Óbuda University

PhD thesis

Two Applications of Biostatistics in the Analysis of Pathophysiological Processes

by

Tamás Ferenci

Supervisor:

Levente Kovács

Applied Informatics Doctoral School

Budapest, 2013

Acknowledgments to the external advisers

I am deeply indebted to my external thesis supervisors, *Balázs Benyó* and *J Geoffrey Chase* for their initiation of and active participation in the research for the second part of this dissertation, and for their kindness for making it possible to join their ongoing research. Without their continuous support, I would have never reached the results presented in the second thesis group.

I would also likely to say thanks to *Zsuzsanna Almássy*, who was my external thesis supervisor for the first thesis group.

Finally, although not being a thesis supervisor, I feel it almost obligatory to say thanks to *Zoltán Benyó*, who was one of the pioneers of biomedical engineering in Hungary.

Contents

Acknowledgments

First of all, I would like to say thanks to my doctoral supervisor, *Levente Kovács*, with whom I first started to investigate a biostatistical problem in 2007 at the Budapest University of Technology and Economics (BUTE). In addition to his continuous support, without which I would possibly never became a biostatistician, he was also instrumental in keeping me in the academic sphere. Also, he always did his best to pave the way amongst the different administrative and technical difficulties.

Next, I would like to mention a few of my teachers, from whom I learnt a lot about probability theory, statistics, and mathematics in general. *Balázs Kotosz*, my teacher at the Corvinus University of Budapest (CUB) was the one with whom I first started researching theoretical statistics. As far as CUB is concerned, I would also like to say special thanks to *László Hunyadi*, who was the reviewer of some of my early papers. He has really set a standard for me with his selfless, wholehearted support as evidenced by the amount of work he dedicated to provide helpful reviews for papers of students whom he has never even met before. Also, it was an honor working with *Jenő Reiczigel*, president of the Hungarian Clinical Biostatistical Society. I am also grateful to *Márton Balázs* from the Department of Stochastics at the BUTE. Although he is primarily engaged in researching probability theory, for me, his delightful classes laid the solid foundations for statistics as well.

Finally, I would like to say thanks to my family, especially for raising me in a way that I am always in pursuit of causes and explanations.

I am indebted to *Zoli*, without whom I would perhaps never got interested in medicine – which was a turning point in my life. Finally, I would like to say thanks to *Ildi* – this dissertation could have never been written without her patience and support.

List of Figures

List of Tables

Abstract

The application of biostatistical tools is indispensable in many current medical research; so is informatics, which makes the use of many of these tools feasible.

As medicine became more and more empirically oriented in the last centuries, and as it became more and more model-oriented in the last decades, the mathematical and – specifically – biostatistical methods received special attention. The application of such apparatus is necessary to the precise investigation of many questions, and can also help to raise new ones.

The present dissertation shows two examples for this. The first thesis group deals with the topic of obesity, more specifically, pediatric obesity. Nowadays this issue – due to the worrisome epidemiological data – gets emphasized public health focus. The concrete question I investigated was how obesity affects laboratory parameters – which, in some sense, gives insight into how obesity affects the human body. Within this thesis group, I have developed a biostatistical framework, which makes the comprehensive analysis of this question possible, both in uni- and in multivariate sense.

The second thesis group also considers a problem of an intensively researched topic: it deals with the objective evaluation and examination of the so-called tight glycemic control protocols that are used in critical care. One of the key tasks of such protocols is the prediction of the patients' insulin sensitivity. Within this thesis group, I have developed a biostatistical method, which makes it possible to model the evolution of a patient's insulin sensitivity in the context of the predictions provided by the protocol. The method explicitly incorporates the patient's diagnosis and the length-of-stay in the intensive care unit, which can fundamentally influence the evolution of the insulin sensitivity. The method thus makes it possible to quantitatively assess the protocol, furthermore it can also provide (even clinical) suggestions on how to improve the protocol, considering different goals.

Absztrakt

A biostatisztikai módszerek alkalmazása manapság megkerülhetetlen része számos orvosi kutatásnak, hasonlóan az informatikához, mely számos ilyen módszer használatát teszi gyakorlatban is kivitelezhetővé.

Azáltal, hogy az orvostudomány egyre inkább empirikusan orientálttá vált az elmúlt néhány évszázadban, illetve azáltal, hogy egyre inkább modell-alapúvá fejlődött az elmúlt néhány évtizedben, jelentősen felértékelődtek a matematikai, illetve – ezen belül – a biostatisztikai módszerek. Az ilyen eszköztár alkalmazása elengedhetetlen számos kérdés precíz vizsgálatához, illetve sok esetben új kérdések felvetését is nagyban segíti.

A jelen disszertáció erre mutat két példát. Az első téziscsoport az elhízás, ezen belül a gyermekkori elhízás problémakörével foglalkozik. Ez manapság – az aggasztó epidemiológiai adatok miatt – egyre nagyobb népegészségügyi jelentőséget kap. A kérdésfelvetés azt vizsgálja, hogy az elhízás milyen hatással van a rutinszerűen vizsgált laborparaméterekre – ez egyfajta betekintést ad abba, hogy az elhízás hogyan hat az emberi szervezetre. A téziscsoport keretében kidolgoztam egy biostatisztikai keretrendszert, mely lehetővé teszi e kérdés átfogó vizsgálatát, mind egy- mind többváltozós értelemben.

A második téziscsoport egy szintén intenzíven kutatott terület egy problémáját vizsgálja: az intenzív ellátásban használatos ún. szoros vércukorszint-kontroll protokollok objektív minősítésével és vizsgálatával foglalkozik. E protokollok egy kulcsfontosságú feladata a betegek inzulin-szenzitivitásának előrejelzése. A téziscsoporton belül kidolgoztam egy biostatisztikai módszert, mely lehetővé teszi az inzulin-szenzitivitás alakulásának modellezését, adott protokoll által szolgáltatott előrejelzések fényében. A módszer explicite beépíti a beteg diagnózisát, és az intenzív terápiás osztályon eltöltött időt, melyek alapvetően befolyásolhatják az inzulin-szenzitivitás alakulását. Az eljárás ezáltal lehetővé teszi, hogy a protokollt kvantitatíve minősítsük, sőt, akár klinikai javaslatokat is tud adni lehetséges javításokra, különböző célok figyelembevételével.

List of abbreviations

Abbreviation	Meaning
AMISE	Asymptotic mean integrated squared error
CA	Cluster analysis
CDC	Centers for Disease Control and Prevention
cdf	Cumulative distribution function
ecdf	Empirical cumulative distribution function
ICU	Intensive Care Unit
iid	Independent and identically distributed
KDE	Kernel density estimation
MISE	Mean integrated squared error
ML	Maximum likelihood
NHANES	National Health and Nutrition Examination Survey
PCA	Principal components analysis
pdf	Probability density function
SI	Insulin sensitivity
SPRINT	Specialized Relative Insulin and Nutrition Tables
TGC	Tight glycemic control

1. Introduction

Modern medical research can hardly be imagined without the active participation – or at least support – of biostatistics.

This is not surprising if we consider a few tendencies of the development of medical science. First, the *empirical orientation* became more and more pronounced during the last centuries, and downright determinant during the 20th century. While anecdotal recollections of empirically driven medical researches date back to Biblical times, we can not speak of systematic, empirical thinking in the context of medicine before the 18th century. (It was in 1747 that James Lind performed his sometimes questioned, but nevertheless celebrated clinical trial (Milne 2012), in which he demonstrated that scurvy can be treated with citrus fruits.)

This is not independent of the fact that this was the time when tenable knowledge started to accumulate about both the function of healthy body and its diseases. As for the former, this is connected to progresses in anatomy, especially due to the results of autopsies. (The only part that might be shocking is how long it took (relative to the human history) to correctly describe even the most basic physiological functions; for instance, the first essentially correct description of human blood circulation was given only in 1628 (Sloan 1978) by William Harvey.) As far as the latter is concerned, it was also the time when more modern, science-based theories started to replace those earlier ideas that are rather ridiculous by today's standards about the causes of diseases (bad air, revenge of gods, imbalance of humors etc.).

From this point on, empirical orientation just grew stronger and stronger, and it is perhaps no exaggeration to say that this empirical orientation ended in *Evidence Based Medicine* (Sackett et al. 1996) in the second half of the 20th century. (Whose crucial idea is to base clinical decision making on the collection and critical evaluation of the best available scientific evidences – which are mostly the results of empirical investigations.)

It should be noted that not only the importance of empirical results increased (in general), but within it, specifically the importance of *quantitative* results as well. This was also reinforced by the fact that the last decades of the 20th century brought a previously unthinkable computational capacity that can be employed to both data processing and

storage.

The other factor we should consider is the medical science's progress towards *model-based* approach. Traditionally, medical diagnosis was distinguished from engineering diagnosis in that the latter has an exact model of how the failing object is *supposed to* operate, in contrast to medicine, where we do not have an arbitrarily detailed description of the "good" state. This is less and less so: due to the advancements in physiology, in many fields of the medical science, models of such precision were developed that can be applied clinically.

My dissertation shows two examples for these: two "case studies" for the application of biostatistical methods in medical research. The first thesis group investigates questions related to obesity (Chapter 2) which is one the leading concerns of public health in the developed countries nowadays. The second thesis group deals with a special aspect of human blood glucose regulation and its abnormal states: tight glycemic control in intensive care units (Chapter 3).

Both thesis groups (but especially the first) includes methodological development. In introducing these, I will assume the usual preliminaries from mathematical statistics (Stuart and Ord 2009) and biostatistics (Armitage, Berry, and Matthews 2008).

It should be emphasized how significant impact does informatics, applied informatics has on modern biostatistics. It has at least three concrete aspects.

First, and perhaps the most "mechanical" support that computers can give is the automatic performing of routine numerical calculations (such as the calculation of a mean, or the performing of a test). Although many statistics courses still educate students how to perform these "by hand" (primarily to facilitate the understanding of the details of the calculations), in practice every mechanical calculation is carried out by computers nowadays.

Computers can also support the work of the statisticians in a more general way. By aiding the handling of large databases (filtering, ordering, searching etc.), transforming the data (encoding variables, applying functions etc.), calculating statistics, visualizing data and so on, they also help a more creative, more effective work. (Partly by reducing or almost eliminating the time consumption of routine tasks, hence helping the statistician to concentrate on the essence of the problem, and partly by giving such support that would be impossible without computers, for example, by drawing interactive three-dimensional graphics.)

Advances in artificial intelligence (Russell and Norvig 2010) made even the automatic information extraction (learning) from data possible, termed machine learning (Witten, Frank, and M. Hall 2011). When combined with large databases (Berman 2013), and

the intention to discover new, previously hidden information (instead of learning known properties), this leads to the concept of data mining (Han, Kamber, and Pei 2011; Hand, Mannila, and Smyth 2001).

Finally, there are certain methods which would be not only hard, but downright impossible without computers. These are the so-called computationally intensive procedures (such as resampling methods and algorithmic models), which have enormous computational requirements, and hence they are as old as computers, because without computers, their development and – especially – meaningful application is unthinkable (Good 2006; Good 2000; Shao and Tu 2012).

2. Effect of Obesity on Laboratory Parameters

My first thesis group addresses problems related to obesity. Obesity is one of the leading concerns of public health in many developed countries. The prevalence of obesity is sharply rising, with serious comorbidities linked casually to obesity, including cardiovascular diseases which are among the leading causes of death in many countries.

Obesity affects the human body in complex ways, impacting many aspects of homeostasis. The deeper understanding of these changes may help us to achieve a better prevention and treatment of this disease.

One manifestation of the effect of obesity is the systematical change in many blood chemistry parameters. Several such laboratory parameter was investigated in relation to obesity, but not comprehensively and with different methodologies.

My aim in this thesis will be to provide a uniform framework which allows the investigation of the effects of obesity on laboratory parameters, even for children, where the growth seriously complicates the definition of overweight and obesity. I will provide a methodology (and an associated computer program, which implements this methodology) to fulfill this aim, and also present concrete results obtained by the application of this methodology on a representative international survey and a non-representative Hungarian survey (which is, however, the first to address this question on Hungarian population).

The rest of this thesis is organized as follows. In Section 2.1 I give a very concise clinical introduction to obesity, focusing on those aspects that will bear relevance for the further discussion. Section 2.2 introduces my first thesis: a new methodology to investigate the effect of obesity on laboratory parameters. This thesis also involves the actual implementation of this methodology as a computer program to provide informatics support in applying this methodology to real-life databases. Finally, in Section 2.3 I present my second thesis, which is essentially the application of this methodology to two databases which will give rise to interesting novel observations about pediatric obesity. This thesis group is summarized in Section 2.4.

2.1. Clinical Introduction

In this Section, I first present results which underline the public health significance of obesity. I will introduce the basic facts about its epidemiology, and the most important clinical consequences that are linked to obesity.

After that, I start to narrow the focus to arrive to my direct aim: to investigate the relationship between obesity and laboratory parameters. Here, I only state the most basic observations on this topic, with the details elaborated later on.

Finally I will outline the directions and goals of the present research.

2.1.1. Epidemiology and Public Health Significance of Obesity

Obesity (Andersen 2003) is considered an epidemic in most parts of the developed world. As an example: it has been long time since overweight and obese people became the majority in the United States' population; according to the latest data, the prevalence of overweight is 34.2%, the prevalence of obesity and extreme obesity is 39.5% among adults aged 20 and over (Ogden and Carrol 2010b). The speed of progress is even more frightening, especially as far as obesity is concerned: the same prevalence was only 14.3% in 1960 (Ogden and Carrol 2010b).

Situation is similar in Hungary: the prevalence of overweight is 34.1%, the prevalence of obesity is 19.5% (Organization for Economic Co-operation and Development 2012).

The same applies to pediatric obesity as well, although the available information is less detailed (Wang and Lobstein 2006; Ogden, Yanovski, et al. 2007). In the United States, the prevalence of obesity among children and adolescents aged 2-19 is 16.9% (Ogden and Carrol 2010a), in Hungary, the same prevalence is estimated to be about 5-10% (Kern 2007; Antal et al. 2009). Due to its public health importance, many review is available on the epidemiology of obesity in children and adolescents (Moreno, Ahrens, and Pigeot 2011).

Obesity is in the focus of public health for decades, as – in addition to its continuously increasing prevalence – it also increases all-cause morbidity and mortality (Flegal et al. 2013; Visscher and Seidell 2001; Pi-Sunyer 2009). Type 2 diabetes (formerly known as non-insulin dependent diabetes, which is typically adult-onset), various cardiovascular diseases (including ischaemic heart disease), asthma, gallbladder disease, various malignant tumors are examples for diseases with increased occurrence casually linked to obesity(Guh et al. 2009). These have been described in children too (Burke 2006; Nyberg et al. 2011).

2.1.2. Obesity and Laboratory Parameters

It is well-known that obesity, and even overweight, causes systematical changes in the laboratory results. The reasons of these changes are complex. On one hand, many change is a more or less direct consequence of the manifestly altered homeostatic equilibrium induced by obesity, like elevated serum alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels found in obese adults (Ruhl and Everhart 2003), and in children as well (Dubern, Girardet, and Tounian 2006).

However, in some cases, the change in the laboratory parameters can not be attributed to a single physiological alteration, or even to any well-defined alteration that causes manifest obesity-related finding at all at the moment the laboratory parameter is already changed. A notable example is C-reactive protein (CRP) which is used even predicatively (Bo et al. 2009; Juonala et al. 2011; Ong et al. 2011) because of this reason.

2.1.3. Directions and Goals of my Research

Previous researches in this topic mostly focused on univariate questions (as exemplified by the above citations). In other words, they were were rather association-oriented findings, i.e. they described changes of a certain laboratory result in obese subjects (as opposed to the healthy state). To my best knowledge, no investigation addressed the question how obesity affects the laboratory results from a multivariate perspective (i.e. what is the effect of obesity if not only individual changes, but also alterations in the correlation structure of the laboratory results is considered), especially not in children.

Therefore, my primary aim was to investigate how pediatric obesity influences the uni- and multivariate structure of common laboratory parameters in a precise, uniform way for all parameters.

The principal novelty of my research lies in the fact that I present a methodology that integrates the handling of different levels of overweight and obesity using advanced statistical apparatus.

Detailed references to what is already known in the literature in this topic from previous researches will be given in Section 2.3.

2.2. Methodology to Assess the Effect of Obesity on Laboratory Parameters

I have developed a biostatistical methodology (and an associated computer program) to investigate the effect of obesity on laboratory parameters. This methodology provides a way to analyze both the uni- and the multivariate structure of the laboratory parameters, making the effect of obesity explicit during the process.

The methodology is specifically designed to deal with the data of children and adolescents, where the growth of the subjects is non-negligible (complicating the comparison of individuals of different age).

As already noted, I actually implemented this methodology to provide informatics support to its application. Full source code of the implementation is listed in Appendix A.

Relevant own publications pertaining to this thesis: [F-3; F-15; F-9; F-1; F-4; F-21; F-2; F-11; F-5; F-12; F-8; F-7; F-6; F-13; F-18; F-19; F-20; F-17].

Figure 2.1. illustrates the whole workflow of my methodology. I will introduce each step in detail in the followings.

First, some preliminary question will be discussed in Subsection 2.2.1, after which the programming environment that was used in the implementation of the methodology will be introduced in Subsection 2.2.2.

The methodology will be presented in two parts: first the univariate part is introduced (Subsection 2.2.3), and then apparatus for the multivariate investigations is discussed (Subsection 2.2.4).

2.2.1. Preliminaries

The developed methodology requires a database where both the investigated laboratory parameters and some indicator that assess the degree of overweight and obesity is measured on sufficient number of individuals (sampled representatively in optimal case).

The most popular choice as ”indicator of overweight/obesity” is the Body Mass Index (BMI), that is, body mass (measured in kilograms) divided by the square of body height (measured in centimeters) (Eknoyan 2008). Although drawbacks of BMI for that end are well-known (Romero-Corral et al. 2008; Okorodudu et al. 2010), it is still the most widely used simple indicator of the degree of overweight/obesity (World Health Organization 2013).

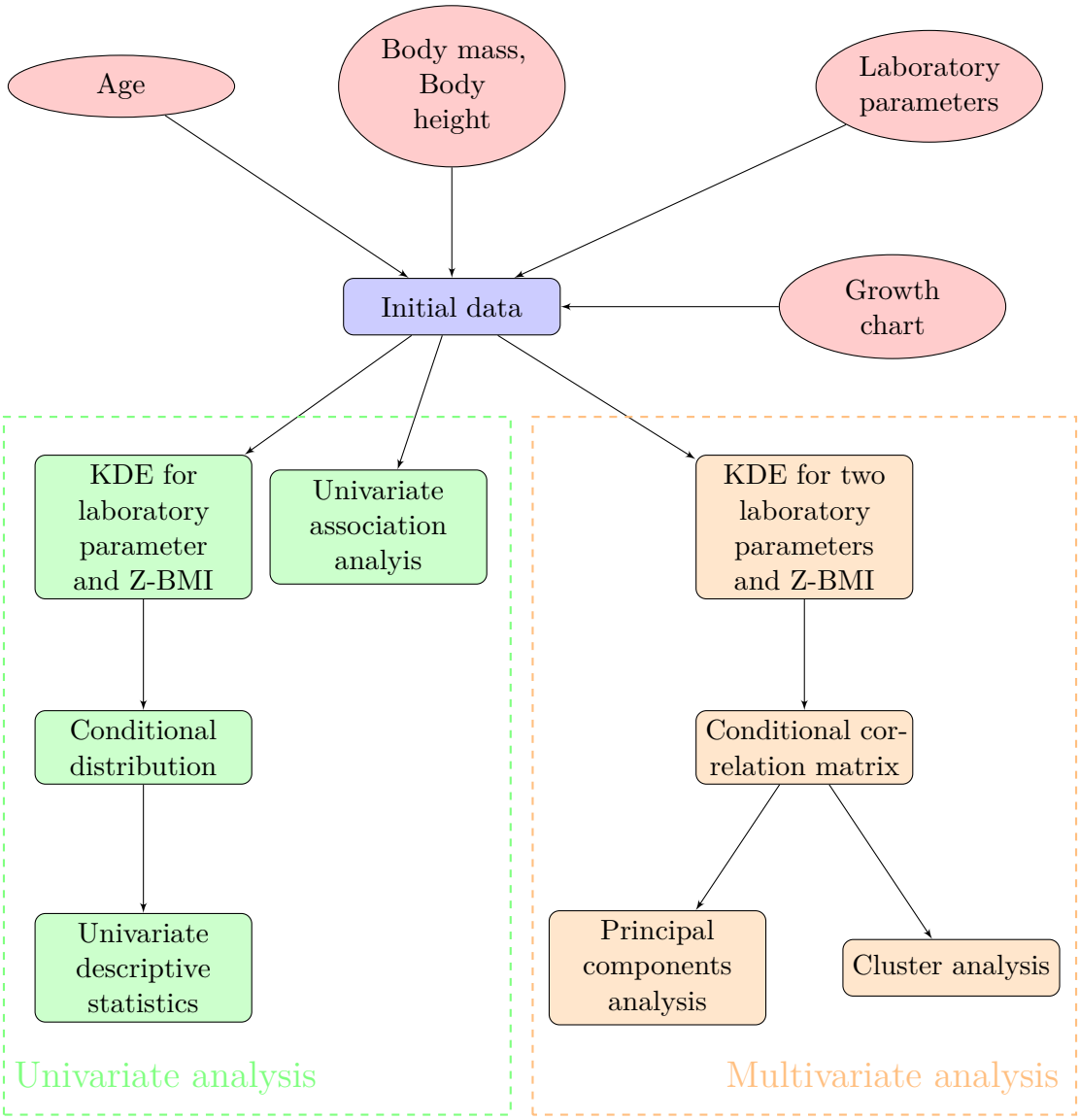


Figure 2.1.: Workflow of the method I developed for the investigation of the effect of obesity on laboratory parameters.

Another factor that has to be specifically handled is the growth of the subjects. In practice, one can only utilize databases that contain the data of differently aged individuals. This poses no problem for adults where the same BMI represents same state of obesity (save for the aforementioned limitations) irrespectively of actual age (due to the lack of growth). However, in settings were the growth of the subjects can not be neglected (that is, children and adolescents) BMI can not be used as an indicator of overweight or obesity as even the same BMI can represent different degree of overweight or obesity (in addition to the limitations of the BMI), as physiological growth has a systematical impact on the distribution of BMI.

To account for this impact, the developed methodology employs a measure that takes the children's growth (i.e. age) into account. Standardized BMI (BMI z-score, or Z-BMI (Cole et al. 2005)) was chosen, which is essentially the deviation of the child's BMI from the mean BMI of the child's age and sex, measured in standard deviation units. This can be calculated based on sex-specific growth charts; the one from Centers for Disease Control and Prevention (CDC) was used in this research (Centers for Disease Control and Prevention 2013). Figure 2.2. shows the 3rd, 10th, 50th, 90th and 97th percentile both for boys and girls according to this growth chart.

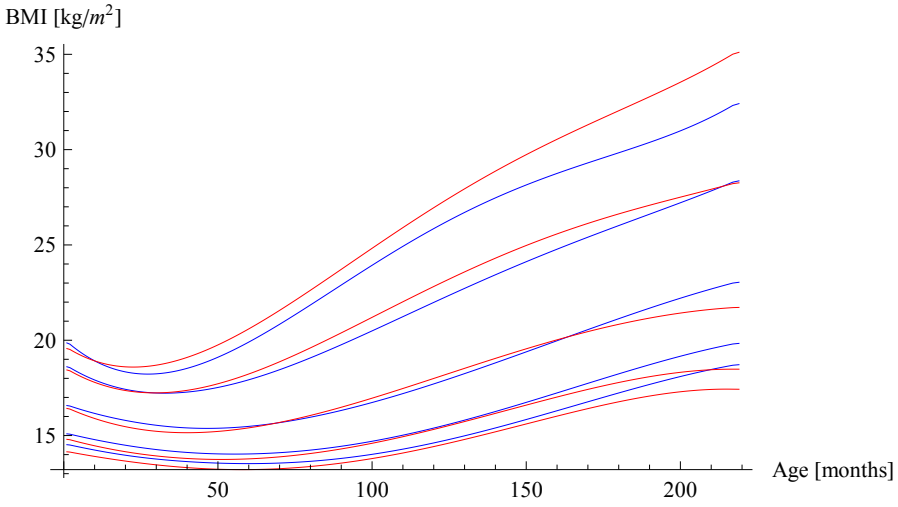


Figure 2.2.: The employed growth chart (Centers for Disease Control and Prevention 2013) depicted with the 3rd, 10th, 50th, 90th and 97th percentile both for boys (blue) and girls (red).

This growth chart also includes the necessary L-M-S parameters (Cole 1990) to calculate

Z-BMI. Namely, Z-BMI is

$$\frac{\left(\frac{BMI}{M}\right)^{L}-1}{L\cdot S} \tag{2.1}$$

if $L \neq 0$, else Z-BMI is

$$\frac{\ln\left(\frac{BMI}{M}\right)}{S}. \tag{2.2}$$

Note that extreme percentiles (i.e. Z-BMI values lower than -2 or higher than $+2$) should be handled with care (Kuczmarski et al. 2002) due to the extrapolation. (In the research, the CDC Growth Chart was used because the Hungarian Growth Charts (Joubert et al. 2006) unfortunately do not include the necessary L-M-S parameters.) Note that the developed methodology employes no hard threshold to specify overweight or obesity, instead, it relies on Z-BMI as a continuous (scale) indicator (proxy) of the degree of overweight and obesity.

Another question is the issue of sexes. Sex has a profound impact on many laboratory parameter (see (Kelly and Munan 1977; Rodger et al. 1987; Taylor et al. 1997) for hematologic examples), and would introduce complex interactions that are hard to interpret from the current point of view, so the most practical form of sex-matching was simply the complete separation of the analysis according to sex. Hence, sexes were separately analyzed in my methodology.

As far as missing values are considered, only subjects that have at most 5 missing laboratory result, and no missing value in BMI, sex and age were reatined. A laboratory parameter is retained only if missing values do not exceed 10% of the sample size. Missing values for the retained subjects and laboratory parameters were univariately imputed with sample median value from the same sex (Enders 2010).

2.2.2. Programming Environment

Statistical analysis was principally performed under the `R` statistical program package (R Core Team 2013), version 3.0.0. The source code of the developed program is given in Appendix A. Libraries `ks` (Duong 2013), `psych` (Revelle 2013) and `weights` (Pasek and Tahk 2012) were used.

`R` is a programming environment specifically for statistical computing and visualization. It is free and open source (available through the GNU General Public License), and is one of the most widely used statistical environments in the academic sphere.

It was developed in the early 1990s as a variant of the – proprietary – `S` programming language (also designed for statistical computing), influenced by the `Scheme` language.

As a programming language, R – in which most of the R environment has been written – supports the procedural programming paradigm and is object-oriented with dynamic typing.

The support for statistical computing is mostly achieved through the extreme diversity of the so-called packages, which extend the capabilities of R. As of early 2013, over 4 500 package is available (with typically dozens added weekly) from 'Accurate, Adaptable, and Accessible Error Metrics for Predictive Models' to ' Zhang–Yue–Pilon trends'. These cover almost every area of modern statistics, including many particular specialty as well. Many novel statistical procedure is first implemented under R.

By default, R has effectively no graphical user interface or integrated development environment, it instead relies on a command line interface.

2.2.3. Univariate Analysis

Univariate analysis will only consider one variable at a time, i.e. it neglects the connections between different variables. In other words, the focus will now be on understanding each laboratory parameter in itself (but including the relationship with obesity).

First, descriptive statistics will be provided. This is complicated by the fact that the current aim to take the effect of obesity into account by calculating the statistics for different levels of overweight and obesity. After this, an analysis that specifically investigates the relationship between overweight/obesity and the laboratory parameter under study will be performed.

Univariate Descriptive Statistics

Univariate reference values of the laboratory parameters for different degrees of overweight and obesity, namely Z-BMI=+1, +2 and +3 are given in my methodology, segregated according to sex. Classical descriptive statistics of mean and standard deviation, and more robust alternatives of median and interquartile range (IQR) were used (Armitage, Berry, and Matthews 2008). The usage of robust statistics is justified by the well-known fact that the distribution of many laboratory parameters is skewed, sometimes highly (Armitage, Berry, and Matthews 2008), for example CRP (Yamada et al. 2001; König et al. 1999) which is known to follow log-normal distribution (Limpert, Stahel, and Abbt 2001).

The question arises how to define these descriptors for a given Z-BMI value (e.g. for Z-BMI=+1). As Z-BMI is a continuous variable, there is no point in calculating an average value (or any other statistic) for subjects that have exactly Z-BMI=+1 (possible there is not even a single subject in the database with a Z-BMI of exactly +1). The problem is

obviously that only a finite sample drawn from an otherwise continuous distribution is available. One solution would be the binning of Z-BMI values, i.e. to give the average value for subjects having 0.5<BMI<1.5 (instead of Z-BMI=1). While this method is quite robust, the drawback is that information is lost (by grouping everyone from Z-BMI=0.5 to Z-BMI=1.5 to the same category, regardless of the subject's actual Z-BMI), hence losing possible tendencies within the 0.5<Z-BMI<1.5 group. Therefore an other alternative was chosen: we tried to reconstruct the – continuous – distribution based on the sample. What is needed is the joint distribution of the Z-BMI and the investigated laboratory parameter: from this, the (conditional) distribution of the laboratory parameter for any given Z-BMI value can be obtained. Given this conditional distribution, any statistic (mean, median, standard deviation etc.) of the investigated laboratory parameter for the exact Z-BMI value on which we conditioned (like Z-BMI=+1) can be numerically calculated, just as it was set forth.

This is essentially a joint probability density function (pdf) estimation task, which was solved by employing kernel density estimation (KDE). To introduce KDE, first a few results from the theory of distribution estimation is re-iterated.

It is well-known that the plug-in estimator of the cumulative distribution function (cdf), the empirical cumulative distribution function (ecdf) has very advantageous statistical properties. In fact, if the sample is an independent and identically distributed (iid) sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a common background distribution with cdf F_X (in this iid case, the sample's ecdf is often denoted with $\widehat{F}_n$), the empirical cdf converges to this cdf pointwise both in strong (and hence in weak) sense, by the Strong and Weak Law of Large Numbers, respectively:

$$\forall x : \widehat{F}_n(x) \xrightarrow{as} F_X(x) \quad \text{that is} \quad \forall x : \mathbb{P}\left(\lim_{n\to\infty} \widehat{F}s_n(x) = F_X(x)\right) = 1. \tag{2.3}$$

(This is a simple consequence of the fact that for any *given* x (hence the pointwise convergence), the distribution of $I_{\{x_i<x\}}$ will be Bernoulli distribution with parameter $F_X(x)$, and these indicators will be independent due to the iid sampling.) This establishes $\widehat{F}_n(x)$ as an unbiased and consistent estimator of $F_X(x)$ (given the fact that the mean of $\mathrm{Bern}(p)$ is p and its variance is $p(1-p)$, hence $\mathbb{E}\widehat{F}_n(x) = \frac{n \cdot F_X(x)}{n} = F_X(x)$ and $\mathbb{D}^2\widehat{F}_n(x) = \frac{n \cdot F_X(x)\left(F_X(x)\right)}{n^2} = \frac{F_X(x)\left(F_X(x)\right)}{n}$).

In addition to that, it can be shown (Glivenko–Cantelli theorem or "central theorem of statistics", 1933) that this convergence is not only pointwise, but also uniform. Specifically, it asserts a convergence in sup-norm (although other reasonable norms can be used as

well):

$$\left\|\widehat{F}_n - F_X\right\|_\infty := \sup_x \left|\widehat{F}_n(x) - F_X(x)\right| \xrightarrow[n\to\infty]{} 0. \tag{2.4}$$

(For the proof, see (Pestman 2009, pp. 306–310).)

In other words, it states that not only $\forall x : \mathbb{P}\left(\lim_{n\to\infty} \widehat{F}_n(x) = F_X(x)\right) = 1$ but also $\mathbb{P}\left(\forall x : \lim_{n\to\infty} \widehat{F}_n(x) = F_X(x)\right) = 1$ stands.

In addition to these, $\widehat{F}_n$ is not only unbiased but also efficient, given that it is a function of a complete, sufficient statistic (the order statistics, namely) by the Lehmann-Scheffé theorem (Lehmann and Casella 1998; Casella and Berger 2002).

Now a way to estimate a population cdf from a sample with comfortable statistical properties is established. (Actually this results can be made even sharper, and the rate of convergence can be characterized as well (Vaart 2000), but this will be sufficient for this purpose.)

One, however, often wishes to estimate the underlying distribution's probability density function. (Mostly because many features of the distribution (such as multimodality) can be better perceived with probability density function; especially in the multivariate case.)

One might be tempted to perform the same "plug-in" estimation for pdf, but this is not possible: ecdf is a jump function (regardless of the sample size), hence its derivative is zero almost everywhere, with undefined derivatives at the points of jump. This makes the estimation of a pdf (also called *density estimation*) a much more complicated issue.

One possible way to treat this problem is the application of *parametric density estimation*. When parametrically estimating a density, one *a priori* presumes a structure for the pdf, that is, it assumes a functional form, where uncertainty in the function is reduced to uncertainty in one or more parameters of the function. For example, one might presume that the pdf has a functional form

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}, \tag{2.5}$$

where μ and σ are the unknown parameters. (Normal approximation.)

By imposing such restriction on the pdf, one reduces the density estimation task to a parameter estimation task. (Which has well-known, long-studied solutions, such as the maximum likelihood (ML) principle (Lehmann and Casella 1998; Millar 2011).)

This, however, comes at the price of commitment to a given functional form. To avoid this, one might apply *nonparametric density estimation*, which has no such commitment (Tapia and Thompson 2002; Devroye and Györfi 1985). The problem is that while cdf can be very well (unbiased, efficiently) approximated with ecdf nonparametrically

(while it was never explicitly mentioned, ecdf is, of course, a nonparametric estimator), no such (uniformly minimum variance unbiased) estimator exists for pdf, as shown by Rosenblatt back in 1956 (Rosenblatt 1956). (Although it was apparently Fix and Hodges (1951) who gave the first treatment of this question.) Hence, nonparametric density estimation will always involve compromises.

The most well-known nonparametric density estimator (which is, however, not called by this name in many introductory texts), is perhaps the histogram.

Consider a (real-valued, scalar) statistical sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and a finite, non-degenerate partition $(\mathcal{P})_{i=1}^{k}$ of the real line, or an $[x_{\min}, x_{\max}] \subseteq \mathbb{R}$ interval of it. (That is, $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ if $i \neq j$ and $\bigcup_{i=1}^{k} \mathcal{P}_i = [x_{\min}, x_{\max}]$ with $\mathrm{diam}\mathcal{P}_i \neq 0$.) Then, the *histogram* of the sample, denoted with $\widehat{f}_{n,\mathrm{hist}}$ is the function

$$\widehat{f}_{n,\mathrm{hist}}(x) = \frac{\nu_i}{n h_i}, \quad \text{if } x \in \mathcal{P}_i, \tag{2.6}$$

where $\nu_i := \sum_{j=1}^{n} I_{\{x_j \in \mathcal{P}_i\}}$ and $h_i := \mathrm{diam}\mathcal{P}_i$.

In plain words, one cuts the real line (or an interval of it) to subintervals, counts the number of samples that fall to each subinterval, and plots this quantity (after a normalization) above the interval.

By this "binning", the problem of discreteness is resolved and a practically useful estimator – at least asymptotically – of the underlying distribution's pdf can be obtained (under very mild conditions for the choice of partitioning). Namely, if $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is an iid sample with histogram estimate $\widehat{f}_{n,\mathrm{hist}}$ from an absolutely continuous distribution with pdf f_X, then $\widehat{f}_{n,\mathrm{hist}}$ is a valid pdf, and if $\sup h_i \xrightarrow[n\to\infty]{} 0$ with $n \cdot \inf h_i \xrightarrow[n\to\infty]{} \infty$ at the same time for a sequence of partitions, then the histogram is a consistent estimator of f_X, that is

$$\widehat{f}_{n,\mathrm{hist}}(x) \xrightarrow{\mathbb{P}} f_X(x), \tag{2.7}$$

but it is not unbiased. (The validity of $\widehat{f}_{n,\mathrm{hist}}$ as a pdf is trivial by its definition. For the second part, see (Pestman 2009, pp. 403–404), for the third part see (Hardle 2004).)

Many result is available on how the bias and the variance of a histogram is connected (Hardle 2004). Without technical details, note that there is a trade-off between the two, governed by the bin widths: the smaller the bin widths are, the smaller the bias is, but with higher variance.

However, histogram is still a step function, and is heavily dependent on the choice of the bins. (The choice of bin widths, and also on their actual location. This is of high importance in practice (Hardle 2004), but will not be discuss in detail now.)

An alternative to histogram is the so-called *kernel density estimator* (KDE). To

motivate this, note that ecdf is simply a rescaled sum of cdfs for indicator variables, which are Heaviside step functions (located at the sample points). The root of the problem why ecdf can not be used to define an empirical pdf is that the Heaviside step function can not be derived. A straightforward solution is to replace this step function with a function that has a derivative everywhere – in other words, replace the cdf of the indicator variable with the cdf of a variable that is continuous, such as the cdf of a normal distribution (with $\mu = 0$ and an appropriate σ^2). These can be similarly summed and scaled to obtain an estimated (but continuous) cdf, from which an estimated pdf an be obtained by differentiation.

This is illustrated on Figure 2.3, which directly compares the two methods.

By the linearity of derivative, it can immediately be seen that the pdf estimated this way will be nothing else than the sum of the pdfs of the normal distributions. The application of normal distribution's is not necessary here, any symmetric function with unit integral on the real line is suitable. Such function is called a *kernel*. (Kernel is not equivalent with pdf as there is no restriction on non-negativity. However, the application of a not everywhere non-negative kernel might result in an estimated pdf that can be negative. While there are certain theoretical consideriations which justify these (Silverman 1986, pp. 66-70), no widely used kernel can be negative for this reason, hence they are in fact pdfs of – symmetric – probability distributions.)

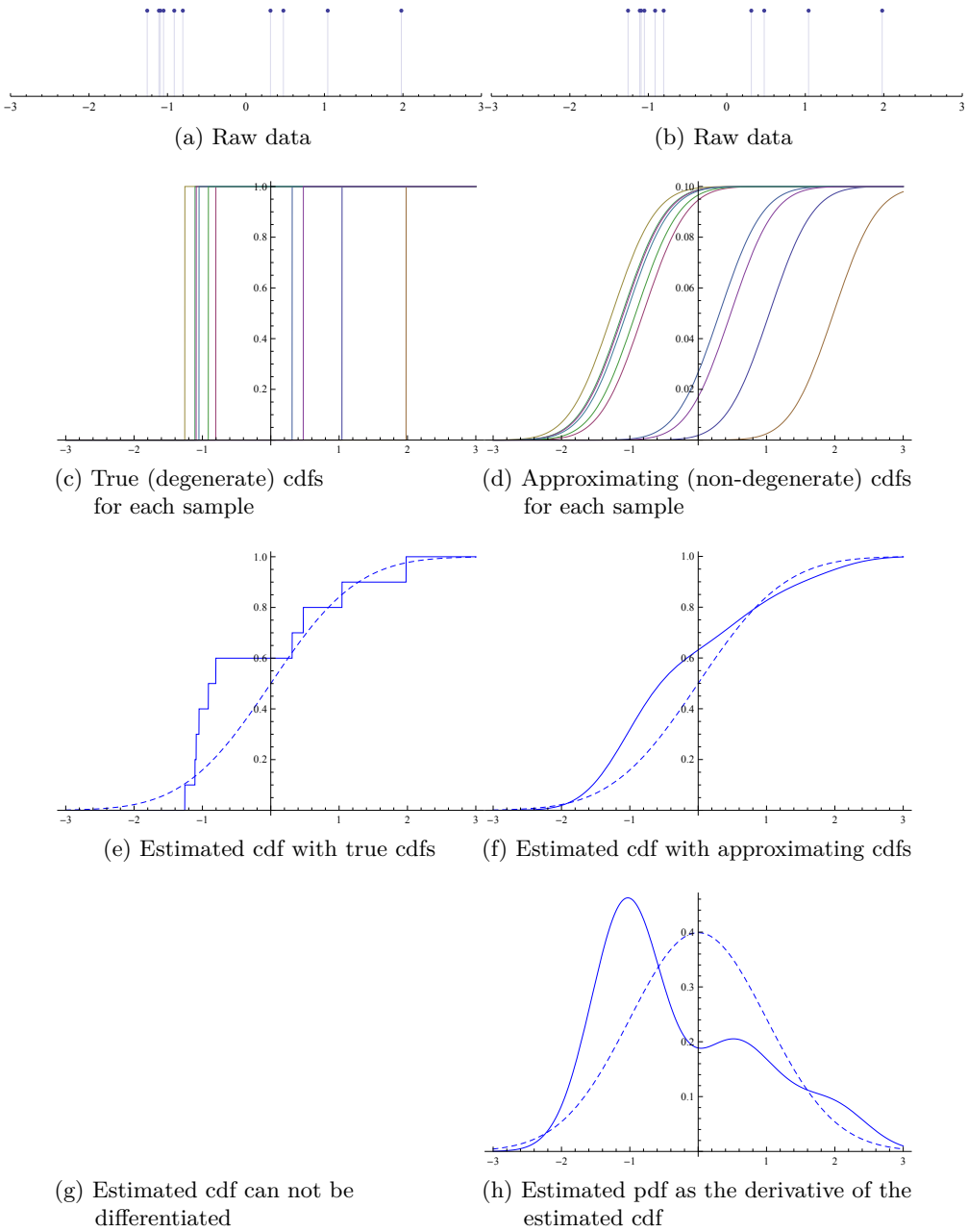
In addition to the choice of the kernel function, there is another free parameter: σ^2 (for normal kernel). The choice of this has a profound impact on the outcome of the estimation as evidenced by Figure 2.4. This also shows each kernel (which are summed to obtain the estimated pdf).

One typically wants to adjust the "steepness" of the kernel function in general, which might be not always possible so simply with a parameter of the kernel function. Hence kernel functions are usually defined for a single "steepness" (or variance, in terms of distribution theory) and are then simply linearly scaled. For the example with the normal distribution it means that the kernel is defined as

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \tag{2.8}$$

and then $K\left(\frac{x}{h}\right)$ is used as the actual kernel (which is equivalent with the above example for $h = \sigma$).

Usually the same kernel is used for every observation with the same h parameter. (This latter might not be rational if the density of the samples shows dramatic differences depending on the value of the sample itself, which gives rise to the so-called adaptive



(a) Raw data

(b) Raw data

(c) True (degenerate) cdfs for each sample

(d) Approximating (non-degenerate) cdfs for each sample

(e) Estimated cdf with true cdfs

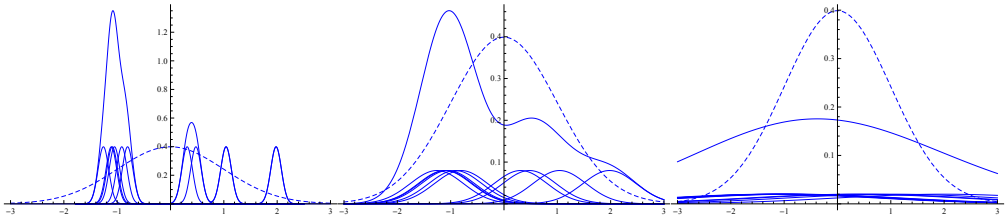
(f) Estimated cdf with approximating cdfs

(g) Estimated cdf can not be differentiated

(h) Estimated pdf as the derivative of the estimated cdf

Figure 2.3.: Kernel density estimation with normal kernels ($\sigma^2 = 1/\sqrt{2}$) for a sample from $\mathcal{N}(0,1)$, sample size $n = 10$. Dashed line shows the true value of the respective curves.

or variable bandwidth KDE, but it will not be discussed here in detail (Terrell and David W. Scott 1992; Sain 1994).)



(a) $\sigma^2 = 1/\sqrt{10}$ (b) $\sigma^2 = 1/\sqrt{2}$ (c) $\sigma^2 = \sqrt{2}$

Figure 2.4.: Effect of σ^2 (i.e. bandwidth) on the KDE. Dashed line is the true pdf, thin lines show the (scaled) kernel functions.

The (univariate) KDE can now be precisely defined. Consider a (real-valued, scalar) statistical sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then, the *kernel density estimator* from that sample, denoted with $\widehat{f}_{n,\text{kernel}}$ is the function

$$\widehat{f}_{n,h,K,\text{kernel}}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.9}$$

where $K(x)$ is an arbitrary symmetric function for which $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$ and $h > 0$.

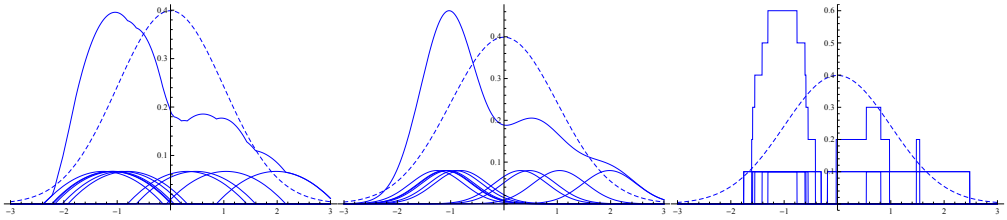
Kernel density estimation was first suggested – independently – by Rosenblatt in 1956 (Rosenblatt 1956) and Parzen in 1962 (Parzen 1962).

The h parameter is usually called the *bandwidth*, it governs the smoothness of the estimate. (It is analogous to the bin width in histogram estimate.) As Figure 2.4. demonstrates, too small bandwidth results in too much "anchoring" to the concrete sample (low bias – high variance, "undersmoothing"), while too high bandwidth causes the estimate to get almost "independent" of the sample (high bias – small variance, "oversmoothing").

The choice of kernel function does not have such profound effect on the result of KDE. There are many kernel functions that have been investigated in the literature in addition to the standard normal, for example the Epanechnikov kernel ($K(x) = \frac{3\sqrt{5}}{4}\left(1 - \frac{1}{5}u^2\right)$ for $|u| < \sqrt{5}$, zero otherwise), the triangular kernel ($K(x) = 1 - |u|$ for $|u| < 1$, zero otherwise), the rectangular kernel ($K(x) = 1/2$ for $|u| < 1$, zero otherwise) and so on (Silverman 1986). A few of them are demonstrated on Figure 2.5.

Theoretically, the Epanechnikov kernel can be shown to be optimal in terms of asymptotic efficiency (Silverman 1986, pp. 41-42), but the differences are rather small, so – especially given the asymptotic nature of this efficiency – one often chooses kernel based on other considerations, such as computational tractability.

The choice of bandwidth is, however, much more challenging. One natural way to



(a) Epanechnikov kernel (b) Normal kernel (c) Rectangular kernel

Figure 2.5.: Estimated pdfs with different kernel function, all for $h = 1/2$.

measure the error of an estimation is the application of the L_2 risk, also termed Mean integrated squared error ($MISE$) in this context:

$$MISE\left(h\right) = \mathbb{E}\int_{-\infty}^{\infty}\left[\widehat{f}_{n,h,K,\text{kernel}}\left(x\right) - f_X\left(x\right)\right]^2 \mathrm{d}x. \tag{2.10}$$

The notation $MISE\left(h\right)$ was used to emphasize that now everything will be considered fixed, except the bandwidth.

To demonstrate how this depends on h, one can show by Taylor-expansion and rearranging of terms that under mild assumptions, the following holds (M. Wand and C. Jones 1995):

$$MISE\left(h\right) = AMISE\left(h\right) + o\left(\frac{1}{nh^4} + h^4\right), \tag{2.11}$$

where $AMISE\left(h\right)$ is

$$AMISE\left(h\right) = \frac{1}{nh}R\left(K\right) + \frac{1}{4}h^4\mu_2\left(K\right)^2 R\left(f''\right) \tag{2.12}$$

standing for "asymptotic MISE". Here $R\left(K\right) = \int_{-\infty}^{\infty} K\left(z\right)^2 \mathrm{d}z$ (which is $\frac{1}{2\sqrt{\pi}}$ for the normal kernel) and $\mu_2\left(K\right) = \int_{-\infty}^{\infty} x^2 K\left(x\right) \mathrm{d}x$ (which is 1 for the normal kernel). The meaning of "asymptotic" is now clear: $AMISE\left(h\right)$ can be used instead of $MISE\left(h\right)$ if the sample size is sufficiently large (and, hence, small h can be chosen). Not only is it possible, but it is also worthy to use $AMISE\left(h\right)$ as it is dramatically easier to handle analytically. In fact, after derivation it turns out that the optimal value of h (i.e. the one that minimizes $AMISE\left(h\right)$) is

$$h^*_{AMISE} = \left[\frac{R\left(K\right)}{\mu_2\left(K\right)^2 R\left(f''\right) n}\right]^{1/5}. \tag{2.13}$$

The problem is that while $R\left(K\right)$ and $\mu_2\left(K\right)^2$ only depends on the kernel used (indeed,

they were explicitly given above for the normal kernel), $R\left(f''\right)$ also depends on the – unknown – true pdf, so this criterion cannot be directly used to estimate optimal h. Instead, other, data-driven methods are employed.

Before proceeding, let us note that the convergence rate of $AMISE\left(h\right)$ is of order $O\left(n^{-4/5}\right)$, which is better than that of histogram ($O\left(n^{-2/3}\right)$), furthermore, it can be shown that no nonparametric method can outperform KDE under mild assumptions (Wahba 1975). (Parametric methods can, of course, provide better (for example $O\left(n^{-1}\right)$) convergence, but this comes at the price of *a priori* commitment to a certain function form, as already discussed.)

The investigated problem consisted of estimating the joint distribution of a laboratory result and Z-BMI, i.e. to estimate a multivariate pdf. The theory introduced above can be directly extended to accommodate this case as well (D. W. Scott 1992). Consider a (real-valued, d-dimensional) statistical sample $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. Then, the *kernel density estimator* from that sample, denoted with $\widehat{f}_{n,\text{kernel}}$ is the function

$$\widehat{f}_{n,h,K,\text{kernel}}\left(\mathbf{x}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{K\left[\mathbf{H}^{-1/2}\left(\mathbf{x}-\mathbf{x}_i\right)\right]}{\left|\mathbf{H}\right|^{1/2}}, \tag{2.14}$$

where $K\left(\mathbf{x}\right)$ is an arbitrary symmetric function for which $\int_{\mathbb{R}^d} K\left(\mathbf{x}\right)\mathrm{d}\mathbf{x} = 1$ and $\mathbf{H}$ is symmetric and positive definite matrix.

Multivariate extension of KDE was first suggested by Cacoullos in 1964 (Cacoullos 1966).

Similarly to the one-dimensional case, the K kernel is typically chosen to be strictly non-negative (i.e. valid pdf).

The extension is straightforward: the kernel is now a multivariate function, and the parameter is not a scalar, but rather a $d \times d$ matrix. The logic, however, is unchanged: "small distributions" are placed around each observation, and these are summed to obtain the final estimate (visually: the surfaces are added and normalized).

The (multivariate) kernel function is often constructed of univariate kernel functions either by taking the product of d one-dimensional kernel functions (so-called product kernel) or applying the one-dimensional kernel function to $\mathbf{x}^T\mathbf{x}$ (so-called spherically or radially symmetric kernel) and then normalizing (M. Wand and C. Jones 1995). However, the choice of kernel function does not have a profound effect on the estimation, and as computational aspects are of even greater importance in multivariate case, K is often chosen to be (multivariate) standard normal density.

The problem of estimating a single bandwidth parameter h is now replaced by estimating

$\frac{1}{2}d\left(d-1\right)$ free parameters of the matrix $\mathbf{H}$. This is increasingly problematic in higher dimensions, hence the space of the bandwidth matrices is sometimes reduced from symmetric, positive definite to positive definite, diagonal or positive definite, scalar. Now only two- and (later) three-dimensional estimates will be needed, hence the search space will not be restricted. (Especially because while this restriction indeed makes estimation more feasible (for instance, using scalar matrix reduces the dimensionality to 1, *irrespectively* of the actual d), but can be shown to significantly degrade the performance of KDE (Chacón 2009; M. P. Wand and M. C. Jones 1993).)

Obtaining a reasonable estimate for $\mathbf{H}$ again starts with expressing $MISE$ and $AMISE$. It can be shown that under weak assumptions (Chacón, Duon, and M. P. Wand 2009)

$$MISE\left(\mathbf{H}\right)=AMISE\left(\mathbf{H}\right)+o\left(\frac{\mathrm{tr}\mathbf{H}^{-1}}{n\left|\mathbf{H}\right|^{1/2}}+\mathrm{tr}^{2}\mathbf{H}\right) \tag{2.15}$$

and

$$AMISE\left(\mathbf{H}\right)=\frac{R\left(K\right)}{n\left|\mathbf{H}\right|^{1/2}}+\frac{1}{4}\mu_{2}\left(K\right)^{2}\left(\mathrm{vec}^{T}\mathbf{H}\right)\mathbf{\Psi}_{4}\left(\mathrm{vec}\mathbf{H}\right), \tag{2.16}$$

where $R\left(K\right)=\int_{\mathbb{R}^{d}}K\left(\mathbf{z}\right)^{2}\mathrm{d}\mathbf{z}$ (which is $\left(4\pi\right)^{-d/2}$ for the normal kernel), $\mu_{2}\left(K\right)=\int_{\mathbb{R}^{d}}z_{i}^{2}K\left(\mathbf{z}\right)\mathrm{d}\mathbf{z}$ (for any i, in other words $\int_{\mathbb{R}^{d}}\mathbf{z}\mathbf{z}^{T}K\left(\mathbf{z}\right)\mathrm{d}\mathbf{z}=\mu_{2}\left(K\right)\mathbf{I}$; this is $\mathbf{I}$ for the normal kernel), $\mathbf{\Psi}_{4}=\int_{\mathbb{R}^{d}}\left[\mathrm{vec}D^{2}f_{\mathbf{X}}\left(\mathbf{x}\right)\right]\left[\mathrm{vec}^{T}D^{2}f_{\mathbf{X}}\left(\mathbf{x}\right)\right]\mathrm{d}\mathbf{x}$ with $D^{2}f_{\mathbf{X}}\left(\mathbf{x}\right)$ being the $d\times d$ Hessian of the second order partial derivatives of $f_{\mathbf{X}}\left(\mathbf{x}\right)$, and vec meaning the vectorization of a matrix by stacking its columns.

Similarly to the univariate case, $R\left(K\right)$ and $\mu_{2}\left(K\right)$ depend only on the kernel used, but $\mathbf{\Psi}_{4}$ also depends on the – unknown – density that is to be estimated. This likewise makes the direct minimization useless, for instance, even if the search is restricted to scalar bandwidth matrices $\mathbf{H}=h\mathbf{I}$ the optimum will be

$$h_{AMISE}^{*}=\left[\frac{d\cdot R\left(K\right)}{\mu_{2}\left(K\right)\int_{\mathbb{R}^{d}}n\left(\nabla^{2}f_{\mathbf{X}}\left(\mathbf{x}\right)\right)^{2}\mathrm{d}\mathbf{x}}\right]^{\frac{1}{d+4}} \tag{2.17}$$

again depending on the unknown density that is to be estimated. (Should we express optimal $AMISE$, we would obtain that its convergence rate is $O\left(n^{\frac{4}{d+4}}\right)$ which is just a manifestation of the well-known curse of dimensionality.)

Like it was mentioned at the univariate case, data-driven methods are needed to estimate the bandwidth matrix. The approach that was now used is the so-called *smoothed cross-validation* (SCV) that was introduced in 1992 (P. Hall, Marron, and Park

1992). SCV is demonstrated to be amongst the most reliable methods for estimating a full bandwidth matrix (Duong and Hazelton 2005).

To demonstrate the whole procedure, an example will be considered that is based on one of the datasets that is to be introduced in Subsection 2.3.1. For the moment, it is only important to know that we have $n = 240$ samples of boys with their Z-BMI and HDL cholesterol levels. To illustrate these parameters, Figure 2.6. shows their scattergram.

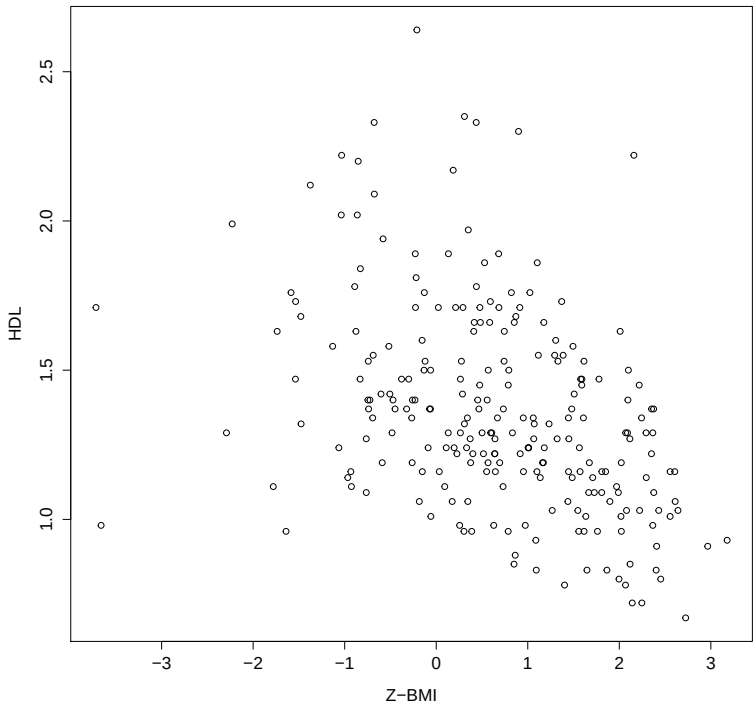


Figure 2.6.: Scattergram of the Z-BMI and HDL cholesterol of the boys from the NHANES study (see Subsection 2.3.1).

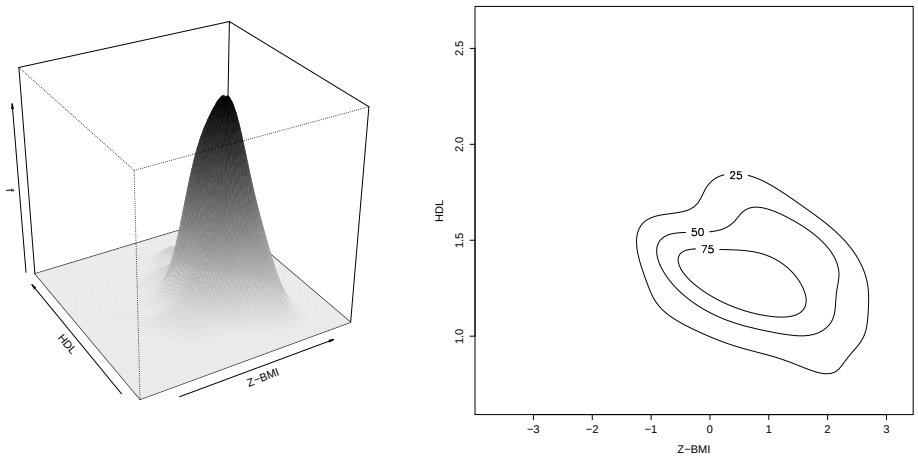
It is immediately obvious that the relationship is negative, i.e. increasing Z-BMI is associated with decreasing HDL cholesterol. (Which, of course, does not imply causation in any direction.) One can observe the difficulties induced by the fact that both variable is continuous: there is no simple way to define any statistics for a given Z-BMI (in order to, for example, characterize the typical HDL for a given Z-BMI level).

Hence the KDE that was extensively described above for $d = 2$ was performed with

normal kernel. The optimal bandwidth matrix, estimated with SCV, was

$$\mathbf{H}_{SCV} = \begin{pmatrix} 0.29056218 & -0.03589191 \\ -0.03589191 & 0.01781749 \end{pmatrix}. \tag{2.18}$$

The density estimate that was obtained using these is shown on Figure 2.7 with 3-dimensional perspective plot and – more perspicuously – contour plot.



(a) Perspective plot (b) Contour plot

Figure 2.7.: Estimated joint pdf of Z-BMI and HDL cholesterol for the boys from the NHANES study.

Once the two-dimensional (BMI vs. investigated variable) joint pdf is estimated, the distribution of the investigated parameter for any given BMI can be obtained by "slicing" the two-dimensional surface perpendicular to the BMI axis at the point of interest (e.g. Z-BMI=+1). The "slice" should be then normalized so that the area under its curve equals to 1; that is, a conditional (one-dimensional) distribution from the two-dimensional joint pdf is obtained, conditioning on Z-BMI. These are illustrated on Figure 2.8.

The required statistic can be then directly computed from the obtained pdf of the conditional distribution (for example by numeric integration in case of mean).

Univariate Association Analysis

The next issue that was addressed was the quantifying of the relationship between Z-BMI and the investigated laboratory parameter. The above examination only calculated

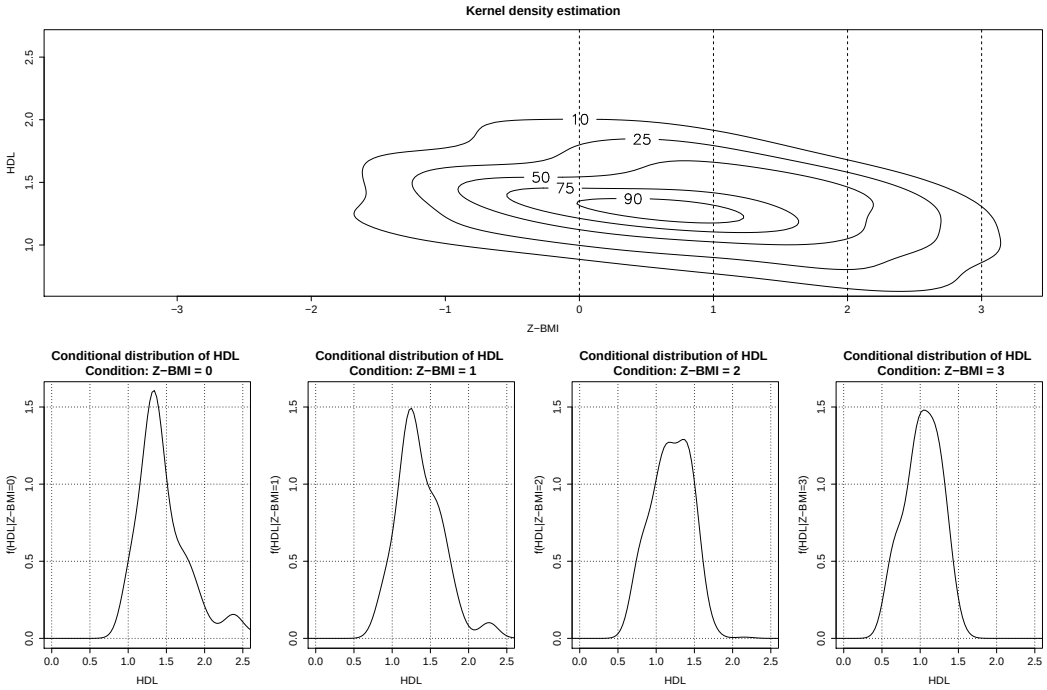


Figure 2.8.: Conditional distribution of HDL cholesterol level for different Z-BMI levels (as conditions). The position of conditions is illustrated on the joint distribution with dashed lines.

certain statistics *separately* for some notable Z-BMI values, but has not dealt with analyzing the statistical relationship between the two variables. To that end, a classical (univariate) analysis on the association of obesity with systematical alteration of different laboratory parameters was performed, parameter-by-parameter.

We had a continuous indicator of the degree of obesity (Z-BMI), therefore, instead of binning the Z-BMI values (i.e. discretizing that variable) and then using either a t-test or an ANOVA-type statistical test, we retained every value of the Z-BMI variable unchanged and calculated its correlation coefficient with the investigated laboratory result. (See the scattergram of Figure 2.6 to have an overall impression on this correlation.) The advantage of binning would have been its ability to detect non-linear relationships as well; to compensate this, Spearman-ρ (Maritz 1995) correlation coefficient was used (instead of the more classical Pearson (product-moment) coefficient), which detects monotone connections in general, and not only linear connections. (At the price of slightly smaller power (Clark-Carter 2009).) This is also justified by the already mentioned non-normality of some of the laboratory parameters, and the possible presence of outliers (Chok 2010).

Spearman-ρ is a rank correlation coefficient, that is, it is based on the ranks of

the observations not their actual value. Spearman-ρ is simply the Pearson correlation coefficient of the ranks of the two samples. In fact, if $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ denotes the two samples, then

$$\rho = \frac{\sum_{i=1}^n \left(x_i^* - \overline{x}\right)\left(y_i^* - \overline{y}\right)}{\sqrt{\sum_{i=1}^n \left(x_i^* - \overline{x}\right)^2}\sqrt{\sum_{i=1}^n \left(x_i^* - \overline{x}\right)^2}}, \tag{2.19}$$

where x_i^* and y_i^* denotes the i-th observation in ascending order within the x and y sample, respectively.

The only question that arises is the definition if ties are present. However, for continuous data (as ours) this usually poses no problem. Also, in this case the above statistic can be simply calculated as $\rho = \frac{6\sum_{i=1}^n \left(x_i^* - \overline{x}\right)^2}{n\left(n^2-1\right)}$.

By the properties of the Pearson correlation coefficient, it is immediately obtain that $-1 \leq \rho \leq +1$, with $|\rho| = 1$ if and only if there is a perfectly monotonic map between the observations of the two samples. Note that the function form itself does not matter (that's why Spearman-ρ is sometimes called non-parametric correlation coefficient), only monotonicity counts. The sign of the correlation coefficient measures the direction of the association, identically to the Pearsonian case.

Significance (for the null hypothesis $H_0 : \rho = 0$) can be calculated either by asymptotic t-approximation using the test statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ (Press 2007) or exactly, using permutation test (for very small samples) and Edgeworth expansion (David, Kendall, and Stuart 1951). Now the latter approach was used, utilizing the algorithm AS89 (Best and Roberts 1975).

For selecting the significant differences, the effects of the multiple comparisons situation (Miller 1981) also have to be taken into account. As several hypothesis testing is run in parallel (and their results are considered disjunctively, i.e. we are looking for any difference) the usage of the pre-specified significance level (e.g. 5%) for every test would result in an experimentwise α far above the significance level. In particular, if k tests are run in parallel, with H_0 being true (in the population) for *every* of them, the probability of finding *at least one* significant test – despite that – is $1 - (1-\alpha)^k$. For $k = 2$ this is 9.75%, for $k = 30$ (which is close to the realistic cases when laboratory results are considered) this is 78.5%, i.e. it is far more likely that we find a – false – significance than not finding such. (Also, consider that the number of expected false positives is αk which is above 1 for $k > 20$ when $\alpha = 0.05$.) This is the phenomenon of α-inflation.

To protect against this, a correction has to be applied. The most straightforward approach is to utilize the so-called Bonferroni- (or Boole-) inequality, which results in that $1 - (1-\alpha)^k \leq \alpha k$. Hence, using corrected significance level $\alpha' = \frac{\alpha}{k}$ guarantees that the probability of finding a false significance (i.e. experimentwise α) can not exceed α,

even if k tests are run in parallel.

This is the so-called Bonferroni-correction (Shaffer 1995). The major drawback of this correction is that it is overly conservative and also makes the detection of true effects extremely hard. (In other words, it will provide a very weak overall test.)

However, there is a way to improve this without requiring further assumptions: this is the so-called Holm–Bonferroni-correction (Holm 1979). (Improvement is understood such that Holm–Bonferroni-method dominates Bonferroni: under no circumstances is it specifically better to use Bonferroni-correction, but in many cases it is better to use Holm–Bonferroni, i.e. Holm–Bonferroni is uniformly more powerful.)

Holm–Bonferroni-correction starts by ordering the p-values and calculating the usual Bonferroni-corrected significance level: $\alpha' = \frac{\alpha}{k}$. If there is no p-value that smaller than that, every hypothesis testing is considered insignificant. However, if there is at least one, than the smallest is considered significant (evidently), but after that, this procedure is repeated with significance level $\alpha'' = \frac{\alpha}{k-1}$ (instead of using again $\alpha' = \frac{\alpha}{k}$ as in Bonferroni-correction). And so on, until a point is reached where no p-value is below the actual significance level. Remaining tests are considered insignificant. It can be demonstrated that this method provides the same strong control on the experimentwise α as the Bonferroni-correction, despite the fact the higher significance levels are used.

This was the correction that was used to judge whether the association of a laboratory parameter with Z-BMI is significant or not.

Other, even more powerful correction methods are available as well, but they either have certain assumptions only under which they are valid (such as Hommel-correction (Hommel 1988)) or do not provide a strong control on the experimentwise α (such as the False Discovery Rate (Benjamini and Hochberg 1995)). Therefore these methods were not used now.

2.2.4. Investigation of the Multivariate Structure

The investigation of the multivariate structure poses a greater challenge, even leaving the problem of varying Z-BMI aside. To practically grab the issue, it is customary to confine ourselves to linear connections, therefore reducing the problem to the investigation of the usual correlation matrix. This, however, is still problematic with the traditional tools of multivariate statistics. This problem will be exposed first.

After that, I will introduce my methodology, which is based on defining correlation matrices for given Z-BMI levels. I will call these ”conditional correlation matrices” – I am not aware of any application of this concept in such context.

Two methods of modern multivariate statistics will be then employed for the actual

analysis: Principal Components Analysis (PCA) and Cluster Analysis (CA). I will also introduce the logic of these methods.

Traditional Approaches' Shortcomings, and their Alternatives

The correlation structure of a database is directly reflected in its correlation matrix (if we confine ourselves to linear connections) and can be visualized by (matrix)scattergrams. Both methods fail to facilitate to understanding of the correlation structure if the number of variables is too high: matrix scattergrams and correlation matrices can be hardly overseen for more than 10 variables (Venables and Ripley 2002). Even for typical laboratory examinations, there are 30 (or even more) variables, so other approaches of multivariate statistics had to be employed. This is demonstrated on Figure 2.9 which shows the correlation matrix of the laboratory variables for boys in the NHANES database (for the database see Subsection 2.3.1), visualized with heatmap.

It can be visually observed how difficult it is to reveal patterns, especially those that involve several variables. Matrixscattergram is virtually impossible to be meaningfully plot for this case.

I employed two modern methods to capture the multivariate structure of such, high-dimensional data: Principal Component Analysis (PCA) and Cluster Analysis (CA). Both methods can be used to ease the understanding and interpretation of the correlation structure of large databases. Note that these methods do not require the database itself, only its correlation matrix, so the first task is to define the correlation matrix, now taking the effect of Z-BMI into account.

Conditional correlation matrices

The first task is therefore to define a correlation matrix between the laboratory parameters for a given Z-BMI. The logic that will be followed will be the same as in the univariate case: density estimation (KDE) will be used to obtain an estimate of the joint pdf, condition on Z-BMI to obtain a conditional pdf and calculate the correlation matrix from this conditional pdf. I will call this "conditional correlation matrix".

The direct attack of this problem is infeasible: the maximum dimensionality of a joint pdf that can be sensibly estimated from realistic sample sizes is about 5 (due to the curse of dimensionality). Having 30-40 or even more laboratory parameters, it is downright impossible to estimate a joint pdf of all laboratory parameters and Z-BMI, from which the correlation matrix (after conditioning) could be calculated. However, a very simple observation helps: a correlation matrix does not need to be estimated

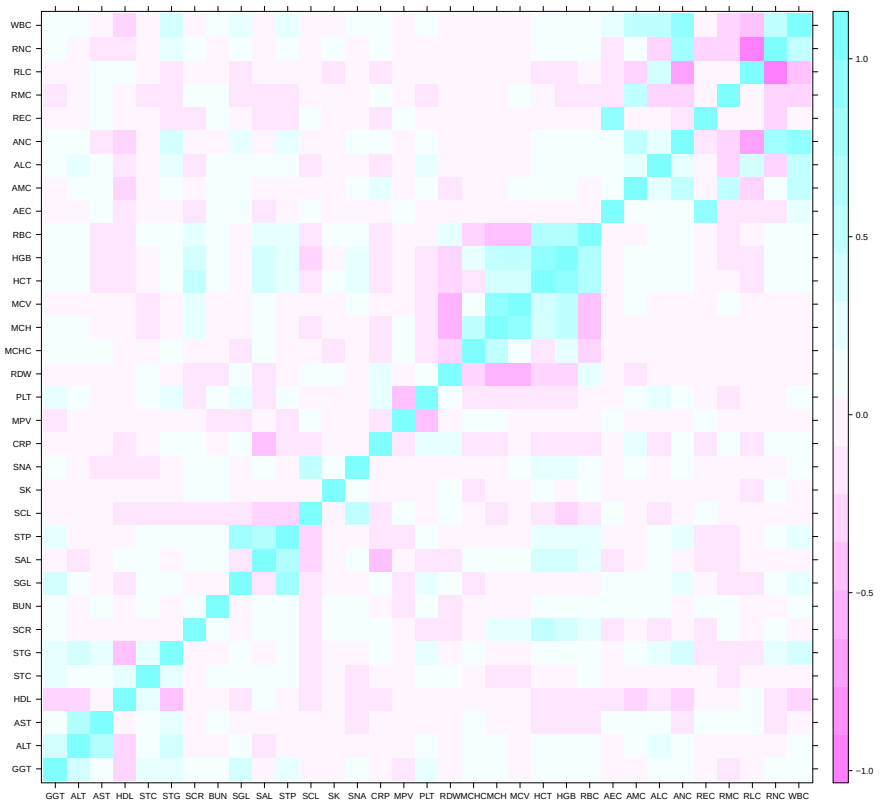


Figure 2.9.: Correlation matrix of the laboratory results the boys from the NHANES study (see Subsection 2.3.1) visualized with heatmap.

"in whole" (by matrix multiplication) – of course, we get to the very same matrix if we estimate it element-by-element (i.e. we estimate the pairwise correlations). Thus, for k laboratory parameter, it is not necessary to estimate a $(k+1)$-dimensional pdf (laboratory parameters and Z-BMI), rather, it is always sufficient to perform a three-dimensional pdf estimation (*irrespective* of k!), from which one element of the conditional correlation matrix can be calculated, and repeat this $\frac{k(k+1)}{2}$ times. With this logic, the curse of dimensionality can be broken in this case.

The only problem with this approach is that this way the elements of the correlation matrix are approximated from different samples, hence, the resulting matrix is not necessarily positive semidefinite (as a correlation matrix should be). This would prevent the performing of PCA (or any other method that expects a valid correlation matrix as input), so smoothing (Wothke 1993; Jäckel 2002) was applied to reconstruct a closely

approximating positive semidefinite matrix from the correlation matrix by eliminating negative eigenvalues and rescaling positive ones. (This is acceptable in this case, because in practice it will only have few negative eigenvalues, with very small absolute values.)

In more detail, we take the approximated correlation matrix of l variables $\mathbf{C}_{l\times l}$ and calculate its spectral decomposition (Poole 2010). (It surely exists as $\mathbf{C}$ is a real and symmetric matrix, even if the above approach is to construct it – these two properties are surely preserved, even under element-by-element reconstruction). The factorization is

$$\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T. \tag{2.20}$$

This form incorporates the fact that the eigenvectors of a real symmetric matrix are orthogonal.

Now, let us detail the eigenvalues:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{n_+} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_{n_+-1} & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_{-n_-+1} & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \lambda_{-n_-} \end{pmatrix}, \tag{2.21}$$

where λ_i denotes a nonnegative eigenvalue if $i > 0$, a negative if $i < 0$, and $n_+ + n_- = l$.

We now try to eliminate the negative eigenvalues. A logical solution is to simply replace every λ_i with 0 for $i < 0$, but this alters the sum of the eigenvalues. (Which is the trace of the original matrix, hence – as it is a correlation matrix – should be fixed, namely the number of variables, l.) To prevent this, we will simply linearly rescale every eigenvalue so that their sum becomes l.

More precisely, the negative eigenvalues are replaced with a very small positive number (instead of zero), so that the matrix will be positive definite and not only positive semidefinite. By denoting this number with ε, a scaling factor of $\frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}$ is obtained. Using this, and the transformation defined above, we get the following modified eigenvalue-

matrix:

$$\widetilde{\mathbf{\Lambda}} = \begin{pmatrix} \frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}\lambda_{n_+} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}\lambda_2 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}\lambda_1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}\varepsilon & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & \frac{l}{\varepsilon n_- + \sum_{i=1}^{n_+} \lambda_i}\varepsilon \end{pmatrix}, \tag{2.22}$$

This will be in turn used to reconstruct an approximating matrix – which, however, will be positive definite. Namely, we will use the matrix:

$$\widetilde{\mathbf{C}} = \mathbf{Q}\widetilde{\mathbf{\Lambda}}\mathbf{Q}^T. \tag{2.23}$$

in the further analyses.

It can be shown (Gill, Murray, and Wright 1981) that the quality of the approximation defined above can be limited in sense of matrix distance, namely

$$\left\|\widetilde{\mathbf{C}} - \mathbf{C}\right\|_F \leq 2\left|\lambda_{-n_-}\right|,$$

where $\|\cdot\|_F$ is the usual Frobenius norm of a matrix (Poole 2010). In the light of the remark that for our case, negative eigenvalues tended to be very small in magnitude, this statement justifies the application of this simple procedure. Note however, that several other, more sophisticated procedure is available for the same end (Higham 2002; Higham 1989; Higham 1988; Knol and Berge 1989; Cheng and Higham 1998).

Finally, let us note that we actually used correlation (and not covariance) matrices because laboratory parameters have different measurement scales. Using correlation matrices is equivalent to standardizing the dataset, which removes their scale-dependence.

Principal Components Analysis

Principal Components Analysis (PCA) is one the most classical tools of multivariate data analysis (Flury 1997). It can be employed (and interpreted) in many different ways; now I will use it as a tool to ease the understanding the structure of a correlation matrix. It should be again emphasized that for our purpose, the whole procedure will be introduced as a method to analyze the correlation (and not the covariance) matrix of the dataset, which is – as it will be shown – equivalent to using the standardized database. This is

necessary as the variables of our database are scale-dependent (with different units of measurement) which should not have an effect on the result.

To have an overall picture of how this is possible, let us first introduce some details of PCA (Jolliffe 2002).

Essentially, PCA forms linear combinations of the original variables so that the resulting variables will be more optimal than the original ones (in some, defined sense of "optimality"). That is: if $\mathbf{X}_{n\times p}$ denotes the data matrix (with n observations each being a p dimensional real vector) a principal component is simply a new vector (variable) $\mathbf{Xv}$. So indeed, a principal component is a (linear) mixture of the original variables.

Those more inclined toward the geometrical interpretations of linear transformations immediately notice that if we prescribe that $\mathbf{v}$ is of unit norm than this is nothing else than projecting the observations from a p-dimensional space to a line in that space (defined by the unit vector v), i.e. we perform a coordinate change. (Every element of $\mathbf{Xv}$ will be a dot product of the observation's p dimensional vector and $\mathbf{v}$.)

As usual, we can make it a complete change of basis by introducing further axes of projection, denote them $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$, which are the column vectors of the matrix $\mathbf{V}$. We get to a usual Cartesian coordinate system if the axes are orthogonal, i.e. every vector is orthogonal, in other words, $\mathbf{V}$ is orthogonal (and also orthonormal). An orthogonal matrix is invertible, so we can change between the coordinate systems in both directions.

Sticking to the graphical interpretation of the above transformation, we can illustrate the rationale of this change of coordinate system. Consider a two-dimensional case where the observations are distributed as shown on Figure 2.10. The original, and the transformed Cartesian coordinate systems are shown in black and red.

When both coordinates are used, the two representations are equivalent. (As evidenced by the invertible transformation matrix.) However, if we are to perform a dimension reduction (in this case: represent the data in one dimension) this is no longer the case! If the observations are projected to the first axis of the original coordinate system, the loss in information will be much greater than by the projection to the first axis of the transformed coordinate system. (I will later precisely define what "loss in information" means, in the meantime, consider its intuitive interpretation.) This is the sense in which the transformed representation is better: it has rearranged the axes so that when dimension reduction is performed, the loss in information can be minimized (as opposed to the original representation).

Note that this transformation implies that the new coordinate system's origin is the same as the original's (i.e. only a rotation of the coordinate system is performed), hence this transformation only makes sense, and the aforementioned optimality can only be

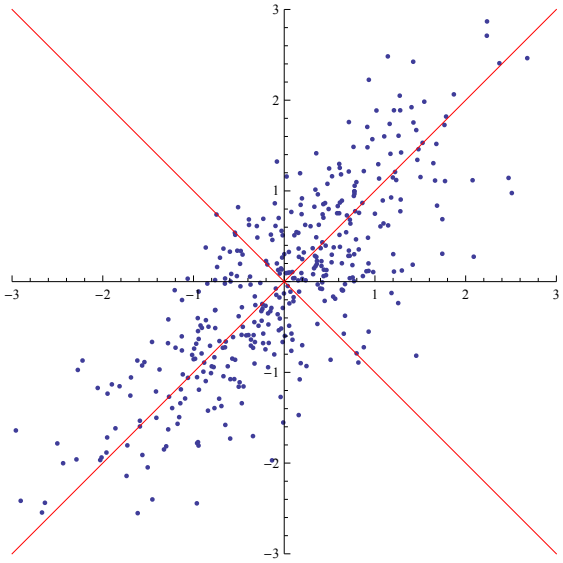


Figure 2.10.: Visual illustration of the logic of Principal Components Analysis.

reached for dimension reduction if the database is centered. (Otherwise, the direction of the principal component might simply depend on the mean of the variables.) Therefore, to perform PCA, the database has to be first centered, i.e. the mean should be subtracted from each variable (so that every variable's mean will be zero).

Now let us more precisely elaborate what is meant by "loss in information". Consider our situation where the database is not only centered, but also standardized (i.e. the mean is subtracted from each variable and divide by the standard deviation), denote this with $\mathbf{X}_z$. By this, variables will be not only of zero mean, but of unit variance – for every axis. However, this only stands for the original axes: we might still find axes for which the projection is not of unit variance, see Figure 2.10. Finding an axis for which the projection has a variance higher than 1 means that the information of the original database is better represented on that axis, because after standardization, variance measures information conveyed by the axis in some sense. (I will shortly make this more precise.) In fact, the first principal component will be the solution of the following optimization problem:

$$\max_{\|\mathbf{v}_1\|=1} \mathbb{D}^2\left(\mathbf{X}_z\mathbf{v}_1\right). \tag{2.24}$$

The solution of this is not especially complicated. As $\frac{1}{n}\mathbf{1}^T\left(\mathbf{X}_z\mathbf{v}_1\right) = \frac{1}{n}\left(\mathbf{1}^T\mathbf{X}_z\right)\mathbf{v}_1 =$

$\mathbf{0}^T\mathbf{v}_1 = 0$ (i.e. the variable is of zero mean), we have

$$\mathbb{D}^2\left(\mathbf{X}_z\mathbf{v}_1\right) = \frac{1}{n}\left(\mathbf{X}_z\mathbf{v}_1\right)^T\left(\mathbf{X}_z\mathbf{v}_1\right) = \frac{1}{n}\mathbf{v}_1^T\mathbf{X}_z^T\mathbf{X}_z\mathbf{v}_1 = \mathbf{v}_1^T\mathbf{C}\mathbf{v}_1, \tag{2.25}$$

where $\mathbf{C}$ denotes the correlation matrix of the original data.

We solve the optimization problem by the method of Lagrange multipliers (Bertsekas 1996). Define

$$\Lambda\left(\mathbf{v}_1, \lambda\right) = \mathbf{v}_1^T\mathbf{C}\mathbf{v}_1 - \lambda\left(\mathbf{v}_1^T\mathbf{v}_1 - 1\right), \tag{2.26}$$

and then calculate its partial derivatives (Petersen and Pedersen 2006):

$$\frac{\partial\Lambda\left(\mathbf{v}_1, \lambda\right)}{\partial\mathbf{v}_1} = 2\mathbf{C}\mathbf{v}_1 - 2\lambda\mathbf{v}_1 \tag{2.27}$$

and

$$\frac{\partial\Lambda\left(\mathbf{v}_1, \lambda\right)}{\partial\lambda} = \mathbf{v}_1^T\mathbf{v}_1 - 1 \tag{2.28}$$

Setting the first to zero, we obtain

$$\left(\mathbf{C} - \lambda\mathbf{I}\right)\mathbf{v}_1 = 0 \tag{2.29}$$

which is simply the eigenvalue/eigenvector problem (Poole 2010) for the matrix $\mathbf{C}$. Thus, λ has to be an eigenvalue of the correlation matrix, while $\mathbf{v}_1$ is an eigenvector.

The only question remains is to decide which eigenvalue/eigenvector to use. For this, note that if $\mathbf{v}_1$ is an eigenvector of $\mathbf{C}$, then the objective function is

$$\mathbf{v}_1^T\mathbf{C}\mathbf{v}_1 = \mathbf{v}_1^T\lambda\mathbf{v}_1 = \lambda, \tag{2.30}$$

by the constraint (partial derivative with respect to λ) with λ being the corresponding eigenvalue.

Therefore the first principal component can be formed by using the (normalized) eigenvector of the correlation matrix corresponding to the largest eigenvalue as a weighting vector. It was also deduced that the variance of this principal component will be just the largest eigenvalue.

Further principal components can be calculated by imposing the additional condition of orthogonality to the previous principal components. (Which is equivalent to saying that the weighting vectors are orthogonal as $\mathbf{v}^T\mathbf{A}\mathbf{v} = 0 \Leftrightarrow \mathbf{v} = 0$ if $\mathbf{A}$ is not singular, which might be presumed for a correlation matrix, by assuming linearly independent variables.) Namely, for the kth principal component, we have to solve the optimization

task

$$\mathbf{v}_k = \max_{\substack{\|\mathbf{v}_k\|=1 \\ \forall j<k: \mathbf{v}_j^T \cdot \mathbf{v}_k = 0}} \mathbb{D}^2 \left(\mathbf{X}\mathbf{v}_k \right). \tag{2.31}$$

The solution essentially goes along the same lines with slightly more complicated algebra. The details will not be discussed here; the results will be analogous: the kth principal component can be formed by the (normalized) eigenvector of the correlation matrix corresponding to the kth largest eigenvalue, and the variance of the principal component will be that eigenvalue.

This procedure can be effectively executed on computer in practice using for example singular value decomposition (Trefethen and Bau 1997).

Accepting that the variance is some way measuring the information, it can be seen that the overall information in the original database is n. Consistently with what was said about the transformed database being equivalent, it is also the information in the transformed variables, as the sum of the eigenvalues in a matrix equals it trace, in this case $\sum_{i=1}^{p} \lambda_p = \mathrm{tr}\mathbf{C} = p$. But, a variable with eigenvalue (variance) larger than 1 carries more information than an original variable; for example, for the first component it is $\frac{\lambda_1}{n}$. This is also called the explained variance of that component. The explained variance of the first two principal components is $\frac{\lambda_1+\lambda_2}{n}$ and so on.

Now, let us return to the more precise definition of information loss in the context of dimension reduction. As already discussed, using the original p-dimensional database and using p principal components is essentially equivalent, as it simply represents a change of basis. Hence, there is no information loss associated with this transform (no matter how it is defined), as exactly the same points are represented by both. Dimension reduction means that we represent the database using $p' < p$ coordinates, with minimizing the "information loss" that is induced by using fewer axes. Now it is time to introduce the precise definition of this "information loss": we will measure this by the average squared Euclidean distance between the original points and the points that were reconstructed using the fewer coordinates. This latter is understood as projecting the observations to the $p' < p$ axes used in dimension reduction, and then expressing their coordinates in the original coordinate system. That is, if we use the notion $\mathbf{V}_{p'}$ for the matrix formed from the first p' columns of $\mathbf{V}$, then the coordinates in the original coordinate system of the database reduced to p' dimensions using PCA is obviously

$$\mathbf{X}_z \mathbf{V}_{p'} \mathbf{V}_{p'}^T. \tag{2.32}$$

Thus, the error using the criterion defined above for $p' < p$ dimensions is

$$C\left(p'\right) = \left\|\mathbf{X}_z - \mathbf{X}_z \mathbf{V}_{p'} \mathbf{V}_{p'}^T\right\|^2. \tag{2.33}$$

And now comes the interesting theorem: it can be shown that PCA is the method that minimizes this criterion for *every* $p' < p$ dimension reduction among *all* linear transformation of the database! In other words, by transforming the database with PCA, we get an optimal representation in a sense that no matter how large dimension reduction we want to achieve the best will be (among linear transformations of the database) to use the first few principal components.

Let us close this discussion with two concluding remarks. First, basing PCA on the correlation matrix and using the standardized database are essentially synonymous. We might perform this spectral decomposition on the covariance matrix, this would be equivalent to using a centered (but not standardized) database. As already mentioned, this only makes sense if the data are measured on the same scale. If not (as in this case) we should use the correlation matrix (i.e. standardize the database) to eliminate the scale-dependence of the variables. Second, let us note that PCA could be introduced on purely probability theory grounds (as opposed to the "sample approach" that was used above). That would mean the investigation of the random variable $X\mathbf{v}$, where X is a p-dimensional vector random variable of some specified distribution (typically: p-variate normal). The results are analogous to those in the above discussion, we will not elaborate this in more detail.

At this point, it is natural to ask how this all is related to understanding a correlation matrix, the task that set forth. The answer is simple: the coefficients of the weighting vectors (i.e. the columns of the $\mathbf{V}$ matrix) shed light on the connections between the original variables. (Precisely: usually not the $\mathbf{V}$ matrix is used, but $\mathbf{V}\mathrm{diag}\left(\langle\lambda_1, \lambda_2, \ldots, \lambda_p\rangle\right)^{-1/2}$, which can be easily demonstrated to contain the correlations between the original variables and the principal components, and is sometimes called the loading matrix.) More specifically: those variables that are highly correlated (in absolute value) with the same principal component, are also correlated among themselves; hence, instead of searching for high correlations within a $p \times p$ matrix (where p denotes the number of laboratory parameters), it is sufficient to consider p values (a column of the loading matrix) at a time and repeat this p times, even if all principal component is considered. This essentially means that the problem can be decomposed into subtasks that are much easier to solve. Furthermore, as principal components are in the order of decreasing importance (based on the error that occurs when the last principal components are omitted in the

reconstruction of the observed variables), it is usually enough to consider the first few columns of the loading matrix in many practical case. This way, information may be obtained even from large databases on what variables exhibit statistically connected behavior, which might indicate causal (physiological, this time) connections between them.

It is clear from the above description – as the medical "meaning" of a principal component is given by those original variables that are highly correlated with that component – that for interpretation purposes, the best is if each column of the loading matrix is the most 'polarized', i.e. the coefficients in it are either close to ± 1 or to 0, but not in between. To achieve that end, a so-called rotation is usually applied. Rotation means the transformation of the principal components with an orthogonal matrix – it does not change the overall variance explained, but redistributes it among the principal components. One of the most popular techniques is the varimax rotation (Kaiser 1958), which has the variance of the squares of the coefficients in the columns of the loading matrix as objective function, and maximizes it by rotation.

Now PCA was used purely to transform the correlation matrix, i.e. it was applied in descriptive sense, with no inductive statistics performed. Because of this, neither the calculation of the KMO measure, nor any hypothesis testing was performed.

PCA was performed for the conditional correlation matrices for Z-BMI=+1, +2 and +3 (obtained as described above), consistently with what was done in case of the univariate descriptors.

To ease the interpretation of the loading matrices, varimax rotation was applied after performing PCA to achieve a well-interpretable component structure. The number of extracted components was set to 13, this was selected to support interpretability and also to ensure that those components are extracted that have an eigenvalue larger than 1 (Kaiser's criterion (Kaiser 1960)).

Cluster Analysis

Cluster Analysis (CA) aims (Tan, Kumar, and Steinbach 2007) to form groups of data objects such that objects within a group are similar to each other, while objects in different groups are dissimilar (according to some prespecified metric of similarity). Such groups are called clusters in the context of CA. Obviously, there is a trade-off between the two considerations: we can maximize the within-group similarity if we consider every object a cluster itself, but this is usually a very poor solution in terms of between-group dissimilarity. On the other hand, between-group dissimilarity is the best if every objects belongs to the very same cluster, but this is usually a very poor solution in terms of

within-group similarity. Thus, a compromise has to be found between the two extrema.

The choice of metrics that define similarity between objects is limited by the information that is available about the objects. For example, if objects are described by real vectors, Euclidean distance or L^p distances in general might be used, for count data (i.e. integer vectors), χ-squared metric is a popular choice, while for binary data, there is plenty of widely used metrics, such as the Jaccard index.

Either way, after applying the similarity metric, we end up with a similarity (or dissimilarity) matrix $\mathbf{D}$, which will we the input of CA. d_{ij} is similarity (or dissimilarity) of the data objects i and j. This matrix can be arbitrary (except, of course, that it has to be a valid distance matrix). The application of similarity or dissimilarity does not mean a profound difference, they can be converted to each other, albeit not unambiguously.

There are several philosophically different approaches to CA, the two classical (and most important) being hierarchical clustering and k-means clustering. Additional modern methods include density-based clustering (Kriegel et al. 2011) and spectral clustering (Luxburg 2007).

Now hierarchical clustering will be used, which is based on the stepwise forming of clusters, with no need to a priori define the number of clusters to be formed.

More specifically, the so-called agglomerative hierarchical clustering will be used, where every data object is considered a cluster in itself at the beginning, then they are merged according to a logic we will immediately describe, until every data object belongs to the same cluster. Hence, we "iterate through" every possible compromise between the two considerations of CA. There is no need to an a priori commitment, the results will include the whole spectrum.

To elaborate the details (B. Everitt and Hothorn 2011; B. S. Everitt et al. 2011): in every step, the algorithm merges the two closest clusters. (A cluster can be either a data object or a group of objects.) At the beginning, every data object is a cluster in itself. The only open question that has to be addressed before the mechanics of the algorithm is fully specified and can be run, is the definition of the "closest" clusters, i.e. the definition of the distance of two clusters (because so far we only defined distances between objects, but not between clusters). This is a trivial problem only in the first step, but after that, the question of defining distance between an object and a true cluster (i.e. several objects) or between two true clusters arises.

Several strategies (called linkage criteria) are used in practice. Two cluster's distance might be called the maximum distance between the objects in the clusters (complete linkage), the minimum distance (single linkage), the average distance (mean linkage), but other, more sophisticated solutions are also possible.

I have already shown [F-8] that the results in this task are not sensitive to the distance metric and the cluster definition that is used, hence, I employed the popular Ward's method as cluster-distance definition and 1 minus the absolute value of correlation (Glynn 2005) as a distance (dissimilarity) measure. (This distance measure is equivalent to the well-known Euclidean distance if standardized variables are used, save for a constant factor depending only on the number of coordinates.) At this point, it worth noting that the only data-specific input that agglomerative hierarchical CA with Ward's method and the above distance metric needs is the **D** matrix, i.e. it does not use the original data itself, conforming to what was already said about currently employed algorithms only requiring a correlation matrix.

While the distance metric is clear from the above definition, the linkage criterion worth describing in more detail. Ward's method was first described in 1963 (Ward 1963). It aims to minimize the within-group variance, or – as it is usually stated – the within-group error sum of squares. This is defined as the sum of the error sum of squares in each cluster, where the latter is understood as the sum for each coordinate:

$$ESS = \sum_c ESS_c = \sum_c \left[\sum_{i=1}^{n_c} \sum_{j=1}^{p} \left(x_{c,i,j} - \overline{x_{c,j}} \right)^2 \right]. \tag{2.34}$$

In other words, the error sum of squares in a cluster is the trace of the covariance matrix of that cluster.

It is obvious that ESS can only increase when clusters are merged (as the new cluster centroid will not be minimizing the sum of squares for the merged clusters). Hence, the precise objective function of Ward's method is to choose that merging which minimizes the increase in ESS.

More importantly, it can be demonstrated that the aforementioned increase is proportional to the squared Euclidean distance between the centroids of the merged clusters. Nevertheless, the method is not equivalent to merging the clusters with the closest centroids (so-called centroid clustering), because Ward's method also implies a weighting of the centroids based on the number of samples in the clusters.

Ward's method can be implemented within the framework of the Lance–Williams-algorithm (Lance and W. T. Williams 1967), which was the approach that was applied in the developed methodology.

Using this apparatus, CA was performed for the conditional correlation matrices for Z-BMI=+1, +2 and +3 (obtained as described above), consistently with what was done in case of the univariate descriptors.

The "data objects" that were clustered were not the cases (as it is typical in clustering), but the variables, i.e. the laboratory parameters. Hence "similarity" means statistically connected behavior (based on the correlation matrix). That is: we are aiming to identify groups of laboratory parameters that exhibit similar behavior. (This task is essentially the same as the one we set forth with PCA, but the approach to solve it is completely different.)

For the current purpose, the most useful representation of the results of an agglomerative hierarchical CA is the dendrogram. Variables are connected with lines on a dendrogram, each connection having a so-called "height" that is measured on the vertical axis. One makes a compromise between within-group similarity and between-group dissimilarity by choosing a "threshold" on the vertical axis: for a given threshold, those variables will be in the same cluster that are connected below the threshold. This reflects what was said about hierarchical clustering not requiring an a priori commitment in the compromise.

This also means that the smaller the height is, the variables are more similar. Thus, the variables or clusters of variables that are more "deeply" connected are more similar. This way, groups of similar variables can be formed for any minimum of similarity (that is reflected as a height at which connections are "cut off", as already described). This way, the dendrogram can be considered as a graphical interpretation of the correlation matrix.

2.3. Clinical Interpretations for the Effects of Obesity on Laboratory Parameters

I have provided clinical interpretations for the effects of obesity on laboratory parameters based on a representative international survey and a non-representative survey that was performed on Hungarian adolescents specifically for the aims of the present investigation. I have discussed results pertaining to both the uni- and the multivariate structure of the investigated variables.

Using the methodology and program developed for this purpose (and described in Section 2.2), real-life data of adolescents were analyzed to investigate the connection between overweight/obesity and laboratory parameters. Findings were then interpreted from physiological and pathophysiological perspective and evaluted clinically.

Relevant own publications pertaining to this thesis: [F-3; F-15; F-9; F-1; F-4; F-21; F-2; F-11; F-5; F-12; F-8; F-7; F-6; F-13; F-18; F-19; F-20; F-17].

To improve the robustness, two datasets were used. One is from a non-representative Hungarian study we performed specifically for this purpose, the other is the appropriate part of the large representative US survey called NHANES. These will be introduced in Subsection 2.3.1.

After the databases are introduced, I present the results first for the univariate analyses of them (Subsection 2.3.2), and then for the analysis of their multivariate structure (Subsection 2.3.3).

The source code presented in Section 2.2 (listed in Appendix, Chapter A) also includes the necessary parts to perform the concrete calculations on the above two databases. Details on the program are given in the referenced Section, it will not be discussed here.

2.3.1. Databases Used

Two independent datasets of adolescents were used, based on which the effect of obesity on laboratory results was investigated using the methodology I developed for this purpose, described in Subsection 2.3.1. Both datasets consist of laboratory results, age, body mass and body height of children aged between 12 and 18 years, as my aim was specifically to address pediatric obesity due to its high public health significance (Ebbeling, Pawlak, and Ludwig 2002; Deckelbaum and C. L. Williams 2001; Must, Strauss, et al. 1999), which I have already discussed in Subsection 2.1.1. The content of the databases conforms the specification prescribed in Subsection 2.2.1.

Two datasets were used to ensure robustness. The first database is part of the representative US survey called NHANES. This was used due to its large sample size, carefully designed sampling plan, and representative nature, which provides us the opportunity to perform an investigation with reliable external validity. I, however, also aimed to investigate Hungarian population as well. Unfortunately, no survey was available which contained the necessary data, so we organized an own survey for this purpose. Due to its non-representative nature, we should be careful when drawing conclusions from this database, nevertheless, I included it here as a pilot study, so that Hungarian population is also addressed.

NHANES

The National Health and Nutrition Examination Survey (NHANES) is a nation-wide survey (Centers for Disease Control and Prevention, National Center for Health Statistics 2013a) performed annually in the United States by the National Center for Health Statistics (NCHS), division of the Centers for Disease Control and Prevention (CDC).

NHANES has a complex survey design to ensure its representativeness for the civilian non-institutionalized US population (Centers for Disease Control and Prevention, National Center for Health Statistics 2006). The amount of collected data is tremendous, including demographic data, physical examination, collection of laboratory parameters, and an extremely thorough questionnaire concentrating on anamnesis and lifestyle.

Results are published on a biannual basis and are publicly available. For the current analysis, data from the 2009-2010 Continuous NHANES cycle's database (Centers for Disease Control and Prevention, National Center for Health Statistics 2013b) was used being in accordance with the time period of the Hungarian measurements. (These are also the latest available NHANES data, as datasets for the 2011-2012 cycle have not yet been published (Centers for Disease Control and Prevention, National Center for Health Statistics 2013c).)

For the purpose of the present investigation, the NHANES dataset was filtered to subjects aged between 12 and 18 years. The so-called 'Laboratory' dataset was used (consisting of the laboratory parameters). Concrete laboratory parameters were selected to match those that were available in the Hungarian study, and included the following: 'Standard Biochemistry Profile' (`BIOPRO`), 'Complete Blood Count with 5-Part Differential in Whole Blood' (`CBC`), 'C-Reactive Protein' (`CRP`), 'HDL-Cholesterol' (`HDL`) and 'Triglycerides and LDL-Cholesterol' (`TRIGLY`). Basic anthropometric data (for BMI) was available from the 'Body Measures' (`BMX`) dataset of the 'Examination data', while basic demographic data (for age) was available from the 'Demographic Variables and Sample Weights' (`DEMO`) dataset of the 'Demographics data'.

After the above filtering, the NHANES dataset had a sample size of $n = 440$ (with 200 females and 200 males). The distribution of their Z-BMI scores for both females and males are shown in Figure 2.11.

The concrete laboratory parameters that were used together with their units of measurement and abbreviations are shown in Table 2.1.

To account for the complex survey design of the NHANES, sample weighting was applied as per the Analytic and Reporting Guidelines of the NHANES (Centers for Disease Control and Prevention, National Center for Health Statistics 2006).

Hungarian study

What I call the „Hungarian study" is a Hungarian cross-sectional, multicenter clinical observation that we arranged specifically for this investigation.

We collected data from subjects aged between 12 and 18 years, including both healthy volunteers and clinically obese ones. Sampling was done independently in the two groups

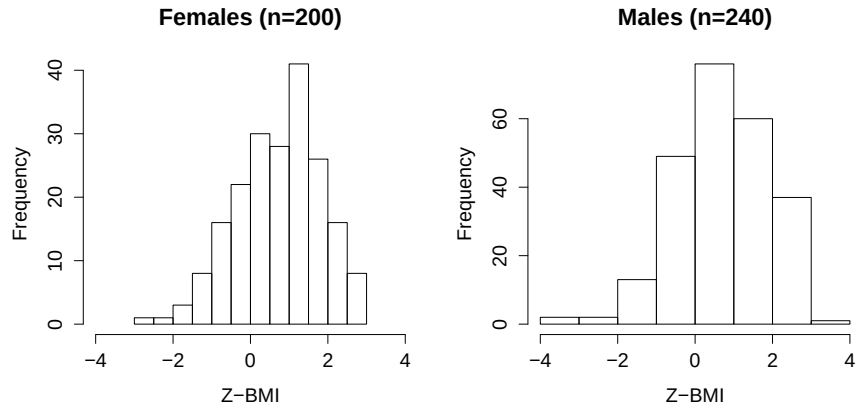


Figure 2.11.: Distribution of the Z-BMI scores in the NHANES database for both females and males.

up to a pre-specified quota, so the sample sizes are not representative for the prevalence of obesity. Clinically obese subjects were intentionally included to oversample the region of severe obesity.

The healthy control group consisted of volunteers from four Hungarian secondary schools, three of them being located in the capital city (Budapest), one in a rural town (Mátészalka). Subjects were selected as a convenience sample, so results are not necessarily representative at national level. Each child participated with full written informed consent from their parents and the study was pre-authorized by the Hungarian Regional Bioethical Commission. The data was collected between April 2008 and May 2009. Examinations of healthy volunteers included anthropometric measurements, body composition analysis (with InBody 3.0 multi-frequency bioelectric impedance analyzer), fasting blood sample drawing for standard laboratory parameters and anamnestic data recording. Measurements were carried out by physicians of the Heim Pál Children's Hospital (Budapest) and results were manually recorded in electronic format [F-4]. From these data, the anthropometric, demographic and laboratory parts will be used now.

The obese group consisted of children treated in the Heim Pál Children's Hospital, with their main diagnosis being E66.9 (according to ICD-10) "Obesity, unspecified" with no significant comorbidity. Data (including laboratory parameters) of the obese children were extracted from the hospital's electronic records with a custom application developed by the authors as discussed in [F-4].

From these data, we again used the laboratory parameters, age, body mass and body height. The concrete laboratory parameters that were used together with their units of

measurement and abbreviations are shown in Table 2.1.

The Hungarian database consists of $n = 183$ subjects (113 males, 70 females). The distribution of the BMI z-scores of these subjects is shown in Figure 2.12. (The oversampling of obese population is obvious.)

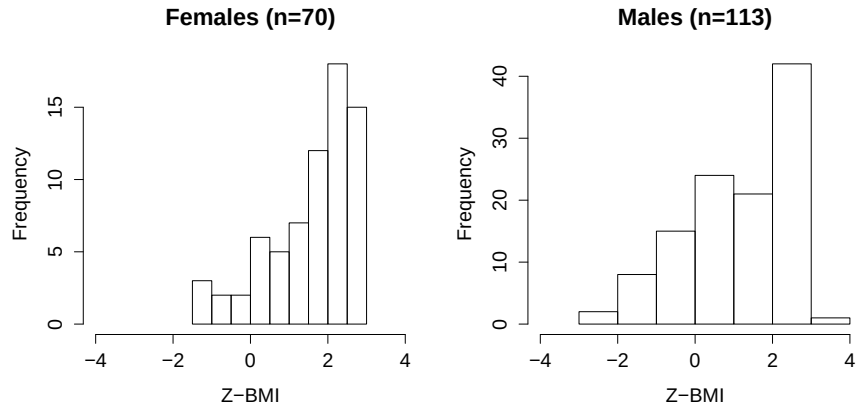


Figure 2.12.: Distribution of the Z-BMI scores in the Hungarian study for both females and males.

2.3.2. Univariate analysis

Most important univariate descriptors of the laboratory results for different levels of obesity (Z-BMI=+1, +2 and +3) and the results of the univariate association analysis, both segregated according to sex are given in Table 2.2 for the NHANES and in Table 2.3 for the Hungarian database.

Table 2.1.: Investigated laboratory parameters with name, abbreviation and unit of measurement.

Laboratory parameter	**Abbr.**	**Unit of measurement**
White blood cell count	WBC	G/l
Relative neutrophil count	RNC	%
Relative lymphocyte count	RLC	%
Relative monocyte count	RMC	%
Relative eosinophil count	REC	%
Absolute neutrophil count	ANC	G/l
Absolute lymphocyte count	ALC	G/l
Absolute monocyte count	AMC	G/l
Absolute eosinophil count	AEC	G/l
Red blood cell count	RBC	T/l
Hemoglobin	HGB	g/l
Hematocrit	HCT	%
Mean corpuscular volume	MCV	fl
Mean corpuscular hemoglobin	MCH	pg
Mean corpuscular hemoglobin concentration	MCHC	g/l
Red blood cell distribution width	RDW	%
Platelet count	PLT	G/l
Mean platelet volume	MPV	fl
C-reactive protein	CRP	mg/l
se Sodium	SNA	mmol/l
se Potassium	SK	mmol/l
se Chloride	SCL	mmol/l
se Total protein	STP	g/l
se Albumin	SAL	g/l
se Globulin	SGL	g/l
Blood urea nitrogen	BUN	mmol/l
se Creatinine	SCR	μmol/l
se Triglycerides	STG	mmol/l
se Total cholesterol	STC	mmol/l
se HDL cholesterol	HDL	mmol/l
Aspartate transaminase	AST	IU/l
Alanine transaminase	ALT	IU/l
Gamma glutamyl transpeptidase	GGT	IU/l

Table 2.2.: Univariate descriptors of the laboratory parameters for different levels of obesity (Z-BMI=+1, +2 and +3), segregated according to sex in Mean (Median) ± SD (IQR) format and the result of the univariate association analysis (ρ Spearman correlation coefficient, and its p-value and Holm-corrected p-value; '***' marks association that is significant at 0.1%, '**' marks association that is significant at 1%, '*' marks association that is significant at 5% and '.' marks association that is significant at 10% for the Holm-correction in every case) for the NHANES.

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
WBC	Male	5.9 (5.7) ± 1.5 (1.9)	6.3 (6.2) ± 1.5 (1.9)	6.8 (6.7) ± 1.7 (1.8)	0.23	0.0007	0.0183	*
	Female	6.5 (6.2) ± 1.9 (2.2)	7.0 (6.7) ± 2.2 (2.6)	7.8 (7.4) ± 2.1 (2.3)	0.18	0.0125	0.2998	
RNC	Male	48.9 (49.7) ± 9.6 (12.2)	52.2 (52.7) ± 8.2 (10.7)	55.0 (55.4) ± 7.4 (9.6)	0.22	0.0009	0.0241	*
	Female	54.9 (55.2) ± 10.6 (14.4)	54.6 (54.6) ± 9.5 (12.4)	57.3 (57.7) ± 7.8 (10.7)	0.22	0.0025	0.0660	.
RLC	Male	38.1 (37.5) ± 8.7 (11.8)	34.3 (33.7) ± 7.2 (8.6)	32.5 (32.3) ± 6.2 (7.4)	−0.17	0.0118	0.2723	
	Female	34.2 (33.9) ± 9.2 (12.8)	34.6 (34.6) ± 7.9 (10.9)	33.9 (33.8) ± 6.5 (9.3)	−0.14	0.0564	1.0000	
RMC	Male	8.9 (8.7) ± 2.2 (3.0)	8.9 (8.6) ± 2.3 (2.8)	8.3 (8.1) ± 2.1 (2.6)	−0.10	0.1571	1.0000	
	Female	7.6 (7.6) ± 2.0 (2.8)	7.4 (7.1) ± 2.1 (2.4)	6.8 (6.6) ± 1.8 (2.2)	−0.30	0.0000	0.0007	***
REC	Male	3.7 (3.1) ± 2.5 (2.7)	4.1 (3.5) ± 2.8 (2.6)	3.8 (3.4) ± 2.5 (2.1)	0.00	0.9831	1.0000	
	Female	2.9 (2.5) ± 1.9 (2.2)	3.0 (2.5) ± 2.1 (2.0)	2.3 (2.0) ± 1.6 (1.5)	−0.04	0.5959	1.0000	
ANC	Male	3.0 (2.8) ± 1.1 (1.4)	3.3 (3.3) ± 1.1 (1.5)	3.8 (3.7) ± 1.3 (1.3)	0.27	0.0001	0.0019	**
	Female	3.7 (3.3) ± 1.7 (1.8)	3.9 (3.7) ± 1.7 (2.0)	4.4 (4.4) ± 1.5 (1.9)	0.26	0.0004	0.0109	*
ALC	Male	2.2 (2.1) ± 0.6 (0.7)	2.1 (2.0) ± 0.5 (0.6)	2.2 (2.2) ± 0.4 (0.5)	0.09	0.1893	1.0000	
	Female	2.1 (2.1) ± 0.5 (0.7)	2.3 (2.2) ± 0.6 (0.7)	2.6 (2.4) ± 0.6 (0.7)	0.05	0.4719	1.0000	
AMC	Male	0.5 (0.5) ± 0.2 (0.2)	0.6 (0.5) ± 0.2 (0.2)	0.5 (0.5) ± 0.2 (0.2)	0.10	0.1212	1.0000	
	Female	0.5 (0.5) ± 0.2 (0.2)	0.5 (0.5) ± 0.2 (0.2)	0.5 (0.5) ± 0.2 (0.2)	−0.07	0.3176	1.0000	
AEC	Male	0.2 (0.2) ± 0.2 (0.2)	0.2 (0.2) ± 0.2 (0.2)	0.2 (0.2) ± 0.2 (0.2)	0.10	0.1469	1.0000	
	Female	0.2 (0.2) ± 0.1 (0.1)	0.2 (0.2) ± 0.1 (0.1)	0.2 (0.2) ± 0.1 (0.1)	0.07	0.3440	1.0000	

Table 2.2 – continued on next page

Table 2.2 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
RBC	Male	5.0 (5.0) ± 0.4 (0.5)	5.1 (5.1) ± 0.3 (0.5)	5.2 (5.2) ± 0.4 (0.5)	0.18	0.0087	0.2081	
	Female	4.5 (4.5) ± 0.4 (0.6)	4.5 (4.4) ± 0.3 (0.4)	4.6 (4.5) ± 0.4 (0.5)	−0.03	0.6688	1.0000	
HGB	Male	148.0 (147.9) ± 11.4 (15.6)	147.0 (146.2) ± 11.4 (14.9)	146.6 (146.9) ± 11.4 (16.8)	0.05	0.4817	1.0000	
	Female	134.7 (135.3) ± 10.3 (12.7)	132.5 (133.0) ± 8.8 (11.4)	126.8 (127.1) ± 8.9 (12.7)	−0.19	0.0105	0.2635	
HCT	Male	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	0.03	0.6477	1.0000	
	Female	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	−0.17	0.0213	0.4899	
MCV	Male	86.8 (87.3) ± 4.5 (5.7)	84.4 (84.6) ± 5.0 (7.2)	83.1 (83.1) ± 5.1 (6.8)	−0.16	0.0142	0.2982	
	Female	87.3 (87.4) ± 4.8 (5.9)	86.8 (87.1) ± 5.1 (6.5)	82.3 (84.6) ± 7.9 (8.7)	−0.11	0.1432	1.0000	
MCH	Male	29.7 (29.9) ± 1.7 (2.2)	29.0 (29.1) ± 2.0 (2.8)	28.5 (28.6) ± 2.0 (2.7)	−0.11	0.0916	1.0000	
	Female	30.1 (30.2) ± 2.0 (2.6)	29.8 (29.9) ± 1.9 (2.4)	28.0 (29.0) ± 3.2 (3.2)	−0.15	0.0411	0.9036	
MCHC	Male	342.7 (342.7) ± 8.6 (11.6)	343.2 (342.9) ± 8.1 (10.3)	342.8 (342.6) ± 8.1 (9.6)	0.09	0.1608	1.0000	
	Female	344.3 (344.1) ± 10.0 (12.2)	343.0 (343.3) ± 9.3 (12.7)	338.5 (339.1) ± 11.4 (16.4)	−0.09	0.2129	1.0000	
RDW	Male	12.5 (12.4) ± 0.6 (0.6)	12.6 (12.5) ± 0.8 (0.8)	12.7 (12.6) ± 0.9 (1.1)	−0.05	0.4542	1.0000	
	Female	12.4 (12.3) ± 0.8 (0.9)	12.6 (12.4) ± 0.9 (0.9)	13.6 (13.0) ± 1.8 (2.6)	0.09	0.2077	1.0000	
PLT	Male	243.8 (240.5) ± 54.3 (75.3)	248.1 (248.1) ± 56.3 (81.8)	270.6 (272.4) ± 56.2 (82.5)	0.10	0.1351	1.0000	
	Female	251.2 (246.5) ± 59.6 (81.8)	267.6 (265.4) ± 65.3 (88.5)	282.8 (277.0) ± 64.8 (76.6)	0.11	0.1285	1.0000	
MPV	Male	7.7 (7.7) ± 0.8 (1.1)	7.7 (7.7) ± 0.8 (0.9)	7.7 (7.7) ± 0.7 (0.8)	−0.05	0.4952	1.0000	
	Female	7.8 (7.7) ± 0.9 (1.3)	8.0 (7.9) ± 0.9 (1.2)	8.2 (8.1) ± 0.9 (1.2)	0.09	0.2181	1.0000	
CRP	Male	0.2 (0.1) ± 0.2 (0.1)	0.2 (0.1) ± 0.1 (0.2)	0.3 (0.2) ± 0.3 (0.2)	0.43	0.0000	0.0000	***
	Female	0.1 (0.1) ± 0.3 (0.1)	0.2 (0.1) ± 0.3 (0.2)	0.4 (0.4) ± 0.3 (0.4)	0.42	0.0000	0.0000	***
SNA	Male	139.6 (139.5) ± 1.6 (2.0)	139.4 (139.3) ± 1.5 (1.8)	139.3 (139.2) ± 1.3 (1.5)	−0.08	0.2311	1.0000	
	Female	139.3 (139.3) ± 1.6 (2.1)	139.5 (139.6) ± 1.7 (2.3)	139.3 (139.3) ± 1.4 (2.0)	0.01	0.8813	1.0000	

Table 2.2 – continued on next page

Table 2.2 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
SK	Male	4.1 (4.1) ± 0.3 (0.4)	4.2 (4.1) ± 0.3 (0.3)	4.2 (4.2) ± 0.3 (0.3)	0.03	0.6378	1.0000	
	Female	4.0 (4.0) ± 0.3 (0.4)	4.0 (4.0) ± 0.3 (0.3)	4.0 (4.0) ± 0.2 (0.3)	0.12	0.0991	1.0000	
SCL	Male	103.9 (103.7) ± 2.1 (2.9)	104.4 (104.3) ± 2.3 (2.8)	104.5 (104.6) ± 2.2 (2.9)	0.01	0.9293	1.0000	
	Female	105.3 (105.4) ± 2.0 (2.7)	105.6 (105.6) ± 2.0 (2.9)	106.0 (105.9) ± 1.6 (2.3)	0.13	0.0849	1.0000	
STP	Male	72.1 (72.0) ± 4.8 (6.7)	72.2 (72.0) ± 4.0 (5.2)	71.5 (71.3) ± 3.3 (4.3)	0.00	0.9677	1.0000	
	Female	71.9 (71.8) ± 4.4 (6.3)	71.2 (71.1) ± 4.0 (5.4)	70.5 (70.5) ± 3.3 (4.6)	−0.12	0.0971	1.0000	
SAL	Male	45.1 (45.0) ± 3.0 (4.2)	44.3 (44.2) ± 3.0 (4.4)	42.9 (42.7) ± 2.8 (3.8)	−0.16	0.0177	0.3543	
	Female	43.5 (43.6) ± 2.6 (3.5)	42.4 (42.6) ± 2.6 (3.4)	39.9 (39.5) ± 2.7 (4.0)	−0.33	0.0000	0.0001	***
SGL	Male	27.1 (26.8) ± 3.8 (5.1)	28.0 (28.0) ± 3.3 (4.4)	28.8 (28.7) ± 2.8 (3.5)	0.17	0.0126	0.2767	
	Female	28.2 (27.9) ± 3.8 (5.3)	28.8 (28.7) ± 3.8 (5.3)	30.8 (30.7) ± 3.7 (4.9)	0.09	0.2258	1.0000	
BUN	Male	3.6 (3.6) ± 1.1 (1.7)	3.7 (3.7) ± 0.9 (1.3)	3.7 (3.7) ± 0.7 (1.0)	−0.05	0.4955	1.0000	
	Female	3.4 (3.4) ± 1.2 (1.5)	3.2 (3.1) ± 1.1 (1.4)	3.2 (3.2) ± 1.1 (1.4)	−0.10	0.1866	1.0000	
SCR	Male	67.3 (66.2) ± 15.9 (19.6)	64.8 (64.3) ± 13.9 (18.5)	61.4 (60.1) ± 13.0 (18.6)	0.02	0.8081	1.0000	
	Female	60.8 (59.4) ± 13.4 (15.9)	56.3 (55.7) ± 11.2 (15.9)	57.2 (56.5) ± 10.0 (14.3)	−0.03	0.6329	1.0000	
STG	Male	0.9 (0.8) ± 0.5 (0.6)	1.1 (0.9) ± 0.6 (0.7)	1.3 (1.0) ± 0.7 (1.1)	0.21	0.0014	0.0369	*
	Female	0.9 (0.8) ± 0.6 (0.5)	0.9 (0.7) ± 0.5 (0.5)	0.9 (0.8) ± 0.3 (0.4)	0.00	0.9930	1.0000	
STC	Male	3.9 (3.8) ± 0.8 (0.9)	4.0 (3.9) ± 0.8 (1.1)	4.2 (4.1) ± 0.9 (1.2)	0.00	0.9836	1.0000	
	Female	4.2 (4.2) ± 0.8 (1.0)	4.2 (4.2) ± 0.6 (0.8)	4.1 (4.1) ± 0.6 (0.8)	0.05	0.4940	1.0000	
HDL	Male	1.4 (1.3) ± 0.3 (0.4)	1.2 (1.2) ± 0.3 (0.4)	1.1 (1.1) ± 0.2 (0.3)	−0.31	0.0000	0.0001	***
	Female	1.5 (1.4) ± 0.3 (0.4)	1.4 (1.3) ± 0.3 (0.4)	1.2 (1.1) ± 0.2 (0.3)	−0.22	0.0019	0.0514	.
AST	Male	25.5 (23.8) ± 8.6 (6.9)	27.5 (26.6) ± 6.9 (7.6)	28.0 (27.3) ± 6.1 (7.0)	0.19	0.0041	0.1016	
	Female	23.3 (21.9) ± 7.2 (6.6)	22.0 (21.0) ± 5.6 (6.3)	22.4 (20.0) ± 8.0 (6.1)	−0.09	0.2390	1.0000	

Table 2.2 – continued on next page

Table 2.2 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
ALT	Male	20.1 (18.0) ± 10.9 (6.7)	28.0 (23.2) ± 15.8 (14.3)	34.9 (31.3) ± 19.0 (15.9)	0.48	0.0000	0.0000	***
	Female	16.6 (15.2) ± 7.3 (4.3)	19.0 (17.5) ± 7.3 (5.5)	21.7 (18.6) ± 12.3 (5.6)	0.30	0.0000	0.0010	***
GGT	Male	14.7 (14.1) ± 5.1 (6.2)	17.3 (16.0) ± 6.5 (7.6)	20.7 (19.1) ± 7.6 (11.6)	0.39	0.0000	0.0000	***
	Female	12.0 (11.1) ± 4.9 (5.1)	15.8 (14.2) ± 6.6 (7.5)	20.0 (17.0) ± 8.9 (13.2)	0.40	0.0000	0.0000	***

Table 2.3.: Univariate descriptors of the laboratory parameters for different levels of obesity (Z-BMI=+1, +2 and +3), segregated according to sex in Mean (Median) ± SD (IQR) format and the result of the univariate association analysis (ρ Spearman correlation coefficient, and its p-value and Holm-corrected p-value; '***' marks association that is significant at 0.1%, '**' marks association that is significant at 1%, '*' marks association that is significant at 5% and '.' marks association that is significant at 10% for the Holm-correction in every case) for the NHANES.

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
WBC	Male	6.8 (6.7) ± 1.4 (1.8)	7.5 (7.4) ± 1.5 (2.0)	8.0 (7.9) ± 1.5 (1.9)	0.53	0.0000	0.0000	***
	Female	7.4 (7.4) ± 1.3 (1.9)	7.9 (7.7) ± 1.7 (2.3)	8.5 (8.4) ± 1.9 (3.3)	0.24	0.0431	0.9041	
RNC	Male	50.5 (50.7) ± 7.3 (9.0)	52.5 (53.2) ± 7.4 (9.1)	54.8 (55.9) ± 8.4 (9.8)	0.26	0.0061	0.1093	
	Female	52.9 (52.2) ± 9.1 (11.8)	54.7 (55.0) ± 9.6 (14.3)	57.1 (57.9) ± 9.5 (14.5)	0.11	0.3641	1.0000	
RLC	Male	36.0 (35.5) ± 7.3 (9.0)	35.0 (34.6) ± 7.0 (9.3)	33.1 (32.5) ± 7.3 (9.2)	−0.21	0.0272	0.3259	
	Female	34.5 (34.6) ± 8.2 (11.8)	34.5 (34.0) ± 8.4 (12.3)	32.8 (32.1) ± 8.0 (11.8)	−0.06	0.6196	1.0000	
RMC	Male	9.6 (9.2) ± 2.2 (2.5)	9.4 (8.7) ± 2.6 (2.8)	9.0 (8.4) ± 2.4 (2.5)	−0.23	0.0123	0.1781	
	Female	8.3 (8.2) ± 1.7 (2.6)	8.2 (8.0) ± 1.6 (2.3)	7.9 (7.8) ± 1.5 (2.1)	−0.22	0.0660	1.0000	
REC	Male	3.2 (2.8) ± 1.8 (1.9)	2.9 (2.6) ± 1.6 (1.7)	3.1 (2.6) ± 2.1 (1.9)	−0.08	0.4155	1.0000	
	Female	3.9 (2.7) ± 3.5 (3.1)	2.9 (2.3) ± 2.2 (2.3)	2.9 (2.4) ± 2.2 (2.4)	−0.10	0.4125	1.0000	
ANC	Male	3.5 (3.4) ± 1.1 (1.3)	4.1 (3.9) ± 1.2 (1.4)	4.4 (4.4) ± 1.2 (1.5)	0.48	0.0000	0.0000	***
	Female	4.1 (4.0) ± 1.2 (1.7)	4.4 (4.2) ± 1.4 (1.9)	5.0 (4.9) ± 1.5 (2.3)	0.23	0.0600	1.0000	
ALC	Male	2.4 (2.3) ± 0.6 (0.8)	2.6 (2.6) ± 0.6 (0.9)	2.7 (2.7) ± 0.6 (0.9)	0.33	0.0004	0.0082	**
	Female	2.5 (2.4) ± 0.6 (0.8)	2.7 (2.6) ± 0.7 (1.0)	2.7 (2.6) ± 0.8 (1.0)	0.16	0.1801	1.0000	
AMC	Male	0.7 (0.6) ± 0.2 (0.2)	0.7 (0.7) ± 0.2 (0.3)	0.7 (0.7) ± 0.2 (0.3)	0.31	0.0009	0.0199	*
	Female	0.6 (0.6) ± 0.1 (0.2)	0.6 (0.6) ± 0.2 (0.2)	0.7 (0.6) ± 0.2 (0.3)	0.07	0.5755	1.0000	
AEC	Male	0.2 (0.2) ± 0.1 (0.1)	0.2 (0.2) ± 0.1 (0.1)	0.2 (0.2) ± 0.2 (0.1)	0.14	0.1495	1.0000	
	Female	0.3 (0.2) ± 0.3 (0.2)	0.2 (0.2) ± 0.2 (0.2)	0.2 (0.2) ± 0.2 (0.2)	−0.02	0.8742	1.0000	

Table 2.3 – continued on next page

Table 2.3 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
RBC	Male	5.2 (5.2) ± 0.3 (0.4)	5.3 (5.3) ± 0.3 (0.5)	5.3 (5.4) ± 0.4 (0.5)	0.13	0.1642	1.0000	
	Female	4.8 (4.7) ± 0.3 (0.4)	4.8 (4.8) ± 0.3 (0.4)	4.9 (4.9) ± 0.2 (0.3)	0.40	0.0006	0.0171	*
HGB	Male	150.6 (150.8) ± 11.5 (16.6)	146.9 (146.2) ± 13.3 (20.4)	146.7 (146.8) ± 13.8 (21.4)	−0.19	0.0422	0.4646	
	Female	135.6 (135.7) ± 8.8 (12.3)	135.0 (134.9) ± 9.8 (14.3)	133.8 (134.4) ± 9.1 (13.5)	−0.07	0.5656	1.0000	
HCT	Male	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.1)	0.4 (0.4) ± 0.0 (0.1)	−0.23	0.0129	0.1781	
	Female	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	0.4 (0.4) ± 0.0 (0.0)	−0.07	0.5900	1.0000	
MCV	Male	85.4 (85.5) ± 4.0 (5.6)	83.1 (83.2) ± 4.2 (5.8)	82.0 (82.0) ± 4.3 (5.7)	−0.45	0.0000	0.0000	***
	Female	87.1 (87.3) ± 4.2 (6.2)	85.6 (85.8) ± 4.1 (6.0)	83.7 (83.8) ± 3.8 (5.4)	−0.47	0.0000	0.0015	**
MCH	Male	28.7 (28.8) ± 1.4 (2.0)	28.0 (28.1) ± 1.7 (2.2)	27.7 (27.7) ± 1.8 (2.4)	−0.36	0.0001	0.0021	**
	Female	28.6 (28.7) ± 1.6 (2.4)	28.2 (28.2) ± 1.6 (2.2)	27.2 (27.3) ± 1.6 (2.2)	−0.46	0.0001	0.0016	**
MCHC	Male	336.8 (335.9) ± 10.4 (13.1)	337.4 (337.3) ± 10.6 (14.3)	337.0 (337.5) ± 10.5 (14.7)	0.03	0.7375	1.0000	
	Female	328.9 (327.1) ± 11.1 (17.7)	329.7 (329.4) ± 10.4 (15.3)	326.2 (326.3) ± 8.8 (12.4)	−0.10	0.4071	1.0000	
RDW	Male	13.5 (13.4) ± 0.7 (0.9)	13.8 (13.7) ± 0.8 (1.1)	13.9 (13.9) ± 0.8 (1.1)	0.24	0.0104	0.1662	
	Female	13.6 (13.6) ± 0.8 (1.1)	13.6 (13.6) ± 0.9 (1.3)	13.9 (13.8) ± 0.9 (1.2)	0.26	0.0277	0.6654	
PLT	Male	257.3 (253.7) ± 58.7 (80.5)	276.5 (277.0) ± 54.1 (70.9)	296.0 (296.2) ± 57.9 (77.0)	0.40	0.0000	0.0003	***
	Female	277.2 (273.8) ± 45.4 (67.4)	292.4 (288.2) ± 48.6 (68.1)	304.2 (299.0) ± 50.3 (71.1)	0.37	0.0018	0.0496	*
MPV	Male	10.7 (10.6) ± 0.8 (1.2)	10.6 (10.6) ± 0.8 (1.2)	10.7 (10.6) ± 0.8 (1.3)	−0.13	0.1683	1.0000	
	Female	10.9 (11.0) ± 0.7 (1.0)	10.7 (10.7) ± 0.6 (0.8)	10.7 (10.7) ± 0.6 (0.9)	−0.12	0.3092	1.0000	
CRP	Male	3.9 (2.1) ± 7.6 (2.6)	5.6 (3.6) ± 9.7 (4.1)	7.0 (4.9) ± 7.1 (5.7)	0.64	0.0000	0.0000	***
	Female	2.2 (1.8) ± 2.4 (2.2)	4.6 (2.8) ± 6.1 (4.0)	5.8 (4.7) ± 4.8 (5.6)	0.61	0.0000	0.0000	***
SNA	Male	138.6 (138.6) ± 1.8 (2.4)	138.3 (138.3) ± 2.1 (2.8)	138.6 (138.5) ± 2.0 (2.7)	−0.08	0.3756	1.0000	
	Female	138.2 (138.3) ± 1.6 (2.0)	138.0 (138.0) ± 1.8 (2.7)	138.3 (138.3) ± 1.8 (2.6)	0.02	0.8765	1.0000	

Table 2.3 – continued on next page

Table 2.3 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
SK	Male	4.2 (4.2) ± 0.3 (0.5)	4.3 (4.3) ± 0.4 (0.5)	4.4 (4.4) ± 0.3 (0.5)	0.25	0.0082	0.1393	
	Female	4.4 (4.4) ± 0.3 (0.5)	4.3 (4.3) ± 0.3 (0.4)	4.4 (4.4) ± 0.3 (0.4)	0.12	0.3204	1.0000	
SCL	Male	102.1 (101.9) ± 2.1 (3.0)	102.5 (102.4) ± 2.1 (3.1)	102.7 (102.5) ± 2.2 (3.2)	0.07	0.4939	1.0000	
	Female	102.6 (102.6) ± 1.6 (2.1)	103.3 (103.2) ± 2.1 (2.8)	103.6 (103.6) ± 2.0 (2.8)	0.19	0.1073	1.0000	
STP	Male	76.6 (76.5) ± 4.2 (5.9)	76.5 (76.5) ± 3.9 (5.5)	75.5 (75.1) ± 4.2 (6.1)	−0.14	0.1432	1.0000	
	Female	76.4 (75.9) ± 4.4 (6.5)	76.5 (76.0) ± 4.8 (7.3)	76.4 (75.8) ± 4.5 (6.3)	0.09	0.4768	1.0000	
SAL	Male	50.2 (50.1) ± 2.6 (3.5)	48.7 (48.5) ± 2.7 (3.8)	47.5 (47.1) ± 2.4 (3.6)	−0.47	0.0000	0.0000	***
	Female	48.5 (48.6) ± 2.4 (3.4)	47.5 (47.6) ± 2.8 (4.0)	46.2 (46.0) ± 2.6 (3.5)	−0.35	0.0028	0.0744	.
SGL	Male	26.6 (26.6) ± 3.7 (4.7)	27.7 (27.5) ± 3.7 (4.6)	28.0 (27.7) ± 3.7 (4.7)	0.19	0.0458	0.4646	
	Female	27.9 (27.3) ± 3.9 (6.3)	29.3 (29.0) ± 3.8 (6.1)	30.2 (30.1) ± 3.6 (5.2)	0.39	0.0009	0.0253	*
BUN	Male	4.6 (4.6) ± 1.2 (1.9)	4.5 (4.4) ± 1.1 (1.7)	4.4 (4.4) ± 1.0 (1.6)	−0.11	0.2377	1.0000	
	Female	4.0 (4.0) ± 0.8 (1.2)	4.2 (4.0) ± 1.0 (1.4)	4.2 (4.0) ± 1.1 (1.7)	−0.13	0.2739	1.0000	
SCR	Male	73.2 (73.7) ± 9.5 (12.6)	68.2 (67.7) ± 9.6 (14.4)	65.1 (64.3) ± 8.7 (13.3)	−0.43	0.0000	0.0001	***
	Female	58.6 (58.6) ± 5.9 (8.4)	60.6 (60.6) ± 7.0 (10.8)	61.1 (61.4) ± 6.4 (8.9)	0.05	0.6798	1.0000	
STG	Male	1.0 (0.9) ± 0.5 (0.6)	1.2 (1.1) ± 0.5 (0.6)	1.2 (1.1) ± 0.6 (0.6)	0.43	0.0000	0.0000	***
	Female	1.2 (1.0) ± 0.6 (0.9)	1.3 (1.1) ± 0.7 (0.7)	1.3 (1.2) ± 0.5 (0.6)	0.34	0.0039	0.1020	
STC	Male	3.9 (3.8) ± 1.0 (1.1)	4.2 (4.1) ± 0.9 (1.0)	4.2 (4.1) ± 0.8 (1.0)	0.29	0.0017	0.0340	*
	Female	4.1 (4.0) ± 0.9 (0.9)	4.3 (4.2) ± 0.8 (1.1)	4.3 (4.2) ± 0.7 (1.0)	0.23	0.0515	1.0000	
HDL	Male	1.3 (1.2) ± 0.2 (0.3)	1.2 (1.1) ± 0.2 (0.3)	1.1 (1.1) ± 0.2 (0.3)	−0.27	0.0043	0.0811	.
	Female	1.3 (1.3) ± 0.2 (0.3)	1.3 (1.3) ± 0.2 (0.3)	1.2 (1.2) ± 0.2 (0.3)	−0.26	0.0327	0.7517	
AST	Male	22.7 (22.2) ± 7.3 (9.3)	25.0 (23.9) ± 8.9 (10.6)	26.2 (24.6) ± 9.3 (10.3)	0.24	0.0119	0.1781	
	Female	21.0 (18.7) ± 7.9 (8.7)	20.4 (19.0) ± 7.4 (9.5)	19.1 (17.9) ± 6.5 (8.3)	−0.16	0.1870	1.0000	

Table 2.3 – continued on next page

Table 2.3 – continued from previous page

Variable	Sex	Z-BMI=+1	Z-BMI=+2	Z-BMI=+3	ρ	p	Corrected p	
ALT	Male	22.0 (19.3) ± 12.8 (11.2)	30.4 (26.1) ± 18.5 (16.3)	36.4 (30.8) ± 20.6 (21.0)	0.56	0.0000	0.0000	***
	Female	17.4 (14.0) ± 10.1 (9.9)	19.5 (16.7) ± 9.5 (12.1)	22.2 (19.7) ± 9.8 (11.6)	0.32	0.0063	0.1567	
GGT	Male	23.5 (21.4) ± 9.3 (14.0)	25.7 (22.8) ± 11.4 (12.0)	32.8 (29.2) ± 13.9 (18.4)	0.52	0.0000	0.0000	***
	Female	15.7 (14.2) ± 6.6 (8.0)	18.5 (16.6) ± 8.7 (8.7)	21.2 (19.4) ± 8.1 (9.0)	0.25	0.0335	0.7517	

These descriptive statistics are important as reference values, and to facilitate further analyses and interpretation of the results.

According to the NHANES, the laboratory parameters that are significantly altered by obesity in both sexes are RNC, ANC, CRP, HDL, ALT and GGT. In addition to that, WBC and STG changes significantly only in males, RMC and SAL changes significantly only in females.

According to the Hungarian study, the laboratory parameters that are significantly altered by obesity in both sexes are MCV, MCH, PLT, CRP and SAL. In addition to that, WBC (with ANC, ALC and AMC), SCR, STG, STC, HDL, ALT and GGT only change significantly in males. RBC and SGL only change significantly in females. The high number of parameters changing significantly only for males is likely attributable in this case to the fact that the sample size was much larger for males in the Hungarian study – it is quite possible that these changes would have been significant for females as well, were we able to use larger female sample. It is also worth noting that results from the Hungarian study should be handled with more caution due to its non-representative nature.

These results can be interpreted together with the descriptors for different Z-BMI values (and the value of the Spearman-ρ correlation coefficient), which shed light on the medical content of differences by revealing the direction and the (clinical) size of the difference.

Most obvious is the presence of inflammation-related changes: elevated levels of the inflammation marker CRP, elevated PLT and elevated WBC and WBC fractions (for males). These can be considered as an empirical confirmation that obesity is associated with inflammation. This has been first noted decades ago, and now hypothesized to be caused by the proinflammatory mediators released by the excessive white adipose tissue (Bastard et al. 2006; Stienstra et al. 2007). These results are consistent with the literature: the idea that obesity can be considered as a systematic, low-level, chronic inflammation state is widely discussed (Ferroni et al. 2004) and was described specifically in children, too (Sacheck 2008). In addition to theoretical results, the elevations of the abovementioned inflammation markers were clinically also observed (Oda and Kawai 2010), specifically in children (Syrenicz et al. 2006; Gilbert-Diamond et al. 2012).

Another important finding is that MCV and MCH both have a significant decreasing tendency as the degree of obesity increases in the Hungarian study. Obesity is known to be associated with (some) features of anemia (Ausk and Ioannou 2008), which is logical if we consider that chronic inflammation is often associated with anemia. However, clinically, we can not speak of anemia as the hemoglobin levels are not decreasing with the

degree of obesity in the Hungarian study. This is consistent with what (Ausk and Ioannou 2008) have found, but they have not reported results for MCV and MCH. Therefore, these findings not only confirm theirs, but also extend it by pointing out that the nature of this phenomenon is hypochromic microcytosis (which was to be expected for chronic inflammation). This effect has already been described in the literature (Tungtrongchitr et al. 2000), with some authors linking it iron deficiency (Solá et al. 2007) (which usually, however, causes depressed HGB as well). Why this was only apparent in the Hungarian study (but not in NHANES) is an open question, the most likely explanation is that it manifests only for extreme levels of obesity – and such subjects were (intentionally) oversampled in the Hungarian study (just detect such alterations).

The elevation of ALT and GGT (dramatically seen in every case, save for females in the Hungarian study) can be considered as an indication of the effect of obesity, especially central obesity on liver function that is relatable to non-alcoholic fatty liver disease (Lam and Mobarhan 2004; Colicchio et al. 2005; Gholam et al. 2007). It is worth noting that some even presume that this effect is mediating between obesity and type 2 diabetes (Lawlor et al. 2005).

Changes in STG (and slightly: STC and HDL) are almost trivial, as they are indicators of fat metabolism state (Vliet et al. 2011).

Decreasing tendency of SAL with the degree of obesity seems to be contradicting to what literature reported, at least with respect to metabolic syndrome (Cho et al. 2012; Ishizaka et al. 2007). Note however that these results pertain to adult population.

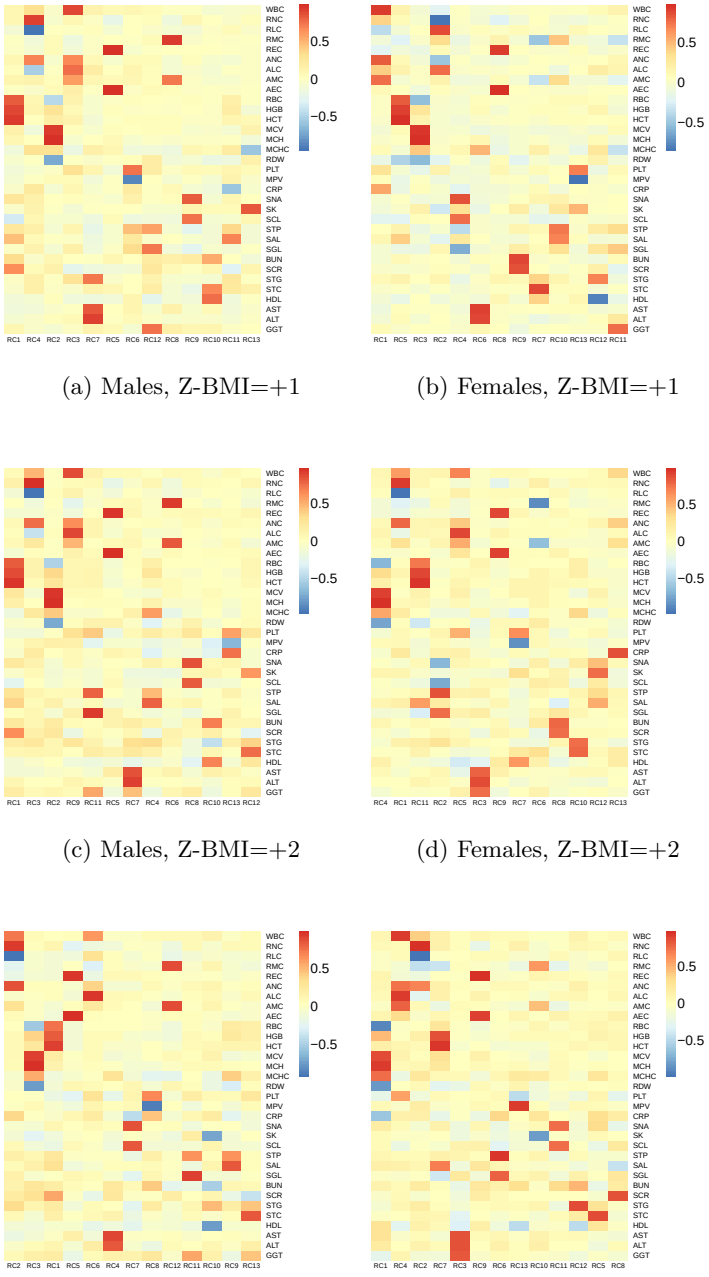
2.3.3. Multivariate analysis

Results of PCA, i.e. the loading matrices (visualized with heatmaps) for different levels of obesity (Z-BMI=+1, +2 and +3) segregated according to sex are shown in Figure 2.13 (NHANES) and Figure 2.14 (Hungarian study).

Dendrograms, obtained with CA are shown in Figure 2.15 (NHANES) and Figure 2.16, again for different levels of obesity (Z-BMI=+1, +2 and +3) segregated according to sex.

It is immediately obvious from the results of the PCA that the structure of the components was largely the same, irrespectively of the degree of obesity (Z-BMI) and sex. (Save for the order of components, which is sometimes varied.)

The medical ”meaning” of a principal component can be given based on those variables that highly correlate with the component (which can be read from the loading matrix). Summing the information from the different loading matrices, the following common components can be identified:



(a) Males, Z-BMI=+1

(b) Females, Z-BMI=+1

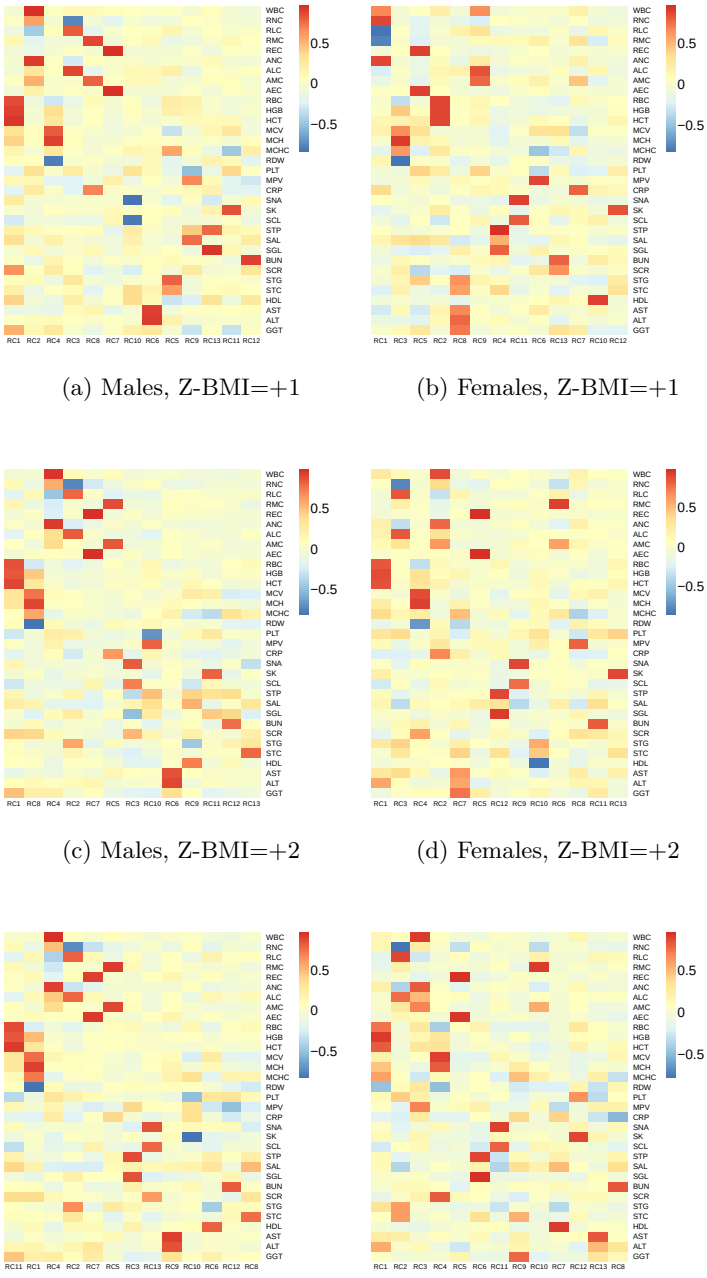
(c) Males, Z-BMI=+2

(d) Females, Z-BMI=+2

(e) Males, Z-BMI=+3

(f) Females, Z-BMI=+3

Figure 2.13.: Results of the PCA (loading matrices visualized with heatmaps) for different Z-BMI levels (Z-BMI=+1, +2 and +3), segregated according to sex for the NHANES.



(a) Males, Z-BMI=+1

(b) Females, Z-BMI=+1

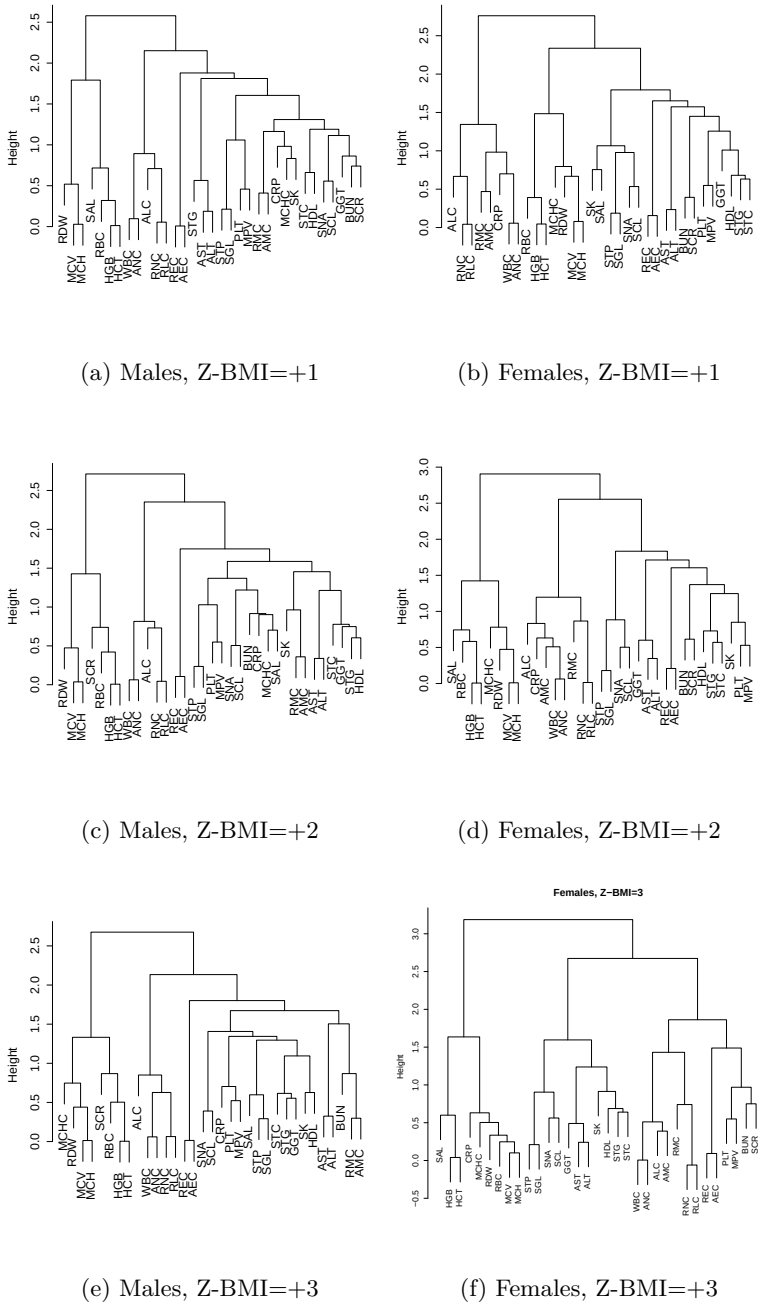
(c) Males, Z-BMI=+2

(d) Females, Z-BMI=+2

(e) Males, Z-BMI=+3

(f) Females, Z-BMI=+3

Figure 2.14.: Results of the PCA (loading matrices visualized with heatmaps) for different Z-BMI levels (Z-BMI=+1, +2 and +3), segregated according to sex for the Hungarian study.



(a) Males, Z-BMI=+1

(b) Females, Z-BMI=+1

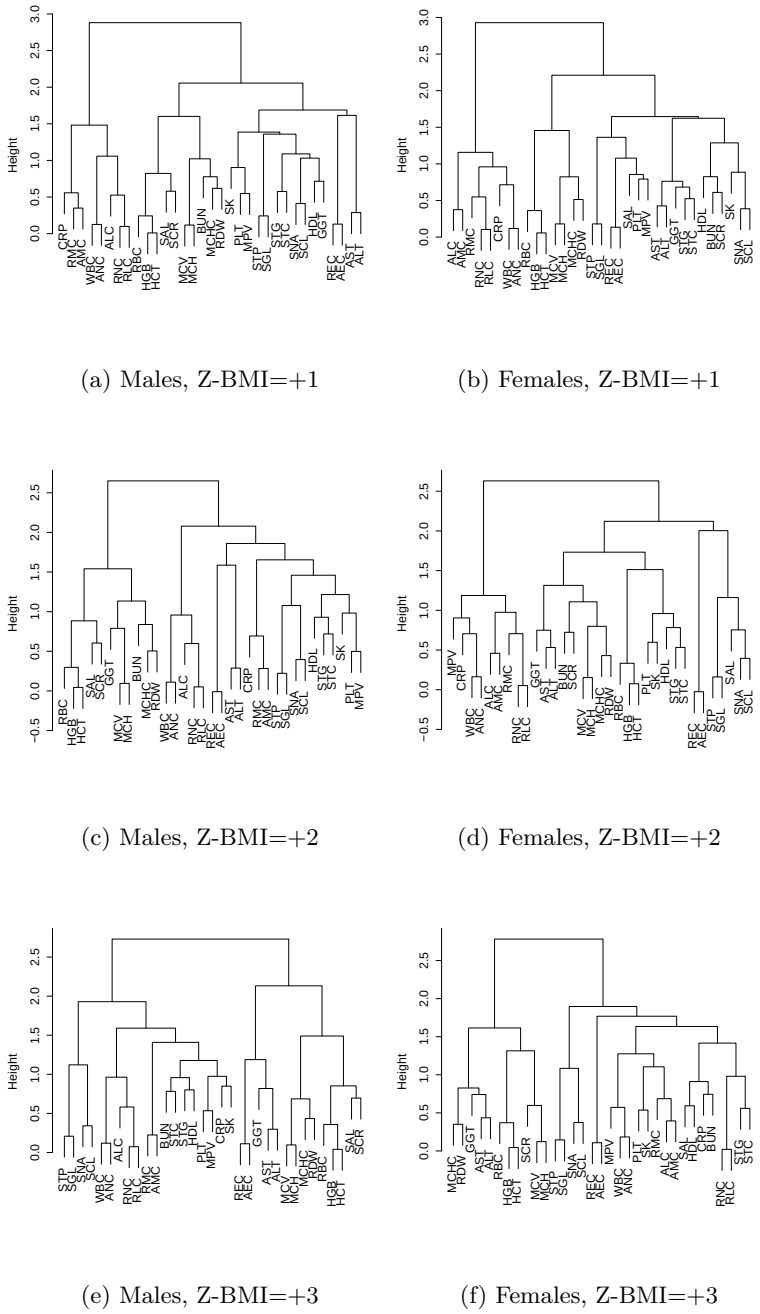
(c) Males, Z-BMI=+2

(d) Females, Z-BMI=+2

(e) Males, Z-BMI=+3

(f) Females, Z-BMI=+3

Figure 2.15.: Results of the CA (visualized with dendrograms) for different Z-BMI levels (Z-BMI=+1, +2 and +3), segregated according to sex for the NHANES.



(a) Males, Z-BMI=+1

(b) Females, Z-BMI=+1

(c) Males, Z-BMI=+2

(d) Females, Z-BMI=+2

(e) Males, Z-BMI=+3

(f) Females, Z-BMI=+3

Figure 2.16.: Results of the CA (visualized with dendrograms) for different Z-BMI levels (Z-BMI=+1, +2 and +3), segregated according to sex for the Hungarian study.

White blood cell components Relative and absolute counts of the same fraction usually form separate components; WBC is typically in the neutrophil component.

Macroscopic red blood cell component Consists of RBC, HGB and HCT, all being positively correlated with the component.

Microscopic red blood cell component Consists of MCV, MCH and MCHC (positively correlated with the component) and RDW (negatively correlated).

Platelet component Consists of PLT (positively correlated) and MPV (negatively correlated).

Liver enzymes components Consists of ALT and AST (positively correlated). GGT's association is less pronounced, but sometimes observable.

Inorganic constituents of the serum component Consists of SNA and SCL (positively correlated). SK forms separate component.

Organic constituents of the serum component Consists of STP and SGL (positively correlated).

Blood lipids component Consists of STG (positively correlated) and HDL (negatively correlated). STC usually forms separate component.

These findings are consistent with the dendrograms of the CA: variables that belong to the same component are usually found to be closely connected on the dendrogram. (For example, the deepest connection is between HGB and HCT (which is followed by RBC) almost invariably.)

CA also confirms that the correlation structure is largely independent from both the sex and the degree of obesity. Although dendrograms are varied, structural blocks (like the already mentioned HGB-HCT-RBC trio) can be identified that are largely the same in every case.

2.4. Conclusion

Univariate examination of laboratory results sheds light on the pathophysiological alterations that are associated with obesity. While these changes were mostly already well-known for particular parameters, I now performed a comprehensive, uniform investigation for 33 routinely measured blood tests.

The analysis of the multivariate structure of the laboratory results reveals groups of variables that exhibit similar stochastic behavior, pointing to shared physiological background. On the other hand, this analysis also demonstrated that the correlation structure of the laboratory parameters is largely unaffected by the degree of obesity and sex.

The method I proposed for the analysis of the multivariate structure (obtaining conditional correlation matrices through KDE element-by-element with smoothing being applied afterwards, and the analysis of these matrices with PCA or CA) lived up to expectations and was demonstrated to be a useful tool in similar tasks.

These results can be used to deepen our understanding of the pathophysiology of overweight and obesity, and how these diseases affect the human body. Such understanding can be then in turn used to optimize prevention and therapy, which has a direct significance from the public health point of view.

Thesis group 1: Effects of obesity on laboratory parameters.

Thesis 1.1:

> ***Thesis 1.1***
> **I have developed a biostatistical methodology (and an associated computer program) to investigate the effect of obesity on laboratory parameters. This methodology provides a way to analyze both the uni- and the multivariate structure of the laboratory parameters, making the effect of obesity explicit during the process.**

Thesis 1.2:

> ***Thesis 1.2***
> **I have provided clinical interpretations for the effects of obesity on laboratory parameters based on a representative international survey and a non-representative survey that was performed on Hungarian adolescents specifically for the aims of the present investigation. I have discussed results pertaining to both the uni- and the multivariate structure of the investigated variables.**

Relevant own publications pertaining to this thesis group: [F-3; F-15; F-9; F-1; F-4; F-21; F-2; F-11; F-5; F-12; F-8; F-7; F-6; F-13; F-18; F-19; F-20; F-17].

3. Modeling and Evaluating the Performance of Tight Glycemic Control Protocols

My second thesis investigates a particular question about human blood glucose regulation. This system is one the most well-known and most deeply studied human regulatory systems, especially because of its huge clinical significance due to diabetes.

However, it became recently clear that the blood glucose regulation has clinical significance outside diabetology as well. One example arises in critical care: patients admitted to intensive care units are known to exhibit stress-induced, non-physiological glycemic excursions. Several studies linked this to adverse outcomes, hence, efforts were made to control this.

Tight glycemic control (TGC) is an approach to address this. Numerous protocols have been developed to realize TGC in the hope of improving outcome for the critically ill. One crucial point is the handling of insulin sensitivity, which is dependent on the diagnosis and also evolves over time.

In this thesis, I provide a biostatistical model to quantitatively evaluate TGC protocols in terms of modeling insulin sensitivity. The model explicitly incorporates diagnosis and length of stay.

The rest of this thesis is organized as follows. In Section 3.1 I introduce TGC in more detail, and describe the state-of-the-art in the research question set forth above. In Section 3.2 I summarize the concrete aims and directions of my research and in Section 3.3 I detail the biostatistical methodology that I used for the investigation. Section 3.4 presents the results, while 3.5 gives a discussion of them and also details their clinical relevance and applicability. This thesis group is summarized in Section 3.6.

3.1. Significance of Tight Glycemic Control in Critical Care: Literature Review and Background of my Research

Stress induced hyperglycemia is a significant issue in critical care, affecting up to 30-50% of patients and increasing morbidity and mortality (Krinsley 2003; McCowen, Malhotra, and Bistrian 2001). Controlling glycemia has proved difficult due to the associated risk of hypoglycemia when highly dynamic patients are treated with exogenous insulin (Griesdale et al. 2009). Both extremes, as well as glycemic variability, have been independently linked to increased morbidity and mortality (Bagshaw et al. 2009; Egi et al. 2006; Krinsley 2008), creating a difficult clinical problem.

More specifically, inter- and intra- patient metabolic variability drive outcome glycemic variability and hypoglycemic risk (Chase, Compte, et al. 2011) making good control difficult. In particular, sudden and large rises in insulin sensitivity can result in a hypoglycemic event when exogenous insulin is given over a typical 3-4 hour measurement interval. It is critical to determine the size and likelihood of these intra-patient variations, to enable a more complete understanding of the inherent risks in glycemic control.

Very few studies have examined time-varying evolution of insulin sensitivity and its variability in the critically ill. Langouche et al. (2007) noted that insulin sensitivity rose between days 1 and 5 over their large cohorts, but provided no daily or diagnostic specific evolution. Lin et al. (2008) showed that hour to hour changes for a clinically validated model-based insulin sensitivity indicator could be quite large as a function of current insulin sensitivity level for a medical Intensive Care Unit (ICU) cohort that covered all diagnostic categories and days of ICU stay. However, no studies to date have explicitly described the evolution of intra-patient insulin sensitivity and its variability on a daily basis, or for different diagnostic categories.

Such information would provide insight into the risk of hypoglycemia by diagnostic category and day of ICU stay. Additionally, insight into the likelihood of glycemic variability resulting from greater or lesser intra-patient variability of insulin sensitivity could be attained.

This thesis presents the first rigorous statistical analysis of inter- and intra- patient insulin sensitivity variability as a function of diagnostic category and day of stay. It is also the first to examine the long-term behavior of insulin sensitivity.

The significance of these can be understood in the light of glycemic control, especially tight glycemic control (TGC). TGC protocols aim to address specifically this issue. Glycemic control can reduce negative outcomes (Krinsley 2004; Chase, Shaw, et al. 2008; Van den Berghe et al. 2001), but has proven difficult (Casaer et al. 2011; Brunkhorst

et al. 2008; Finfer and The NICE-SUGAR Study Investigators 2009). Only Chase, Shaw, et al. (2008) reduced both mortality and hypoglycemia.

3.2. Directions and Goals of my Research

I have developed a novel methodology to evaluate and model the insulin sensitivity variability and its evolution over time for patients in different diagnosis groups. This also makes the more thorough investigation of the performance of tight glycemic control protocols possible.

I actually implemented this methodology to provide informatics support to its application. Full source code to perform this analysis is listed in Appendix B.

Relevant own publications pertaining to this thesis: [F-14; F-10; F-16].

3.3. Materials and Methods of Investigation

This Section introduces the methodology that was applied during the present investigation.

I first present the patient data in Subsection 3.3.1. After that, I discuss the questions associated with defining appropriate indicators for measuring patient variability (Subsection 3.3.2) and analyzing this variability (Subsection 3.3.3). I then introduce the methods used for the statistical evaluation and modeling (Subsection 3.3.4), and conclude this Section by exposing the data processing methods that were applied (Subsection 3.3.5).

3.3.1. Patient Data

Clinical data from $n = 390$ patients (47 836 hours) in the SPRINT medical ICU cohort (Chase, Shaw, et al. 2008) are used to identify hourly, model-based insulin sensitivity (SI) values ($SI\,(n)$). SPRINT (Specialized Relative Insulin and Nutrition Tables) is a model-based, clinically validated tight glycemic control (TGC) protocol that provides explicit control for both nutrition intake and insulin input (Chase, Shaw, et al. 2008).

Hour-to-hour changes are evaluated for the cohort over all days of ICU stay using a stochastic model (Lin et al. 2008) that provides kernel density estimation-based distributions of $SI\,(n+1)$ values (in terms of predicted distribution, i.e. $\widehat{F}_{SI_{n+1}}$) for each current $SI\,(n)$ value using all 47 836 data points.

Table 3.1 shows the patient demographic details, including diagnostic categories. These were created based on the APACHE III codes, and consist of operative and non-operative groups for cardiac, gastric and all other patients (with abbreviations OpC, NOpC, OpG, NOpG, OpO and NOpO, respectively). For the daily statistics, only patients who had at least 24 hours of glycemic control and ICU stay were used.

Table 3.1.: The distribution (according to length-of-stay and diagnosis group) and the most important demographic indicators of the patients. Data are shown in an n, age, percentage of females format, with age statistics arranged in Mean (Median) ± SD (IQR) manner. Columns indicate minimum (and not exact) length-of stay, so the same patient may appear in several cells.

Group	**Day 1**			**Day 2**		
	n	Age	Sex	n	Age	Sex
NOpC	28	59.5 (62) ± 16.5 (24)	35.7	18	58.4 (60) ± 16.1 (19)	38.9
OpC	35	72.9 (73) ± 7.12 (11)	22.9	21	72.9 (73) ± 6.54 (10)	23.8
NOpG	16	64.3 (67) ± 12.8 (15)	25.0	13	64.4 (71) ± 14.2 (19)	23.1
OpG	42	67.9 (72) ± 12.4 (13)	35.7	29	69.9 (72) ± 10.8 (11)	27.6
NOpO	119	54.7 (59) ± 18.0 (27)	46.2	101	54.5 (59) ± 18.0 (28)	42.6
OpO	21	50.8 (56) ± 19.2 (31)	38.1	16	54.9 (58) ± 18.5 (31)	43.8
Group	**Day 3**			**Day 4+**		
	n	Age	Sex	n	Age	Sex
NOpC	11	64.2 (63) ± 10.6 (16)	18.2	11	64.2 (63) ± 10.6 (16)	18.2
OpC	18	73.2 (74) ± 6.46 (9)	27.8	18	73.2 (74) ± 6.46 (9)	27.8
NOpG	13	64.4 (71) ± 14.2 (19)	23.1	13	64.4 (71) ± 14.2 (19)	23.1
OpG	23	69.2 (71) ± 9.46 (12)	26.1	23	69.2 (71) ± 9.46 (12)	26.1
NOpO	88	54.2 (58) ± 17.9 (27)	45.5	88	54.2 (58) ± 17.9 (27)	45.5
OpO	15	54.7 (57) ± 19.1 (34)	40.0	15	54.7 (57) ± 19.1 (34)	40.0

The Upper South Regional Ethics Committee, New Zealand, granted ethics approval for the audit, analysis, and publication of these data. Data collection is described in detail in (Chase, Shaw, et al. 2008).

3.3.2. Measuring Variability

Actual $SI\,(n+1)$ values for each day of ICU stay and each diagnostic category (cardiac, gastric, all other, both operative and non-operative in all three types) are compared

to the distributions provided by the stochastic model of Lin et al. (2008) that covers all diagnostic categories and all days of ICU stay. The results thus show the relative and absolute evolution of SI variability ($SI\,(n) \rightarrow SI\,(n+1)$) for a given diagnostic category over time, relative to all patients and days of stay, which should highlight times or diagnostic groups with greater or lesser than average risk.

The percentile of the actual $SI\,(n+1)$ values on their predicted distribution will be illustrated with histograms. If the prediction is perfect (that is, the distribution of actual values is identical to the predicted distribution), every 10% wide interval of the histogram contains 10% of the measurements. This ideal case therefore corresponds to a flat distribution. Kurtic distributions are seen when the actual values were more concentrated at the median than the predicted distribution, suggesting confidence bands could have been tightened. In contrast, U-shaped distributions indicate cases where confidence bands should be widened due to increased variability.

As already mentioned, the investigations for SI variability will be based on the accuracy of these predictions, i.e. we will call a patient variable if the predictions are not accurate (the actual values are not following the predicted distribution). First, the present insulin sensitivity ($SI\,(n)$) is identified, then, the cohort model is used to predict the distribution of insulin sensitivity at the next time-point ($\widehat{F}_{SI(n+1)}$) for the given $SI\,(n)$. The actual (identified) $SI\,(n+1)$ value might be away from the median of this distribution, and this difference over time going forward is the variability in which we are interested. For this end, predicted SI distribution ($\widehat{F}_{SI_{n+1}}$) will be confronted with actual SI of the next hour (SI_{n+1}). Thus, variability was defined by the position of the realized eventual $SI\,(n+1)$ value relative to its predicted distribution $\widehat{F}_{SI(n+1)}$.

More precisely, two indicators will be defined to assess variability for each patient over a given day, and results are aggregated by diagnostic category.

First, a quadratic indicator is defined as the average of squared deviations of the percentile of the actual $SI\,(n+1)$ value on its predicted distribution (from the overall cohort model) from the ideal 50th percentile:

$$QUAD\,(n+1) = \left[\widehat{F}_{SI_{n+1}}\,(SI_{n+1}) - 0.5\right]^2. \tag{3.1}$$

This value increases the more variable a given patient. The quadratic indicator thus measures overall intra-patient variability.

Second, a one-sided threshold indicator counts the number of $SI\,(n+1)$ values for a given patient that exceed the 90th percentile of $SI\,(n+1)$ in the whole-cohort model of

Lin et al. (2008):

$$OST\,(n+1) = I_{\left\{\widehat{F}_{SI_{n+1}}(SI_{n+1})>0.9\right\}}. \tag{3.2}$$

This indicator thus counts the number of large positive changes in $SI\,(n+1)$ that would induce large drops in glucose level on dosing exogenous insulin based on the $SI\,(n)$ value. A value greater than 10% for a given patient, day or diagnostic category indicates a greater risk for these changes compared to the overall cohort on all days of ICU stay. This indicator thus specifically assesses hypoglycemic risk due to intra-patient variability in insulin sensitivity and its daily evolution.

Hence, these two indicators measure overall variability and hypoglycemic risk from variability. Clinically, the quadratic measure is one of risk to glycemic control performance and outcome arising due to variability in insulin sensitivity, and the one-sided threshold assesses risk to patient safety in glycemic control.

These indicators are illustrated on Figure 3.1, which shows the evolution of the insulin sensitivity of a 67 years old male patient (FT5002) with septic shock principal diagnosis (all other, non-operative category) through 162 hours. Each patient has such a trajectory. For every hour, the distribution of $SI\,(n+1)$ was predicted based on $SI\,(n)$ using the model of Lin et al. (2008), which is illustrated with the underlying colormap representing the cumulative distribution function of the predicted distribution. 50th percentile (i.e. median) of this predicted distribution of $SI\,(n+1)$ is explicitly shown. The Figure also illustrates how these indicators are calculated, showing the predicted distribution and the actual SI for a given hour.

3.3.3. Analysis of Variability

An overall variability score can be calculated for a given diagnosis group by averaging the overall variability scores for patients belonging to that group. However, if the individual length of stay differs, simple arithmetic averaging would assign unequal weights for each patient's measurements. To avoid the problems associated with unequal weighting due to patient discharge, only series of equal length were averaged. In particular, results and analysis were divided by the first 24 hours ("day 1"), second 24 hours ("day 2"), third 24 hours ("day 3"), and remaining time in ICU ("day 4+"). Thus only complete 24 hour intervals were used (except for day 4+, of course) to avoid bias.

Per-patient average penalty score distributions by diagnosis group each day are shown using violin plots (Hintze and Nelson 1998). Violin plots bear similarities to boxplots, but use kernel density estimation to directly convey information on the shape of the distribution for more accurate comparison.

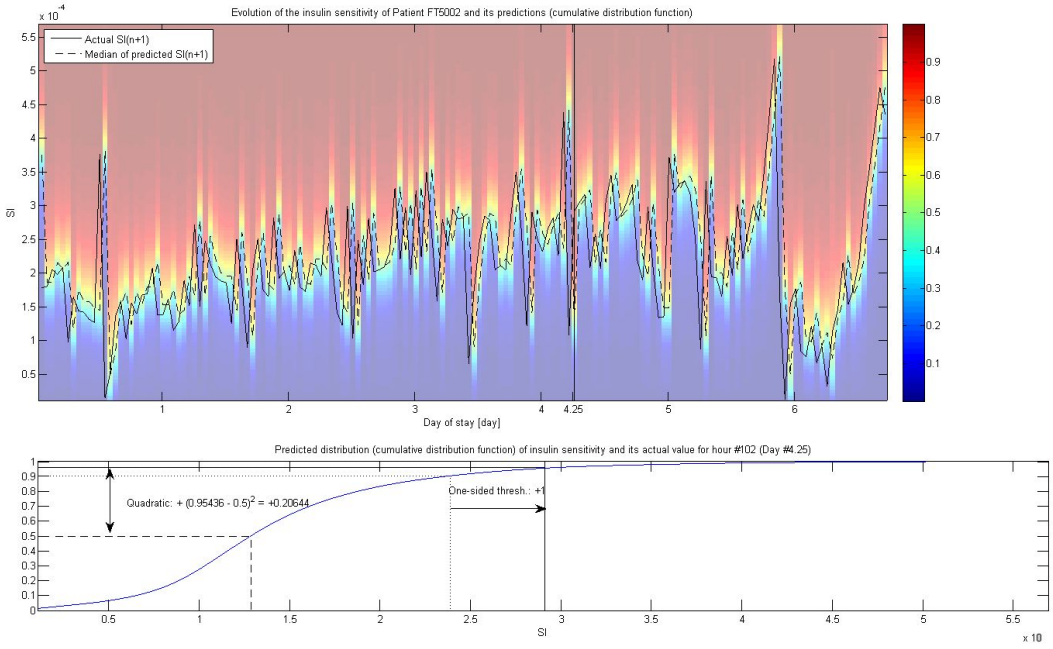


Figure 3.1.: Illustration of the evolution of SI for a given patient (FT5002). Background colors represent the cumulative distribution function of the prediction for $SI\,(n+1)$ based on $SI\,(n)$ using the whole cohort; its 25th, 50th (i.e. median) and 75th percentile is explicitly shown. Lower part of the Figure highlights the calculation of the two indicators using Hour #102 (Day #4.25, marked on the upper part) as an example.

3.3.4. Statistical Methods

To have an overall impression on the effect of the time spent in ICU on the SI variability, a LOWESS estimator (Cleveland 1979) was plotted for the scatterplot of quadratic indicator and time spent (in minutes) per diagnosis group on Figure 3.2. (Plotting the scattergrams itself would have been useless due to the high number of points.) Note that this presentation neglects the dependence between the measurements for the same patient, so it can only be used to give an overall picture of the tendencies.

It is immediately obvious that time has a complex effect on SI variability, which exhibits a biphasic behavior in most of the cases: there is an initial phase with decreasing variability, then a breakpoint comes, and the variability is either decreasing in a drastically slower rate, or stagnates, or – in some cases – it even starts a pronounced elevation. This is worthy of pursuit, despite the fact that the estimation at long length of stays is unreliable due to relatively lower sample size.

I will return to this question in Subsection 3.3.6, but apart from that, let us confine

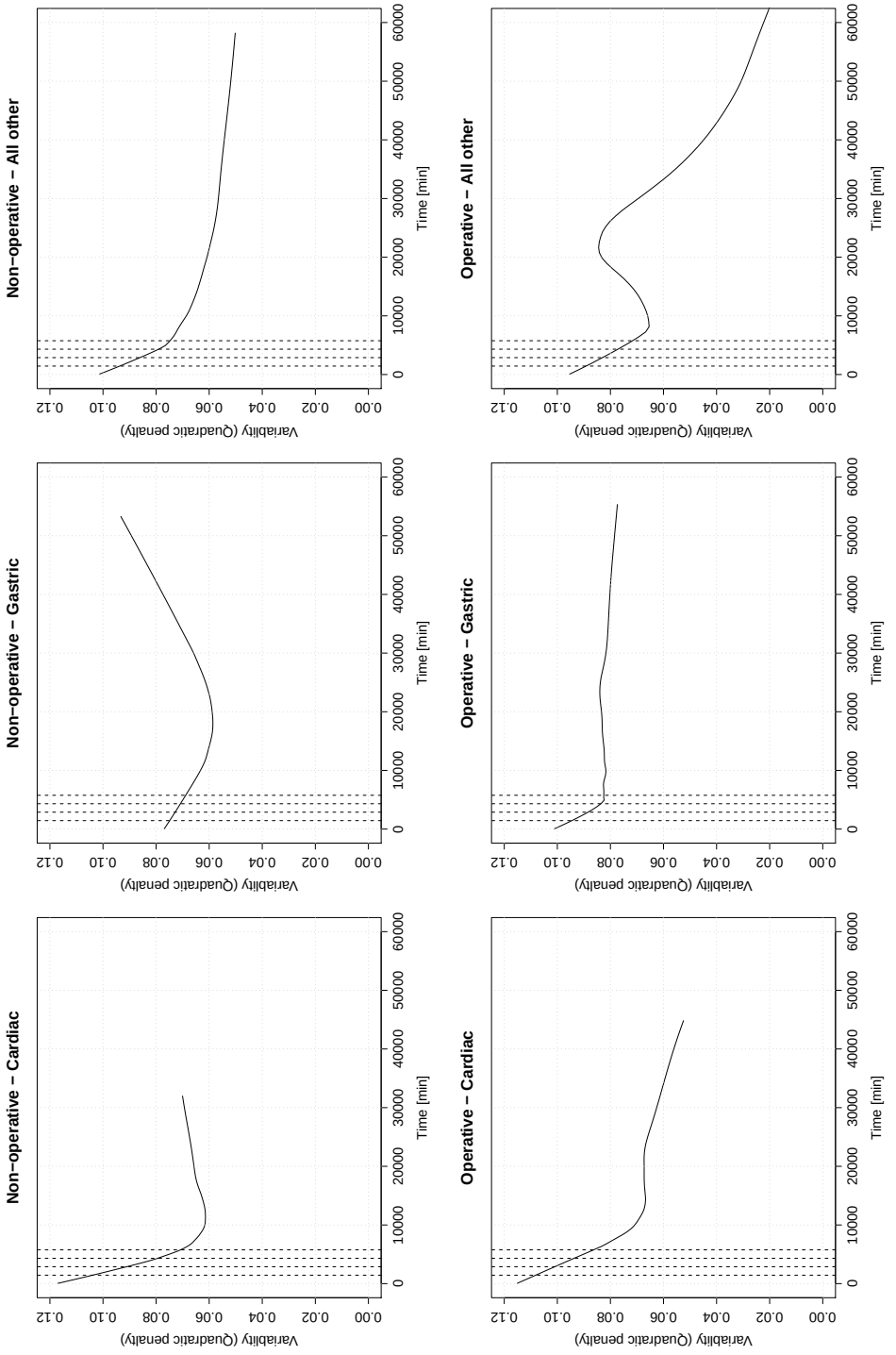


Figure 3.2.: LOWESS estimators for the scatterplot between minute-precision length of stay and quadratic indicator of SI variability, segregated according to diagnosis group. Dashed vertical lines indicate the end of the first four days.

our attention to investigate the early, seemingly mostly linear response of the first few days. (To illustrate this, the first four day is marked on Figure 3.2.) The database was restricted to observations having Time $< 8\,000$ minutes (i.e. the first 5.5 days of stay) for the estimation of the forthcoming models, hence limiting it to the "linearity region" of the SI variability vs. time function, as evidenced by Figure 3.2. The first few days are the most relevant from the clinical management point of view. The linear functional form is also more tractable and easier to estimate, so to perform this "short-term" modeling, linear functional form will be used

As a preliminary investigation, it was first examined whether the differences in SI variability between diagnosis groups are significant, if the database is simply split according to days, and perform separate analyses. For that end, Kruskal–Wallis-test was used, as there was no a priori information on the normality of the data (Dalgaard 2008).

However, to account for the grouping of the data and to explicitly incorporate time, linear mixed-effects modeling was used (J. C. Pinheiro and D. M. Bates 2000; Brown and Prescott 2006). The aim was to find significant differences in SI variability indicators between diagnosis groups and/or days. The (longitudinal) data were arranged in a two-way classification, with time a within-subject factor and diagnosis group considered a between-subject factor. In the developed model, the fixed effects were the Time (time spent in ICU in minutes as a continuous variable) and the Diagnosis (as a nominal factor with 6 levels) without intercept ("cell means coding"). Minute-precision length-of-stay (Time) was used for measuring time to make the estimation of the mixed-effects model possible. The random effect was added with per-patient grouping, with both random intercept and random slope permitted with respect to time, both of which was deemed necessary with LR-test ($p < 0.001$ for both quadratic and one-sided penalty) (Fox and Weisberg 2011). The inclusion of an AR(1) autocorrelation of the within-subject errors was not found to be necessary for the quadratic penalty ($p = 0.9961$) (Fox and Weisberg 2011). The fixed effects interaction terms between Time and Diagnosis were found to be insignificant ($p = 0.8227$ for quadratic penalty, $p = 0.2077$ for one-sided penalty) showing that that the slope with respect to the time spent in ICU does not depend on the diagnosis group, and were thus eliminated. (Effect of Diagnosis was significant ($p < 0.0001$ for both penalty), so the intercept does depend on the diagnosis group.) The resulting statistical model for the quadratic penalty of SI variability was therefore the following:

$$\begin{aligned} Variability_{i,j} = (\beta_{0,NOpC} \cdot Class_{i,NOpC} + \beta_{0,NOpG} \cdot Class_{i,NOpG} + \ldots + \\ + \ldots + \beta_{0,OpO} \cdot Class_{i,OpO} + b_{0,i}) + (\beta_1 + b_{1,i}) \cdot Time_{i,j} + \varepsilon_{i,j}, \end{aligned} \tag{3.3}$$

where i identifies the patient, j identifies the measurement (i.e. $Time_{i,j}$ is the time of the jth measurement on patient i), $Class_{i,C}$ is the indicator variable for Class C (i.e. takes the value of 1 if patient i is in class C, 0 otherwise). For the one-sided threshold penalty – as the response is essentially binary – generalized linear mixed effects (GLME) modeling (Fritzmaurice, Laird, and Ware 2004) was used instead of the traditional linear mixed effects (LME) modeling. The link function was chosen to be logistic, and the distribution family was binomial. For the quadratic penalty, LME modeling was used, but the penalty score was (monotonically) logit-transformed beforehand to map the skewed distribution on $[0, 0.25]$ to an approximately normal one on the real line (Fox and Weisberg 2011). This sacrifies the interpretability of the coefficients for the correct specification of the model, but the former was of little concern, as the numerical values of the coefficients will not be used for further analysis. Linearity for the transformed data was still feasible.

The coefficients are denoted with β for the fixed, and with b for the random effects. The fixed effects coefficient of Time characterizes – for the whole population – how variability changes over time, with positive value implying increasing variability, negative implying decreasing variability, and the absolute value showing the size of this effect. The fixed effects coefficients of diagnosis groups show the estimated variability of a patient in the given diagnosis group when admitted to the ICU.

Restricted maximum likelihood (REML) was used for the estimation of LME models and Laplace-approximation for GLME. Residual variance was rather high in both cases, indicating that the models were only able to capture a small part of the variation – but this is to be expected, given that no information was available other than time spent in ICU and diagnosis group.

After performing ANOVA to assess the significance of main effects, post-hoc testing on significant effects was carried out using Tukey's Honestly Significant Differences (HSD) method (Hsu 1996), providing the correction that takes the multiple comparisons situation into account.

3.3.5. Data Processing

Data processing was done using Mathworks `Matlab` (MATLAB 2009a) (version 2009a). Statistical analysis was performed under the `R` statistical program package (R Core Team 2013), version 3.0.0 with `nlme` package for LME modeling (J. Pinheiro et al. 2013) and `lme4` package for GLME modeling (D. Bates, Maechler, and Bolker 2013). For the details, see Subsection 2.2.2.

Full source code for performing the analysis is given in Appendix B.

Libraries `multcomp` (Hothorn, Bretz, and Westfall 2008), `nlme` (J. Pinheiro et al. 2013),

`lme4` (D. Bates, Maechler, and Bolker 2013) were used.

3.3.6. Long-term analysis

As already discussed, linear approximation is only adequate for "short-term" analysis (i.e. the investigation of the first few days). While this is the most important clinically, we might be also interested in providing a "long-term" model, which should be – therefore – non-linear (Ritz and Streibig 2008).

However, the longer horizon we investigate, the less data we will have to estimate a model. Hence, for the long-term analysis a classical, fixed-effects model was used, and this approach was demonstrated only on the quadratic penalty.

To grasp the long-term evolution of SI variability, a simple piecewise linear regression (with two linear segments and a break point at a random position) seems to be an adequate approximation based on the inspection of Figure 3.2. This, however, has the drawback the the log-likelihood of a regression model will not be differentiable at the break point, which might induce complications for certain models (including ours (Gallop et al. 2011)). To work-around this problem, we instead used a model which connects the linear segments with a small, smooth section, thereby producing a log-likelihood function that is differentiable everywhere. This function form was described by Bacon and Watts (1971), and can be written as follows:

$$f(t) = \alpha_0 + \alpha_1 (t - t_0) + \alpha_2 (t - t_0) \tanh\left(\frac{t - t_0}{\gamma}\right), \tag{3.4}$$

where $\alpha_0, \alpha_1, \alpha_2$ are the parameters of the two segments (α_0 is the function value at the break point, t_0 is the position of the break point, $\alpha_1 - \alpha_2$ is the asymptotical slope before, and $\alpha_1 - \alpha_2$ is the asymptotic slope after the break point) and γ adjusts the smoothness of transition.

3.4. New Scientific Results

In this Section I present the results obtained with the methods described above. The vast majority of the Section deals with the results of short-term modeling (Subsection 3.4.1), but I also cover long-term modeling (Subsection 3.4.2).

3.4.1. Short-term modeling

Figure 3.3 shows the distribution of the percentile of actual $SI\,(n+1)$ on its predicted distribution for different days and diagnosis groups.

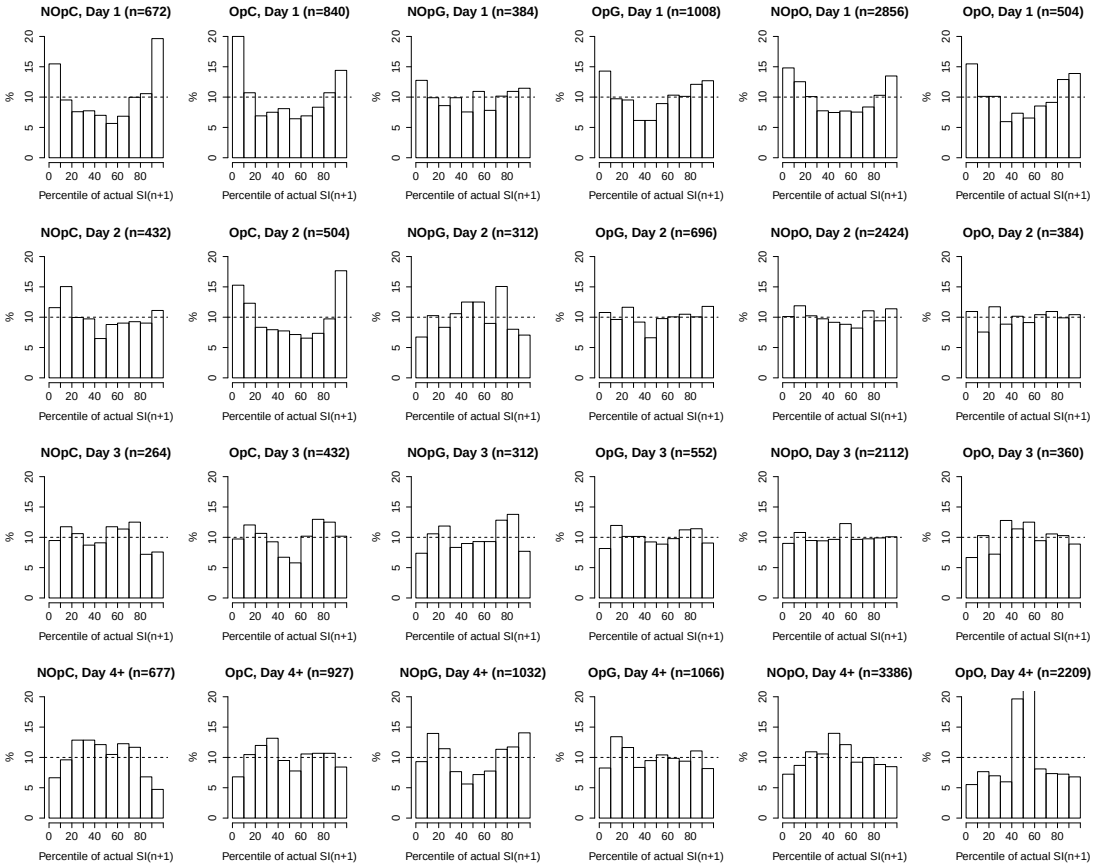


Figure 3.3.: Histograms of the percentile of actual $SI\,(n+1)$ values on their predicted distribution grouped according to day (rows) and diagnosis group (columns). Dashed line indicates the ideal (uniform) case of perfect prediction. The number of hourly measurements which was used to construct the histogram is shown in the title.

The distributions in Figure 3.3 suggest poor coverage of the whole-cohort model on day 1, almost ubiquitously across diagnosis groups. On day 2, every diagnosis group "flattens", except for Operative - Cardiac. On day 3, the predictions are acceptable in every diagnosis group in that the actual distribution of $SI\,(n+1)$ largely matches the whole cohort-predicted distribution. Finally, on day 4 and onwards the coverage is very over-conservative in the Operative - All other category.

Figure 3.4 shows the violin plot of the distributions of per-patient overall variability indicators in different diagnosis groups, segregated according to ICU day and diagnosis group.

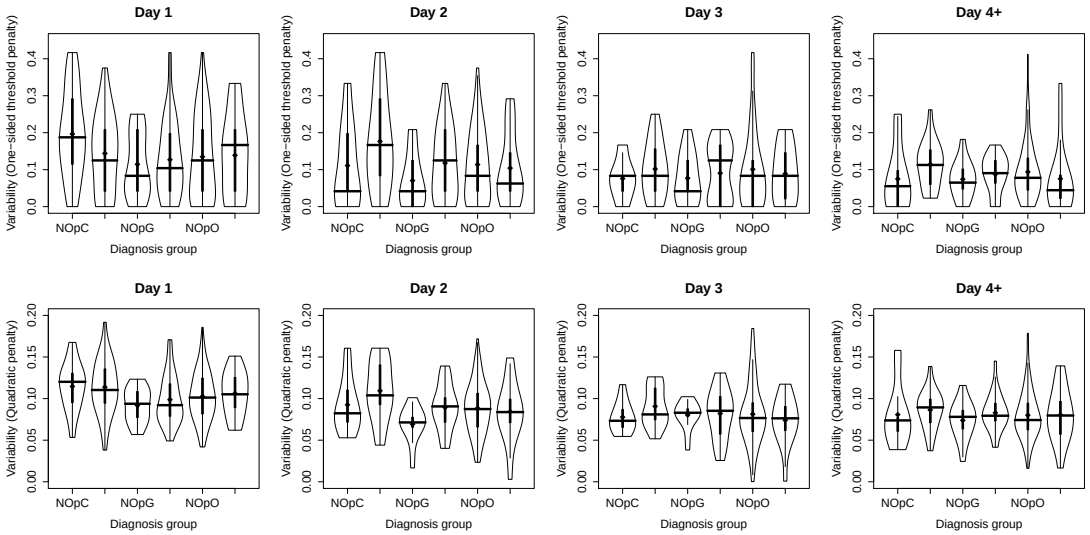


Figure 3.4.: Violin plots of per-patient overall variability scores segregated according to day and diagnosis group. Upper row shows one-sided threshold penalty, while lower row shows the quadratic penalty. Thick vertical lines indicate the interquartile range, the crossing horizontal line is at the median. Dots indicate the mean.

Figure 3.4 (top row) suggests that one-sided threshold penalties exhibit much larger, typically positively-skewed variations. There is a slight trend in the central tendency, as median variability in this indicator appears to decrease as time increases. A trend towards reduced spread in this (one-sided) variability over time is more pronounced, indicating decreasing risk of hypoglycemia over time when all else is equal.

In contrast, quadratic penalties are much more centrally concentrated, and have a smaller coefficient of variation. The continuous lowering of variability over time in every group is also seen, but a reduction in spread is not as pronounced. The two indicators are consistent in assigning ”higher” and ”lower” variabilities similarly over time and diagnostic group, albeit on different scales.

Significance of the between-diagnosis group differences per-day according to both variability indicators is shown on Table 3.2.

It can be seen there are no significant differences in SI variability according to diagnosis group on day 3 and after, no matter which indicator is used. There are no significant

Table 3.2.: p-values of Kruskal–Wallis-test for the equality of average SI variability across diagnosis groups segregated according to day.

Day	**One-sided threshold**	**Quadratic**
Day 1	0.1809	0.02234
Day 2	0.1814	0.02094
Day 3	0.9702	0.6884
Day 4 and onwards	0.1352	0.6499

differences at all (on either day) according to the one-sided threshold penalty, however, there are significant differences on day 1 and on day 2 when the quadratic penalty is employed. (The former observation can be explained by the higher spread of per-patient variability indicators as seen on Figure 3.4.)

For the two cases, where significant difference was detected (day 1 and day 2 with quadratic penalty) post hoc testing was employed. Results are shown on Table 3.3.

For day 1, no significant pairwise difference can be detected, on day 2, Non-operative Gastro and Operative Cardio was significantly different ($p = 0.00472$), while Non-operative All other and Operative Cardio was very close to significance ($p = 0.06305$).

Turning now to the more advanced modeling, parameters of the fitted GLME model (for one-sided threshold penalty) and LME model (for quadratic penalty) are shown in Table 3.4.

As can be seen from Table 3.4, time trend was significant ($p < 0.0001$) with a coefficient of -0.1234/day for the one-sided threshold penalty, and -0.1810/day for the (transformed) quadratic penalty, indicating the decreasing variability over time in both cases. These results also imply a decreasing risk of hypoglycemia inducing variability in insulin sensitivity over time, matching trends in Figure 3.4.

Post-hoc testing for diagnosis groups also revealed significant differences. Using Tukey's HSD method (see Table 3.5), Non-operative – Cardiac group had significantly ($p = 0.0175$) higher variability than Non-operative – Gastric for the one-sided threshold penalty. Non-operative – All other category also exhibited marginally significantly ($p = 0.0832$) lower SI variability than Non-operative - Cardiac patients. The Operative – Cardiac exhibited significantly ($p = 0.0444$) higher variability than Non-operative Gastric for the (transformed) quadratic penalty. These results suggest that the Non-operative – Gastric group is amongst the least variable groups, while the Cardiac groups exhibit the highest variability irrespective of day. These results are consistent with Figure 4, though it is worth noting that cardiac patients "change place" from day 1 to day 2 irrespective

Table 3.3.: p-values for the post-hoc testing of the significant differences (Day 1 and Day 2 with quadratic penalty).

Compared pair	**One-sided threshold**		**Quadratic**	
	Estimate	p	Estimate	p
OpC - NOpC	-0.000871	1.000	0.0167845	0.53874
NOpG - NOpC	-0.022773	0.124	-0.0232665	0.30979
OpG - NOpC	-0.015855	0.219	-0.0031855	0.99934
NOpO - NOpC	-0.011849	0.371	-0.0040197	0.99571
OpO - NOpC	-0.008392	0.914	-0.0081472	0.97213
NOpG - OpC	-0.021902	0.125	-0.0400510	0.00472
OpG - OpC	-0.014984	0.212	-0.0199700	0.22025
NOpO - OpC	-0.010978	0.357	-0.0208042	0.06305
OpO - OpC	-0.007521	0.933	-0.0249317	0.15341
OpG - NOpG	0.006918	0.964	0.0200809	0.37789
NOpO - NOpG	0.010924	0.711	0.0192468	0.28641
OpO - NOpG	0.014381	0.661	0.0151193	0.77735
NOpO - OpG	0.004006	0.971	-0.0008342	0.99999
OpO - OpG	0.007463	0.927	-0.0049617	0.99542
OpO - NOpO	0.003457	0.996	-0.0041275	0.99617

of penalty: Non-operative - Cardiac patients are more variable than Operative – Cardiac group on day 1, but this order is reversed from day 2 onwards.

3.4.2. Long-term modeling

As already discussed, fixed effects modeling was only performed for the long-term analysis.

Fixed effects modeling is possible via non-linear least squares regression (Gallant 2009) using the Bacon–Watts function specified in (3.4). This was done separately for diagnostic groups, with regressions fit for every patient. (Fitting was done through Levenberg-Marquard (Gallant 2009) to optimize convergence.) Results (distribution of the coefficients) are shown on Figure 3.5.

One can see that break points are estimated to be at a short length of stay compared to Figure 3.2, but the direction of the slopes matches the expectations: before the breakpoints, the slopes are typically negative, while after the breakpoints they are of much smaller magnitude on the one hand, and are also closer to zero, or even positive. (For

Table 3.4.: Summary of the estimated fixed effect coefficients of the LME model for (logit-transformed) quadratic penalty and the GLME model for the one-sided threshold penalty, and the p-value for the test of significance for Time. The coefficient of Time is given both per minute and per day ($24 \cdot 60 = 1440$ times the former).

Variable	**One-sided pen.**	**(Transformed) Quadratic pen.**
Non-operative – Cardiac	-1.5807	-0.5033
Operative – Cardiac	-1.9092	-0.4427
Non-operative – Gastric	-2.3532	-1.0480
Operative – Gastric	-1.8791	-0.6922
Non-operative – All other	-1.9903	-0.7350
Operative – All other	-2.0911	-0.8467
Time (per minute)	-0.00008571	-0.0001257
Time (per day)	-0.1234224	-0.1810
	$p < 0.0001$	$p < 0.0001$

example, in the Non-operative Gastric group which also completely matches Figure 3.2.)

3.5. Discussion and Practical Applicability of the Results

Clinically, those results indicate a decreasing likelihood of hypoglycemia induced by large rises (variations) in insulin sensitivity over short measurement and intervention intervals as days of ICU stay increase based on the one-sided threshold results. The overall risk of increased variability of both forms (one-sided and quadratic indicators) by diagnostic category is highest for Cardiac patient groups.

This latter observation is matching the increased hypoglycemia observed in glycemic control studies in these cohorts (e.g. (Preiser et al. 2009)). The highest variability on day 1 is consistent with the increased hypoglycemia and range observed in the first 24 hours in the study by Bagshaw et al. (2009), which was associated with increased risk of death. The overall higher variability (quadratic measure) on day 1 in all groups is also reflective of increased hypoglycemia and variability reported in most glycemic control studies irrespective of cohort (Griesdale et al. 2009; Bagshaw et al. 2009).

The major strength of this approach is that it also provides a rigorous statistical framework, which makes the quantification of these effects possible. It is, however, limited in some sense because it is inherently linked to the SPRINT protocol (as it

Table 3.5.: Estimates of differences and the p-values for the test of their significance (using Tukey-HSD post hoc testing for the multiple comparisons situation) for the pairwise comparison of diagnostic categories.

Comparison	**One-sided penalty**		**(Transformed) Quadratic penalty**	
	Estimate	p	Estimate	p
OpC – NOpC	-0.3285	0.4188	0.0606	0.9992
NOpG – NOpC	-0.7724	0.0172	-0.5451	0.1505
OpG – NOpC	-0.2984	0.5130	-0.1889	0.8637
NOpO – NOpC	-0.4096	0.0835	-0.2317	0.6190
OpO – NOpC	-0.5104	0.1438	-0.3434	0.5038
NOpG – OpC	-0.4440	0.3607	-0.6057	0.0444
OpG – OpC	0.0300	1.0000	-0.2495	0.4946
NOpO – OpC	-0.0811	0.9890	-0.2923	0.1525
OpO – OpC	-0.1819	0.9335	-0.4040	0.2077
OpG – NOpG	0.4740	0.2765	0.3563	0.5179
NOpO – NOpG	0.3628	0.5024	0.3135	0.5799
OpO – NOpG	0.2621	0.9034	0.2017	0.9539
NOpO – OpG	-0.1112	0.9503	-0.0428	0.9992
OpO – OpG	-0.2120	0.8732	-0.1545	0.9518
OpO – NOpO	-0.1008	0.9919	-0.1117	0.9817

interprets variability as the deviation of the actual SI from its prediction provided by the particular algorithm in that protocol).

The physiological causes of this variability have links to the counter-regulatory and oxidative stress responses, and inflammatory acute immune response typically seen in hyperglycemic critically ill patients. That the variability declines over days 1-4 as the acute phase passes also matches expectations and physiological observations. Drug therapies, such as glucocorticoid or inotrope use (Pretty et al. 2011) among others, may also be implicated as a causative factor. However, the high level of patient-specificity observed within any group makes determining specific causes or magnitude of effect difficult.

For glycemic control, high levels of variability combined with infrequent blood glucose measurement are a major disincentive to higher insulin doses and/or low glycemic targets. The only study to reduce both mortality and hypoglycemia (Chase, Shaw, et al. 2008) was notable in modulating both insulin and nutrition inputs to achieve good control with

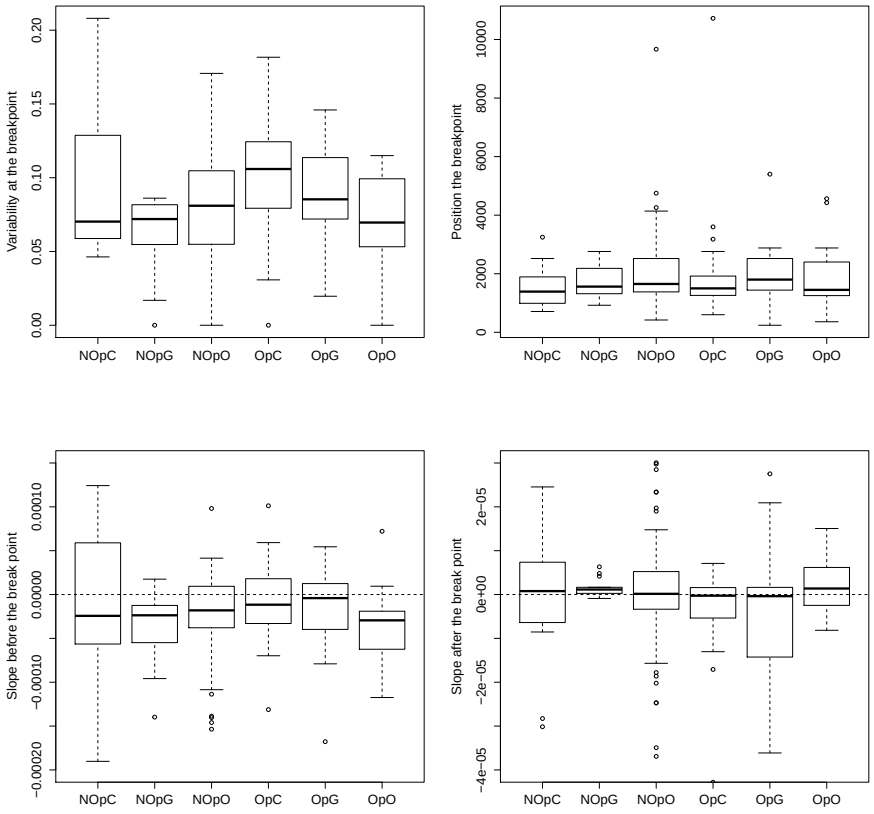


Figure 3.5.: Distribution of the parameters for the per-patient non-linear regression by diagnosis group.

lesser insulin and thus reduce hypoglycemic risk. Hence, either higher targets (Moghissi et al. 2009) and/or adding nutritional intake into consideration in providing glycemic control (Suhaimi et al. 2010) must be considered for at least some diagnostic groups (e.g Cardiac patients) and days of ICU stay (day 1) based on these results.

While on the short-term, linear models seem to provide an adequate fit for the SI variability, yielding the results discussed above, the long-term modeling can only be done in a manner that incorporates the biphasic nature of the insulin sensitivity variability. The results show a possible way: fixed-effects modeling using the non-linear Bacon–Watts function form (which is closely piecewise linear, but with a differentiable log-likelihood everywhere) provides a proper way to capture the nature of the evolution of SI.

This demonstrated that the long-term evolution is indeed biphasic in most of the cases. The early phase response (decreasing variability) is analyzed in detail above, while in the

long run, this variability stalls, or even starts to increase. The developed model permits not only to qualitatively assess this, but also to quantify these tendencies.

3.6. Conclusion

Inter-patient variability in insulin sensitivity peaks on day 1 across diagnostic groups and indicators. Operative – All other patients are more predictable after day 4 than an all patients and days of stay model accounted for, shown by conservative coverage. The distribution of overall intra-patient variability assessed per-patient and the mixed-effects model shows there are distinctive differences between diagnosis groups, irrespective of the time spent in the ICU. In particular, the Non-operative – Gastric group exhibits the smallest variability, while Cardiac groups are amongst the most variable. Clinically, these results show decreasing risk of hypoglycemia as length of stay increases, as well as some reduction in glycemic variability when all else is equal. The overall results can be used to guide the design and implementation of glycemic management specific to diagnosis group and ICU day of stay to improve control and reduce risk.

Thesis 2. Modeling and Evaluating the Performance of Tight Glycemic Control Protocols.

Thesis 2

I have developed a novel methodology to evaluate and model the insulin sensitivity variability and its evolution over time for patients in different diagnosis groups. This also makes the more thorough investigation of the performance of tight glycemic control protocols possible.

Relevant own publications pertaining to this thesis group: [F-14; F-10; F-16].

4. Conclusion

This dissertation presented two applications of biostatistics in the analysis of pathophysiological processes.

The first thesis group investigated questions about obesity, which is the in focus of public health for decades. I now examined the effects of obesity on the human body by analyzing how laboratory parameters are altered by overweight and obesity. To my recent knowledge, this was the first investigation to comprehensively address every routinely used laboratory parameters and to address their multivariate structure. For that end, I developed a novel methodology that provides a complete framework for such investigations. I implemented this methodology as well to provide informatics support for the real-life application of my approach. This treatment also included the analysis of a non-representative Hungarian study, which was performed specifically for this purpose, and – to my best knowledge – is the first study to address this question on Hungarian adolescents.

Nevertheless, there is still room for improvement. By using databases that include adults as well, it is possible to base on larger sample size, on the one hand, and also to make inference on the effect of age on the investigated questions. As far as the Hungarian database is concerned, its convenience sample nature limits the inferences we can draw from it. It would be greatly beneficial from the public health point of view to perform a representative Hungarian study that includes demographic, anthropometric and laboratory parameters (and, perhaps, other relevant indicators as well). Such study would be useful outside our question as well.

The other thesis group described a problem about tight glycemic control protocols. I developed a statistical method that provides objective, quantitative evaluation of how well the protocol predicts the insulin sensitivity of a patient (which is one of the critical steps for such protocols). The model considers both the patient's diagnosis group, and the evolution of his/her state over time. In addition to the objective assessment, my model can formulate advices, down to the clinical level, on how to improve such protocols.

One of the main development possibilities here is the extending to other TGC protocols. Here I only analyzed the SPRINT protocol, while many other is also available. Analyzing

further protocols would be especially interesting as it would create a possibility to compare different protocols to each other, and draw objective conclusions on their effectiveness.

A common possibility for improvement is the inclusion of, and application of biostatistics on control engineering which is already extensively used in modeling (Mandal 2006; Ogata 2010), specifically in the problems of public health too (Kovács, Szalay, Tamás Ferenci, Sápi, et al. 2012; Makroglou, Li, and Kuang 2006; Cobelli et al. 2009).

Both thesis groups involved the development of computer programs that implemented the introduced methodologies and statistical models. I laid emphasis on this to show how modern applied informatics supports the work of biostatisticians, as discussed in the Introduction.

Bibliography

References

Andersen, R. (2003). *Obesity: Etiology, Assessment, Treatment, and Prevention*. Human Kinetics. ISBN: 9780736003285.

Antal, Magda, Szabolcs Péter, Lajos Biró, Katalin Nagy, Andrea Regöly-Mérei, Györgyi Arató, Csaba Szabó, and Eva Martos (2009). "Prevalence of underweight, overweight and obesity on the basis of body mass index and body fat percentage in Hungarian schoolchildren: representative survey in metropolitan elementary schools". In: *Annals of Nutrition and Metabolism* 54.3, pp. 171–176. ISSN: 1421-9697.

Armitage, P., G. Berry, and J.N.S. Matthews (2008). *Statistical Methods in Medical Research*. Wiley. ISBN: 9780470775349.

Ausk, Karlee J. and George N. Ioannou (2008). "Is Obesity Associated With Anemia of Chronic Disease? A Population-based Study". In: *Obesity* 16.10, pp. 2356–2361. ISSN: 1930-739X. DOI: 10.1038/oby.2008.353. URL: http://dx.doi.org/10.1038/oby.2008.353.

Bacon, David W and Donald G Watts (1971). "Estimating the transition between two intersecting straight lines". In: *Biometrika* 58.3, pp. 525–534.

Bagshaw, S, R Bellomo, M Jacka, M Egi, G Hart, C George, and t. A. C. M. Committee (2009). "The impact of early hypoglycemia and blood glucose variability on outcome in critical illness". In: *Crit Care* 13, R91.

Bastard, Jean-Philippe, Mustapha Maachi, Claire Lagathu, Min Ji Kim, Martine Caron, Hubert Vidal, Jacqueline Capeau, and Bruno Feve (2006). "Recent advances in the relationship between obesity, inflammation, and insulin resistance." In: *European Cytokine Network* 17.1, pp. 4–12. ISSN: 1148-5493.

Bates, Douglas, Martin Maechler, and Ben Bolker (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2. URL: http://CRAN.R-project.org/package=lme4.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal*

Statistical Society. Series B (Methodological) 57.1, pp. 289–300. ISSN: 00359246. DOI: 10.2307/2346101.

Berman, J.J. (2013). *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Elsevier Science & Technology Books. ISBN: 9780124045767.

Bertsekas, D.P. (1996). *Constrained optimization and Lagrange multiplier methods*. Optimization and neural computation series. Athena Scientific. ISBN: 9781886529045.

Best, D. J. and D. E. Roberts (1975). "Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.3, pp. 377–379. ISSN: 00359254.

Bo, S., R. Rosato, G. Ciccone, R. Gambino, M. Durazzo, L. Gentile, M. Cassader, P. Cavallo-Perin, and G. Pagano (2009). "What predicts the occurrence of the metabolic syndrome in a population-based cohort of adult healthy subjects?" In: *Diabetes/Metabolism Research and Reviews* 25.1, pp. 76–82. ISSN: 1520-7560. DOI: 10.1002/dmrr.910. URL: http://dx.doi.org/10.1002/dmrr.910.

Brown, H and R Prescott (2006). *Applied Mixed Models in Medicine*. New York: Wiley.

Brunkhorst, Frank M., Christoph Engel, Frank Bloos, Andreas Meier-Hellmann, Max Ragaller, Norbert Weiler, Onnen Moerer, Matthias Gruendling, Michael Oppert, Stefan Grond, Derk Olthoff, Ulrich Jaschinski, Stefan John, Rolf Rossaint, Tobias Welte, Martin Schaefer, Peter Kern, Evelyn Kuhnt, Michael Kiehntopf, Christiane Hartog, Charles Natanson, Markus Loeffler, and Konrad Reinhart (2008). "Intensive Insulin Therapy and Pentastarch Resuscitation in Severe Sepsis". In: *New England Journal of Medicine* 358.2, pp. 125–139. DOI: 10.1056/NEJMoa070716. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMoa070716. URL: http://www.nejm.org/doi/full/10.1056/NEJMoa070716.

Burke, Valerie (2006). "Obesity in childhood and cardiovascular risk". In: *Clinical and Experimental Pharmacology and Physiology* 33.9, pp. 831–837. ISSN: 1440-1681. DOI: 10.1111/j.1440-1681.2006.04449.x. URL: http://dx.doi.org/10.1111/j.1440-1681.2006.04449.x.

Cacoullos, Theophilos (1966). "Estimation of a multivariate density". In: *Annals of the Institute of Statistical Mathematics* 18.1, pp. 179–189. ISSN: 0020-3157. DOI: 10.1007/BF02869528. URL: http://dx.doi.org/10.1007/BF02869528.

Casaer, Michael P., Dieter Mesotten, Greet Hermans, Pieter J. Wouters, Miet Schetz, Geert Meyfroidt, Sophie Van Cromphaut, Catherine Ingels, Philippe Meersseman, Jan Muller, Dirk Vlasselaers, Yves Debaveye, Lars Desmet, Jasperina Dubois, Aime Van Assche, Simon Vanderheyden, Alexander Wilmer, and Greet Van den Berghe (2011). "Early versus Late Parenteral Nutrition in Critically Ill Adults". In: *New*

England Journal of Medicine 365.6, pp. 506–517. DOI: 10.1056/NEJMoa1102662. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMoa1102662. URL: http://www.nejm.org/doi/full/10.1056/NEJMoa1102662.

Casella, G. and R.L. Berger (2002). *Statistical inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning. ISBN: 9780534243128.

Centers for Disease Control and Prevention (2013). *Growth Chart*. http://www.cdc.gov/growthcharts/. [Online; accessed 26. 03. 2013.] URL: http://www.cdc.gov/growthcharts/.

Centers for Disease Control and Prevention, National Center for Health Statistics (2006). *Analytic and reporting guidelines, The National Health and Nutrition Examination Survey (NHANES)*. http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf. [Online; accessed 21. 04. 2013.] URL: http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf.

— (2013a). *National Health and Nutrition Examination Survey*. http://www.cdc.gov/nchs/nhanes.htm. [Online; accessed 21. 04. 2013.] URL: http://www.cdc.gov/nchs/nhanes.htm.

— (2013b). *National Health and Nutrition Examination Survey, NHANES 2009-2010*. http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx. [Online; accessed 21. 04. 2013.] URL: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx.

— (2013c). *National Health and Nutrition Examination Survey, NHANES 2011-2012*. http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx. [Online; accessed 21. 04. 2013.] URL: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx.

Chacón, José E. (2009). "Data-driven choice of the smoothing parametrization for kernel density estimators". In: *Canadian Journal of Statistics* 37.2, pp. 249–265. ISSN: 1708-945X. DOI: 10.1002/cjs.10016. URL: http://dx.doi.org/10.1002/cjs.10016.

Chacón, José E., T. Duon, and M. P. Wand (2009). "Asymptotics for general multivariate kernel density derivative estimators". In: URL: http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1058&context=cssmwp.

Chase, J G, Aaron J. Le Compte, Fatanah Suhaimi, Geoffrey M. Shaw, Adrienne Lynn, Jessica Lin, Christopher G. Pretty, Normy Razak, Jacquelyn D. Parente, Christopher E. Hann, Jean-Charles Preiser, and Thomas Desaive (2011). "Tight glycemic control in critical care – The leading role of insulin sensitivity and patient variability: A review and model-based analysis". In: *Computer Methods and Programs in Biomedicine*

102.2, pp. 156–171. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2010.11.006. URL: http://www.sciencedirect.com/science/article/pii/S0169260710002828.

Chase, J G, G Shaw, A Le Compte, T Lonergan, M Willacy, X W Wong, J Lin, T Lotz, D Lee, and C Hann (2008). "Implementation and evaluation of the SPRINT protocol for tight glycaemic control in critically ill patients: a clinical practice change". In: *Crit Care* 12, R45.

Cheng, S. and Nicholas J. Higham (1998). "A Modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization". In: *SIAM Journal on Matrix Analysis and Applications* 19.4, pp. 1097–1110. DOI: 10.1137/S0895479896302898. eprint: http://epubs.siam.org/doi/pdf/10.1137/S0895479896302898. URL: http://epubs.siam.org/doi/abs/10.1137/S0895479896302898.

Cho, Hye Min, Hyeon Chang Kim, Ju-Mi Lee, Sun Min Oh, Dong Phil Choi, and Il Suh (2012). "The association between serum albumin levels and metabolic syndrome in a rural population of Korea". In: *Journal of Preventive Medicine and Public Health* 45.2, pp. 98–104. ISSN: 2233-4521.

Chok, Nian Shong (2010). "Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data". BSc Thesis. University of Pittsburgh.

Clark-Carter, D. (2009). *Quantitative Psychological Research: The Complete Student's Companion*. Taylor & Francis. ISBN: 9780203870709.

Cleveland, W S (1979). "Robust locally weighted regression and smoothing scatterplots". In: *J Amer Statist Assoc* 74, pp. 829–836.

Cobelli, C., C. Dalla Man, G. Sparacino, L. Magni, G. De Nicolao, and B.P. Kovatchev (2009). "Diabetes: Models, Signals, and Control". In: *IEEE Reviews in Biomedical Engineering* 2, pp. 54–96. ISSN: 1937-3333. DOI: 10.1109/RBME.2009.2036073.

Cole, T. J. (1990). "The LMS method for constructing normalized growth standards." In: *European Journal of Clinical Nutrition* 44.1, pp. 45–60. ISSN: 0954-3007.

Cole, T. J., M. S. Faith, A. Pietrobelli, and M. Heo (2005). "What is the best measure of adiposity change in growing children: BMI, BMI %, BMI z-score or BMI centile?" In: *European Journal of Clinical Nutrition* 59.3, pp. 419–25. ISSN: 0954-3007.

Colicchio, P., G. Tarantino, F. del Genio, P. Sorrentino, G. Saldalamacchia, C. Finelli, P. Conca, F. Contaldo, and F. Pasanisi (2005). "Non-alcoholic fatty liver disease in young adult severely obese non-diabetic patients in South Italy". In: *Annals of Nutrition and Metabolism* 49.5, pp. 289–95.

Dalgaard, P. (2008). *Introductory Statistics with R*. Statistics and Computing. Springer. ISBN: 9780387790534.

David, S. T., M. G. Kendall, and A. Stuart (1951). "Some Questions of Distribution in the Theory of Rank Correlation". In: *Biometrika* 38.1/2, pp. 131–140. ISSN: 00063444.

Deckelbaum, Richard J. and Christine L. Williams (2001). "Childhood Obesity: The Health Issue". In: *Obesity Research* 9.S11, 239S–243S. ISSN: 1550-8528. DOI: 10.1038/oby.2001.125. URL: http://dx.doi.org/10.1038/oby.2001.125.

Devroye, L. and L. Györfi (1985). *Nonparametric density estimation: the L1 view*. Wiley series in probability and mathematical statistics. Wiley. ISBN: 9780471816461.

Dubern, Beatrice, Jean-Philippe Girardet, and Patrick Tounian (2006). "Insulin resistance and ferritin as major determinants of abnormal serum aminotransferase in severely obese children". In: *International Journal of Pediatric Obesity* 1.2, pp. 77–82. ISSN: 1747-7174. DOI: 10.1080/17477160600569594. URL: http://dx.doi.org/10.1080/17477160600569594.

Duong, Tarn (2013). *ks: Kernel smoothing*. R package version 1.8.12. URL: http://CRAN.R-project.org/package=ks.

Duong, Tarn and Martin L. Hazelton (2005). "Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation". In: *Scandinavian Journal of Statistics* 32.3, pp. 485–506. ISSN: 1467-9469. DOI: 10.1111/j.1467-9469.2005.00445.x. URL: http://dx.doi.org/10.1111/j.1467-9469.2005.00445.x.

Ebbeling, Cara B., Dorota B. Pawlak, and David S. Ludwig (2002). "Childhood obesity: public-health crisis, common sense cure". In: *The Lancet* 360.9331, pp. 473–482. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(02)09678-2. URL: http://www.sciencedirect.com/science/article/pii/S0140673602096782.

Egi, M, R Bellomo, E Stachowski, C J French, and G Hart (2006). "Variability of blood glucose concentration and short-term mortality in critically ill patients". In: *Anesthesiology* 105, pp. 244–252.

Eknoyan, Garabed (2008). "Adolphe Quetelet (1796–1874)–the average man and indices of obesity". In: *Nephrology Dialysis Transplantation* 23.1, pp. 47–51. DOI: 10.1093/ndt/gfm517. eprint: http://ndt.oxfordjournals.org/content/23/1/47.full.pdf+html. URL: http://ndt.oxfordjournals.org/content/23/1/47.abstract.

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

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis*. Wiley series in probability and statistics. Wiley. ISBN: 9780470978443.

Everitt, Brian and Torsten Hothorn (2011). "Cluster Analysis". English. In: *An Introduction to Applied Multivariate Analysis with R*. Use R. Springer New York, pp. 163–200. ISBN: 978-1-4419-9649-7. DOI: 10.1007/978-1-4419-9650-3_6. URL: http://dx.doi.org/10.1007/978-1-4419-9650-3_6.

Ferroni, Patrizia, Stefani Basili, Angela Falco, and Giovanni Davi (2004). "Inflammation, insulin resistance, and obesity". In: *Current Atherosclerosis Reports* 6 (6), pp. 424–431. ISSN: 1523-3804. DOI: 10.1007/s11883-004-0082-x. URL: http://dx.doi.org/10.1007/s11883-004-0082-x.

Finfer, S. and The NICE-SUGAR Study Investigators (2009). "Intensive versus Conventional Glucose Control in Critically Ill Patients". In: *New England Journal of Medicine* 360.13, pp. 1283–1297. DOI: 10.1056/NEJMoa0810625. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMoa0810625. URL: http://www.nejm.org/doi/full/10.1056/NEJMoa0810625.

Fix, E. and J. L. Hodges (1951). *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. Tech. rep. Project 21-49-004, Report Number 4. USAF School of Aviation Medicine, Randolf Field, Texas, pp. 261–279.

Flegal, K. M., B. K. Kit, H. Orpana, and B. I. Graubard (2013). "Association of all-cause mortality with overweight and obesity using standard body mass index categories: A systematic review and meta-analysis". In: *JAMA* 309.1, pp. 71–82. DOI: 10.1001/jama.2012.113905. eprint: /data/Journals/JAMA/926163/jrv120009_71_82.pdf. URL: +%20http://dx.doi.org/10.1001/jama.2012.113905.

Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer Texts in Statistics. Springer. ISBN: 9780387982069.

Fox, J and S Weisberg (2011). *An R Companion to Applied Regression*. Thousand Oaks: Sage.

Fritzmaurice, G M, N M Laird, and J H Ware (2004). *Applied Longitudinal Analysis*. Hoboken: Wiley-Interscience.

Gallant, A.R. (2009). *Nonlinear Statistical Models*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780470317372.

Gallop, Robert J, Sona Dimidjian, David C Atkins, and Vito Muggeo (2011). "Quantifying treatment effects when flexibly modeling individual change in a nonlinear mixed effects model". In: *Journal of Data Science* 9, pp. 221–241.

Gholam, Pierre M., Louis Flancbaum, Jason T. Machan, Douglas A Charney, and Donald P. Kotler (2007). "Nonalcoholic fatty liver disease in severely obese subjects". In: *Am J Gastroenterol* 102.2, pp. 399–408. ISSN: 0002-9270.

Gilbert-Diamond, D., A. Baylin, M. Mora-Plazas, and E. Villamor (2012). "Chronic inflammation is associated with overweight in Colombian school children". In: *Nutr Metab Cardiovasc Dis* 22.3, pp. 244–51. ISSN: 1590-3729.

Gill, P.E., W. Murray, and M.H. Wright (1981). *Practical optimization*. Academic Press. ISBN: 9780122839504.

Glynn, E. F. (2005). *Correlation 'Distances' and Hierarchical Clustering*. http://research.stowers-institute.org/efg/R/Visualization/cor-cluster/index.htm. [Online; accessed 26. 03. 2013.] URL: http://research.stowers-institute.org/efg/R/Visualization/cor-cluster/index.htm.

Good, P.I. (2000). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer series in statistics. Springer. ISBN: 9780387988986.

— (2006). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser Boston. ISBN: 9780817643867.

Griesdale, D E, R J de Souza, R M van Dam, D K Heyland, D J Cook, A Malhotra, R Dhaliwal, W R Henderson, D R Chittock, S Finfer, and D Talmor (2009). "Intensive insulin therapy and mortality among critically ill patients: a meta-analysis including NICE-SUGAR study data". In: *CMAJ* 180, pp. 821–827.

Guh, Daphne, Wei Zhang, Nick Bansback, Zubin Amarsi, C Laird Birmingham, and Aslam Anis (2009). "The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis". In: *BMC Public Health* 9.1, p. 88. ISSN: 1471-2458. DOI: 10.1186/1471-2458-9-88. URL: http://www.biomedcentral.com/1471-2458/9/88.

Hall, Peter, J. S. Marron, and Byeong U. Park (1992). "Smoothed cross-validation". In: *Probability Theory and Related Fields* 92 (1), pp. 1–20. ISSN: 0178-8051. DOI: 10.1007/BF01205233. URL: http://dx.doi.org/10.1007/BF01205233.

Han, J., M. Kamber, and J. Pei (2011). *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. ISBN: 9780123814807.

Hand, D.J., H. Mannila, and P. Smyth (2001). *Principles of data mining*. MIT Press. ISBN: 9780262082907.

Hardle, W. (2004). *Nonparametric and semiparametric models*. Springer Series in Statistics Series. Springer-Verlag GmbH. ISBN: 9783540207221.

Higham, Nicholas J. (1988). "Computing a nearest symmetric positive semidefinite matrix". In: *Linear Algebra and its Applications* 103, pp. 103–118. ISSN: 0024-3795. DOI: 10.1016/0024-3795(88)90223-6. URL: http://www.sciencedirect.com/science/article/pii/0024379588902236.

— (1989). "Matrix Nearness Problems and Applications". In: *Applications of Matrix Theory*. Oxford University Press, pp. 1–27.

— (2002). "Computing the nearest correlation matrix – a problem from finance". In: *IMA Journal of Numerical Analysis* 22.3, pp. 329–343. DOI: 10.1093/imanum/22.3.329.

eprint: http://imajna.oxfordjournals.org/content/22/3/329.full.pdf+html. URL: http://imajna.oxfordjournals.org/content/22/3/329.abstract.

Hintze, J L and R D Nelson (1998). "Violin Plots: A Box Plot-Density Trace Synergism". In: *Amer Statistician* 52, pp. 181–184.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics* 6, pp. 65–70.

Hommel, G. (1988). "A stagewise rejective multiple test procedure based on a modified Bonferroni test". In: *Biometrika* 75.2, pp. 383–386. DOI: 10.1093/biomet/75.2.383. eprint: http://biomet.oxfordjournals.org/content/75/2/383.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/75/2/383.abstract.

Hothorn, Torsten, Frank Bretz, and Peter Westfall (2008). "Simultaneous Inference in General Parametric Models". In: *Biometrical Journal* 50.3, pp. 346–363.

Hsu, J (1996). *Multiple Comparisons: Theory and Methods*. Boca Raton: Chapman and Hall/CRC.

Ishizaka, Nobukazu, Yuko Ishizaka, Ryozo Nagai, Ei-Ichi Toda, Hideki Hashimoto, and Minoru Yamakado (2007). "Association between serum albumin, carotid atherosclerosis, and metabolic syndrome in Japanese individuals". In: *Atherosclerosis* 193.2, pp. 373–379. ISSN: 0021-9150. DOI: 10.1016/j.atherosclerosis.2006.06.031. URL: http://www.sciencedirect.com/science/article/pii/S0021915006003972.

Jäckel, P. (2002). *Monte Carlo Methods in Finance*. The Wiley Finance Series. Wiley. ISBN: 9780471497417.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer. ISBN: 9780387954424.

Joubert, Kálmán, Sarolta Darvay, Gyula Gyenis, ödön Éltetö, Kornélia Mag, van't Hof Martin, and Rózsa Ágfalvi (2006). *Results of the National Longitudinal Child Growth Survey from Birth to the Age of 18 Years [in Hungarian]*. Tech. rep. Központi Statisztikai Hivatal Népességtudományi Kutatóintézet.

Juonala, Markus, Jonna Juhola, Costan G. Magnussen, Peter Würtz, Jorma S. A. Viikari, Russell Thomson, Ilkka Seppälä, Jussi Hernesniemi, Mika Kähönen, Terho Lehtimäki, Mikko Hurme, Risto Telama, Vera Mikkilä, Carita Eklund, Leena Räsänen, Mirka Hintsanen, Liisa Keltikangas-Järvinen, Mika Kivimäki, and Olli T. Raitakari (2011). "Childhood Environmental and Genetic Predictors of Adulthood Obesity: The Cardiovascular Risk in Young Finns Study". In: *Journal of Clinical Endocrinology & Metabolism* 96.9, E1542–E1549. DOI: 10.1210/jc.2011-1243. eprint: http://jcem.endojournals.org/content/96/9/E1542.full.pdf+html. URL: http://jcem.endojournals.org/content/96/9/E1542.abstract.

Kaiser, Henry F. (1958). "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23 (3), pp. 187–200. ISSN: 0033-3123. DOI: 10.1007/BF02289233. URL: http://dx.doi.org/10.1007/BF02289233.

— (1960). "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* 20.1, pp. 141–151. eprint: http://epm.sagepub.com/content/20/1/141.full.pdf+html. URL: http://epm.sagepub.com/content/20/1/141.short.

Kelly, Anthea and Louis Munan (1977). "Haematologic Profile of Natural Populations: Red Cell Parameters". In: *British Journal of Haematology* 35.1, pp. 153–160. ISSN: 1365-2141. DOI: 10.1111/j.1365-2141.1977.tb00570.x. URL: http://dx.doi.org/10.1111/j.1365-2141.1977.tb00570.x.

Kern, Boglárka (2007). "The Prevalence of Overweight and Obesity in Hungarian Children". In: *Intensive course on biological anthropology, 1st Summer School of the European Anthropological Association*. Vol. 1. EAA Summer School eBook, pp. 181–186.

Knol, Dirk L. and Jos M.F. Berge (1989). "Least-squares approximation of an improper correlation matrix by a proper one". English. In: *Psychometrika* 54.1, pp. 53–61. ISSN: 0033-3123. DOI: 10.1007/BF02294448. URL: http://dx.doi.org/10.1007/BF02294448.

König, Wolfgang, Malte Sund, Margit Fröhlich, Hans-Günther Fischer, Hannelore Löwel, Angela Döring, Winston L. Hutchinson, and Mark B. Pepys (1999). "C-Reactive Protein, a Sensitive Marker of Inflammation, Predicts Future Risk of Coronary Heart Disease in Initially Healthy Middle-Aged Men: Results From the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) Augsburg Cohort Study, 1984 to 1992". In: *Circulation* 99.2, pp. 237–242. DOI: 10.1161/01.CIR.99.2.237. eprint: http://circ.ahajournals.org/content/99/2/237.full.pdf+html. URL: http://circ.ahajournals.org/content/99/2/237.abstract.

Kriegel, Hans-Peter, Peer Kröger, Jörg Sander, and Arthur Zimek (2011). "Density-based clustering". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3, pp. 231–240. ISSN: 1942-4795. DOI: 10.1002/widm.30. URL: http://dx.doi.org/10.1002/widm.30.

Krinsley, J S (2003). "Association between hyperglycemia and increased hospital mortality in a heterogeneous population of critically ill patients". In: *Mayo Clin Proc* 78, pp. 1471–1478.

— (2004). "Effect of an Intensive Glucose Management Protocol on the Mortality of Critically Ill Adult Patients". In: *Mayo Clinic Proceedings* 79.8, pp. 992–1000. ISSN:

0025-6196. DOI: 10.4065/79.8.992. URL: http://www.sciencedirect.com/science/article/pii/S002561961162572X.

Krinsley, J S (2008). "Glycemic variability: a strong independent predictor of mortality in critically ill patients". In: *Crit Care Med* 36, pp. 3008–3013.

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, and Clifford L. Johnson (2002). "2000 CDC Growth Charts for the United States: methods and development". In: *Vital Health Stat 11* 246, pp. 1–190. ISSN: 0083-1980.

Lam, Gregory M. and Sohrab Mobarhan (2004). "Central Obesity and Elevated Liver Enzymes". In: *Nutrition Reviews* 62.10, pp. 394–399. ISSN: 1753-4887. DOI: 10.1111/j.1753-4887.2004.tb00010.x. URL: http://dx.doi.org/10.1111/j.1753-4887.2004.tb00010.x.

Lance, G. N. and W. T. Williams (1967). "A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems". In: *The Computer Journal* 9.4, pp. 373–380. DOI: 10.1093/comjnl/9.4.373. URL: http://dx.doi.org/10.1093/comjnl/9.4.373.

Langouche, L, S Vander Perre, P J Wouters, A D'Hoore, T K Hansen, and G Van den Berghe (2007). "Effect of intensive insulin therapy on insulin sensitivity in the critically ill". In: *J Clin Endocrinol Metab* 92, pp. 3890–3897.

Lawlor, Debbie A., Naveed Sattar, George Davey Smith, and Shah Ebrahim (2005). "The Associations of Physical Activity and Adiposity with Alanine Aminotransferase and Gamma-Glutamyltransferase". In: *American Journal of Epidemiology* 161.11, pp. 1081–1088. DOI: 10.1093/aje/kwi125. eprint: http://aje.oxfordjournals.org/content/161/11/1081.full.pdf+html. URL: http://aje.oxfordjournals.org/content/161/11/1081.abstract.

Lehmann, E.L. and G. Casella (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer. ISBN: 9780387985022.

Limpert, E., W. A. Stahel, and M. Abbt (2001). "Log-normal Distributions across the Sciences: Keys and Clues". In: *BioScience* 51.5, pp. 341–352. ISSN: 0006-3568.

Lin, J, D Lee, J G Chase, G M Shaw, A Le Compte, T Lotz, J Wong, T Lonergan, and C E Hann (2008). "Stochastic modelling of insulin sensitivity and adaptive glycemic control for critical care". In: *Comput Methods Programs Biomed* 89, pp. 141–152.

Luxburg, Ulrike (2007). "A tutorial on spectral clustering". English. In: *Statistics and Computing* 17.4, pp. 395–416. ISSN: 0960-3174. DOI: 10.1007/s11222-007-9033-z. URL: http://dx.doi.org/10.1007/s11222-007-9033-z.

Makroglou, Athena, Jiaxu Li, and Yang Kuang (2006). "Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview".

In: *Applied Numerical Mathematics* 56.3-4, pp. 559–573. ISSN: 0168-9274. DOI: 10.1016/j.apnum.2005.04.023. URL: http://www.sciencedirect.com/science/article/pii/S0168927405000929.

Mandal, A.K. (2006). *Introduction to Control Engineering: Modeling, Analysis and Design*. New Age International Pvt. Limited, Publishers. ISBN: 9788122418217.

Maritz, J. S. (1995). *Distribution-free Statistical Methods*. Monographs on Statistics & Applied Probability. Chapman & Hall. ISBN: 9780412552601.

MATLAB (2009a). *version 7.8 (R2009a)*. Natick, Massachusetts: The MathWorks Inc.

McCowen, K C, A Malhotra, and B R Bistrian (2001). "Stress-induced hyperglycemia". In: *Crit Care Clin* 17, pp. 107–124.

Millar, R.B. (2011). *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. Statistics in Practice. Wiley. ISBN: 9781119977711.

Miller, R. G. (1981). *Simultaneous statistical inference*. Springer series in statistics. Springer-Verlag. ISBN: 9780387905488.

Milne, Iain (2012). "Who was James Lind, and what exactly did he achieve". In: *Journal of the Royal Society of Medicine* 105.12, pp. 503–508. DOI: 10.1258/jrsm.2012.12k090. eprint: http://jrs.sagepub.com/content/105/12/503.full.pdf+html. URL: http://jrs.sagepub.com/content/105/12/503.short.

Moghissi, E S, M T Korytkowski, M DiNardo, D Einhorn, R Hellman, I B Hirsch, S E Inzucchi, F Ismail-Beigi, M S Kirkman, and G E Umpierrez (2009). "American Association of Clinical Endocrinologists and American Diabetes Association consensus statement on inpatient glycemic control". In: *Diab Care* 32, pp. 1119–1131.

Moreno, L. A., W. Ahrens, and I. Pigeot (2011). *Epidemiology of Obesity In Children and Adolescents: Prevalence and Etiology*. Springer Series on Epidemiology and Public Health, 2. Springer New York. ISBN: 9781441960399.

Must, Aviva, Richard S Strauss, et al. (1999). "Risks and consequences of childhood and adolescent obesity". In: *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity* 23, S2–11.

Nyberg, Gisela, Ulf Ekelund, Tülay Yucel-Lindberg, Thomas Modeér, and Claude Marcus (2011). "Differences in metabolic risk factors between normal weight and overweight children". In: *International Journal of Pediatric Obesity* 6.3-4, pp. 244–252. ISSN: 1747-7174. DOI: 10.3109/17477166.2011.575226. URL: http://dx.doi.org/10.3109/17477166.2011.575226.

Oda, Eiji and Ryu Kawai (2010). "Comparison between high-sensitivity C-reactive protein (hs-CRP) and white blood cell count (WBC) as an inflammatory component

of metabolic syndrome in Japanese". In: *Internal Medicine* 49.2, pp. 117–124. ISSN: 1349-7235.

Ogata, K. (2010). *Modern control engineering*. Instrumentation and controls series. Prentice Hall. ISBN: 9780136156734.

Ogden, Cynthia L. and Margaret D. Carrol (2010a). *Prevalence of Obesity Among Children and Adolescents: United States, Trends 1963-1965 Through 2007-2008*. NCHS Health E-Stats.

— (2010b). *Prevalence of Overweight, Obesity, and Extreme Obesity Among Adults: United States, Trends 1960-1962 Through 2007-2008*. NCHS Health E-Stats.

Ogden, Cynthia L., Susan Z. Yanovski, Margaret D. Carroll, and Katherine M. Flegal (2007). "The Epidemiology of Obesity". In: *Gastroenterology* 132.6, pp. 2087–2102. ISSN: 0016-5085. DOI: 10.1053/j.gastro.2007.03.052. URL: http://www.sciencedirect.com/science/article/pii/S0016508507005793.

Okorodudu, D. O., M. F. Jumean, V. M. Montori, A. Romero-Corral, V. K. Somers, P. J. Erwin, and F. Lopez-Jimenez (2010). "Diagnostic performance of body mass index to identify obesity as defined by body adiposity: a systematic review and meta-analysis". In: *Int J Obes (Lond)* 34.5, pp. 791–799. ISSN: 1476-5497.

Ong, K. L., A. W. K. Tso, A. Xu, L. S. C. Law, M. Li, N. M. S. Wat, K. A. Rye, T. H. Lam, B. M. Y. Cheung, and K. S. L. Lam (2011). "Evaluation of the combined use of adiponectin and C-reactive protein levels as biomarkers for predicting the deterioration in glycaemia after a median of 5.4 years". In: *Diabetologia* 54 (10), pp. 2552–2560. ISSN: 0012-186X. DOI: 10.1007/s00125-011-2227-0. URL: http://dx.doi.org/10.1007/s00125-011-2227-0.

Organization for Economic Co-operation and Development (2012). *Factbook 2011-2012, Economic, Environmental and Social Statistics*. Organization for Economic Co-operation and Development.

Parzen, Emanuel (1962). "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076.

Pasek, Josh and Alex Tahk (2012). *weights: Weighting and Weighted Statistics*. R package version 0.75. URL: http://CRAN.R-project.org/package=weights.

Pestman, W.R. (2009). *Mathematical Statistics*. De Gruyter Textbook. De Gruyter. ISBN: 9783110208535.

Petersen, Kaare Brandt and Michael Syskind Pedersen (2006). *The matrix cookbook*.

Pinheiro, J C and D M Bates (2000). *Mixed Effects Models in S and S-Plus*. New York: Springer.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-109.

Poole, D. (2010). *Linear Algebra: A Modern Introduction*. Available 2011 Titles Enhanced Web Assign Series. BROOKS COLE Publishing Company. ISBN: 9780538735452.

Preiser, J C, P Devos, S Ruiz-Santana, C Mélot, D Annane, J Groeneveld, G Iapichino, X Leverve, G Nitenberg, P Singer, J Wernerman, M Joannidis, A Stecher, and R Chioléro (2009). "A prospective randomised multi-centre controlled trial on tight glucose control by intensive insulin therapy in adult intensive care units: the Glucontrol study". In: *Intensive Care Med* 35, pp. 1738–1748.

Press, W.H. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press. ISBN: 9780521880688.

Pretty, C, J G Chase, J Lin, G M Shaw, A Le Compte, N Razak, and J D Parente (2011). "Impact of glucocorticoids on insulin resistance in the critically ill". In: *Comput Methods Programs Biomed* 102, pp. 172–180.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Revelle, William (2013). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.3.2. Northwestern University. Evanston, Illinois. URL: http://CRAN.R-project.org/package=psych.

Ritz, C. and J.C. Streibig (2008). *Nonlinear Regression with R*. Use R! Springer. ISBN: 9780387096155.

Rodger, R. S. C., K. Fletcher, B. J. Fail, H. Rahman, L. Sviland, and P. J. Hamilton (1987). "Factors influencing haematological measurements in healthy adults". In: *Journal of Chronic Diseases* 40.10, pp. 943–947. ISSN: 0021-9681. DOI: 10.1016/0021-9681(87)90144-5. URL: http://www.sciencedirect.com/science/article/pii/0021968187901445.

Romero-Corral, A., V. K. Somers, J. Sierra-Johnson, R. J. Thomas, M. L. Collazo-Clavell, J. Korinek, T. G. Allison, J. A. Batsis, F. H. Sert-Kuniyoshi, and F. Lopez-Jimenez (2008). "Accuracy of body mass index in diagnosing obesity in the adult general population". In: *Int J Obes (Lond)* 32.6, pp. 959–966. ISSN: 1476-5497.

Rosenblatt, Murray (1956). "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3, pp. 832–837. ISSN: 0003-4851. DOI: 10.1214/aoms/1177728190.

Ruhl, Constance E. and James E. Everhart (2003). "Determinants of the association of overweight with elevated serum alanine aminotransferase activity in the United

States". In: *Gastroenterology* 124.1, pp. 71–79. ISSN: 0016-5085. DOI: 10.1053/gast.2003.50004. URL: http://www.sciencedirect.com/science/article/pii/S0016508503500208.

Russell, S.J. and P. Norvig (2010). *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Pearson Education/Prentice Hall. ISBN: 9780136042594.

Sacheck, J. (2008). "Pediatric Obesity: An Inflammatory Condition?" In: *Journal of Parenteral and Enteral Nutrition* 32.6, pp. 633–637. DOI: 10.1177/0148607108324876. eprint: http://pen.sagepub.com/content/32/6/633.full.pdf+html. URL: http://pen.sagepub.com/content/32/6/633.abstract.

Sackett, David L, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson (Jan. 1996). "Evidence based medicine: what it is and what it isn't". In: *BMJ* 312.7023, pp. 71–72. DOI: 10.1136/bmj.312.7023.71.

Sain, Stephan R. (1994). "Adaptive Kernel Density Estimation". PhD thesis. Houston, Texas: Rice University.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. ISBN: 9780471547709.

Shaffer, J. P. (1995). "Multiple hypothesis-testing". In: *Annual review of psychology* 46, pp. 561–584. ISSN: 0066-4308. DOI: {10.1146/annurev.psych.46.1.561}.

Shao, J. and D. Tu (2012). *The Jackknife and Bootstrap*. Springer Series in Statistics Series. Springer London, Limited. ISBN: 9781461269038.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall. ISBN: 9780412246203.

Sloan, AW (1978). "William Harvey, physician and scientist." In: *South African medical journal* 54.6, p. 247.

Solá, E, A Vayá, M L Santaolaria, A Hernández-Mijares, E Réganon, V Vila, V Martínez-Sales, and D Corella (2007). "Erythrocyte deformability in obesity measured by ektacytometric techniques." In: *Clin Hemorheol Microcirc* 37.3, pp. 219–27. ISSN: 1386-0291.

Stienstra, Rinke, Caroline Duval, Michael Müller, and Sander Kersten (2007). "PPARs, Obesity, and Inflammation." In: *PPAR Research* 2007, p. 95974. ISSN: 1687-4757.

Stuart, A. and K. Ord (2009). *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory*. Kendall's Library of Statistics. Wiley. ISBN: 9780340614303.

Suhaimi, F, A Le Compte, J C Preiser, G M Shaw, P Massion, R Radermecker, C G Pretty, J Lin, T Desaive, and J G Chase (2010). "What makes tight glycemic control

tight? The impact of variability and nutrition in two clinical studies". In: *J Diabetes Sci Technol* 4, pp. 284–298.

Pi-Sunyer, Xavier (2009). "The medical risks of obesity". In: *Postgraduate Medicine* 121.6, pp. 21–33. ISSN: 1941-9260.

Syrenicz, Anhelli, Barbara Garanty-Bogacka, Malgorzata Syrenicz, Aneta Gebala, and Mieczyslaw Walczak (2006). "Low-grade systemic inflammation and the risk of type 2 diabetes in obese children and adolescents". In: *Neuro Endocrinology Letters* 27.4, pp. 453–458. ISSN: 0172-780X.

Tan, Ping-Nan, Vipin Kumar, and Michael Steinbach (2007). *Introduction To Data Mining*. Pearson Education. ISBN: 9788131714720.

Tapia, R.A. and J.R. Thompson (2002). *Nonparametric Probability Density Estimation*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press. ISBN: 9780801869761.

Taylor, M. R. H., C. V. Holland, R. Spencer, J. F. Jackson, G. I. O'Connor, and J. R. O'Donnell (1997). "Haematological reference ranges for schoolchildren". In: *Clinical & Laboratory Haematology* 19.1, pp. 1–15. ISSN: 1365-2257. DOI: 10.1046/j.1365-2257.1997.00204.x. URL: http://dx.doi.org/10.1046/j.1365-2257.1997.00204.x.

Terrell, George R. and David W. Scott (1992). "Variable Kernel Density Estimation". In: *The Annals of Statistics* 20.3, pp. 1236–1265.

Trefethen, L.N. and D.A. Bau (1997). *Numerical Linear Algebra*. Cambridge University Press. ISBN: 9780898713619.

Tungtrongchitr, R., P. Pongpaew, B. Phonrat, S. Tribunyatkul, D. Viroonudomphol, V. Supawan, P. Jintaridhi, A. Lertchavanakul, N. Vudhivai, and F. P. Schelp (2000). "Leptin concentration in relation to body mass index (BMI) and hematological measurements in Thai obese and overweight subjects." In: *Southeast Asian J Trop Med Public Health* 31.4, pp. 787–94. ISSN: 0125-1562.

Vaart, A.W. van der (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN: 9780521784504.

Van den Berghe, Greet, Pieter Wouters, Frank Weekers, Charles Verwaest, Frans Bruyninckx, Miet Schetz, Dirk Vlasselaers, Patrick Ferdinande, Peter Lauwers, and Roger Bouillon (2001). "Intensive Insulin Therapy in Critically Ill Patients". In: *New England Journal of Medicine* 345.19, pp. 1359–1367. DOI: 10.1056/NEJMoa011300. eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMoa011300. URL: http://www.nejm.org/doi/full/10.1056/NEJMoa011300.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Statistics and Computing. Springer. ISBN: 9780387954578.

Visscher, T. L. S. and J. C. Seidell (2001). "The public health impact of obesity". In: *Annual review of public health* 22, pp. 355–375. ISSN: 0163-7525. DOI: {10.1146/annurev.publhealth.22.1.355}.

Vliet, Mariska van, Martijn Heymans, Ines von Rosenstiel, Desiderius Brandjes, Jos Beijnen, and Michaela Diamant (2011). "Cardiometabolic risk variables in overweight and obese children: a worldwide comparison". In: *Cardiovascular Diabetology* 10.1, p. 106. ISSN: 1475-2840. DOI: 10.1186/1475-2840-10-106. URL: http://www.cardiab.com/content/10/1/106.

Wahba, Grace (1975). "Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation". In: *The Annals of Statistics* 3.1, pp. 15–29. URL: http://www.jstor.org/discover/10.2307/2958077.

Wand, M. P. and M. C. Jones (1993). "Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation". In: *Journal of the American Statistical Association* 88.422, pp. 520–528. DOI: 10.1080/01621459.1993.10476303. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1993.10476303. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476303.

Wand, M.P. and C. Jones (1995). *Kernel smoothing*. Monographs on Statistics and Applied Probability. Chapman and Hall. ISBN: 9780412552700.

Wang, Youfa and Tim Lobstein (2006). "Worldwide trends in childhood overweight and obesity". In: *International Journal of Pediatric Obesity* 1.1, pp. 11–25. ISSN: 1747-7174. DOI: 10.1080/17477160600586747. URL: http://dx.doi.org/10.1080/17477160600586747.

Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244. ISSN: 01621459. DOI: 10.2307/2282967. URL: http://dx.doi.org/10.2307/2282967.

Witten, I.H., E. Frank, and M.A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. ISBN: 9780080890364.

World Health Organization (2013). *Global Database on Body Mass Index – an interactive surveillance tool for monitoring nutrition transition*. http://apps.who.int/bmi/. [Online; accessed 27. 03. 2013.] URL: http://apps.who.int/bmi/.

Wothke, W. (1993). "Nonpositive definite matrices in structural modeling". In: *Testing structural equation models*. Ed. by K. A. Bollen and J. S. Long. Newbury Park, CA: SAGE, pp. 256–293.

Yamada, Seishi, Tadao Gotoh, Yoshiyuki Nakashima, Kazunori Kayaba, Shizukiyo Ishikawa, Naoki Nago, Yosikazu Nakamura, Yoshihisa Itoh, and Eiji Kajii (2001). "Distribution of Serum C-Reactive Protein and Its Association with Atherosclerotic Risk Factors in a Japanese Population: Jichi Medical School Cohort Study". In: *American Journal of Epidemiology* 153.12, pp. 1183–1190. DOI: 10.1093/aje/153.12.1183. eprint: http://aje.oxfordjournals.org/content/153/12/1183.full.pdf+html. URL: http://aje.oxfordjournals.org/content/153/12/1183.abstract.

Own Publications Pertaining to Theses

F-1 Almássy, Zsuzsanna, Levente Kovács, Tamás Ferenci, Antal Czinner, and Zoltán Benyó (2009). *Prediabeteses állapot megelőzésére szolgáló módszer kidolgozása.* Presentation, a Magyar Gyermekorvosok Társasága és a Magyar Diabétesz Társaság XXVI. Gyermekdiabétesz tudományos ülése, Gödöllő, Hungary.

F-2 — (2010). "Veszélyeztetettség előrejelzésére alkalmas szűrővizsgálat egészséges és elhízott gyermekeken". In: *Diabetologica Hungarica* 18.Supplementum 1, pp. 55–56.

F-3 Almássy, Zsuzsanna, Levente Kovács, Tamás Ferenci, Zsolt Vajda, and Adalbert Kovács (2008). *Tömegszűrésre alkalmas prediktív módszer veszélyeztetett gyermekek esetén?* Presentation, a Magyar Gyermekorvosok Társasága és a Magyar Diabétesz Társaság XXV. Gyermekdiabétesz tudományos ülése, Kiskőrös, Hungary.

F-4 Ferenci, Tamás (2009a). "Kiskorú magyar populáció obesitassal összefüggő paramétereinek biostatisztikai elemzése". MSc thesis. Budapest, Hungary: Budapest University of Technology and Economics.

F-5 — (2010b). "Kiskorú magyar populáció paramétereinek biostatisztikai modellezése az obesitas rizikófaktorainak elemzésére". MSc thesis. Budapest, Hungary: Budapest University of Technology and Economics.

F-6 — (2011b). *Elhízás hatása a laboreredményekre: Többváltozós elemzési és modellezési lehetőségek.* Presentation, IX. Magyar Biometriai, Biomatematikai és Bioinformatikai Konferencia, Budapest, Hungary.

F-7 Ferenci, Tamás, Zsuzsanna Almássy, Adalbert Kovács, and Levente Kovács (2011a). "Effects of Obesity: A Multivariate Analysis of Laboratory Parameters". In: *Scientific bulletin of politechnica university of timisoara transactions on automatic control and computer science* 56.70, pp. 145–152.

F-8 Ferenci, Tamás, Zsuzsanna Almássy, Adalbert Kovács, and Levente Kovács (2011b). "Effects of obesity: A multivariate analysis of laboratory parameters". In: *2011*

6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 629–634. DOI: 10.1109/SACI.2011.5873079.

F-9 Ferenci, Tamás, Zsuzsanna Almássy, Zoltán Merkei, Adalbert Kovács, and Levente Kovács (2008). "Cluster analysis of obesity-related parameters of Hungarian children". In: *Proc. of BUDAMED*, pp. 33–37.

F-10 Ferenci, Tamás, Balázs Benyó, Levente Kovács, Liam Fisk, Geoffrey M. Shaw, and J. Geoffrey Chase (2013). "Daily Evolution of Insulin Sensitivity Variability with Respect to Diagnosis in the Critically Ill". In: *PLoS ONE* 8.2, e57119. DOI: 10.1371/journal.pone.0057119. URL: http://dx.doi.org/10.1371/journal.pone.0057119.

F-11 Ferenci, Tamás, Levente Kovács, Zsuzsanna Almássy, László Szilágyi, Balázs Benyó, and Zoltán Benyó (2010). "Differences in the laboratory parameters of obese and healthy Hungarian children and their use in automatic classification". In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 3883–3886. DOI: 10.1109/IEMBS.2010.5627672.

F-12 Ferenci, Tamás, Levente Kovács, Zsuzsanna Almássy, László Szilágyi, and Zoltán Benyó (2011). "Automatic Classification of Obesity in Teenage Population based on Laboratory Results". In: *MACRo 2011 – 3d International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics*, pp. 293–304. ISBN: 9789731970547.

F-13 Ferenci, Tamás, Levente Kovács, Balázs Benyó, and Adalbert Kovács (2012). "Using Support Vector Machines to Recognize Changes Characteristic to Obesity in Laboratory Results". In: *5th European Conference of the International Federation for Medical and Biological Engineering*. Ed. by Ákos Jobbágy. Vol. 37. IFMBE Proceedings. Springer Berlin Heidelberg, pp. 215–218. ISBN: 978-3-642-23507-8. DOI: 10.1007/978-3-642-23508-5_57. URL: http://dx.doi.org/10.1007/978-3-642-23508-5_57.

F-14 Ferenci, Tamás, Levente Kovács, Balázs Benyó, Aaron Le Compte, Geoffrey Shaw, and J G Chase (2012). "Effect of diagnosis on variability of ICU patients in insulin sensitivity". In: *IFAC BMS 2012 - 8th IFAC Symposium on Biological and Medical Systems*. Ed. by Balazs Benyó, Steen Andreassen, David Dagan Feng, Ewart Carson, J Geoffrey Chase, and Levente Kovács. IFAC, pp. 462–466.

F-15 Ferenci, Tamás and Zoltán Merkei (2008). "Az elhízás epidemiológiája: kiskorú magyar populáció adatainak többszempontú statisztikai elemzése – egy antropometriai megközelítés". TDK. Budapest, Hungary: Budapest University of Technology and Economics.

F-16 Kovács, Levente, Tamás Ferenci, Balázs Benyó, Adalbert Kovács, and J G Chase (2012). “Short- and long-term evolution of insulin sensitivity variability in critically ill patients”. In: *ICMA 2012 – 13th International Conference on Mathematics and its Applications Conference*. Ed. by I Cuculescu, J Jaric, P Gavruta, I Golet, and L Cadariu. Timisoara, Romania: Editura Politehnica, pp. 329–334.

F-17 Kovács, Levente, Tamás Ferenci, Johanna Sápi, and Péter Szalay (2012). “Népegészségügyi problémák számítógépes modellezése”. In: *IME – Informatika és menedzsment az egészségügyben* 11.8, pp. 49–55.

F-18 Kovács, Levente, Tamás Ferenci, Péter Szalay, and Zsuzsanna Almássy (2011). *Kiskorú magyar populáció paramétereinek biostatisztikai modellezése az obesitas rizikófaktorainak elemzésére*. Presentation, MTA konferencia, Aktuális Orvosbiológiai Mérnöki Kutatások, Budapest, Hungary.

F-19 Kovács, Levente, Péter Szalay, Tamás Ferenci, Dániel Drexler, Johanna Sápi, Istvá Harmati, and Zoltán Benyó (2011). “Modeling and optimal control strategies of diseases with high public health impact”. In: *Intelligent Engineering Systems (INES), 2011 15th IEEE International Conference on*, pp. 23–28. DOI: 10.1109/INES.2011.5954713.

F-20 Kovács, Levente, Péter Szalay, Tamás Ferenci, Johanna Sápi, Péter Sas, Dániel Drexler, István Harmati, Zoltán Benyó, and Adalbert Kovács (2012). “Model-based control algorithms for optimal therapy of high-impact public health diseases”. In: *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*, pp. 531–536. DOI: 10.1109/INES.2012.6249892.

F-21 Merkei, Zoltán and Tamás Ferenci (2009). “Új módszer egyes laboreredmények obesitast előrejelző erejének számszerű jellemzésére”. TDK. Budapest, Hungary: Semmelweis University.

Own Publications Not Pertaining to Theses

Fx-1 Ferenci, Tamás (2009b). “Kismintás biostatisztikai vizsgálatok néhány módszertani kérdése”. TDK. Budapest, Hungary: Corvinus University of Budapest.

Fx-2 — (2010a). “A Semmelweis Egyetem 2009. évi TDK-konferenciájára benyújtott dolgozatok biostatisztikai színvonalának vizsgálata”. TDK. Budapest, Hungary: Semmelweis University.

Fx-3 Ferenci, Tamás (2011a). *Biostatisztikai alapok – egy belgyógyászati esettanulmány példáján*. Mérési útmutató.

Fx-4 Ferenci, Tamás and Balázs Kotosz (2010a). “Nemnormális, parametrizált eloszlású valószínűségi változók”. In: *Statisztikai Szemle* 88.7–8, pp. 803–832.

Fx-5 — (2010b). *Statisztikai tesztek robusztusságának vizsgálata GP-GPU számítási módszerrel*. Presentation, ”Válság az oktatásban? Oktatás a válságban!” – a Budapest Corvinus Egyetem Közgadaságtudományi Karának konferenciája.

Fx-6 — (2010c). “Using Massively Parallel Processing in the Testing of the Robustness of Statistical Tests with Monte Carlo Simulation”. In: *Proceedings of the Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference*, pp. 1343–1366. ISBN: 9789630695589.

Fx-7 Kapelner, Tamás, László Madarász, and Tamás Ferenci (2013). “A független komponens analízis és empirikus vizsgálata”. In: *Statisztikai Szemle* 91.3, pp. 253–286.

Fx-8 Kotosz, Balázs and Tamás Ferenci (2010). *Nemnormális, parametrizált eloszlású véletlen változók generálása*. Presentation, ”Válság az oktatásban? Oktatás a válságban!” – a Budapest Corvinus Egyetem Közgadaságtudományi Karának konferenciája.

Fx-9 Reiczigel, Jenő and Tamás Ferenci (2012). *On the validity of power simulation based on Fleishman distributions*. Poster, P21.18, 33rd Annual Conference of the International Society for Clinical Biostatistics, Bergen, Norway.

A. Program for Effect of Obesity on Laboratory Parameters

```
1  library( foreign )
2  library( weights )
3  library( ks )
4  library( psych )
5
6  GrowthChart <- read.csv2( "bmiagerev.csv" )
7  source( "bmizscore.R" )
8
9  ###BEGIN --- Load NHANES #
10 #####
11 nhanes <- read.xport( "./nhanes20092010/DEMO_F.XPT" )
12 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/BIOPRO_F.XPT" ) )
13 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/CBC_F.XPT" ) )
14 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/CRP_F.XPT" ) )
15 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/HDL_F.XPT" ) )
16 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/TRIGLY_F.XPT" ) )
17 nhanes <- merge( nhanes, read.xport( "./nhanes20092010/BMX_F.XPT" ) )
18
19 nhanes <- nhanes[ !is.na( nhanes$RIDAGEEX ), ]
20 nhanes <- nhanes[ nhanes$RIDAGEEX / 12 >= 12 & nhanes$RIDAGEEX / 12 <= 18, ]
21 nhanes <- nhanes[ !is.na( nhanes$BMXBMI ), ]
22 nhanes <- nhanes[ !is.na( nhanes$RIAGENDR ), ]
23
24 nhanes <- data.frame( WBC = nhanes$LBXWBCSI, RNC = nhanes$LBXNEPCT,
25                        RLC = nhanes$LBXLYPCT, RMC = nhanes$LBXMOPCT,
26                        REC = nhanes$LBXEOPCT, ANC = nhanes$LBDNENO,
27                        ALC = nhanes$LBDLYMNO, AMC = nhanes$LBDMONO,
28                        AEC = nhanes$LBDEONO, RBC = nhanes$LBXRBCSI,
29                        HGB = nhanes$LBXHGB * 10, HCT = nhanes$LBXHCT / 100,
30                        MCV = nhanes$LBXMCVSI, MCH = nhanes$LBXMCHSI,
31                        MCHC = nhanes$LBXMC * 10, RDW = nhanes$LBXRDW,
32                        PLT = nhanes$LBXPLTSI, MPV = nhanes$LBXMPSI,
33                        CRP = nhanes$LBXCRP, SNA = nhanes$LBXSNASI,
34                        SK = nhanes$LBXSKSI, SCL = nhanes$LBXSCLSI,
35                        STP = nhanes$LBDSTPSI, SAL = nhanes$LBDSALSI,
36                        SGL = nhanes$LBDSGBSI, BUN = nhanes$LBDSBUSI,
37                        SCR = nhanes$LBDSCRSI, STG = nhanes$LBDTRSI,
38                        STC = nhanes$LBDSCHSI, HDL = nhanes$LBDHDDSI,
39                        AST = nhanes$LBXSASSI, ALT = nhanes$LBXSATSI,
```

```
                          GGT = nhanes$LBXSGTSI, BMI = nhanes$BMXBMI,
                          WEIGHT = nhanes$WTSAF2YR, GENDER = nhanes$RIAGENDR,
                          AGEMOS = nhanes$RIDAGEEX)
nhanes$ZBMI <- mapply( BMIZScore, nhanes$BMI, nhanes$GENDER,
                        nhanes$AGEMOS, MoreArgs =
                          list( GrowthChart = GrowthChart ) )
nhanesMales <- nhanes[ nhanes$GENDER == 1, ]
nhanesFemales <- nhanes[ nhanes$GENDER == 2, ]

colnames( nhanesMales )
apply( nhanesMales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( nhanesMales )[ 1 ]
apply( nhanesMales, 1, function( x ) sum( is.na( x ) ) )
rowSel <- apply( nhanesMales, 1, function( x ) sum( is.na( x ) ) ) == 0
nhanesMales <- nhanesMales[ rowSel, ]
apply( nhanesMales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( nhanesMales )[ 1 ]
apply( nhanesMales, 1, function( x ) sum( is.na( x ) ) )
dim( nhanesMales )

colnames( nhanesFemales )
apply( nhanesFemales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( nhanesFemales )[ 1 ]
apply( nhanesFemales, 1, function( x ) sum( is.na( x ) ) )
rowSel <- apply( nhanesFemales, 1, function( x ) sum( is.na( x ) ) ) == 0
nhanesFemales <- nhanesFemales[ rowSel, ]
apply( nhanesFemales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( nhanesFemales )[ 1 ]
apply( nhanesFemales, 1, function( x ) sum( is.na( x ) ) )
dim( nhanesFemales )

nhanesMales$WEIGHT <- nhanesMales$WEIGHT /
  sum( nhanesMales$WEIGHT ) * dim( nhanesMales )[ 1 ]
nhanesFemales$WEIGHT <- nhanesFemales$WEIGHT /
  sum( nhanesFemales$WEIGHT ) * dim( nhanesFemales )[ 1 ]
#####
### END -- Load NHANES #

###BEGIN -- Load Hungarian study #
#####
hun <- read.csv2( "beo7.csv" )
hun$BMI <- hun$BodyMass / ( hun$BodyHeight / 100 )^2
hun$ZBMI <- mapply( BMIZScore, hun$BMI, hun$Sex, hun$AgeMos,
                    MoreArgs = list( GrowthChart = GrowthChart ) )
hun <- hun[ hun$AgeMos / 12 <= 18 & hun$AgeMos / 12 >= 12, ]
hunMales <- hun[ hun$Sex == 1, ]
hunFemales <- hun[ hun$Sex == 2, ]

colnames( hunMales )
apply( hunMales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( hunMales )[ 1 ]
```

```
apply( hunMales[ , 7:42 ], 1, function( x ) sum( is.na( x ) ) )
rowSel <- apply( hunMales[ , 7:42 ], 1,
                 function( x ) sum( is.na( x ) ) ) < 6
apply( hunMales[ rowSel, ], 2, function( x ) sum( is.na( x ) ) ) /
  dim( hunMales[ rowSel, ] )[ 1 ]
colSel <- apply( hunMales[ rowSel, ], 2, function( x )
  sum( is.na(x  ) ) ) / dim( hunMales[ rowSel, ] )[ 1 ] < 0.2
hunMales <- hunMales[ rowSel, colSel ]
apply( hunMales, 2, function( x ) sum( is.na( x ))  ) /
  dim( hunMales )[ 1 ]
apply( hunMales, 1, function( x ) sum( is.na( x ) ) )
dim( hunMales )
colnames( hunMales )
colnames( hunMales )[ 7:39 ] <- c( "WBC", "RNC", "RLC", "RMC", "REC",
                                   "ANC", "ALC", "AMC", "AEC", "RBC",
                                   "HGB", "HCT", "MCV", "MCH", "MCHC",
                                   "RDW", "PLT", "MPV", "CRP", "SNA",
                                   "SK", "SCL", "STP", "SAL", "SGL",
                                   "BUN", "SCR", "STG", "STC", "HDL",
                                   "AST", "ALT", "GGT" )
colnames( hunMales )

colnames( hunFemales )
apply( hunFemales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( hunFemales )[ 1 ]
apply( hunFemales[ , 7:42 ], 1, function( x ) sum( is.na( x ) ) )
rowSel <- apply( hunFemales[ , 7:42 ], 1,
                 function( x ) sum( is.na( x ) ) ) < 6
apply( hunFemales[ rowSel, ], 2, function( x ) sum( is.na( x ) ) ) /
  dim( hunFemales[ rowSel, ] )[ 1 ]
colSel <- apply( hunFemales[ rowSel, ], 2, function( x )
  sum( is.na(x  ) ) ) / dim( hunFemales[ rowSel, ] )[ 1 ] < 0.2
hunFemales <- hunFemales[ rowSel, colSel ]
apply( hunFemales, 2, function( x ) sum( is.na( x ))  ) /
  dim( hunFemales )[ 1 ]
apply( hunFemales, 1, function( x ) sum( is.na( x ) ) )
dim( hunFemales )
colnames( hunFemales )
colnames( hunFemales )[ 7:39 ] <- c( "WBC", "RNC", "RLC", "RMC", "REC",
                                     "ANC", "ALC", "AMC", "AEC", "RBC",
                                     "HGB", "HCT", "MCV", "MCH", "MCHC",
                                     "RDW", "PLT", "MPV", "CRP", "SNA",
                                     "SK", "SCL", "STP", "SAL", "SGL",
                                     "BUN", "SCR", "STG", "STC", "HDL",
                                     "AST", "ALT", "GGT" )
colnames( hunFemales )

for ( i in 7:39 ) {
  hunMales[ , i ] <- replace( hunMales[ , i ], is.na( hunMales[ , i ] ),
                              median( hunMales[ , i ], na.rm = TRUE ) )
  hunFemales[ , i ] <- replace( hunFemales[ , i ], is.na( hunFemales[ , i ] ),
```

```
                                    median( hunFemales[ , i ], na.rm = TRUE ) )
}

apply( hunMales, 2, function( x ) sum( is.na( x ) ) )  /
  dim( hunMales )[ 1 ]
dim(hunMales)
apply( hunFemales, 2, function( x ) sum( is.na( x ) ) ) /
  dim( hunFemales )[ 1 ]
dim( hunFemales )

hunMales$WEIGHT <- rep( 1, dim( hunMales )[ 1 ] )
hunFemales$WEIGHT <- rep( 1, dim( hunFemales )[ 1 ] )
#####
### END -- Load Hungarian study #

###BEGIN -- Distribution of Z-BMI scores  #
#####
par( mfrow = c( 1, 2 ) )
hist( c( hunMales$ZBMI, hunMales$ZBMI ), xlab = "Z-BMI",
      main = "Hungarian study", xlim = c( -3 , 4 ) )
hist( c( nhanesMales$ZBMI, nhanesMales$ZBMI ), xlab = "Z-BMI",
      main = "NHANES", xlim = c( -3, 4 ) )

par( mfrow = c( 1, 2 ) )
hist( nhanesFemales$ZBMI, xlab = "Z-BMI",  main = "Females (n=200)", xlim = c( -4,
    4 ) )
hist( nhanesMales$ZBMI, xlab = "Z-BMI",  main = "Males (n=240)", xlim = c( -4, 4 )
    )

par( mfrow = c( 1, 2 ) )
hist( hunFemales$ZBMI, xlab = "Z-BMI",  main = "Females (n=70)", xlim = c( -4, 4 )
    )
hist( hunMales$ZBMI, xlab = "Z-BMI",  main = "Males (n=113)", xlim = c( -4, 4 ) )
#####
### END -- Distribution of Z-BMI scores  #

###BEGIN -- Illustration of univariate descriptive statistics  #
#####
DataMatrix <- as.matrix( data.frame( nhanesMales$ZBMI, nhanesMales$HDL ) )
HscvDM <- Hscv( DataMatrix )
ResKde <- kde( DataMatrix, HscvDM, w = nhanesMales$WEIGHT )
conts <- c( 10, 25, 50, 75, 90 )
cuts <- c( 0, 1, 2, 3 )
ylm <- c( -5, 10 )
n <- 500

layout( matrix( c( 1, 1, 2, 2, 3, 4, 5, 6 ), 2, 4, byrow = TRUE) )
plot( nhanesMales$ZBMI, nhanesMales$HDL, xlab = "Z-BMI",ylab = "HDL",
      sub = "A", main="Scattergram")
plot( ResKde, cont = conts, display = "slice", xlab = "Z-BMI",
      ylab = "HDL",sub="B", main="Kernel density estimation")
```

```
abline( v = c( 0, 1, 2, 3 ), lty = 2 )
ResEval0 <- kde( DataMatrix, HscvDM, eval.points = matrix(
  c( rep( 0, n ), seq( ylm[ 1 ], ylm[ 2 ], ( ylm[ 2 ] -
    ylm[ 1 ] ) / ( n - 1 ) ) ), n, 2 ), w = nhanesMales$WEIGHT )
ResEval1 <- kde( DataMatrix, HscvDM, eval.points = matrix(
  c( rep( 1, n ), seq( ylm[ 1 ], ylm[ 2 ], ( ylm[ 2 ] -
    ylm[ 1 ] ) / ( n - 1 ) ) ), n, 2 ), w = nhanesMales$WEIGHT )
ResEval2 <- kde( DataMatrix, HscvDM, eval.points = matrix(
  c( rep( 2, n ), seq( ylm[ 1 ], ylm[ 2 ], ( ylm[ 2 ] -
    ylm[ 1 ] ) / ( n - 1 ) ) ), n, 2 ), w = nhanesMales$WEIGHT )
ResEval3 <- kde( DataMatrix, HscvDM, eval.points = matrix(
  c( rep( 3, n ), seq( ylm[ 1 ], ylm[ 2 ], ( ylm[ 2 ] -
    ylm[ 1 ] ) / ( n - 1 ) ) ), n, 2 ), w = nhanesMales$WEIGHT )
plot( ResEval0$eval.points[ , 2 ], ResEval0$estimate /
  ( sum( ResEval0$estimate ) * ( ylm[2] - ylm[1] )/ ( n - 1 ) ), "l",
      xlim = c( 0, 2.5 ), ylim = c( 0, 1.6 ), xlab = "HDL", sub = "C1",
      ylab = "f(HDL|Z-BMI=0)",
      main = "Conditional distribution of HDL \n Condition: Z-BMI = 0" )
grid( col = "black" )
plot( ResEval1$eval.points[ , 2 ], ResEval1$estimate /
  ( sum( ResEval1$estimate ) * ( ylm[2] - ylm[1] )/ ( n - 1 ) ), "l",
      xlim = c( 0, 2.5 ), ylim = c( 0, 1.6 ), xlab = "HDL", sub = "C2",
      ylab = "f(HDL|Z-BMI=1)",
      main = "Conditional distribution of HDL \n Condition: Z-BMI = 1" )
grid( col = "black" )
plot( ResEval2$eval.points[ , 2 ], ResEval2$estimate /
  ( sum( ResEval2$estimate ) * ( ylm[2] - ylm[1] )/ ( n - 1 ) ), "l",
      xlim = c( 0, 2.5 ), ylim = c( 0, 1.6 ), xlab = "HDL", sub = "C3",
      ylab = "f(HDL|Z-BMI=2)",
      main = "Conditional distribution of HDL \n Condition: Z-BMI = 2" )
grid( col = "black" )
plot( ResEval3$eval.points[ , 2 ], ResEval3$estimate /
  ( sum( ResEval3$estimate ) * ( ylm[2] - ylm[1] )/ ( n - 1 ) ), "l",
      xlim = c( 0, 2.5 ), ylim = c( 0, 1.6 ), xlab = "HDL", sub = "C4",
      ylab = "f(HDL|Z-BMI=3)",
      main = "Conditional distribution of HDL \n Condition: Z-BMI = 3" )
grid( col = "black" )
#####
### END -- Illustration of univariate descriptive statistics  #

###BEGIN -- Univariate descriptive statistics, function def  #
#####
desc <- function( data, bmi, weight, cut, n ) {
  DataMatrix <- as.matrix( data.frame( bmi, data ) )
  ResEval <- kde( DataMatrix, Hscv( DataMatrix ), w = weight,
                  eval.points = matrix(
                    c( rep( cut, n ), seq( min( data ), max( data ),
                                           ( max( data ) - min( data ) ) /
                                             ( n - 1 ) ) ), n, 2 ) )
  m <- sum( ResEval$eval.points[ , 2 ] * ResEval$estimate /
    sum( ResEval$estimate ) )
```

```
  res<-data.frame(
    Mean = m,
    Median = tail(
      ResEval$eval.points[ , 2 ][ cumsum( ResEval$estimate ) /
        sum( ResEval$estimate ) < 0.5 ], 1 ),
    SD = sqrt( sum( ( ResEval$eval.points[ , 2 ] - m )^2 *
      ResEval$estimate / sum( ResEval$estimate ) ) ),
    IQR = tail(
      ResEval$eval.points[ , 2 ][ cumsum( ResEval$estimate ) /
        sum( ResEval$estimate ) < 0.75 ], 1 ) - tail(
          ResEval$eval.points[ , 2 ][ cumsum( ResEval$estimate ) /
            sum( ResEval$estimate ) < 0.25 ], 1 )
  )
  return( res )
}
# Example: desc( nhanesMales$HDL, nhanesMales$ZBMI,
# nhanesMales$WEIGHT, 1, 1000)
descAll <- function( database, zbmi, weight, n, file ) {
write.csv2( t( matrix( unlist( apply( database, 2, function (x) {
  data.frame( ZBMI1 = desc( x, zbmi, weight, 1, n), ZBMI2 =
    desc( x, zbmi, weight, 2, n), ZBMI3 =
    desc( x, zbmi, weight, 3, n ) ) } ) ), nr = 12 ) ), file = file )
}
# Example: descAll( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
# nhanesMales$WEIGHT, 1000, "NHANES3DescMale.csv" )
#####
### END -- Univariate descriptive statistics, function def  #

###BEGIN -- Univariate descriptive statistics  #
#####
descAll( hunMales[ , 7:39 ], hunMales$ZBMI, hunMales$WEIGHT,
         1000, "HUN3DescMale.csv" )
descAll( hunFemales[ , 7:39 ], hunFemales$ZBMI, hunFemales$WEIGHT,
         1000, "HUN3DescFemale.csv" )
descAll( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
         nhanesMales$WEIGHT, 1000, "NHANES3DescMale.csv" )
descAll( nhanesFemales[ , 1:33 ], nhanesFemales$ZBMI,
         nhanesFemales$WEIGHT, 1000, "NHANES3DescFemale.csv" )
#####
### END -- Univariate descriptive statistics  #

###BEGIN -- Univariate association analysis, function def #
#####
assocAll <- function( database, zbmi, weight, file ) {
write.csv2( data.frame( Rho =
  apply( database, 2, function ( x ) {
    wtd.cor( rank( zbmi ), rank( x ), weight = weight )[ 1 ]
    } ), P = apply( database, 2, function ( x ) { wtd.cor(
      rank( zbmi ), rank( x ), weight = weight )[ 4 ] } ),
                          Pcorr = p.adjust(
              apply( database, 2, function ( x ) { wtd.cor(
```

```
292                 rank( zbmi ), rank( x ), weight = weight )[ 4 ] } ),
293             method = "holm" ) ), file = file )
294 }
295 # Example: assocAll( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
296 # nhanesMales$WEIGHT, "NHANES3AssocMale.csv")
297 #####
298 ### END -- Univariate association analysis, function def #
299
300 ###BEGIN -- Univariate association analysis #
301 #####
302 assocAll( hunMales[ , 7:39 ], hunMales$ZBMI, hunMales$WEIGHT,
303           "HUN3AssocMale.csv")
304 assocAll( hunFemales[ , 7:39 ], hunFemales$ZBMI, hunFemales$WEIGHT,
305           "HUN3AssocFemale.csv")
306 assocAll( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
307           nhanesMales$WEIGHT, "NHANES3AssocMale.csv")
308 assocAll( nhanesFemales[ , 1:33 ], nhanesFemales$ZBMI,
309           nhanesFemales$WEIGHT, "NHANES3AssocFemale.csv")
310 #####
311 ### END -- Univariate association analysis #
312
313 ###BEGIN -- Multivariate structure, function def #
314 #####
315 CondCov <- function( x, y, zbmi, condition, weight, grd ) {
316   DataMatrix <- as.matrix( data.frame( zbmi, x, y ) )
317   xlm <- c( min( x ) * 0.9 , max( x ) * 1.1 )
318   ylm <- c( min( y ) * 0.9 , max( y ) * 1.1 )
319   ResEval <- kde( DataMatrix, Hscv( DataMatrix, pilot = "dscalar" ),
320                   w = weight, eval.points =
321                     matrix( c( rep( condition, grd * grd ), rep(
322                       seq( xlm[ 1 ], xlm[ 2 ], ( xlm[ 2 ] - xlm[ 1 ] ) /
323                         ( grd - 1 ) ), each = grd ), rep( seq(
324                           ylm[ 1 ], ylm[ 2 ], ( ylm[ 2 ] - ylm[ 1 ] ) /
325                             ( grd - 1 ) ), grd ) ), grd * grd, 3 ) )
326   MeanX <- sum( ResEval$eval.points[ , 2 ] * ResEval$estimate /
327     sum( ResEval$estimate ) )
328   MeanY <- sum( ResEval$eval.points[ , 3 ] * ResEval$estimate /
329     sum( ResEval$estimate ) )
330   MeanXY <- sum( ResEval$eval.points[ , 2 ] *
331     ResEval$eval.points[ , 3 ] * ResEval$estimate / sum( ResEval$estimate ) )
332   return(MeanXY-MeanX*MeanY)
333 }
334 CondVar <- function( x, zbmi, condition, weight, grd ) {
335   DataMatrix <- as.matrix( data.frame( zbmi, x ) )
336   xlm <- c( min( x ) * 0.9, max( x ) * 1.1 )
337   ResEval <- kde( DataMatrix, Hscv( DataMatrix, pilot = "dscalar" ),
338                   w = weight, eval.points =
339                     matrix( c ( rep( condition, grd ), seq(
340                       xlm[ 1 ], xlm[ 2 ], ( xlm[ 2 ] - xlm[ 1 ] ) /
341                         ( grd - 1 ) ) ), grd, 2 ) )
342   MeanX2 <- sum( ResEval$eval.points[ ,2 ]^2 * ResEval$estimate /
```

```
343     sum( ResEval$estimate ) )
344   MeanX <- sum( ResEval$eval.points[ , 2 ] * ResEval$estimate /
345     sum( ResEval$estimate ) )
346   return(MeanX2-MeanX^2)
347 }
348 CondCovMat <- function( database, zbmi, condition, weight, grd) {
349   n <- dim( database )[ 2 ]
350   res <- matrix( rep( 0, n * n ), nc = n )
351   for ( i in 1:( n - 1 ) )
352     for ( j in ( i + 1 ):n ) {
353       print( c( i, j ) )
354       res[ i, j ] <- CondCov( database[ , i ], database[ , j ],
355                               zbmi, condition, weight, grd )
356     }
357   for ( i in 1:n )
358     res[ i, i ] <- CondVar( database[ , i ], zbmi, condition,
359                             weight, grd )
360   for ( i in 1:( n - 1 ) )
361     for ( j in ( i + 1 ):n )
362       res[ j, i ] <- res[ i, j ]
363   return(res)
364 }
365 #####
366 ### END -- Multivariate structure, function def #
367
368 ###BEGIN -- Multivariate structure, correlation matrices #
369 #####
370 write.csv2( cov2cor( CondCovMat( hunMales[ , 7:39 ], hunMales$ZBMI,
371                                  1, hunMales$WEIGHT, 50 ) ),
372             "hunMales1.csv", row.names = FALSE )
373 write.csv2( cov2cor( CondCovMat( hunMales[ , 7:39 ], hunMales$ZBMI,
374                                  2, hunMales$WEIGHT, 50 ) ),
375             "hunMales2.csv", row.names = FALSE )
376 write.csv2( cov2cor( CondCovMat( hunMales[ , 7:39 ], hunMales$ZBMI,
377                                  3, hunMales$WEIGHT, 50 ) ),
378             "hunMales3.csv", row.names = FALSE )
379 write.csv2( cov2cor( CondCovMat( hunFemales[ , 7:39 ], hunFemales$ZBMI,
380                                  1, hunFemales$WEIGHT, 50 ) ),
381             "hunFemales1.csv", row.names = FALSE )
382 write.csv2( cov2cor( CondCovMat( hunFemales[ , 7:39 ], hunFemales$ZBMI,
383                                  2, hunFemales$WEIGHT, 50 ) ),
384             "hunFemales2.csv", row.names = FALSE )
385 write.csv2( cov2cor( CondCovMat( hunFemales[ , 7:39 ], hunFemales$ZBMI,
386                                  3, hunFemales$WEIGHT, 50 ) ),
387             "hunFemales3.csv", row.names = FALSE )
388 write.csv2( cov2cor( CondCovMat( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
389                                  1, nhanesMales$WEIGHT, 50 ) ),
390             "nhanesMales1.csv", row.names = FALSE )
391 write.csv2( cov2cor( CondCovMat( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
392                                  2, nhanesMales$WEIGHT, 50 ) ),
393             "nhanesMales2.csv", row.names = FALSE )
```

```
write.csv2( cov2cor( CondCovMat( nhanesMales[ , 1:33 ], nhanesMales$ZBMI,
                                 3, nhanesMales$WEIGHT, 50 ) ),
            "nhanesMales3.csv", row.names = FALSE )
write.csv2( cov2cor( CondCovMat( nhanesFemales[ , 1:33 ], nhanesFemales$ZBMI,
                                 1, nhanesFemales$WEIGHT, 50 ) ),
            "nhanesFemales1.csv", row.names = FALSE )
write.csv2( cov2cor( CondCovMat( nhanesFemales[ , 1:33 ], nhanesFemales$ZBMI,
                                 2, nhanesFemales$WEIGHT, 50 ) ),
            "nhanesFemales2.csv", row.names = FALSE )
write.csv2( cov2cor( CondCovMat( nhanesFemales[ , 1:33 ], nhanesFemales$ZBMI,
                                 3, nhanesFemales$WEIGHT, 50 ) ),
            "nhanesFemales3.csv", row.names = FALSE )

hunMales1 <- read.csv2( "hunMales1.csv", row.names =
  colnames( hunMales )[ 7:39 ], col.names =
  colnames( hunMales )[ 7:39 ] )
hunMales2 <- read.csv2( "hunMales2.csv", row.names =
  colnames( hunMales )[ 7:39 ], col.names =
  colnames( hunMales )[ 7:39 ] )
hunMales3 <- read.csv2( "hunMales3.csv", row.names =
  colnames( hunMales )[ 7:39 ], col.names =
  colnames( hunMales )[ 7:39 ] )
hunFemales1 <- read.csv2( "hunFemales1.csv", row.names =
  colnames( hunFemales )[ 7:39 ], col.names =
  colnames( hunFemales )[ 7:39 ] )
hunFemales2 <- read.csv2( "hunFemales2.csv", row.names =
  colnames( hunFemales )[ 7:39 ], col.names =
  colnames( hunFemales )[ 7:39 ] )
hunFemales3 <- read.csv2( "hunFemales3.csv", row.names =
  colnames( hunFemales )[ 7:39 ], col.names =
  colnames( hunFemales )[ 7:39 ] )
nhanesMales1 <- read.csv2( "nhanesMales1.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
nhanesMales2 <- read.csv2( "nhanesMales2.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
nhanesMales3 <- read.csv2( "nhanesMales3.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
nhanesFemales1 <- read.csv2( "nhanesFemales1.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
nhanesFemales2 <- read.csv2( "nhanesFemales2.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
nhanesFemales3 <- read.csv2( "nhanesFemales3.csv", row.names =
  colnames( nhanesMales )[ 1:33 ], col.names =
  colnames( nhanesMales )[ 1:33 ] )
#####
### END -- Multivariate structure, correlation matrices #
```

```
445
446 ###BEGIN -- Principal Components Analysis, function def #
447 #####
448 pcaScree <- function( CorMat1, CorMat2, CorMat3, main, ylim ) {
449   plot( principal( CorMat1 )$values, xlab = "Index", type = "l",
450         ylab = "Eigenvalue", main = main, ylim = ylim,
451         col = "green")
452   lines( principal( CorMat2 )$values, col = "orange" )
453   lines( principal( CorMat3 )$values, col = "red" )
454   grid()
455   abline( h = 1, lty = "dashed" )
456 }
457 # Example: pcaScree( nhanesMales1, nhanesMales2, nhanesMales3,
458 # "Males, NHANES", c( 0, 7 ) )
459 pcaHm <- function( CorMat, nf, main, colnames ) {
460   fit <- principal( CorMat, nfactors = nf, rotate = "varimax" )
461   aheatmap( Colv = NA, Rowv = NA, fit$loading, revC = TRUE,
462             labRow = colnames, main = main, legend = FALSE )
463   return(fit)
464 }
465 # Example: pcaHm( nhanesMales1, 15, main = "Z-BMI=1",
466 # colnames( nhanesMales )[ 1:33 ] )
467 #####
468 ### END -- Principal Components Analysis, function def #
469
470 ###BEGIN -- Principal Components Analysis #
471 #####
472 dev.off()
473 pdf("NEWfig3.pdf")
474 par( mfrow = c( 2, 2), oma = c( 3, 0, 0, 0 ), usr = c( 0, 1, 0, 1 ) )
475 pcaScree( hunMales1, hunMales2, hunMales3,
476           "Males, Hungarian study", c( 0, 7 ) )
477 pcaScree( hunFemales1, hunFemales2, hunFemales3,
478           "Females, Hungarian study", c( 0, 7 ) )
479 pcaScree( nhanesMales1, nhanesMales2, nhanesMales3,
480           "Males, NHANES", c( 0, 7 ) )
481 pcaScree( nhanesFemales1, nhanesFemales2, nhanesFemales3,
482           "Females, NHANES", c( 0, 7 ) )
483 par( xpd = NA )
484 legend( -25, -4, c( "Z-BMI=1", "Z-BMI=2", "Z-BMI=3" ),
485         fill = c( "green", "orange", "red" ), horiz=TRUE )
486 dev.off()
487
488 dev.off()
489 pdf( "NEWfig4b.pdf", height = 20 )
490 par( mfrow = c( 3, 2) )
491 pcaHm( hunMales1, 13, main = "Males, Z-BMI=1",
492        colnames( hunMales )[ 7:39 ] )
493 pcaHm( hunFemales1, 13, main = "Females, Z-BMI=1",
494        colnames( hunMales )[ 7:39 ] )
495 pcaHm( hunMales2, 13, main = "Males, Z-BMI=2",
```

```
496         colnames( hunMales )[ 7:39 ] )
497 pcaHm( hunFemales2, 13, main = "Females, Z-BMI=2",
498         colnames( hunMales )[ 7:39 ] )
499 pcaHm( hunMales3, 13, main = "Males, Z-BMI=3",
500         colnames( hunMales )[ 7:39 ] )
501 pcaHm( hunFemales3, 13, main = "Females, Z-BMI=3",
502         colnames( hunMales )[ 7:39 ] )
503 dev.off()
504
505 dev.off()
506 pdf( "NEWfig5b.pdf", height = 20 )
507 par( mfrow = c( 3, 2) )
508 pcaHm( nhanesMales1, 13, main = "Males, Z-BMI=1",
509         colnames( nhanesMales )[ 1:33 ] )
510 pcaHm( nhanesFemales1, 13, main = "Females, Z-BMI=1",
511         colnames( nhanesMales )[ 1:33 ] )
512 pcaHm( nhanesMales2, 13, main = "Males, Z-BMI=2",
513         colnames( nhanesMales )[ 1:33 ] )
514 pcaHm( nhanesFemales2, 13, main = "Females, Z-BMI=2",
515         colnames( nhanesMales )[ 1:33 ] )
516 pcaHm( nhanesMales3, 13, main = "Males, Z-BMI=3",
517         colnames( nhanesMales )[ 1:33 ] )
518 pcaHm( nhanesFemales3, 13, main = "Females, Z-BMI=3",
519         colnames( nhanesMales )[ 1:33 ] )
520 dev.off()
521 #####
522 ### END -- Principal Components Analysis#
523
524 ###BEGIN -- Cluster Analysis #
525 #####
526 par( mfrow = c( 3, 2 ) )
527 plot( hclust( as.dist( 1 - abs( hunMales1 ) ),
528                 method = "ward" ), xlab = "", sub = "",
529       main = "Males, Z-BMI=1" )
530 plot( hclust( as.dist( 1 - abs( hunFemales1 ) ),
531                 method = "ward" ), xlab = "", sub = "",
532       main = "Females, Z-BMI=1" )
533 plot( hclust( as.dist( 1 - abs( hunMales2 ) ),
534                 method = "ward" ), xlab = "", sub = "",
535       main = "Males, Z-BMI=2" )
536 plot( hclust( as.dist( 1 - abs( hunFemales2 ) ),
537                 method = "ward" ), xlab = "", sub = "",
538       main = "Females, Z-BMI=2" )
539 plot( hclust( as.dist( 1 - abs( hunMales3 ) ),
540                 method = "ward" ), xlab = "", sub = "",
541       main = "Males, Z-BMI=3" )
542 plot( hclust( as.dist( 1 - abs( hunFemales3 ) ),
543                 method = "ward" ), xlab = "", sub = "",
544       main = "Females, Z-BMI=3" )
545
546 par( mfrow = c( 3, 2 ) )
```

```
plot( hclust( as.dist( 1 - abs( nhanesMales1 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Males, Z-BMI=1" )
plot( hclust( as.dist( 1 - abs( nhanesFemales1 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Females, Z-BMI=1" )
plot( hclust( as.dist( 1 - abs( nhanesMales2 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Males, Z-BMI=2" )
plot( hclust( as.dist( 1 - abs( nhanesFemales2 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Females, Z-BMI=2" )
plot( hclust( as.dist( 1 - abs( nhanesMales3 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Males, Z-BMI=3" )
plot( hclust( as.dist( 1 - abs( nhanesFemales3 ) ),
                method = "ward" ), xlab = "", sub = "",
      main = "Females, Z-BMI=3" )
#####
### END -- Cluster Analysis #
```

B. Program for Modeling and Evaluating the Performance of Tight Glycemic Control Protocols

```
library(gdata)
library(lme4)
library(nlme)
library(multcomp)
library(vioplot)

result <- read.xls("result.xls")
result[ result==-1 ] <- NA
resultOST <- result[, c(1,4,15,16,8,10,12,14) ]
resultQUAD <- result[, c(1,4,15,16,7,9,11,13) ]
resultOSTlme <- make.rm( 1:4,5:8, resultOST )
resultQUADlme <- make.rm( 1:4,5:8, resultQUAD )
resultOSTlme$Class <- as.factor( resultOSTlme$Class )
resultQUADlme$Class <- as.factor( resultQUADlme$Class )
colnames(resultOSTlme)[1] <- "Patient"
colnames(resultOSTlme)[5] <- "Variability"
colnames(resultOSTlme)[6] <- "Day"
colnames(resultQUADlme)[1] <- "Patient"
colnames(resultQUADlme)[5] <- "Variability"
colnames(resultQUADlme)[6] <- "Day"
resultOSTlme$Day <- factor( resultOSTlme$Day,
                            labels = c(1,2,3,4), ordered = F )
resultQUADlme$Day <- factor( resultQUADlme$Day,
                             labels = c(1,2,3,4), ordered = F )
classlabels <- c("NOpC","OpC","NOpG","OpG","NOpO","OpO")
resultOSTlme$Class <- factor( resultOSTlme$Class,
                              labels = classlabels,
                              ordered = F)
resultQUADlme$Class <- factor( resultQUADlme$Class,
                               labels = classlabels,
                               ordered = F)
resultOSTlmeNA <- resultOSTlme[ !is.na(resultOSTlme$Variability), ]
resultQUADlmeNA <- resultQUADlme[ !is.na(resultQUADlme$Variability), ]
mains <- c( "Day 1", "Day 2", "Day 3", "Day 4+")

resultRE3 <- read.csv( "resultRE3.csv", header = F )
colnames(resultRE3) <- c("Patient", "Class", "Minute", "Estimate")
```

```
38 resultRE3$Patient <- as.factor( resultRE3$Patient )
39 resultRE3$Class <- factor( resultRE3$Class, labels = classlabels, ordered = F )
40 resultRE3$QUAD <- ( resultRE3$Estimate - 0.5 )^2
41 resultRE3$OST <- resultRE3$Estimate > 0.9
42
43 par(mfrow=c(2,3))
44 yl <- c( 0.0, 0.12 )
45 xl <- c( 0, 60000 )
46 plot(lowess(x=resultRE3[resultRE3$Class=="NOpC",]$Minute,
47              y=resultRE3[resultRE3$Class=="NOpC",]$QUAD),xlab="Time [min]",
48      ylab="Variablity (Quadratic penalty)",main="Non-operative - Cardiac",type="l"
         ,
49      ylim=yl,xlim=xl)
50 grid()
51 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
52 plot(lowess(x=resultRE3[resultRE3$Class=="NOpG",]$Minute,
53              y=resultRE3[resultRE3$Class=="NOpG",]$QUAD),xlab="Time [min]",
54      ylab="Variablity (Quadratic penalty)",main="Non-operative - Gastric",type="l"
         ,
55      ylim=yl,xlim=xl)
56 grid()
57 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
58 plot(lowess(x=resultRE3[resultRE3$Class=="NOpO",]$Minute,
59              y=resultRE3[resultRE3$Class=="NOpO",]$QUAD),xlab="Time [min]",
60      ylab="Variablity (Quadratic penalty)",main="Non-operative - All other",
61      type="l",ylim=yl,xlim=xl)
62 grid()
63 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
64 plot(lowess(x=resultRE3[resultRE3$Class=="OpC",]$Minute,
65              y=resultRE3[resultRE3$Class=="OpC",]$QUAD),xlab="Time [min]",
66      ylab="Variablity (Quadratic penalty)",main="Operative - Cardiac",type="l",
67      ylim=yl,xlim=xl)
68 grid()
69 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
70 plot(lowess(x=resultRE3[resultRE3$Class=="OpG",]$Minute,
71              y=resultRE3[resultRE3$Class=="OpG",]$QUAD),xlab="Time [min]",
72      ylab="Variablity (Quadratic penalty)",main="Operative - Gastric",
73      type="l",ylim=yl,xlim=xl)
74 grid()
75 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
76 plot(lowess(x=resultRE3[resultRE3$Class=="OpO",]$Minute,
77              y=resultRE3[resultRE3$Class=="OpO",]$QUAD),xlab="Time [min]",
78      ylab="Variablity (Quadratic penalty)",main="Operative - All other",type="l",
79      ylim=yl,xlim=xl)
80 grid()
81 abline( v = c( 1, 2, 3, 4 ) * 24 * 60, lty = "dashed" )
82 dev.off()
83
84 resultRE3 <- resultRE3[ resultRE3$Minute < 8000, ]
85
86 resultRaw <- read.csv("resultRaw.csv", header = F)
```

```
87
88 par( mfrow = c( 4, 6 ) )
89 for (i in 1:4)
90   for (j in 1:6) {
91     hist( resultRaw[ resultRaw[,1] == i & resultRaw[,2] == j, ][,3], ylab = "%",
92           main = paste( classlabels[ j ], ", ", mains[ i ] ,
93                         " (n=", length( resultRaw[
94                           resultRaw[,1] == i & resultRaw[,2] == j, ][,3] ), ")",
95                         sep = "" ), xlab = "Percentile of actual SI(n+1)",
96           xlim = c( 0, 1 ), ylim = c( 0, 2), breaks = 10, freq = F, axes = F )
97     abline( h = 1, lty = 2)
98     axis(1, at = seq( 0, 1, 0.1 ), labels = seq( 0, 100, 10 ) )
99     axis(2, at = seq( 0, 2, 0.5 ), labels = seq( 0, 20, 5 ) )
100  }
101 dev.off()
102
103 par(mfrow=c(2,4))
104 for (i in 1:4) {
105   vioplot.formula( Variability ~ Class,
106                    data = resultOSTlmeNA[ resultOSTlmeNA$Day == i, ],
107                    ylim = c( 0, 0.45), col = "white", colMed="black" )
108   title( main = mains[ i ], xlab = "Diagnosis group",
109          ylab = "Variability (One-sided threshold penalty)" )
110   means <- with( resultOSTlmeNA[ resultOSTlmeNA$Day == i, ],
111                  tapply( Variability, Class, mean ) )
112   points( means, pch=18, cex = 1.5 )
113 }
114 for (i in 1:4) {
115   vioplot.formula( Variability ~ Class,
116                    data = resultQUADlmeNA[ resultQUADlmeNA$Day == i, ],
117                    ylim = c( 0, 0.2), col = "white", colMed="black" )
118   title( main = mains[ i ], xlab = "Diagnosis group",
119          ylab = "Variability (Quadratic penalty)" )
120   means <- with( resultQUADlmeNA[ resultQUADlmeNA$Day == i, ],
121                  tapply( Variability, Class, mean ) )
122   points( means, pch=18, cex = 1.5 )
123 }
124 dev.off()
125
126 kruskal.test(Variability ~ Class,
127              data = resultOSTlmeNA[ resultOSTlmeNA$Day == 1, ])$p.value
128 kruskal.test(Variability ~ Class,
129              data = resultQUADlmeNA[ resultQUADlmeNA$Day == 1, ])$p.value
130 kruskal.test(Variability ~ Class,
131              data = resultOSTlmeNA[ resultOSTlmeNA$Day == 2, ])$p.value
132 kruskal.test(Variability ~ Class,
133              data = resultQUADlmeNA[ resultQUADlmeNA$Day == 2, ])$p.value
134 kruskal.test(Variability ~ Class,
135              data = resultOSTlmeNA[ resultOSTlmeNA$Day == 3, ])$p.value
136 kruskal.test(Variability ~ Class,
137              data = resultQUADlmeNA[ resultQUADlmeNA$Day == 3, ])$p.value
```

```
kruskal.test(Variability ~ Class,
             data = resultOSTlmeNA[ resultOSTlmeNA$Day == 4, ])$p.value
kruskal.test(Variability ~ Class,
             data = resultQUADlmeNA[ resultQUADlmeNA$Day == 4, ])$p.value

KWD1QUAD <- aov( Variability ~ Class,
                 data = resultQUADlmeNA[ resultQUADlmeNA$Day == 1, ] )
KWD2QUAD <- aov( Variability ~ Class,
                 data = resultQUADlmeNA[ resultQUADlmeNA$Day == 2, ] )
summary( glht( KWD1QUAD, linfct = mcp( Class = "Tukey" ) ) )
summary( glht( KWD2QUAD, linfct = mcp( Class = "Tukey" ) ) )

fmQUAD<-lme(QUAD~Class+Minute-1, random = ~Minute|Patient,data=resultRE3)
fmQUADb <- update(fmQUAD, random = ~1|Patient)
anova(fmQUAD,fmQUADb)
fmQUADc <- update(fmQUAD, random = ~Minute-1|Patient)
anova(fmQUAD,fmQUADc)
fmQUADcor <- update(fmQUAD,correlation = corCAR1(form = ~Minute |Patient))
anova(fmQUAD,fmQUADcor)
fmQUADcor2 <- update(fmQUAD,correlation =
                       corCompSymm(form = ~Minute |Patient ))
anova(fmQUAD,fmQUADcor2)
hist(resultRE3$QUAD)
resultRE3$QUAD<-log(4*resultRE3$QUAD/(1-4*resultRE3$QUAD))
hist(resultRE3$QUAD)
summary(fmQUAD)
summary( glht( fmQUAD, linfct = mcp( Class = "Tukey" ) ) )

fmOST<-glmer(OST~Class+Minute-1+(Minute|Patient),data=resultRE3,
             family=binomial(link = "logit"))
fmOSTb<-glmer(OST~Class+Minute+(1|Patient),data=resultRE3,
              family=binomial(link = "logit"),
              control = list(maxIter = 2000),REML=F)
anova(fmOST,fmOSTb)
fmOSTc<-glmer(OST~Class+Minute+(Minute-1|Patient),data=resultRE3,
              family=binomial(link = "logit"))
anova(fmOST,fmOSTc)
fmOSTcor <- update(fmOST,correlation = corCAR1(form = ~Minute |Patient))
anova(fmOST,fmOSTcor)
summary( fmOST )
summary( glht( fmOST, linfct = mcp( Class = "Tukey" ) ) )

hockey <- function(x, alpha1, beta1, beta2, brk, eps = diff(range(x)/10)) {
  x1 <- brk - eps
  x2 <- brk + eps
  b <- (x2*beta1 - x1*beta2)/(x2 - x1)
  cc <- (beta2 - b)/(2*x2)
  a <- alpha1 + beta1*x1 - b*x1 - cc*x1^2
  alpha2 <- - beta2*x2 + (a + b*x2 + cc*x2^2)
  lebrk <- (x <= brk - eps)
  gebrk <- (x >= brk + eps)
```

```
  eqbrk <- (x > brk - eps & x < brk + eps)
  result <- rep(0,length(x))
  result[lebrk] <- alpha1 + beta1*x[lebrk]
  result[eqbrk] <- a + b*x[eqbrk] + cc*x[eqbrk]^2
  result[gebrk] <- alpha2 + beta2*x[gebrk]
  result
}

ta<-tapply(resultRE3$QUAD, resultRE3$Patient, length)>48
resultRE3$ok<-sapply(resultRE3$Patient, function(x){return(ta[x])})

nll1NOpC<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                    data=resultRE3[resultRE3$ok==TRUE,],
                    start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                    control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                    upper=c(Inf,Inf,Inf,Inf),subset=Class=="NOpC")
nll1NOpG<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                    data=resultRE3[resultRE3$ok==TRUE,],
                    start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                    control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                    upper=c(Inf,Inf,Inf,Inf),subset=Class=="NOpG")
nll1NOpO<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                    data=resultRE3[resultRE3$ok==TRUE,],
                    start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                    control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                    upper=c(Inf,Inf,Inf,Inf),subset=Class=="NOpO")
nll1OpC<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                   data=resultRE3[resultRE3$ok==TRUE,],
                   start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                   control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                   upper=c(Inf,Inf,Inf,Inf),subset=Class=="OpC")
nll1OpG<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                   data=resultRE3[resultRE3$ok==TRUE,],
                   start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                   control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                   upper=c(Inf,Inf,Inf,Inf),subset=Class=="OpG")
nll1OpO<-nlsLMList(QUAD~hockey(Minute,a,b,c,brk)|Patient,
                   data=resultRE3[resultRE3$ok==TRUE,],
                   start=list(a=0.1,b=-6e-06,c=6e-06,brk=100),
                   control=list(maxiter=1000),lower=c(0,-Inf,-Inf,0),
                   upper=c(Inf,Inf,Inf,Inf),subset=Class=="OpO")

par(mfrow=c(2,2))
boxplot( list( "NOpC"=coef(nll1NOpC)[,1],
               "NOpG"=coef(nll1NOpG)[,1],
               "NOpO"=coef(nll1NOpO)[,1],
               "OpC"=coef(nll1OpC)[,1],
               "OpG"=coef(nll1OpG)[,1],
               "OpO"=coef(nll1OpO)[,1]
), ylab="Variability at the breakpoint"
)
```

```
240 boxplot( list( "NOpC"=coef(nll1NOpC)[,4],
241                "NOpG"=coef(nll1NOpG)[,4],
242                "NOpO"=coef(nll1NOpO)[,4],
243                "OpC"=coef(nll1OpC)[,4],
244                "OpG"=coef(nll1OpG)[,4],
245                "OpO"=coef(nll1OpO)[,4]
246 ), ylab="Position the breakpoint"
247 )
248 boxplot( list( "NOpC"=coef(nll1NOpC)[,2]-coef(nll1NOpC)[,3],
249                "NOpG"=coef(nll1NOpG)[,2]-coef(nll1NOpG)[,3],
250                "NOpO"=coef(nll1NOpO)[,2]-coef(nll1NOpO)[,3],
251                "OpC"=coef(nll1OpC)[,2]-coef(nll1OpC)[,3],
252                "OpG"=coef(nll1OpG)[,2]-coef(nll1OpG)[,3],
253                "OpO"=coef(nll1OpO)[,2]-coef(nll1OpO)[,3]
254 ), ylab="Slope before the break point", ylim = c(-2e-4,1.5e-4)
255 )
256 abline( h = 0, lty = "dashed" )
257 boxplot( list( "NOpC"=coef(nll1NOpC)[,2]+coef(nll1NOpC)[,3],
258                "NOpG"=coef(nll1NOpG)[,2]+coef(nll1NOpG)[,3],
259                "NOpO"=coef(nll1NOpO)[,2]+coef(nll1NOpO)[,3],
260                "OpC"=coef(nll1OpC)[,2]+coef(nll1OpC)[,3],
261                "OpG"=coef(nll1OpG)[,2]+coef(nll1OpG)[,3],
262                "OpO"=coef(nll1OpO)[,2]+coef(nll1OpO)[,3]
263 ), ylab="Slope after the break point", ylim = c(-4e-5,3e-5)
264 )
265 abline( h = 0, lty = "dashed" )
266 dev.off()
```