathbf{X} - \boldsymbol{\mu}).$$

This is equivalent to performing PCA on correlation matrix rather than covariance matrix. When using the correlation matrix, the total variation of standardised variables is the trace of correlation matrix.

## 3.2 Survival Analysis

Survival analysis is used when modelling the time until an event occurs. This type of modelling has many application areas: it can be used for medical studies to predict patient's survival time or for companies to predict when customer churns. This chapter includes the theoretical overview of survival analysis: survival and censoring times are explained in chapter 3.2.1, time scale in chapter 3.2.2, survival and hazard functions in chapter 3.2.3, Cox proportional hazards model in chapter 3.2.4 and finally the assumption of proportional hazards in chapter 3.2.5. These chapters are based on James et al. (2013) and Collett (2015) unless stated otherwise.

### 3.2.1 Survival and Censoring Times

It is assumed that for each individual there's a true survival time $T$ and censoring time $C$ and that the survival time is independent from the censoring time. The survival time $T$ represents the time when the event of interest occurs. It is also known as failure time or event time depending on what is the event of interest. The censoring time $C$ represents the time when censoring occurs. Censoring occurs when we stop observing the individual and therefore don't know if and when the event occurs.

One of the two times is always observed. If the event occurs before censoring ($T < C$), the true survival time, $T$, is observed. Otherwise, the censoring time, $C$, is observed. Therefore, the random variable

$$Y = \min(T, C)$$

as well as status indicator

$$\delta = \begin{cases} 1, & \text{if } T \leq C \\ 0, & \text{if } T > C \end{cases}$$

are observed.

The most common form of censoring is right censoring, which occurs when the true survival time is at least as large as observed time $Y$.

### 3.2.2 Time Scale

The selection of the time scale is an important part of survival analysis. In epidemiological studies, there are two main options:

- Using time in the study as time scale. In this case, time zero would be the moment when participant joins the study.
- Using participant's age as time scale. In this case, time zero would be the moment when participant is born.

In this thesis, participant's age was used as time scale. Therefore both survival and censoring times reflect the participant's age at these times. Late entry into the risk set, also known as left truncation, is relevant as participants are not followed since birth but from the moment they enter the study (given that the event has not happened before) (Canchola et al., 2003). With this decision, the relationship between age and survival cannot be measured. Then again, it is unnecessary to adjust for age in the analysis.

### 3.2.3 Survival and Hazard Functions

The survival function

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t),$$

where $F(t)$ is the distribution function of T. It measures the probability of surviving past time $t$.

The hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t}$$

measures the event rate on the moment after time $t$, given the survival past that time $t$. The hazard function can be also derived as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}(\log S(t)),$$

where $f(t)$ is the probability density function of $T$.

### 3.2.4 Cox Proportional Hazards Model

Suppose there are $p$ explanatory variables $X_1, ..., X_p$ for $n$ individuals with values $\mathbf{x}_i = (x_{1i}, ..., x_{pi})^{\mathrm{T}}, \quad i = 1, ..., n.$

Let there be a baseline hazard function $h_0(t)$ for the individual who has all the zero values in the explanatory variables, $\mathbf{x} = \mathbf{0}$. The hazard function for the $i$th individual is

$$h_i(t) = \psi(\mathbf{x}_i)h_0(t),$$

where $\psi(\mathbf{x}_i)$ is a function of $\mathbf{x}_i$. The function $\psi(\mathbf{x}_i)$ is the relative hazard between the individual $i$ with explanatory values $\mathbf{x}_i$ and the individual with zero values $\mathbf{x} = \mathbf{0}$. Since the relative hazard can't be negative, $\psi(\mathbf{x}_i)$ can be expressed as $\psi(\mathbf{x}_i) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathrm{T}}$ is the coefficient vector for the $p$ variables. The hazard function can be therefore written as

$$h_i(t) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)h_0(t) = \exp(\beta_1 x_{1i} + ... + \beta_p x_{pi})h_0(t). \tag{1}$$

Model 1 is called Cox proportional hazards model. This is a semi-parametric model as no probability function is assumed for the survival times. The coefficients in the model can be estimated using the maximum likelihood method without the need to estimate the baseline hazard function $h_0(t)$.

Suppose that there are again $n$ individuals, from them $r$ individuals have distinct event times and $n-r$ individuals are right-censored. It's assumed that there are no equal event times. The $r$ ordered event times are $t_{(1)} < ... < t_{(r)}$, meaning $t_{(j)}$ is the $j$th ordered event time. The set of individuals at risk at time $t_{(j)}$ will be denoted by $R(t_{(j)})$ and called "the risk set".

The relevant likelihood function to estimate the coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathrm{T}}$ for model

1 is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_l)},$$

where $\mathbf{x}_{(j)}$ is the vector of covariates for the individual at $j$th ordered event time.

### 3.2.5 Assumption of Proportional Hazards

While fitting Cox proportional hazards models, it is assumed that the hazards are proportional and constant over time. There are different methods for checking the assumption of proportional hazards. Schoenfeld residuals are used in this thesis and further explained in this chapter.

Schoenfeld residuals differ because there is no single residual for each individual, but a set of residuals, one for each explanatory variable fitted in the model. The $i$th Schoenfeld residual for $j$th explanatory variable in the model is given by

$$r_{S_{ji}} = \delta_i \left\{ x_{ji} - \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_l)} \right\},$$

where $\delta_i$ is the indicator function, $x_{ij}$ is the $j$th explanatory variable for the $i$th individual and $R(t_i)$ is the set of individuals at risk at time $t_i$. Let it be noted that non-zero residuals exist only for uncensored observations.

The scaled version of the Schoenfeld residuals has been found more effective. The vector of Schoenfeld residuals for $i$th individual is $\mathbf{r}_{S_i} = (r_{S_{1i}}, ..., r_{S_{pi}})^{\mathrm{T}}$. The scaled Schoenfeld residuals $r^*_{S_{ji}}$ are components of the vector

$$\mathbf{r}^*_{S_i} = d \ \mathrm{var}(\hat{\boldsymbol{\beta}}) r_{S_i},$$

where $d$ is the number of events and $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the parameter estimates of the fitted model.

The expected value of the $i$th Schoenfeld residual for the $j$th explanatory variable $r^*_{S_{ji}}$ is

given by

$$\mathbf{E}(r^*_{S_{ji}}) \approx \beta_j(t_i)\hat{\beta}_j,$$

where $\beta_j(t)$ is the time-varying coefficient of $X_j$, $\beta_j(t_i)$ is the value of this coefficient at the $i$th individual survival time $t_i$ and $\hat{\beta}_j$ is the estimated value of $\beta_j$.

The zph test of proportional hazards assumption is based on testing whether there is a linear relationship between $\mathbf{E}(r^*_{S_{ji}})$ and a function of time. If there is evidence that $\mathbf{E}(r^*_{S_{ji}})$ is time-dependent, the hypothesis of proportional hazards has to be rejected.

For the explanatory variable $X_j$, linear dependence can be expressed by $\beta_j(t_i) = \beta_j + v_j(t_i - \bar{t})$, where $v_j$ is an unknown regression coefficient. This leads to $\mathbf{E}(r^*_{S_{ji}}) = v_j(t_i - \bar{t})$. If the slope $v_j = 0$, then the variable has no dependence on time and the assumption of proportionality stands.

If $\tau_1, ..., \tau_d$ are the $d$ observed event times across all $n$ individuals, then the appropriate test statistic is

$$\frac{\left\{\sum_{i=1}^{d}(\tau_i - \overline{\tau})r^*_{S_{ji}}\right\}^2}{d\ \mathrm{var}(\hat{\beta}_j)\sum_{i=1}^{d}(\tau_i - \overline{\tau})^2},$$

where $\overline{\tau}$ is the sample mean of the observed event times. Under the null hypothesis that $v_j$ is zero, the statistic has $\chi^2$ distribution with 1 degree of freedom. Significantly large values of the statistic lead to rejection of the proportional hazards assumption for this variable. The global test of the proportional hazards assumption across all $p$ explanatory variables included in the Cox proportional hazards model is found by aggregating the individual test statistics.

## 3.3 Multiple Testing Problem

This chapter gives an overview of the multiple testing problem and introduces both the Bonferroni correction and the Benjamini-Hochberg correction. The chapter is based on James et al. (2013).

Multiple testing problem arises when testing a large number of hypotheses. That means

there is a high likelihood of getting some very small p-values by chance. When conducting multiple tests without correcting the significance level, the probability of making Type I errors, rejecting the true null hypotheses, increases. There are several methods for adjusting the significance level.

The Bonferroni correction is one of the easiest and widely used methods. If significance level $\alpha$ and tests for $m$ hypotheses are considered, the Bonferroni corrected significance level is then found as $\alpha/m$. The Bonferroni correction is considered strict when testing for a large number of hypotheses or when the hypotheses are strongly correlated.

Another method is to control the false discovery rate (FDR). The false discovery rate is defined as the estimated ratio of false positives to total positives. If FDR is controlled at level 0.05, then as many null hypotheses as possible are rejected while guaranteeing that no more than 5% of them (on average) are false positives. The Benjamini-Hochberg procedure is used to control the FDR. The algorithm follows the steps:

1. Specify $q$, the level at which to control the FDR.
2. Compute $p$-values, $p_1, ..., p_m$, for $m$ null hypotheses $H_{01}, ..., H_{0m}$.
3. Order the $m$ $p$-values $p_{(1)} \leq ... \leq p_{(m)}$.
4. Define
$$L = \max \left\{ j : p_{(j)} < qj/m \right\}.$$
5. Reject all null hypotheses $H_{0j}$ for which $p_j \leq p_{(L)}$.

When using the Bonferroni correction, the threshold of $\alpha/m$ doesn't depend on data. When using the Benjamini-Hochberg correction, all null hypotheses are rejected for which the $p$-value is less than or equal to the $L$th smallest $p$-value, where $L$ is a function of all $m$ $p$-values. Therefore, the threshold can't be planned in advance as it is based on the data. This method is particularly well-suited for large datasets where a vast number of tests are conducted for exploratory purposes, rather than for confirmatory ones.

# 4 Analysis

The dataset of 1055 metabolites is extensive but as the number of observations is only 885, including 101 incident cases of hypertension, different variable reduction and selection methods were used.

The 'one in ten rule' is widely used in survival models. According to this rule, one variable can be considered in a model for every 10 events. Considering the number of observations in this thesis, maximum 10 variables can be used in a model. If there are more variables included in the model, the issue of overfitting arises. This means that there can be results that actually don't exist in the population. (Chowdhury and Turin, 2020)

The aim of variable reduction is to choose candidate variables first, particularly, if the sample is small. The aim of variable selection is to choose the variables to be included in the models so that most accurate predictions can be made. Variable selection can be focused on both clinical knowledge from previous research and statistical selection methods. (ibid.)

There are different strategies for both dimensionality reduction and variable selection. In this thesis, chapter 4.1 covers how principal component analysis was used for dimensionality reduction. First, metabolites were mapped into human metabolomic pathways, then principal component analysis was performed and the first three components were chosen from each pathway. Antczak et al. (2013) have used the same strategy on a gene expression dataset as it is effective in improving biological interpretability and reducing dimensionality. Chapter 4.2 covers survival modelling with Cox proportional hazards models and variable selection strategies. First, the baseline model was found by eliminating variables in a backward stepwise manner and then the relevant pathway components were chosen in a forward stepwise manner.

## 4.1 Pathway Analysis

Human metabolic pathways were used to reduce the number of variables. KEGG pathways were downloaded from the KEGG website (Kanehisa Laboratories, 2023a) on 19.10.2022 for all KEGG ID-s associated with the Metabolon dataset. These pathways consist of metabolites that are listed with KEGG codes. Metabolic pathway (hsa01100) is the largest pathway that contains 247 metabolites. These metabolites also belong to several other smaller pathways (sub-pathways of metabolic pathway), therefore this superset was left out of the analysis. In addition to this, the analysis excluded nucleotide metabolism (hsa01232) since its two main sub-pathways, purine metabolism (hsa00230) and pyrimidine metabolism (hsa00240), share the same metabolites and were already included. We look to include an optimal number of pathways in relation to the sample size and methods planned (further described in this chapter). There are 9 pathways including at least 20 metabolites, 17 pathways including at least 15 metabolites and 35 pathways including at least 10 metabolites. We choose to include pathways with at least 10 metabolites in the analysis. These 35 pathways have 274 unique metabolites identified with KEGG codes. From these 274 metabolites, there are 218 present in our metabolomics dataset that belong to 34 different pathways. We will move forward with these 34 pathways with the number of metabolites that are available. Drug cytochrome P450 metabolism (hsa00982) does not have any metabolites available and is therefore left out of the analysis.

There are six KEGG code identified metabolites in the chosen pathways that refer to multiple chemical IDs in the dataset. All the chemical IDs relating to the same KEGG code were checked and their correlations calculated (Table 3). As most of them have a moderate correlation, it was decided to leave both chemical IDs in the analysis. It's not in the scope of this thesis to make a biological decision on how to choose between them.

As seen from the table above from the example of ribitol and arabitol/xylitol, there are some chemical IDs that refer to multiple KEGG codes. These are lysine (KEGGs C00739, C00047), glutamate (KEGGs C00025, C00217), allantoin (KEGGs C02348, C02350),

Table 3: KEGG codes with multiple Chemical IDs and their correlations.

| KEGG code | Chemical name | Chemical ID | Correlation |
|---|---|---|---|
| C00379 | ribitol<br>arabitol/xylitol | 100000406<br>100006430 | 0.38 |
| C00486 | bilirubin (Z,Z)<br>bilirubin (E,Z or Z,E)* | 1090<br>100001951 | 0.65 |
| C01152 | 3-methylhistidine<br>1-methylhistidine | 100000042<br>100001051 | 0.44 |
| C01904 | ribitol<br>arabitol/xylitol | 100000406<br>100006430 | 0.38 |
| C04555 | dehydroepiandrosterone sulfate (DHEA-S)<br>epiandrosterone sulfate | 100000792<br>100001287 | 0.75 |
| C18044 | pregnenediol sulfate (C21H34O5S)*<br>pregnenolone sulfate | 100002067<br>100002129 | 0.90 |

ribitol (KEGGs C00379, C01904, C00474, C00532), linolenate [alpha or gamma; (18:3n3 or 6)] (KEGGs C06426, C06427), mannitol/sorbitol (KEGGs C00392, C00794), arabonate/xylonate (KEGGs C00502, C05411), arabitol/xylitol (KEGGs C00379, C01904), gamma-tocopherol/beta-tocopherol (KEGGs C14152, C02483). It is noted that there's a possibility that one chemical ID is listed multiple times in a pathway determined by KEGG codes. In this case, each chemical ID is still used once in the pathway. The number of unique metabolites (listed with chemical IDs) included in chosen pathways by biochemical classes is visible from Table 4.

For each of 34 pathways, principal component analysis was performed with prcomp function in the statistics software R. All the variables were scaled within the function to have unit variance. Samples from all participants were used in this part of the analysis, regardless of the diagnoses.

In addition to these pathways, all available chemical xenobiotics were treated in the

Table 4: Number of metabolites included in the pathways by biochemical classes.

| Biochemical class | Number of metabolites left | Number unique metabolites included in the pathways |
|---|---|---|
| Amino Acid | 196 | 87 |
| Carbohydrate | 23 | 13 |
| Cofactors and Vitamins | 29 | 14 |
| Energy | 9 | 7 |
| Lipid | 402 | 47 |
| Nucleotide | 37 | 22 |
| Partially Characterized Molecules | 20 | 0 |
| Peptide | 25 | 1 |
| Xenobiotics | 109 | 22 |
| Not known | 205 | 0 |
| Total | 1055 | 213 |

same way with PCA performed. There are 24 chemical xenobiotics in this set including for example perfluorooctanesulfonate (PFOS) and perfluorooctanoate (PFOA). These are synthetic chemicals previously widely used within society, however, now categorised as environmental pollutants with negative effects on human health, and their use has become restricted (European Chemicals Agency, 2023). Only one of the metabolites in chemical xenobiotics overlaps with metabolites already used in the rest of the pathways, therefore 23 unique metabolites were added. In total there are 236 metabolites covered with 35 pathways (chemical xenobiotics are treated and called a pathway from now on).

The first three principal components were checked and the cumulative percentage of variance explained is listed in Table 5. If the next component describes cumulatively less than 5% more than the previous one, it was left out of the analysis. In this case, it's considered not useful in explaining the variance of the pathway. Therefore the third component of biosynthesis of unsaturated fatty acids (hsa01040) is removed from the further analysis.

Table 5: Cumulative variance explained by principal components of pathways.

| Pathway | Chem IDs | Cumulative % of variance explained | | | PCs taken |
|---|---|---|---|---|---|
| | | PC1 | PC2 | PC3 | |
| ABC transporters (hsa02010) | 43 | 18.3 | 27.3 | 34.7 | 3 |
| Biosynthesis of aminoacids (hsa01230) | 39 | 23.9 | 33.7 | 41.2 | 3 |
| Biosynthesis of cofactors (hsa01240) | 37 | 13.6 | 23.5 | 31.4 | 3 |
| 2Oxocarboxylic acid metabolism (hsa01210) | 25 | 27.3 | 38.5 | 46.9 | 3 |
| Central carbon metabolism cancer (hsa05230) | 25 | 27.5 | 41.3 | 50.5 | 3 |
| Protein digestion absorption (hsa04974) | 24 | 33.8 | 45.8 | 52.3 | 3 |
| Bile secretion (hsa04976) | 23 | 15.3 | 25.1 | 34.1 | 3 |
| DAmino acid metabolism (hsa00470) | 21 | 29.1 | 42.2 | 50.8 | 3 |
| AminoacyltRNA biosynthesis (hsa00970) | 20 | 37.9 | 51.7 | 58.2 | 3 |
| Purine metabolism (hsa00230) | 16 | 17.4 | 31 | 42.1 | 3 |
| Cysteine methionine metabolism (hsa00270) | 16 | 23.3 | 36.2 | 46.6 | 3 |
| Arginine proline metabolism (hsa00330) | 16 | 21.4 | 31.9 | 41.8 | 3 |
| Biosynthesis of unsaturated fatty acids (hsa01040) | 16 | 80.3 | 86 | 90.2 | 2 |
| Mineral absorption (hsa04978) | 16 | 38.9 | 53.3 | 60.8 | 3 |
| Glycine serine threonine metabolism (hsa00260) | 15 | 19.2 | 31.4 | 42.2 | 3 |
| Caffeine metabolism (hsa00232) | 14 | 63 | 78.3 | 85.4 | 3 |
| Carbon metabolism (hsa01200) | 14 | 21.2 | 36.5 | 46.7 | 3 |
| Alanine aspartate glutamate metabolism (hsa00250) | 13 | 21.6 | 38.5 | 48.9 | 3 |
| Histidine metabolism (hsa00340) | 13 | 23.1 | 37 | 47.6 | 3 |
| Phenylalanine metabolism (hsa00360) | 13 | 23.8 | 40.9 | 49.9 | 3 |

| Pathway | Chem IDs | Cumulative % of variance explained | | | PCs taken |
|---|---|---|---|---|---|
| | | PC1 | PC2 | PC3 | |
| Neuroactive ligandreceptor interaction (hsa04080) | 13 | 24 | 37.6 | 47.3 | 3 |
| Pyrimidine metabolism (hsa00240) | 12 | 17.6 | 30.9 | 43.2 | 3 |
| Taste transduction (hsa04742) | 12 | 26.1 | 39.7 | 50.3 | 3 |
| Arginine biosynthesis (hsa00220) | 11 | 23.7 | 41.6 | 51.3 | 3 |
| Lysine degradation (hsa00310) | 11 | 25.4 | 37.9 | 48.2 | 3 |
| Tryptophan metabolism (hsa00380) | 11 | 30 | 43 | 53.7 | 3 |
| Glyoxylate dicarboxylate metabolism (hsa00630) | 11 | 24.4 | 43.8 | 58.3 | 3 |
| Pantothenate CoA biosynthesis (hsa00770) | 11 | 17.8 | 31.4 | 43.7 | 3 |
| Primary bile acid biosynthesis (hsa00120) | 10 | 32.7 | 47.2 | 57.9 | 3 |
| Steroid hormone biosynthesis (hsa00140) | 10 | 44.7 | 61.3 | 74.9 | 3 |
| Valine leucine isoleucine biosynthesis (hsa00290) | 10 | 40.5 | 59.3 | 72 | 3 |
| Tyrosine metabolism (hsa00350) | 10 | 22.1 | 35.9 | 47.5 | 3 |
| BetaAlanine metabolism (hsa00410) | 10 | 21 | 35.2 | 47.6 | 3 |
| Pentose glucuronate interconversions (hsa00040) | 9 | 28.2 | 42.4 | 54.6 | 3 |
| Chemical xenobiotics | 24 | 13.4 | 25.8 | 34.4 | 3 |

As known from the PCA theory, all principal components from the same pathway are uncorrelated. As PCA was conducted on each pathway separately, principal components from different pathways can still be strongly correlated as they might have overlapping metabolites. Therefore correlation analysis was conducted on the remaining 104 principal components. The components were systematically looked through starting from the largest pathway and its first principal component. If there was a correlation of 0.9 or higher with one or several smaller components, the smaller component was immediately left out of the analysis. After that, the next components were checked and all the smaller highly correlated components removed. In total, 10 components are excluded from the analysis as visible in Table 6. Therefore 94 principal components remain and are used in further modelling. Highly correlated variables measure essentially the same information therefore removing one of them should not affect the performance of the model (Chowdhury and Turin, 2020).

Table 6: Highly correlated principal components.

| Larger component | Smaller highly correlated component that is removed from the analysis | Correlation |
|---|---|---|
| ABC transporters (hsa02010) PC1 | Biosynthesis of aminoacids (hsa01230) PC1 | 0.92 |
| | Central carbon metabolism cancer (hsa05230) PC1 | 0.92 |
| | Protein digestion absorption (hsa04974) PC1 | 0.90 |
| | DAmino acid metabolism (hsa00470) PC1 | 0.90 |
| Central carbon metabolism cancer (hsa05230) PC2 | DAmino acid metabolism (hsa00470) PC2 | 0.94 |
| Protein digestion absorption (hsa04974) PC2 | AminoacyltRNA biosynthesis (hsa00970) PC2 | 0.99 |
| | Mineral absorption (hsa04978) PC2 | 0.93 |
| Bile secretion (hsa04976) PC1 | Primary bile acid biosynthesis (hsa00120) PC1 | 0.98 |
| AminoacyltRNA biosynthesis (hsa00970) PC1 | Mineral absorption (hsa04978) PC1 | 0.99 |
| Carbon metabolism (hsa01200) PC1 | Alanine aspartate glutamate metabolism (hsa00250) PC1 | 0.92 |

## 4.2 Survival Analysis

The final dataset is assembled from the dataset of 94 principal components from 35 different pathways covering 236 metabolites; sample data including age at sample and time of day; hypertension diagnosis data including incidence and prevalence of the cases, age at diagnosis; death data including age at death; censoring data including age at the end of follow-up period on 31.12.2021; questionnaire data including sex, BMI, smoking status, education class and region type.

Cox proportional hazards models were used to find the pathway components associated with the risk of hypertension. R software package survival and function coxph were used for modelling. The survival object was created with function Surv(start, stop, event) where age was used as time scale. This corresponds to adjusting for age (Fischer et al., 2014). Therefore start refers to age at sample; stop refers to age at diagnosis, age at death or age at the end of follow-up period whichever is the minimum; event refers to whether the participant developed an incident case of hypertension during the period (coded as 1) or not (coded as 0). Participants with prevalent cases of hypertension diagnosed before the time of the sample were left out of the analysis. This leaves a dataset of 551 participants with 101 incident cases and 450 controls.

### 4.2.1 Baseline Model

First, Cox proportional hazards model is fitted with sex, BMI, smoking status, education class, region type and time of day included as predictors. The reference levels for categorical variables are never for smoking, male for sex, rural area for region type, low for education class and before 10 for time of day. The significance level of 0.05 is used. The found model is visible in Table 7.

In Model 1, where sex, BMI, smoking status, education class, region type and time of day are included as predictors, the only variable of statistical importance is BMI with p-value of $7.03 \cdot 10^{-10}$ under the significance level of 0.05. The function drop1 with Chi-Square test was used to find the variable which, when left out, will leave the best model with the lowest Akaike's information criterion (AIC). AIC is commonly used for model comparisons, lower value of the statistic indicates the better model (Collett, 2015). In the backward stepwise selection manner, the following variables are left out of the analysis: region type, smoking, time of day, sex and finally education class. Model 2 with only one predictor BMI remains. The higher is participant's BMI, the higher is the risk of getting diagnosed with hypertension. If the BMI is higher by one point, the risk of being diagnosed with hypertension increases 1.10 times. If the BMI is higher by five points, the risk of

being diagnosed with hypertension increases $1.10^5 = 1.61$ times. The model improved with AIC dropping from 820.66 to 808.71. The assumption of proportionality was checked with function cox.zph. For model 1, the assumption held for all the variables with p-values over 0.05 and the p-value of 0.43 for global test. For model 2, the assumption held with p-value of 0.67 for BMI. Model 2 is used as a baseline model for pathway discovery.

Table 7: Hazard ratio estimations and p-values for the fitted Cox proportional hazards baseline models.

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | exp(coef) | p-value | exp(coef) | p-value |
| sex: Female | 0.95 | 0.83 | | |
| BMI | 1.11 | $7.03 \cdot 10^{-10}$ | 1.10 | $1.34 \cdot 10^{-9}$ |
| smoking: Current | 1.30 | 0.29 | | |
| smoking: Former | 0.86 | 0.61 | | |
| education class: Intermediate | 0.73 | 0.47 | | |
| education class: High | 0.54 | 0.18 | | |
| region type: City | 1.16 | 0.53 | | |
| region type: Town | 0.87 | 0.72 | | |
| region type: Unknown | 0.63 | 0.32 | | |
| time of day: After 15 | 0.77 | 0.38 | | |
| time of day: Between 10 and 15 | 0.73 | 0.22 | | |
| AIC | 820.66 | | 808.71 | |

### 4.2.2 Pathway Discovery

For the pathway discovery, a multivariate model was found in the forward stepwise selection manner. Firstly, 94 models were created, each including BMI and one of the 94 pathway components as predictors. The pathway component leading to the smallest Benjamini-Hochberg corrected p-value in the Cox proportional hazards model was chosen and included in the prediction model. The process was repeated until no additional pathway components were significant at the Benjamini-Hochberg corrected significance level with FDR controlled at level 0.05.

The first pathway component added to the model was PC2 of carbon metabolism (hsa01200). After that, no other pathway components proved significant. Therefore the final model consists of BMI and the PC2 of carbon metabolism as seen in Table 8. The BMI remains important in the model. If the BMI of the participant is higher by one point, the risk of being diagnosed with hypertension increases 1.09 times. The values of PC2 of carbon metabolism range from -5.79 to 6.33. If the PC2 of carbon metabolism is higher by one point, the risk of being diagnosed with hypertension increases 1.31 times. The model improved compared to the baseline model and has an AIC of 795.84. For model 3, the assumption of proportionality held for BMI with p-value of 0.44 and for PC2 of carbon metabolism with p-value of 0.14. The global test had a p-value of 0.23 and therefore the assumption of proportionality stands.

Table 8: Hazard ratio estimations and p-values for the final Cox proportional hazards model.

| | Model 3 | |
|---|---|---|
| | exp(coef) | p-value |
| BMI | 1.09 | $1.84 \cdot 10^{-7}$ |
| Carbon metabolism (hsa01200) PC2 | 1.31 | $8.92 \cdot 10^{-5}$ |
| AIC | 795.84 | |

Therefore, a component of carbon metabolism pathway can help predict hypertension even years before the onset of the disease. The metabolites included in this pathway and their importance in the component were further examined. There are 14 metabolites in the carbon metabolism pathway. These are listed in Table 9 with the respective biochemical class, subclass and loading of the metabolite in second principal component. The loadings show how big is the weight of each metabolite in the component. It can be interpreted as the correlation between the metabolite and the component, therefore loadings can be both positive and negative, also indicating the direction of the relationship.

Carbon metabolism includes 14 metabolites from 4 different biochemical classes and 7

Table 9: Metabolites in carbon metabolism pathway.

| Metabolite | Biochemical class | Biochemical subclass | Loading |
|---|---|---|---|
| Glycine | Amino Acid | Glycine, Serine and Threonine Metabolism | -0.51 |
| Serine | Amino Acid | Glycine, Serine and Threonine Metabolism | -0.51 |
| Citrate | Energy | TCA Cycle | -0.32 |
| Alanine | Amino Acid | Alanine and Aspartate Metabolism | -0.29 |
| Glutamate | Amino Acid | Glutamate Metabolism | 0.27 |
| Cysteine | Amino Acid | Methionine, Cysteine, SAM and Taurine Metabolism | -0.26 |
| Alpha-ketoglutarate | Energy | TCA Cycle | 0.24 |
| Gluconate | Xenobiotics | Food component / plant | -0.15 |
| Aspartate | Amino Acid | Alanine and Aspartate Metabolism | 0.14 |
| Fumarate | Energy | TCA Cycle | -0.14 |
| Glycerate | Carbohydrate | Glycolysis, Gluconeogenesis and Pyruvate Metabolism | -0.12 |
| Malate | Energy | TCA Cycle | -0.10 |
| 2-keto-3-deoxy-gluconate | Xenobiotics | Food component / plant | 0.07 |
| Pyruvate | Carbohydrate | Glycolysis, Gluconeogenesis and Pyruvate Metabolism | -0.07 |

different biochemical subclasses. The largest biochemical class is amino acids with 6 metabolites. The largest subclass is TCA cycle, which belongs to the second largest biochemical class energy with four metabolites. Amino acids glycine and serine have the biggest weight in the second principal component with loadings of -0.51. Citrate, which is part of the TCA cycle, follows with a loading of -0.32 and amino acids alanine, glutamate and cysteine with respective loadings of -0.29, 0.27 and -0.26. The following chapter covers further discussion on these results with validation from the previous findings.

# 5 Discussion

Hypertension has been previously studied using different approaches and datasets. The research has been conducted in both humans and animal models. (Chakraborty et al., 2020) Animal models are animals, such as rats and mice, that are used in research. There are remarkable anatomical and physiological similarities between humans and animals, therefore the animal models help to study the shared mechanisms of the diseases. The results obtained on animals, however, might not necessarily be confirmed in further human studies. (Barré-Sinoussi and Montagutelli, 2015) Therefore human studies provide important insight that can't be achieved through animal models. Both prevalent and incident disease can be studied in humans. Studying people with prevalent disease means exploring differences between people with an existing diagnose to people without. For better understanding of the disease mechanisms, studying incident disease is especially important. It makes it possible to find risk factors that are detectable even years before the diagnosis.

Julkunen et al. (2023) used nuclear magnetic resonance (NMR) biomarker data of 249 metabolites for 118 461 participants in the UK Biobank to find associations between metabolites and prevalence, incidence, and mortality of over 700 common diseases. The incidence of primary hypertension, hypertensive heart disease and hypertensive chronic kidney disease were included in the analysis. Also, the mortality of primary hypertension and hypertensive heart disease, as well as the prevalence of primary hypertension and hypertensive chronic kidney disease. Unlike the definition in this thesis, the diagnoses of hypertension were addressed separately. Each metabolite was analysed for association with the diagnosis using logistic regression for prevalent endpoints or Cox regression for incident and mortality endpoints. The models were adjusted for sex, UK Biobank assessment center, and age. In cases of Cox regression, age was used as the time scale. The extensive collection of results is made available in form of an online atlas of biomarker-disease associations (Nightingale Health, 2023). This work highlights the value of metabolic biomarker profiling in large biobanks. Biobanks make it possible to

follow participants over longer periods and study outcomes that are not present in the participants upon enrollment.

In this thesis, incident hypertension was modelled using principal components of 35 human metabolomic pathways covering 236 metabolites. In addition to this, sex, BMI, smoking status, education class, region type and time of day were fitted for the baseline model. Age was used as time scale in all of the Cox proportional hazards models, meaning adjusting for age. In the final model, BMI and the second principal component of carbon metabolism remained as significant predictors. Carbon metabolism includes 14 metabolites from 4 different biochemical classes: glycine, serine, alanine, glutamate, cysteine and aspartate from amino acids; citrate, alpha-ketoglutarate, fumarate and malate from energy metabolites; gluconate and 2-keto-3-deoxy-gluconate from xenobiotics; glycerate and pyruvate from carbohydrates. Results of this thesis are in accordance with previous studies.

Hypertension has been studied using the Dahl salt-sensitive and salt-resistant rat models, which are genetically hypertensive and normotensive, respectively. They were expected to have differences in blood pressure but not in metabolism. However, it was found that there are deficiencies in several components of TCA cycle, which leads to higher levels of TCA metabolites fumarate, succinate, iso-citrate, aconitate, citrate, pyruvate. Therefore the elevated levels of these metabolites are associated with salt-sensitive hypertension. In addition to this, lower levels of aspartate, arginine, nitric oxide and malate were also found to be significant in contributing to hypertension. (Chakraborty et al., 2020) There are several overlapping metabolites with the results achieved in this thesis. Carbon metabolism includes citrate, alpha-ketoglutarate, fumarate and malate, which are part of the TCA cycle. Moreover, aspartate and pyruvate have been in both cases associated with the development of hypertension.

There have been several studies connecting hypertension and amino acids. Arjmand et al. (2023) used logistic regression models in the Iranian population to find predictors for hypertension from targeted metabolomics data including 30 acylcarnitines and 20 amino

acids. They found that after adjusting for age, sex and BMI, four amino acids alanine, valine, glycine and serine remained significant in predicting stage 2 hypertension. After also adjusting for lipid profile, FPG, use of oral glucose-lowering drugs, statins and anti-hypertensive drugs, only one amino acid glycine remained important. Serine and glycine were combined into an index that also showed significant results in predicting stage 2 hypertension after adjusting for all covariates. The lower values of both glycine separately and the index combined from glycine and serine, lead to higher odds of the development of hypertension. These results are consistent with the findings in this thesis as glycine and serine were the metabolites having the biggest weight in carbon metabolism pathway. Furthermore, Stamler et al. (2013) conducted a cross-sectional epidemiologic study measuring blood pressure and dietary data from people of China, Japan, the United Kingdom, and the United States. They found that the dietary intake of glycine is significantly related to blood pressure in linear regression models adjusted to confounders. Lin et al. (2022) found that genetically predicted higher circulating glycine was associated with a lower risk of hypertension among the UK Biobank participants.

In this thesis, BMI remained as an important predictor for hypertension in all of the models. This is supported by previous research indicating that BMI is an independent risk factor for hypertension (Landi et al., 2018). Neither sex, smoking status, education class, region type nor time of day showed any importance in predicting hypertension. These results are somewhat surprising as men are known to have higher blood pressure than women regardless of race and ethnicity, furthermore across species including dogs, rats and mice (Sandberg and Ji, 2012). It is possible that some of these effects were described by age and BMI and therefore did not become relevant after adjusting for these variables. In the further studies, it is possible to include polygenic risk scores in the models as Estonian Biobank has genotypes for all participants. In this way, the genetic information would also be incorporated in the analysis.

One of the strengths of this thesis lies in the rich untargeted metabolomics dataset of 1505 metabolites. Although many of these metabolites were not used as they did not map into KEGG human metabolic pathways, a significant number of metabolites still

remained. In addition to this, to tackle the possibility of losing valuable information, all available chemical xenobiotics were added as one of the pathways. The total number of 236 metabolites in 35 pathways was used. As a result, a valuable dataset of pathway components was created, which can be used in the future for the analysis of other phenotypes or the same phenotype with new data from linked electronic health records. As an alternative, instead of using KEGG pathways, network modules could have been defined on the dataset. This would have given an opportunity to include all metabolites, but on the other hand, would have complicated the biological interpretation.

One of the limitations in this thesis is the small number of participants in the dataset. This limited the methods available as it was decided to stick to the 'one in ten rule' to avoid overfitting. Considering the number of incident cases of hypertension among participants, a maximum of 10 variables were allowed in a model. To tackle this, principal component analysis was used for variable reduction and capturing the pathways. Forward stepwise selection was used for pathway component selection. In the future, it is possible to use different variable selection strategies and methods to model incident hypertension (Spooner et al., 2020). Additionally, the results of this thesis need validation using an independent dataset. There are several metabolomics datasets in the Estonian Biobank that can be used for this purpose, covering the carbon metabolism pathway at least partly.

# Conclusion

Hypertension is a common chronic medical condition. Furthermore, the most important preventable risk factor for cardiovascular disease, mortality and all-cause morbidity. Hypertension has been extensively researched to identify its causes, risk factors, and potential treatments. There has been an increased interest in incorporating metabolomics in the recent years. Studying connections between hypertension and metabolites can provide valuable insight into the underlying mechanisms of hypertension and also help identify the biomarkers relevant in the development of hypertension for earlier diagnosis.

In this thesis, an extensive mass spectrometry dataset of 1505 metabolites was used to find relevant predictors for incident hypertension. From this dataset, 236 metabolites were mapped to 35 human metabolomic pathways with a goal to both reduce the number of variables, but also to study the effects of these pathways. Principal component analysis was performed on all pathways and first three principal components were selected for each of them. Correlation analysis was performed to remove highly correlated components. This left 94 principal components that were further used in the analysis. The samples belong pseudonymised participants of the Estonian Biobank that have additionally filled questionnaires for background data and are also linked periodically to electronic health records. For modelling incident hypertension, a dataset of 551 participants, with 101 incident cases and 450 controls, was used. It was done using Cox proportional hazards models with age as the time scale and accounting for right censored and left truncated data. The baseline model was found after fitting the model with sex, BMI, smoking status, education class, region type and time of day as predictors. Irrelevant variables were removed in backward stepwise manner, and baseline model was achieved with only BMI remaining as predictor. Finally, relevant pathway components were explored in forward stepwise manner. The second principal component of carbon metabolism pathway proved significant with FDR controlled at level 0.05 when adjusted for BMI. After inclusion of this component, no other components proved significant. The assumption of proportionality held for both BMI and the second principal component of

carbon metabolism. Higher values of BMI and the second principal component of carbon metabolism indicate higher risk of being diagnosed with hypertension. Therefore, both BMI and metabolites from carbon metabolism pathway can help predict hypertension even years before the onset of the disease.

The results were validated against and supported by previous research. Previous studies have consistently identified BMI as an independent risk factor for hypertension (Landi et al., 2018). According to Chakraborty et al. (2020), several metabolites in carbon metabolism are linked to hypertension, including citrate, alpha-ketoglutarate, fumarate, and malate, which are involved in the TCA cycle. In addition, the study found that aspartate and pyruvate are also associated with the development of hypertension. Stamler et al. (2013) and Lin et al. (2022) have found glycine to be associated with hypertension. Arjmand et al. (2023) have concluded the same, furthermore, showed that the index combined from glycine and serine showed significant results in predicting hypertension. This is consistent with the findings in this thesis as glycine and serine have the biggest weight in carbon metabolism pathway.

The findings from this thesis should be further validated in other datasets available in the Estonian Biobank. Additionaly, a valuable dataset of human metabolic pathway components was created, which can be used in the Estonian Biobank for the analysis of other phenotypes.

The author would like to express sincere gratitude to the supervisors, Jaanika Kronberg and Krista Fischer, for their invaluable advice. Furthermore, the author would like to express their gratitude to prof. Tõnu Esko for providing the data.

# References


Antczak, P., H. J. Jo, S. Woo, L. Scanlan, H. Poynton, A. Loguinov, S. Chan, F. Falciani, and C. Vulpe (2013). "Molecular toxicity identification evaluation (mTIE) approach predicts chemical exposure in Daphnia magna". In: *Environmental science & technology* 47.20, pp. 11747–11756.

Arjmand, B., H. Dehghanbanadaki, M. Yoosefi, N. Rezaei, S. Mohammadi Fateh, R. Ghodssi-Ghassemabadi, N. Najjar, S. Hosseinkhani, A. Tayanloo-Beik, H. Adibi, et al. (2023). "Association of plasma acylcarnitines and amino acids with hypertension: A nationwide metabolomics study". In: *PloS one* 18.1, e0279835.

Barré-Sinoussi, F. and X. Montagutelli (2015). "Animal models are essential to biological research: issues and perspectives". In: *Future science OA* 1.4.

Buergel, T., J. Steinfeldt, G. Ruyoga, M. Pietzner, D. Bizzarri, D. Vojinovic, J. Upmeier zu Belzen, L. Loock, P. Kittner, L. Christmann, et al. (2022). "Metabolomic profiles predict individual multidisease outcomes". In: *Nature Medicine* 28.11, pp. 2309–2320.

Buford, T. W. (2016). "Hypertension and aging". In: *Ageing research reviews* 26, pp. 96–111.

Canchola, A. J., S. L. Stewart, L. Bernstein, D. W. West, R. K. Ross, D. Deapen, R. Pinder, P. Reynolds, W. Wright, H. Anton-Culver, et al. (2003). "Cox regression using different time-scales". In: *Western Users of SAS Software. San Francisco, California*.

Chakraborty, S., J. Mandal, T. Yang, X. Cheng, J.-Y. Yeo, C. G. McCarthy, C. F. Wenceslau, L. G. Koch, J. W. Hill, M. Vijay-Kumar, et al. (2020). "Metabolites and hypertension: insights into hypertension as a metabolic disorder: 2019 Harriet Dustan Award". In: *Hypertension* 75.6, pp. 1386–1396.

Chowdhury, M. Z. I. and T. C. Turin (2020). "Variable selection strategies and its importance in clinical prediction modelling". In: *Family medicine and community health* 8.1.

Collett, D. (2015). *Modelling survival data in medical research*. 3rd ed. CRC press.

Dona, A. C., M. Kyriakides, F. Scott, E. A. Shephard, D. Varshavi, K. Veselkov, and J. R. Everett (2016). "A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments". In: *Computational and Structural Biotechnology Journal* 14, pp. 135–153.

European Chemicals Agency (2023). *Per- and polyfluoroalkyl substances (PFAS)*. URL: https://echa.europa.eu/hot-topics/perfluoroalkyl-chemicals-pfas (visited on 01/05/2023).

Fischer, K., J. Kettunen, P. Würtz, T. Haller, A. S. Havulinna, A. J. Kangas, P. Soininen, T. Esko, M.-L. Tammesoo, R. Mägi, et al. (2014). "Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons". In: *PLoS medicine* 11.2, e1001606.

Izenman, A. J. (2008). *Modern multivariate statistical techniques*. Vol. 1. Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. 2nd ed. Vol. 112. Springer.

Julkunen, H., A. Cichońska, M. Tiainen, H. Koskela, K. Nybo, V. Mäkelä, J. Nokso-Koivisto, K. Kristiansson, M. Perola, V. Salomaa, et al. (2023). "Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank". In: *Nature Communications* 14.1, p. 604.

Kanehisa Laboratories (2023a). *KEGG Mapper – Search*. URL: https://www.genome.jp/kegg/mapper/search.html.

– (2023b). *Kyoto Encyclopedia of Genes and Genomes web page*. URL: https://www.genome.jp/kegg/ (visited on 12/02/2023).

Landi, F., R. Calvani, A. Picca, M. Tosato, A. M. Martone, E. Ortolani, A. Sisto, E. D'angelo, E. Serafini, G. Desideri, et al. (2018). "Body mass index is strongly associated with hypertension: Results from the longevity check-up 7+ study". In: *Nutrients* 10.12, p. 1976.

Leitsalu, L., T. Haller, T. Esko, M.-L. Tammesoo, H. Alavere, H. Snieder, M. Perola, P. C. Ng, R. Mägi, L. Milani, et al. (2015). "Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu". In: *International journal of epidemiology* 44.4, pp. 1137–1147.

Lin, C., Z. Sun, Z. Mei, H. Zeng, M. Zhao, J. Hu, M. Xia, T. Huang, C. Wang, X. Gao, et al. (2022). "The causal associations of circulating amino acids with blood pressure: a Mendelian randomization study". In: *BMC medicine* 20.1, pp. 1–11.

Metabolon, Inc (2023). *Metabolon web page*. URL: https://www.metabolon.com/ (visited on 06/04/2023).

Nightingale Health (2023). *Nightingale Biomarker-Disease Atlas*. URL: https://biomarker-atlas.nightingale.cloud/.

Oparil, S., M. C. Acelajado, G. L. Bakris, D. R. Berlowitz, R. Cífková, A. F. Dominiczak, G. Grassi, J. Jordan, N. R. Poulter, A. Rodgers, et al. (2018). "Hypertension". In: *Nature Reviews: Disease Primers* 4.1, p. 18014.

Reinhold, D., H. Pielke-Lombardo, S. Jacobson, D. Ghosh, and K. Kechris (2019). "Pre-analytic considerations for mass spectrometry-based untargeted metabolomics data". In: *High-Throughput Metabolomics: Methods and Protocols*, pp. 323–340.

Sandberg, K. and H. Ji (2012). "Sex differences in primary hypertension". In: *Biology of sex differences* 3.1, pp. 1–21.

Sato, S., E. B. Parr, B. L. Devlin, J. A. Hawley, and P. Sassone-Corsi (2018). "Human metabolomics reveal daily variations under nutritional challenges specific to serum and skeletal muscle". In: *Molecular metabolism* 16, pp. 1–11.

Spooner, A., E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty (2020). "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction". In: *Scientific reports* 10.1, pp. 1–10.

Stamler, J., I. J. Brown, M. L. Daviglus, Q. Chan, K. Miura, N. Okuda, H. Ueshima, L. Zhao, and P. Elliott (2013). "Dietary glycine and blood pressure: the International Study on Macro/Micronutrients and Blood Pressure". In: *The American journal of clinical nutrition* 98.1, pp. 136–145.

Sun, K., D. Lin, M. Li, Y. Mu, J. Zhao, C. Liu, Y. Bi, L. Chen, L. Shi, Q. Li, et al. (2022). "Association of education levels with the risk of hypertension and hypertension control: a nationwide cohort study in Chinese adults". In: *J Epidemiol Community Health* 76.5, pp. 451–457.

The R Foundation (2023). *The R Project for Statistical Computing*. URL: https://www.r-project.org/.

Thirumurugan, D., A. Cholarajan, S. S. Raja, and R. Vijayakumar (2018). "An Introductory Chapter: Secondary Metabolites". In: *Secondary Metabolites - Sources and Applications*. InTech. DOI: 10.5772/intechopen.79766.

Virdis, A., C Giannarelli, M Fritsch Neves, S Taddei, and L Ghiadoni (2010). "Cigarette smoking and hypertension". In: *Current pharmaceutical design* 16.23, pp. 2518–2525.

World Health Organization (June 9, 2021). *Obesity and overweight*. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

– (2023). *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. URL: https://icd.who.int/browse10/2019/en#/I10-I15 (visited on 06/04/2023).

Yu et al. (2023). "Plasma metabolic outliers identified in Estonian human knockouts". Manuscript in preparation.

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Liis Hiie,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Human Metabolic Pathways as Predictors for Hypertension Based on Estonian Biobank Data, mille juhendajad on Jaanika Kronberg ja Krista Fischer, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Liis Hiie
16.05.2023