University of Potsdam

Departement of Bioinformatics

Master Thesis

in order to obtain the Degree of

Master of Science (M. Sc.)

Predictive Analysis of Metabolic and Preventive Patient Data

Graduate

Sören Matzk

Supervisors

Prof. Dr. Joachim Selbig, Potsdam

Dr. med. Helena Orfanos-Boeckel, Berlin

Potsdam, May 23, 2016

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-406103
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-406103

Abstract

Every day huge amounts of medical records are stored by means of hospitals' and medical offices' software. These data are generally unconsidered in research.

In this work anonymized everyday medical records ascertained in a physician's office, covering holistic internal medicine in combination with orthomolecular medicine, are analyzed. Due to the lack of cooperation by the provider of the medical practice software a selection of diagnoses and anthropometric parameters was extracted manually. Information about patients' treatment are not available in this study. Nevertheless, data mining approaches including machine learning techniques are used to enable research, prevention and monitoring of patients' course of treatment.

The potential of these everyday medical data is demonstrated by investigating co-morbidity and pyroluria which is a metabolic dysfunction indicated by increased levels of hydroxyhemopyrrolin-2-one (HPL). It points out that the metabolic syndrome forms a cluster of its components and cancer, as well as mental disorders are grouped with thyroid diseases including autoimmune thyroid diseases. In contrast to prevailing assumptions in which it was estimated that approximately 10 % of the population show increased levels of HPL, in this analysis 84.9 % of the tested patients have an increased concentration of HPL.

Prevention is illustrated by using decision tree models to predict diseases. Evaluation of the obtained model for Hashimoto's disease yield an accuracy of 87.5 %. The model generated for hypothyroidism (accuracy of 60.9 %) reveals shortcomings due to missing information about the treatment.

Dynamics in the biomolecular status of 20 patients who have visited the medical office at least one time a year between 2010 and 2014 for laboratory tests are visualized by STATIS, a consensus analysis based on an extension to principal component analysis. Thereby, one can obtain patterns which are predestinated for specific diseases as hypertension.

This study demonstrates that these often overlooked everyday data are challenging due to its sparsity and heterogeneity but its analysis is a great possibility to do research on disease profiles of real patients.

Contents

List of Figures

List of Tables

List of Abbreviations

HPL Hydroxyhemopyrrolin-2-one

AITD Autoimmune Thyroid Disease

STATIS Structuration des Tableaux à Trois Indices de la Statistique

PCA Principal Component Analysis

MetS Metabolic Syndrome

CVD Cardiovascular Disease

HDL-C High Density Lipoprotein Cholesterol

LDL-C Low Density Lipoprotein Cholesterol

GPT (ALAT) Alanine Transaminase

Gamma-GT Gamma-Glutamyl Transferase

HbA1c Glycated Hemoglobin

TSH Thyroid Stimulating Hormone

TPO-Ab Thyroid Peroxidase Antibody

TG Thyroglobulin

TSH-R Thyroid Stimulating Hormone Receptor

HT Hashimoto's Thyroiditis

GD Graves' Disease

M Less Sensitive Patients

H Highly Sensitive Patients

KNN k-Nearest Neighbor

SQL Structured Query Language

CV Cross-Validation

CVP Cross-Validation Performance

BMI Body Mass Index

FT3 Free Triiodthyronine

FT4 Free Thyroxine

CRP C-Reactive Protein

HOMA Homeostatic Model Assessment

Chapter 1

Introduction

The development in information technology has influenced the way medical records are documented. Every day huge amounts of medical records are stored by means of hospitals' and medical offices' software which are frequently customized for the user. Thus, medical records containing the same information are differently structured in different institutions [1]. Hence, automatic sharing and integration of these data across institutions are impossible which hinders optimal healthcare. Scandinavian countries are in the vanguard of a patient-centered healthcare system. In Denmark, a national centralized computer database containing electronic medical records such as laboratory data and prescribing information. This database is accessible for patients and physicians, but also target of research projects investigating, inter alia, co-morbidities. So, one can study disease profiles of real patients instead of studying people disease by disease as it has been done in the past [2]. Contrarily to common case-control or cohort studies, one deals with the challenging analysis of heterogeneous data. The number of observed patients and their features is high dimensional but sparse. Furthermore, due to the investigation of patients instead of test persons the data does not contain a control group. [3, 4]

In this work anonymized everyday medical records ascertained in a physician's office covering holistic internal medicine in combination with orthomolecular medicine are analyzed. Orthomolecular therapy correct imbalances of substances that are normally present in the body such as vitamins, hormones, minerals, trace elements, macronutrients [5, 6]. In 2012 approximately 82 % of the medical offices in Germany storing their patients' medical records by a medical practice software. [7]. These medical practice softwares meet the requirements of basic digitization and support administration functions, but the analysis of patients' medical records is not their focal point [8]. Due to the lack of cooperation by the provider of the medical practice software a selection of diagnoses and anthropometric parameters was extracted manually. The database was extended by laboratory values which are digitally provided by two laboratories. Unfortunately, a lot of actually existing records, as saliva tests and stool samples as well as information about the treatment, are not available in this study.

The aim of this work is to demonstrate the great potential of these generally overlooked data. Therefore, data mining approaches including machine learning are used to enable research, prevention and monitoring of patients' course of treatment.

The next chapter introduces the data and their preparation. Strategies dealing with missing values are depicted in chapter 3. In chapter 4, the relevance of hydroxyhemopyrrolin-2-one (HPL) for autoimmune thyroid disease (AITD) and mental disorders is investigated. Additionally, medical records are used to discover co-morbidities and disease correlations. In chapter 5, decision trees generated by Quinlan's C5.0 algorithm are used to predict diagnoses and show patterns comprising laboratory tests and anthropometric parameters. Decision trees are highly interpretable nonparametric classifiers which are able to deal noisy and incomplete data [9, 10]. The ability of STATIS[1], allowing simultaneous investigation of multiple tables which is not readily achievable by principal component analysis (PCA), to monitor the course of treatment is discussed in chapter 6. Finally, the outcomes of this work are summarized in chapter 7.

1.1 Medical Background

The following sections briefly introduce diseases of special interest in this study.

1.1.1 Metabolic Syndrome

The metabolic syndrome (MetS) is a cluster of cardiovascular disease (CVD) risk factors that include glucose intolerance, hyperinsulinaemia, dyslipidaemia, hypertension, visceral obesity, hypercoagulability and microalbuminuria [11]. It was also called syndrome X [12], the insulin resistance syndrome [13] and the deadly quartet [14]. As a result of a global epidemic of obesity and diabetes [11], the MetS has become one of the major health challenge in developed countries over the last three decades [15]. Because the MetS consists of various CVD risk factors it can be used as an indicator for type 2 diabetes and CVD [16]. A uniform definition of the MetS does not exist. Several expert groups, including the World Health Organization Diabetes Group [17], the European Group for the Study of Insulin Resistance [18] and the United States National Cholesterol Education Program [19], have attempted to produce diagnostic criteria. The most recent approach by the International Diabetic Federation [20] focuses on a definition which is applicable to different ethnic groups. Their definition for Europeans is shown in the appendix in table A.1 and includes waist circumference, triglyceride, high density lipoprotein cholesterol (HDL-C), fasting plasma glucose and blood pressure. As a result of everyday medical data the definition of the MetS is partially different in this study, see table 1.1. Several studies have shown that increased

[1] acronym for the french expression *'Structuration des Tableaux à Trois Indices de la Statistique'*

Table 1.1. Definition of the MetS. Each tile represents one risk factor for the MetS. One speaks of the MetS if a patient fulfills one of the criteria of at least three tiles.

At least one of the following criteria		
- Raised Waist circumference		
	men	≥ 94 cm
	women	≥ 80 cm
- Raised BMI		≥ 28 kg/m^2
At least one of the following criteria		
- Raised triglyceride		> 150 mg/dL
- Specific treatment for this lipid abnormality		
At least one of the following criteria		
- Reduced HDL-C		
	men	< 40 mg/dL
	women	< 50 mg/dL
- Specific treatment for this lipid abnormality		
At least one of the following criteria		
- Raised blood pressure		
	Systolic	≥ 130 mm Hg
	Diastolic	≥ 85 mm Hg
- Treatment of previously diagnosed hypertension		
At least one of the following criteria		
- Raised glycated hemoglobin		$\geq 5.8\%$
- Previously diagnosed type 2 diabetes		
- Raised alanine transaminase		≥ 30 U/I
- Raised gamma-glutamyl transferase		≥ 35 U/I
-Raised uric acid		≥ 6.5 mg/dL

alanine transaminase (ALAT) [21], gamma-glutamyl transferase (Gamma-GT) [22], uric acid [23] and glycated hemoglobin (HbA1c) [24] are associated with the MetS. In comparison to fasting plasma glucose which underlies diurnal variation [25, 26], HbA1c is more reliable because erythrocytes survive approximately 115 days in the circulation [27]. Additionally, HbA1c is more practical for clinical routine because there are no requirements, such as fasting, for patients.

1.1.2 Autoimmune Thyroid Disease

In the case of an AITD, the immune system attacks the body's own thyroid gland, which can lead to the destruction of thyroid tissue over time [28]. AITD is mainly associated with the three antibodies: thyroid peroxidase antibody(TPO-Ab), thyroglobulin antibodies (TG) and thyroid stimulating hormone receptor antibodies (TSH-R). Each of these antibodies affects the production of the thyroid hormones differently by which AITD can be classified in subgroups such as Hashimoto's thyroiditis (HT) and Graves' disease (GD). HT is characterized by the presence of TPO-Ab as well as TG and leads to symptoms of hypothyroidism. On the contrary, GD is caused by TSH-R and leading to hyperfunction of the thyroid gland. In some cases one can observe TPO-Ab and TG as well as TSH-R causing a switch of GD to HT and vice versa [29]. [30]

It is estimated that about 5 % of the general population suffer from AITDs. The prevalence of AITD is four to ten times higher in women than in men, especially in women between 30 and 50 years. The causes comprise complex interaction of several factors such as genetic and environmental factors. [30]

1.1.3 Pyroluria

Pyroluria is a controversially discussed metabolic dysfunction which is not considered by conventional medicine. Pyroluria is indicated by increased levels of HPL which is a metabolite of the heme synthesis. It is assumed that HPL inhibits the heme synthesis. Heme is necessary for energy production and additionally it is required for detoxification and antioxidant defense. [31–34]

In 1961 Irvine et al. discovered the mauve factor, a pyrrole with increased prevalence in schizophrenics [35]. Research in the 1970s showed that the mauve factor is the hemopyrrole derivative HPL [36, 37]. Furthermore, patients with HPL show deficiencies in in the levels of zinc and vitamin B6, as evinced, inter alia, by Pfeiffer [38]. The connection between HPL and deficiencies in essential metals, as well as the association of HPL with mental illness, are controversially discussed in a review by the Robert-Koch-Institute. In addition, the chemical identity of HPL and its analysis remain contentiously since the 1980s where research on HPL reached its peak and ended without approved results. [39]

In this study, HPL is assumed to be a common factor in the development of MetS as well as in mental disorders and HT. The second hypothesis includes that patients suffering from MetS, mental disorders and HT show an increased level of HPL (> 1.0 nmol/L). Increased levels of HPL in less sensitive patients can result in MetS. Considerably increased levels of HPL (> 2.0 nmol/L) is more prevalent in highly sensitive patients who are more frequently affected by mental disorders and HT. Thus, patients suffering from mental disorders may have somatic dysfunction.

1.2 Machine Learning

In this section supervised and unsupervised learning algorithms are informally introduced. Formal definitions for the applied methodologies follow in chapters 3, 5 and 6. In general, machine learning includes computer programs that automatically learn to detect complex patterns on the basis of given data. The idea is based on the assumption that the future will be similar to the past when sample data was collected and thus, recognized regularities and predictions can also be expected to be correct. Machine learning has a widespread scope of application comprising, inter alia, prediction of medical diagnoses. This work is based on supervised and unsupervised learning algorithms. In supervised learning the training procedure depends on labeled input data such that the model is built by the entire knowledge. Supervision is obtained through labeling. Here, decision trees, boosting and k-Nearest-Neighbor (KNN) approaches are used as representatives of supervised learning. As the name implies unsupervised learning comprises the detection of regularities using an unlabeled input data set. Due to missing labels these models do not explain the semantic meaning of computed regularities. In chapter 6 STATIS is introduced. This method is part of principle component analysis (PCA) family and therefore an unsupervised learning algorithm. [40, 41]

Chapter 2

Data

This chapter briefly explains the origin of the analyzed data and their systematic management in a database system. More details about the creation of this database can be taken from preliminary work [42].

2.1 Data Source

The basis of this study is everyday data of a medical office in Berlin, Germany. Thus, one has to deal with complex and heterogeneous information with respect to laboratory tests and diagnosis of patients. As already mentioned in the introduction it was not possible to automatically extract information from the medical practice software due to the lack of cooperation by its provider. Therefore, it was necessary to minimize manual workload while maximizing the diversity of information, i.e. only a part of the total amount of characteristics was extracted. This includes the smoker status, laboratory results for HPL and anthropometric data (height, weight) and 35 diagnosis and disorders.

Among these characteristics only time points for measured weights was extracted. Smoker status is splitted in four groups: non-smoker, occasional smoker (smokes few times a year), ex-smoker (stopped smoking at least five years ago) and smoker. Laboratory results of the medical office's most relevant laboratory were provided digitally as csv files.

2.2 Data Preparation

All data are stored in a relational database because the analysis requires integrity of patient data and the possibility of simply creating specific datasets. Figure 2.1 shows the underlying entity relationship diagram which consists of the six tables `'Patient'`, `'Diagnosis'`, `'Weight'`, `'Laboratory Test'`, `'Test'` and `'Name of Test'`. Unless table `'Diagnosis'` all tables are normalized to second normal form.

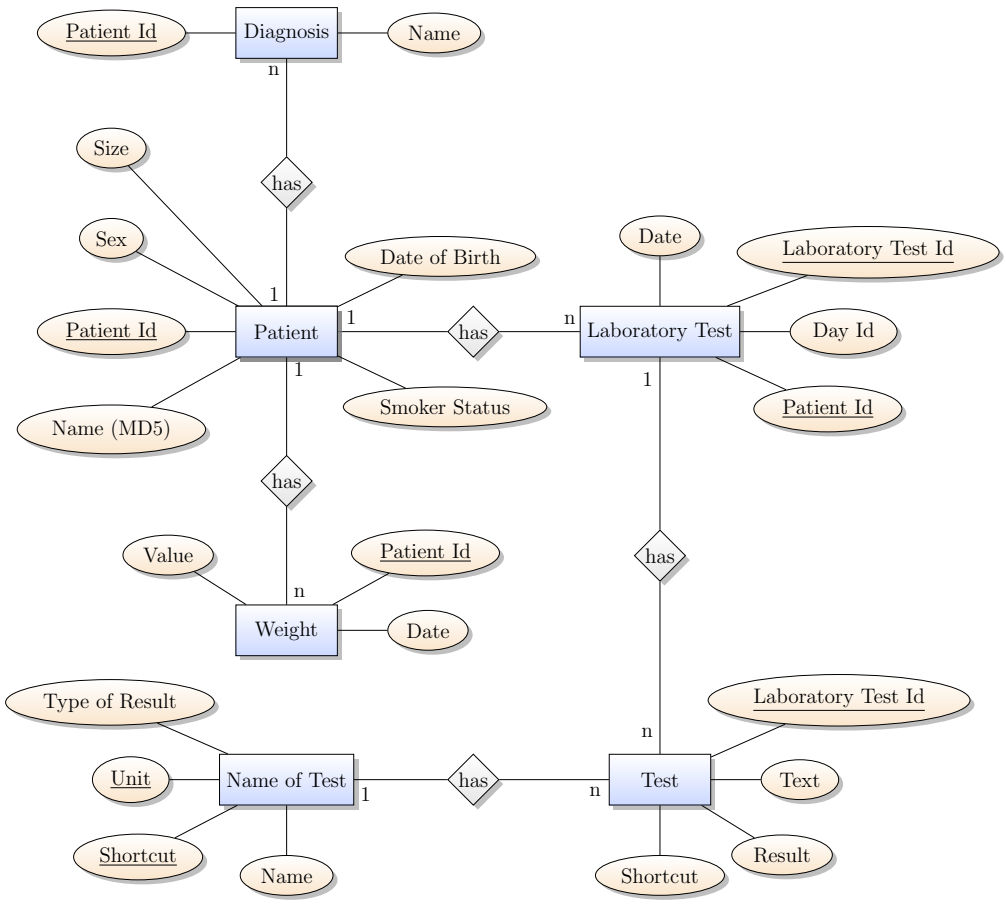

Figure 2.1. Entity-relationship diagramm of the underlying database model. It illustrates the linkage and cardinalities of the contained entities. The model only consists of one to many relations such that each patient has many laboratory tests but each laboratory test belongs to one patient only.

The data model is implemented in SQLite, an open source relational management system. SQLite delivers a serverless, zero-configuration and platform independent Structured Query Language (SQL) database engine which can be manipulated and accessed by the statistical software `R` [43] with the package `RSQLite` [44]. Since October 2008 the digitally provided laboratory results contain names and dates of birth as identification keys. Hence, only patients with laboratory tests since October 2008 are included in the database. Thus, one has to keep in mind that these laboratory data may not contain information about the first medical consultation. Further extensions of this database can only be achieved by manual work because data integrity is not guaranteed in the event of typos or change of name (e.g. due to marriage).

2.3 Database Content

As of September 2014, the study population included 1565 patients (68.6% females, 31.4% males) with an average age of 50.4 ± 15.4 years (range 9 - 101 years).

A basic characteristic of everyday medical office's data is missing values. In this study the data contain $680,298$ (84 %) missing values that are spread over anthropometric data, smoker status and laboratory tests. On the one hand, this is due to cost-effectiveness and on the other to individual organization of medical appointments. The latter leads to numerous missing series of measurements that cannot be described by an exact number because it depends on the selected interval and time period.

Chapter 3

Handling Missing Data

Real-world data are generally more affected by missing values than clinical trials which can be confirmed by section 2.3. Because statistical approaches mainly require data sets comprising no, or at least small numbers of missing data, handling missing values is an important topic in the field of statistics.

3.1 Methods of Resolution

Han et al. mentioned various strategies in their introduction to data mining [41] as:

1) Exclude objects with missing values from analysis

2) Fill in the missing value with a constant

3) Fill in the missing value with a measure of central tendency (e.g. mean or median) of the attribute

   (a) considering all classes

   (b) for each class separately

4) Fill in the missing value using the most probable one computed by machine learning approaches (e.g. KNN)

Each approach has advantages and disadvantages as for instance, 1) is very strict and in case of numerous missing values the resulting data set is highly reduced or even empty. But, exclusion of missing values does not bias the data as it is possibly due to incorrect estimates computed in 2) - 4). These three imputation strategies are inappropriate in the case of a systematic lack of information. Whereas 2) - 4) are methods of choice, if missing values are randomly distributed. Especially 4) has a greater chance to yield more accurate results because it considers the most information compared to the others.

In this study, a less restrictive variant of 1) is used to reduce missing values and strategies 3a) and 4) are used for imputing missing values. Both approaches are described in the following sections. For better illustration, suppose, a $n \times m$ matrix X comprising n objects L described by m attributes. An element of X is denoted as $x_{i,j}$ where $i \in \{1, 2, \dots n\}$ and $j \in \{1, 2, \dots m\}$.

3.1.1 Reduce Missing Data

Considering the data of this study strategy 1) would yield an empty data set and due to the amount of missing values 2) - 4) are not appropriate, as well. Therefore, a less restrictive method, using an arbitrary threshold for determining if an object (row) or attribute (column) should be excluded, is used to reduce missing data. These two steps are depicted in figure 3.1 and can be combined iteratively such that the resulting data table contains a certain proportion of missing values. This dimensionally reduced table can be used for imputation of missing values.

3.1.2 Imputation through k-Nearest-Neighbor Algorithm

As the name implies, the missing value is estimated by the non-missing values of k closest neighbors. The neighborhood of an object L is calculated by a distance metric. Here, the Euclidean distance is used because it shows promising results. The Euclidean distance d is defined as

$$d(L_r, L_s) = \sqrt{\sum_{k=1}^{m} (x_{rk} - x_{sk})^2}, \tag{3.1}$$

where $1 \leq r, s \leq n$. The smaller the distance between two objects, the more closer they are. Then, a weighted average of k nearest neighbors is used for imputation, in which the weights correspond to the closeness. In the case of missing values, $x_{rk} - x_{sk}$ is assumed to be the maximum possible difference. [41, 45]

In this study, a modified approach of KNN is used to obtain more accurate estimates. Due to the fact that the majority of patients received multiple laboratory tests, one can use previously and upcoming laboratory tests for imputation. Thereby, the imputed value w_{imp} is calculated by

$$w_{\text{imp}} = \text{mean}(w_{\text{prev}}, w_{\text{KNN}}, w_{\text{next}}), \tag{3.2}$$

where w_{KNN} denotes the value which is estimated by KNN and w_{Kprev} as well as w_{next} denote the previously measured test result and next measured test result, respectively. If the database contains neither w_{pref} nor w_{next} then w_{imp} is equal to w_{KNN}.

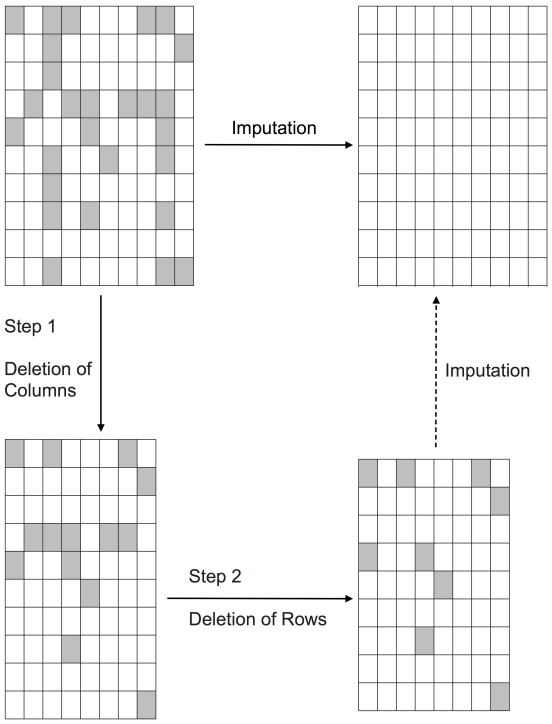

Figure 3.1. Data cleaning including reduction and imputation of missing values. White colored cells represent non-missing attributes. Steps 1 and 2 show deletion of objects or attributes containing more than 50 % missing data. The resulting dimensionally reduced table is more appropriate for imputation approaches because of its proportion of missing values.

Chapter 4

Disease Network

4.1 Introduction

One of the most challenging problems in biomedical research includes the investigation of co-morbidity, i.e. exploration of diseases and disorders that co-occur in one patient because of the associations among these diseases [46]. For instance, these associations can occur because diseases are associated with the same gene [47] or proteins [48]. The number of co-morbid diseases has a significant influence on overall survival of patients [49]. These data are commonly visualized as networks [46–50] which are highly interpretable. Among the aforementioned studies, only the work of Hidalgo et al. considered everyday data recorded by hospitals and insurance programs. By means of these everyday data one is able to investigate disease prevalence and dynamics referring to several aspects as ethnic groups or gender. Thereby, this methodology can help to detect diseases at the earliest detectable phase. [50]

The aims of this chapter include the demonstration of disease networks' (DN) usability in visualizing everyday medical records and the investigation of the influence of HPL levels on MetS, AITD and mental disorder.

4.2 Theoretical Background

4.2.1 Jaccard Index

The determination of similarity between two sets is obtained by the application of metrics. One of these is the Jaccard index [51] which provides a measure for the similarity of binary sets D_i and D_j. It is defined as the size of the intersection of those two sets divided by the union of them:

$$Jac(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}, \tag{4.1}$$

where D_i,$D_j \in \{0,1\}$, $\forall\ i,j \in \{1,2,\ldots,l\}$. Each coefficient $Jac(D_i,D_j)$ is then stored in a symmetric similarity matrix M:

$$M = \begin{pmatrix} 1 & Jac(D_1,D_2) & \cdots & Jac(D_1,D_L) \\ Jac(D_2,D_1) & 1 & \cdots & Jac(D_2,D_L) \\ \vdots & \vdots & \ddots & \vdots \\ Jac(D_L,D_1) & Jac(D_L,D_2) & \cdots & 1 \end{pmatrix} \tag{4.2}$$

A maximal Jaccard index of $Jac(D_i,D_j) = 1$ indicates identity of D_i and D_j, whereas $Jac(D_i,D_j) = 0$ shows that the sets have no element in common.

4.3 Methodology

4.3.1 Correlations between Diseases

For this analysis only diseases occurring more than 20 times are considered. Thus a $k \times l$ binary matrix D is the basis of this approach where k denotes the number of patients and l the number of diseases. Further, a symmetric similarity matrix M is calculated by means of D, where each entry m_{ij} reflects Jaccard index between disease d_i and d_j. Thus, M contains information about the relation between all diseases.

The interpretation of M requires an estimated significance threshold s in order to differentiate between potential real associations and those which might occur by chance [52]. Therefore, one has to generate a matrix D' by shuffling all elements of D which should remove underlying causality. Then a symmetric similarity matrix M' is calculated on the basis of D', analog to M. The 95% confidence level is used which is arbitrary, but commonly used in determining significance. Here, the particular significance threshold s is obtained by calculating the 95^{th} percentile of M'.

Finally, s can be used as a threshold for M' such that

$$m'_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq s \\ 0 & \text{otherwise} \end{cases}, \text{ where } i \in \{1,2,\ldots,k\} \text{ and } j \in \{1,2,\ldots,l\}, \tag{4.3}$$

which leads to a DN.

4.3.2 Correlations between Diseases and HPL

In this part, the relation of HPL to certain diseases is investigated. Therefore, only patients with a measured HPU level are taken into account. The originally continuous test results

Table 4.1. Filter for the three groups: `M`, `H` and `Healthy`.

Group	Rule
`H`	Hashimoto's disease *OR* HPL > 1.0 *AND* BMI < 27 *AND* (stress *AND* exhaustion) *AND* (depression *OR* depressive disorder *OR* somatoform disorder *OR* psychosomatic disorder)
`M`	(MetS *AND* BMI ≥ 30) *OR* ((hypertension *OR* (diabetes mellitus *OR* insulin resistance)) *AND* BMI ≥ 30)
Healthy	at least 5 laboratory tests *AND* non of the following diagnoses: insulin resistance, diabetes mellitus, cancer, psychological disorder, somatoform disorder, hypothyroidism, AITD, MetS, chronic renal insufficiency, depression, depressive disorder, migraine, adrenal weakness, exhaustion, stress asthma, osteoporosis, adaptive disorder, rheumatoid arthritis, anxiety disorder

of HPL are binned in the two categories `low` and `high`, such that `low` includes patients with a HPL level < 1.0 nmol/L and `high` includes patients with a HPL level ≥ 1.0 nmol/L, respectively.

The resulting categorical variable is considered as a disease and thus, the remaining procedure is analog to the above explained calculation of correlations between diseases.

4.3.3 `M-Group` vs. `H-Group`

Alternatively, HPL's influence on AITD and MetS can be investigated by comparing highly sensitive patients (`M`) and less sensitive patients (`H`) with healthy patients. The subgroups are generated by applying logical filters (table 4.1). Healthy patients are filtered due to an arbitrary threshold of five laboratory test. Therewith, one reduces the risk that patients are rated as healthy because of the lack of medical consultations for diagnosing.

4.3.4 Statistical Analysis

Differences in proportions were investigated using χ^2-tests. Wilcoxon tests were used to analyze differences in the mean of measured test results and their corresponding standard values. All alternative hypothesis are two-sided and p-values < 0.05 are considered as statistical significant.

4.4 Results

The following results are based on laboratory tests and diagnosed diseases of 1565 patients containing approximately two-thirds women and one-third men, respectively. Figure 4.1

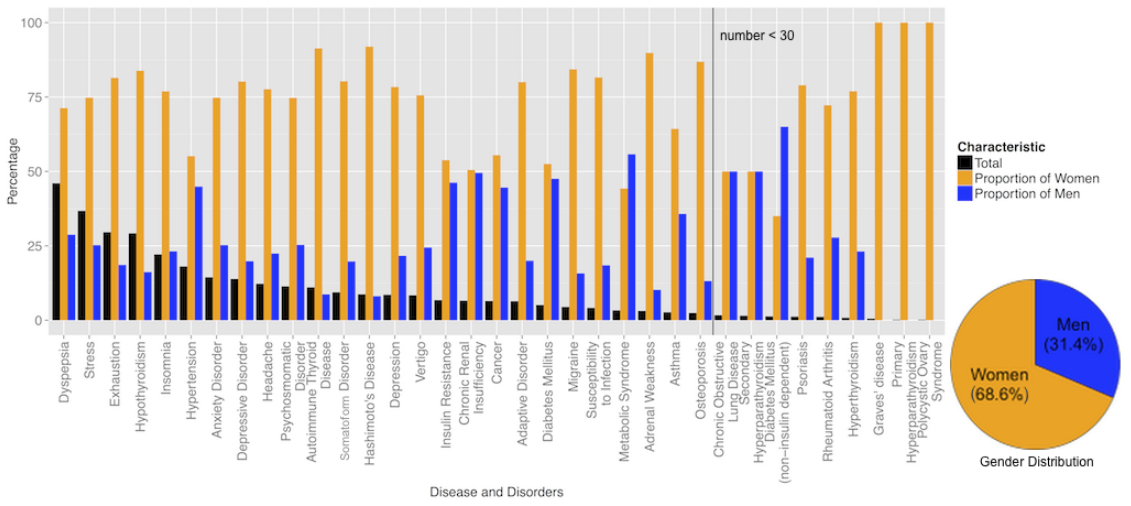

Figure 4.1. Frequencies of Diseases and Disorders. Black bars reflect the proportion of diseases according to all patients. The distribution of women and men in a specific disease is shown in orange and blue bars, respectively. The gender distribution according to the entire data is demonstrated in the pie plot.

shows the occurrence's frequency of 35 diseases and disorders in the study population. One can detect a higher incidence of mental disorders in comparison to organic diseases which is even present in the top five consisting of dyspepsia (46 %), stress (36 %), exhaustion (29 %), hypothyroidism (29 %) and insomnia (22 %). The absolute number of occurrence of each disease is shown in table B.1 in the appendix. Women more frequently suffer from hypothyroidism, AITD and mental disorders than men. Whereas, chronic renal insufficiency, cancer, chronic obstructive lung disease, secondary hyperpathyroidism, as well as MetS and its components are more or less equally distributed.

4.4.1 Correlations between Diseases

Associations between all diseases expressed in the study population can be summarized by constructing a DN. Figure 4.2a depicts the DN according to all patients in which the nodes represent disease phenotypes and edges connect phenotypes that show significant co-occurrence regarding the Jaccard similarity index. The corresponding heatmap is dipslayed in figure B.1a in the appendix. It points out that MetS forms a cluster of its components and cancer, as well as mental diseases are grouped with thyroid diseases. Hypertension does often occur in the context of MetS and cancer, but it is also associated with mental diseases and thyroid diseases. Asthma, adrenal weakness and osteoporosis do not significantly co-occur with any other disease.

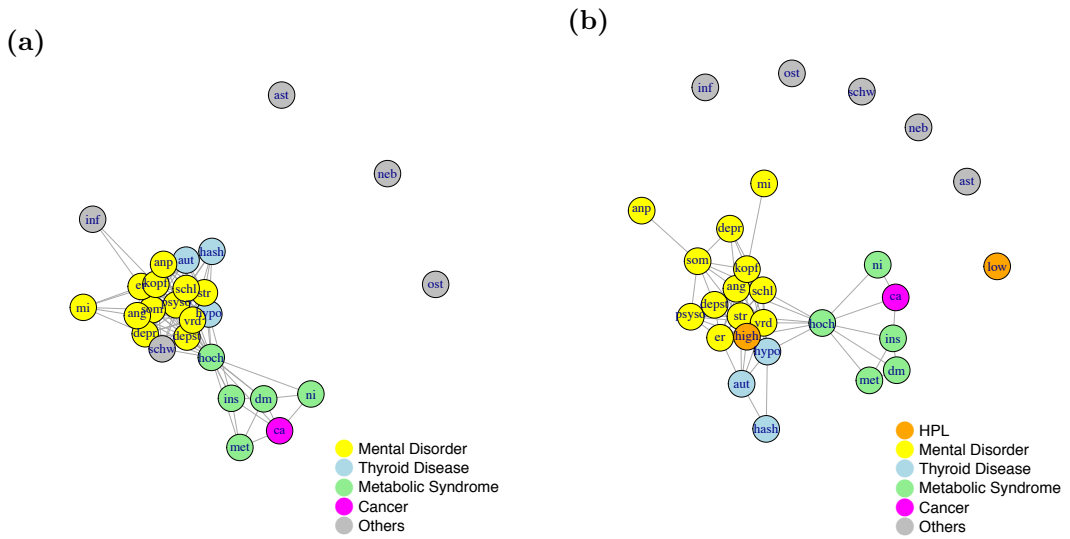

Figure 4.2. Disease networks (DNs). Colored nodes reflect diseases and link represent significant correlations. The shortcuts for the diseases are explained in table B.1 in the appendix. (a) DN constructed using all patients. (b) DN constructed using patients with measured HPL.

4.4.2 Correlations between Diseases and HPL

Additionally, the reliability and co-morbidity of increased HPL levels are investigated by a DN comprising diagnoses of 465 patients with measured HPL levels. The resulting network (figure 4.2b) has a similar structure as the previously described DN including all patients, e.g. the conserved clique consisting of MetS and its components. Among the 395 patients with an increased HPL level of ≥ 1 nmol/L, mental disorder and thyroid disease are more frequent. The node representing the remaining 70 patients with HPL levels smaller than 1 nmol/L does not show any significant linkage. Table 4.2 shows only the laboratory results where the means differ significantly regarding these two subgroups `Low` and `High`.

4.4.3 Group `M` vs. Group `H`

The selection described in table 4.1 yields three groups of similar size, i.e. 65 patients in `M`, 51 patients in `H` and 44 healthy patients. Note, only characteristics that are not included in the logical filter are tested on statistical significant differences. In comparison to the members of group `Healthy`, group `M`'s patients are significantly older which is shown in Table 4.3 including typical characteristics of MetS and AITD. Patients belonging to group `M` differ tremendously in factors associated with MetS compared to the other groups. Differences between healthy patients and group `H` arise from thyroid levels whereas, group `H` has similar thyroid levels as `M`'s patients.

Table 4.2. Characteristics of the two groups `Low` (HPL < 1.0 nmol/L) and `High` (HPL ≥ 1.0 nmol/L) which differ significantly in their means. Numerical variables are described by *mean*±*standard deviation* whereas absolute numbers are used in the case of binary attributes. Percentages represent proportion of the attribute within each group. Statistical significance is indicated by an asterisk.

Characteristics	Low	High
HPL (nmol/L)	0.8 ± 0.1	2.4 ± 1.7
TPO-Ab* (U/mL)	$19.2 \pm 34.0(47.1\%)$	$45.8 \pm 107.1(60.3\%)$
Vitamin D* (ng/mL)	$21.1 \pm 11.4(72.9\%)$	$24.4 \pm 12.0(83.8\%)$
Homocysteine* (μmol/L)	$11.7 \pm 3.1(74.3\%)$	$11.2 \pm 3.2(80.5\%)$

4.5 Discussion

Assuming that diseases' appearances are not affected by gender, their proportional occurrence in women and men should be similar to the gender distribution in figure 4.1. This assumption does not hold. Studies approved that hypothyroidism [53], AITD [30], osteoporosis [54], primary hyperparathyroidism [55], and mental disorder [56] are more prevalent among women than men. These results are confirmed by the disease frequencies in this study shown in figure 4.1. Due to human anatomy it is clear that only women suffer from the polycystic ovary syndrome and hence, it is missing in the aforementioned listing. This study's observations confirm the prevalent assumption that men are more affected by chronic renal insufficiency, cancer, and MetS and its components than women. However, the proportional differences in women and men are minor in diabetes mellitus and hypertension.

4.5.1 Correlations between Diseases

The diseases in the network in figure 4.2a are structured as expected, i.e. MetS is directly linked with its components, and stress as well as exhaustion is highly connected with mental disorders. This DN does not allow to conclude about the directionality of disease progression due to the lack of information about diagnoses made during each visit. However, this network confirms study results referring to MetS and AITD (section 1.1.1 and 1.1.2). The higher prevalence of hypertension in patients suffering from hypothyroidism or vice versa has been investigated in the work of Saito et al.. They discovered that hypothyroidism causes secondary hypertension [57,58]. This connection can be further confirmed by comparing conditional probabilities. Approximately 40 % of patients suffering from hypertension also suffer from hypothyroidism, while 25 % of patients with hypothyroidism suffer from hypertension.

Note, that this network represents the entire study population. Data sets controlled for sex or age can have different structures in DNs. This can be demonstrated by osteoporosis which is well known to be more prevalent in elderly people [54]. In this study osteoporosis does

Table 4.3. Characteristics of the three groups `M` and `H` and `Healthy`. Only characteristics that are not considered in the filter are tested on significant differences. Numerical variables are described by *mean*±*standard deviation* whereas absolute numbers are used in the case of binary attributes. Percentages represent proportion of the attribute within each group. Statistical significance is indicated by an asterisk or cross, respectively.

Characteristics	M $n = 65$	Healthy $n = 44$	H n = 51
Gender			
Men	23(35.4%)	20(45.5%)	$7^{\dagger}$(13.7%)
Women	42(64.6%)	24(54.5%)	$44^{\dagger}$(86.2%)
Age	$57.5 \pm 13.1^{*}$(100%)	44.9 ± 14.9(100%)	46.2 ± 10.4(100%)
Body Mass Index	34.2 ± 4.5(100%)	23.0 ± 3.0(100%)	21.7 ± 2.5(100%)
Hashimoto's Thyroiditis	8(12.3%)	0(0.0%)	18(35.3%)
Levels			
TSH (μIU/mL)	$2.0 \pm 1.3^{*}$(75.4%)	1.4 ± 0.6(84.1%)	$2.0 \pm 1.1^{\dagger}$(74.5%)
FT3 (pg/mL)	3.1 ± 0.4(55.4%)	3.1 ± 0.3(79.5%)	$2.9 \pm 0.4^{\dagger}$(58.8%)
FT4 (ng/dL)	1.2 ± 0.2(56.9%)	1.2 ± 0.1(79.5%)	1.2 ± 0.2(58.8%)
TPO-Ab (U/mL)	31.7 ± 71.4(47.7%)	10.6 ± 8.9(59.1%)	$70.9 \pm 135.5^{\dagger}$(54.9%)
Selen (μg/L)	85.3 ± 21.7(55.4%)	86.2 ± 26.5(90.9%)	87.7 ± 11.6(64.7%)
Metabolic Syndrome	23(35.4%)	0(0%)	1(2.0%)
Hypertension	50(76.9%)	0(0%)	10(19.6%)
Diabetes Mellitus	17(26.2%)	0(0%)	1(2.0%)
Insulin Resistence	42(64.6%)	0(0%)	2(3.9%)
Levels			
HDL-C (mg/dL)	$49.6 \pm 10.5^{*}$(58.5%)	68.1 ± 14.2(38.6%)	65.5 ± 17.3(29.4%)
LDL-C (mg/dL)	136.8 ± 37.5(60.0%)	118.8 ± 33.9(36.4%)	119.5 ± 34.0(31.4%)
GPT (ALAT) (U/L)	$32.2 \pm 32.5^{*}$(61.5%)	23.0 ± 18.1(77.3%)	21.1 ± 9.3(54.9%)
Gamma-GT (U/L)	$31.9 \pm 21.9^{*}$(63.1%)	23.6 ± 34.9(75.0%)	18.5 ± 11.1(54.9%)
HbA1c (%)	$6.0 \pm 0.6^{*}$(69.2%)	5.5 ± 0.26(70.5%)	5.5 ± 0.3(51.0%)
Other Diseases			
Psychosomatics	9(13.8%)	0(0.0%)	25(49.0%)
Stress	34(52.3%)	0(0.0%)	51(100%)
Exhaustion	28(43.1%)	12(27.3%)	51(100%)
Insomnia	20(30.8%)	6(13.6%)	24(47.1%)
Dyspepsia	37(56.9%)	26(59.1%)	39(76.5%)
Headache	8(12.3%)	3(6.8%)	16(31.4%)
Other Parameters			
Vitamin D (ng/mL)	20.5 ± 10.7(64.6%)	20.9 ± 10.1(93.2%)	23.3 ± 12.0(66.7%)
Apr/May/Jun/Jul/Aug/Sep	22	18	16
Oct/Nov/Dec/Jan/Feb/Mar	20	23	18
HPL (nmol/L)	2.5 ± 2.4(26.2%)	1.8 ± 1.0(47.7%)	2.8 ± 3.0(84.3%)
Homocysteine (μmol/L)	11.9 ± 3.2(63.1%)	11.0 ± 3.7(86.4%)	10.1 ± 2.1(66.7%)
Kreatinin (g/L)	0.9 ± 0.3(67.7%)	0.8 ± 0.2(70.5%)	0.8 ± 0.2(56.9%)

not show any significant correlation. But, considering only patients with osteoporosis one obtains the conditional probability of approximately 45 % that one suffers from hypertension. Furthermore, the Jaccard similarity index is affected by the size influence associated with the frequency of occurrence [59]. This is encouraged by the fact that the three unconnected nodes show the lowest number of occurrence.

4.5.2 Correlations between Diseases and HPL

The DN in figure 4.2b can be seen as a map of the phenotypic space including HPL levels categorized as `low` and `high`, respectively. The structure of the network with all patients is preserved and therefore, one can suppose that the reduced sample size (30 %) is appropriate to represent the study population. The presumed association between increased HPL levels and mental disorder is confirmed by this network as well as the connection between HPL and AITD. The Jaccard similarity index of increased HPL levels and MetS is not significant and therefore, they are not connected. The link between increased HPL levels and hypertension may indicate the association of HPL with MetS. Nevertheless, this linkage could be induced by the correlation between HPL and hypothyroidism. Partial correlations could be used to overcome this problem.

The connection of HPL and AITD can be confirmed by the TPO-Abs which are significantly higher in patients with increased HPL levels (table 4.2). However, referring to conditional probabilities included in table B.3 in the appendix, the linkage between HPL and mental illness as well as the association of HPL with MetS remains controversial. Excluding dyspepsia, table B.3 demonstrates no significant difference in prevalence of diagnoses among the two groups high and low. These results are potentially biased by the threshold applied for the binning of HPL levels. Additionally, HPL is assumed to be more prevalent in women and therefore a network controlled for gender might be more appropriate.

4.5.3 Group `M` vs. Group `H`

Characteristics of `M` are typical for the MetS, especially those laboratory values which are not contained by the logical filter for M confirm the definition in table 1.1. The increased age in `M` confirms research results that the number of components of MetS correlates with the years of age [60].

As presumed, the thyroid levels including TSH, TPO-Ab and free triiodothyronine (FT3) of group `H` are distinguishing. Additionally, group H depicts reduced homocysteine levels and standard values referring to laboratory values associated with MetS. On the one hand, these results indicate the healthy lifestyle of group H and on the other hand, it demonstrates the problems of these patients: They are ill, despite their healthy lifestyle, including conscious

nutrition and sports. Due to the lack of information, other characteristics of H as for example serotonin cannot be demonstrated.

HPL levels are increased in groups M and H underpinning the previously proposed theory about HPL, MetS and AITD. However, there is no significant difference in HPL levels of healthy patients. Referring to the previously used threshold of 1.0 nmol/L HPL is also increased in healthy patients and thus, HPL does not seem to be characteristic for either group M or H. These results are possibly influenced by the small number of test results for HPL in healthy patients and group M.

According to guidelines for vitamin D all patients suffer from deficiency in vitamin D which is associated with MetS, AITD and mental illness [61–63]. Table 4.3 shows that the proportion of these measurements is approximately equal referring to periods of six months. Hence, one can exclude seasonal influence as a possible factor for variation in vitamin D levels [64].

The obtained results are limited because healthy people are not part of this study, everybody is ill. A more appropriate approach would include standard values for laboratory values instead of real patients.

4.6 Conclusion

DNs are able to visualize correlations in digital medical records. Here, their power is demonstrated by confirming well known results and thereby, validate the data. Further, the results obtained by the DN including HPL are noteworthy. But, considering test statistics referring to MetS, AITD and mental disorder these results remain controversial. However, this analysis show that 84.9 % of the tested patients have an increased concentration of HPL which dissents prevailing assumptions by Kamsteeg in which it was estimated that approximately 10 % of the population show increased levels of HPL [31]. DNs could be used to study disease evolution of patients by means of their diagnoses at each medical consultation. The selected patients referring to group M and H show the presumed characteristics.

Future work has to investigate whether group H can be well described by laboratory values, besides thyroid levels. The lack of information about laboratory tests associated with mental disorder can overcome by gaining access to medical office's database. Considering DNs, future approaches should evaluate ϕ-correlation which is less affected by the frequency of occurrence due to implicit centering transformations [59]. Alternatively, one could use conditional probabilities yielding more complex and directed networks which allow conjectures about movement of patients in this network.

Chapter 5

Predicting Diseases

5.1 Introduction

While the amount of data has increased considerably, medical data mining including machine learning algorithms has drawn more and more attention over the last two decades [65]. Thereby, one can discover patterns in knowledge-rich data which can be used to identify individuals who are prone to a special disease [66]. Several studies used classification methods such as neural networks, decision trees and support vector machines to predict diseases or phenotypes [67, 68]. Here, decision trees are chosen because of their ability to be highly interpretable, to be a nonparametric classifier, and to deal with noisy and incomplete data [9,10]. Specifically, Quinlan's C5.0 algorithm is used for model building which already showed promising results in prediction of diseases [69].

The aim of this chapter is to demonstrate the predictive power of decision trees, even in everyday medical records. The main focus lies not on finding the best model overall entries stored in the database, but rather on finding patterns of laboratory tests. Because of the lack of information about the time of the diagnose it is crucial to deal with all patients' laboratory values. In this study it is assumed, that a patient's diagnose belongs to its first recorded laboratory test. Note, this laboratory test may not necessarily be the patient's first at this medical office.

5.2 Theoretical Background

In the following, the main steps of C5.0 are explained which consists of construction, pruning and boosting. Further, tree evaluation criteria and resampling techniques are introduced.

Table 5.1. Contingency table showing the outcome of branching.

	Class A	Class B	
$>$ split	n_{11}	n_{12}	n_{+1}
$\leq$ split	n_{21}	n_{22}	n_{+2}
	n_{1+}	n_{2+}	n

5.2.1 Classification Trees

In this study, the focus is on finding solutions for two class problems. For the sake of simplicity following explanations only cover binary trees. A decision tree is a hierarchical model for supervised learning using a divide-conquer strategy to seperate the input data in smaller more homogeneous subgroups [10,40]. Decision trees are composed of attribute nodes linked to two subtrees and leaves labeled with a class that reflects the decision [70]. The most popular decision-tree algorithms are Iterative Dichotomiser 3 (ID3), C4.5, C5.0 [9,71], Classification And Regression Trees (CART) [72] and Chi-square Automatic Interaction Detectors (CHAID) [73]. Among these Quinlan's C5.0, that is the succeeding version of C4.5, showed promising results, even in comparison with neural network approaches [74]. Hence, it is the method of choice. Due to the lack of literature about C5.0, the insides examined in the following are mainly taken from the work of Kuhn et al. who investigated C5.0's source code.

Branching

The above mentioned splitting algorithm has the aim of finding an optimal attribute-threshold pair which branches the sample in two subgroups while maximizing the purity or homogeneity of the resulting subgroups, i.e. each node contains a larger proportion of one class. The splitting criteria is based on Shannon et al.'s information theory [75].

Suppose one has an input dataset including two classes A and B in which the probability of the class A is denoted as p and $q = 1 - p$ denotes the probability of B, respectively. The level of impurity or entropy of the dataset prior to split can be calculated by

$$H(\textit{prior to split}) = -p\log_2(p) - (1-p)\log_2(1-p), \text{where } \log_2(0) := 0. \tag{5.1}$$

Entropy can be described by a flipped parabola in which $H = 1$ (maximal) if $p = q$. The outcome of a possible separation is represented by a 2×2 contingency table (table 5.1) in which the probabilities of classes A and B are displayed in the last row, such that $p = \frac{n_{1+}}{n}$ and $q = \frac{n_{2+}}{n}$.

The entropy after a possible split in correspondence to an attribute-threshold pair is defined as the sum of the weighted average of the information values from each of the resulting

subgroups:

$$H(greater) = -\left[\frac{n_{11}}{n_{+1}} \times log_2\left(\frac{n_{11}}{n_{+1}}\right)\right] - \left[\frac{n_{12}}{n_{+1}} \times log_2\left(\frac{n_{12}}{n_{+1}}\right)\right] \tag{5.2}$$

$$H(less\ or\ equal) = -\left[\frac{n_{21}}{n_{+2}} \times log_2\left(\frac{n_{21}}{n_{+2}}\right)\right] - \left[\frac{n_{22}}{n_{+2}} \times log_2\left(\frac{n_{22}}{n_{+2}}\right)\right] \tag{5.3}$$

Finally, the separation is evaluated by the information gain which is calculated as

$$I(split) = H(prior\ to\ split) - \underbrace{\left[\frac{n_{+1}}{n}H(greater) + \frac{n_{+2}}{n}H(less\ or\ equal)\right]}_{H(after\ split)} \tag{5.4}$$

This greedy splitting process is repeated within each newly created subgroup until the stopping criteria is met (such as minimum pre-specified number of samples in a node).

The explanation above covers only numerical variables. Categorical predictors can be entered into the model as a single entity (grouped categories), or can be decomposed into binary dummy variables (independent categories) [10]. When dealing with missing data information statistics are calculated using non-missing data. The outcome is scaled by the fraction of non-missing data at the split.

Pruning

The greedy approach of tree construction is prone to overfitting. C5.0 uses a heuristical and pessimistic post-pruning approach that eliminates or replaces subtrees to overcome overfitting the data. Here, the pruning algorithm estimates the expected error rate for a set of branches and their parent node. The subtree is pruned if the pessimistic error rate for the parent is smaller than the combined error for a set of branches, which are scaled by weighting according to the proportion of observations along each branch.

The computation of pessimistic error rate depends on the training data involving a pre-specified confidence interval c. Supposing the number of instances at node V is denoted as N, and E depicts the number of errors at V. Now imagine that the true probability of error at V is q, and that the N instances at V are generated by a Bernoulli process with parameter q, of which E turns out to be errors. For large N the distribution of this random variable converges to the normal distribution and the confidence value z is obtained by

$$Pr\left[\frac{f-q}{\sqrt{q(1-q)/N}} > z\right] = c, \tag{5.5}$$

where $f = E/N$ is the observed error rate. [76] Further, the upper confidence limit is used

as the pessimistic error rate e that is estimated as

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}. \tag{5.6}$$

Boosting

Based on the work of Freund and Shapire, boosting is an ensemble methodology for generating and combining multiple classifiers to improve predictive accuracy. According to Kuhn et al., C5.0 uses an approach similar to the AdaBoost algorithm [77]. Models are fit sequentially and each iteration adjusts the case weights according to the accuracy of a sample's prediction. The following section depicts some essential differences or extensions in comparison to AdaBoost. C5.0 attempts to generate an ensemble of trees in which each model has approximately the same number of terminal nodes as the initial tree. Additionally, the stopping criteria for boosting has been modified such that boosting will automatically stop if either the model fits the data well or the model is highly ineffective. Furthermore, C5.0's boosting algorithm differs in the adjustment of the case weights w_K where $K \in \{1 \dots k \dots K\}$ denotes the number of boosting iterations. At first, one has to compute the midpoint by

$$\frac{1}{2}\left[\frac{1}{2}(S_- + S_+) - S_-\right] = \frac{1}{4}(S_+ - S_-), \tag{5.7}$$

where S_- represents the sum of weights for incorrectly classified samples and S_+ denotes the sum of weights for correctly classified samples, respectively. The *midpoint* is necessary to generate the weights for the correctly classfied samples:

$$w_k = w_{k-1}\frac{S_+ - midpoint}{S_+}. \tag{5.8}$$

Misclassified samples are adjusted with

$$w_k = w_{k-1} + \frac{midpoint}{N_-}, \tag{5.9}$$

where N_- denotes the number of incorrectly classified samples. The applied weighting scheme highly increases the weight if a sample is incorrectly predicted. Whereas, due to the multiplicative nature the weights slowly decrease if a sample is correctly classified.

5.2.2 Tree Evaluation Criteria

In this part, evaluation criteria according to two class problems are introduced. The evaluation of classification models is based on the cross-tabulation of the observed and predicted classes for the data, also called confusion matrix (table 5.2), comprising of the number of

Table 5.2. General contingency table visualizing the performance of classification.

		Observed Condition	
Predicted Condition		TP	FP
		FN	TN

true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). [10]

Accuracy is one of the simplest metrics which is computed as

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{5.10}$$

and refers to how consistent the predicted values are with the observed ones. Other metrics such as sensitivity and specificity measure the accuracy in the event population, or non-event population, respectively [10]. Sensitivity and specificity are obtained by

$$sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.11}$$

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{5.12}$$

5.2.3 Resampling Techniques

The desired strategy of generating and testing classification models includes the separation of the data set into training and test data. Unfortunately, this approach is possibly not feasible due to small sample size. To avoid overfitting while training the model with the total data set several resampling approaches such as cross-validation (CV) and bootstrapping are available. The following paragraphs introduce k-fold CV and repeated k-fold CV.

k-Fold Cross-Validation

As in all CV techniques, the aim is to estimate the model performance on independent data. Therefore, the complete data set is randomly split into k mutually exclusive folds of approximately equal size. Then, the model is trained by all but one fold $(k-1)$ and its performance is achieved by testing with the remaining single fold. The latter step is repeated k times, such that the model is tested on each of the k folds. Finally, the estimated performance denoted as CVP is obtained by averaging all k individual performance measures

P_i such as

$$CVP = \frac{1}{k}\sum_{i=1}^{k} P_i. \tag{5.13}$$

Empirical studies showed that $k = 5$ yields appropriate results. As k increases the difference between the estimated and true values of performance (bias) decreases. [10]

Repeated k-Fold Cross-Validation

As the name implies, this technique is characterized by applying k-fold CV n times. The estimated performance is computed as

$$RCVP = \frac{1}{n}\sum_{i=1}^{n} CVP_i, \tag{5.14}$$

where CVP_i denotes the estimate obtained by a single k-fold CV. Therefore, repeated k-fold CV reduces the variation of regular k-fold CV's performance estimates [78]. Hence, it can be used to effectively increase the precision of the estimates. [10]

5.3 Methodology

5.3.1 Model Building and Evaluation

All tree models in this analysis are computed with the `R`-packages `C50` [79] and `caret` [80] by using input data comprising gender and patients' first laboratory tests. The sparseness of the resulting data table is reduced by iterative deletion of missing values as described in section 3.1.1.

The ratio of the number of patients suffering from a specific disease (positive observation) and the amount of patients without this disease (negative observation) is unbalanced and can lead to an optimistic accuracy estimate. For instance, in the case of osteoporosis a model would reach an accuracy of $\frac{1565-38}{1565} = 97.6\%$ only by predicting all patients as healthy considering osteoporosis. A data table containing the same number of positive and negative observations eliminates this effect. This balanced data set is the input for model building procedure. Hence, one assumes no a priori knowledge about frequency of diseases' occurrence. Note, because some diseases as osteoporosis and asthma occur rarely, the data set is not divided in test and training data which is critical corresponding to overfitting. Here, 10 times repeated 5-fold CV is used to avoid overestimation of the model's accuracy.

As a result of the sampling approach that yields the balanced data set, the model and its accuracy do not represent the whole data. Therefore, taking more data into account one can repeat the resampling and model building process any number of times, here 20 times.

Thereby, one can investigate the influence of the selected training data on the model's accuracy. Finally, this procedure results in 20 ensembles for each disease. For further analysis, the simplest model with accuracy not less than the best one by one standard deviation is chosen. A simpler tree would be expected to capture the structure inherent in the problem more likely [9].

5.3.2 Prediction

The best models referring to each disease are used for prediction using patients' second laboratory tests as input data. Due to the problem of missing values, each model is tested on a different group of patients. A patient is part of the data set only if an arbitrarily chosen proportion of important decision trees' attributes is known. An attributes is important if it is present in the majority of trees contained by the model.

5.4 Results

The estimated accuracy for all tree ensembles is depicted in figure 5.1. It points out that thyroid disease and MetS and its components are more correctly predicted than mental disorders and cancer. The accuracy referring to models covering *other* varies strongly across the different samples. The most robust and correct models concerning thyroid diseases are obtained for AITD and HT. On average, both ensembles reach an accuracy of over 70 %. Concerning the median, accuracy of models dealing with MetS and insulin resistance is even better but the variance during resampling is higher.

The ensemble predicting HT is shown in figure 5.2. In each single tree the root node is presented by TPO-Abs. Only the second trial yields an structurally different tree consisting of the three attributes TPO-Ab, sex and potassium. The first and last one produces exactly the same tree.

The evaluation of models for thyroid diseases and MetS with secondly measured laboratory tests is demonstrated in table 5.3. Each model is tested with a different data set. In addition to the evaluation results, table 5.3 specifies the laboratory parameters of which at least 75 % have to be measured for the patients. For instance, the data set used for testing comprises only patients with known values for TSH and TPO-Ab. Models predicting AITD, HT and diabetes mellitus achieved the best results. The predictions are even more accurate than the estimates obtained by CV. Whereas, the predictive power of models detecting insulin resistance and MetS is weaker than the median of CV estimates for accuracy.

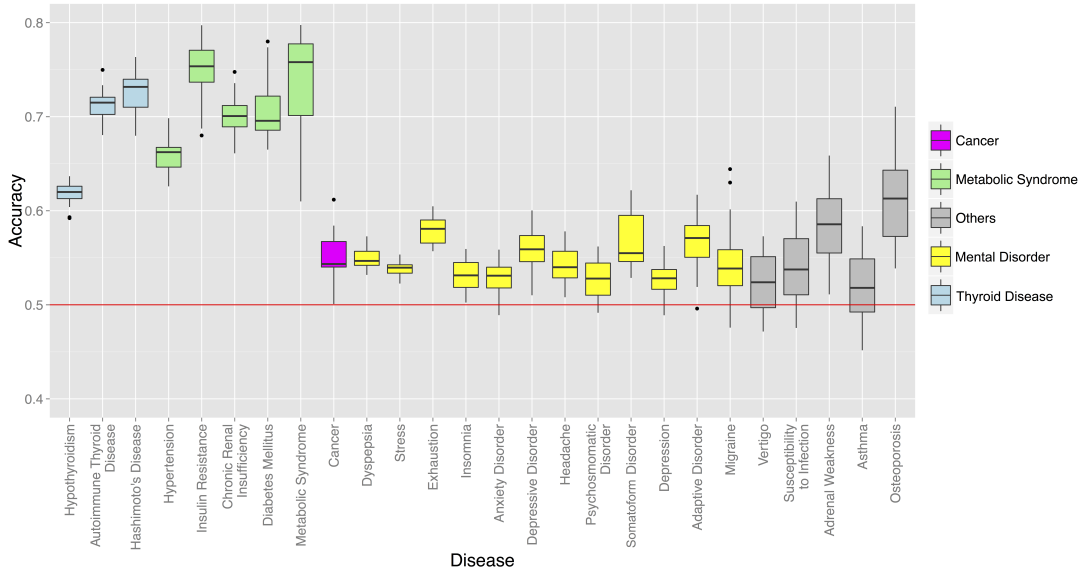

Figure 5.1. Accuracy range estimated by repeated k-fold cross-validation.

Table 5.3. Evaluation of disease prediction concerning secondly measured laboratory tests.

Disease	Proportion	Sensitivity	Specificity	Accuracy	Filter
AITD	12/41	83.3 %	96.6 %	92.7 %	TSH, TPO-Ab
Hashimoto's Disease	16/80	75.0 %	90.6 %	87.5 %	TPO-Ab
Hypothyroidism	80/179	62.5 %	59.6 %	60.9 %	TSH, free T3, freeT4
MetS	11/204	100 %	64.3 %	66.2 %	HbA1c
Diabetes Mellitus	24/204	83.3 %	94.4 %	93.1 %	HbA1c
Hypertension	26/86	73.1 %	76.6 %	61.6 %	HbA1c, cholesterol, CRP quantitative,
Insulin Resistance	1/19	100 %	66.6 %	68.4 %	coenzyme Q10, HDL-C
Chronic Renal Insufficiency	9/79	100 %	62.9 %	67.1 %	creatinine, homocysteine

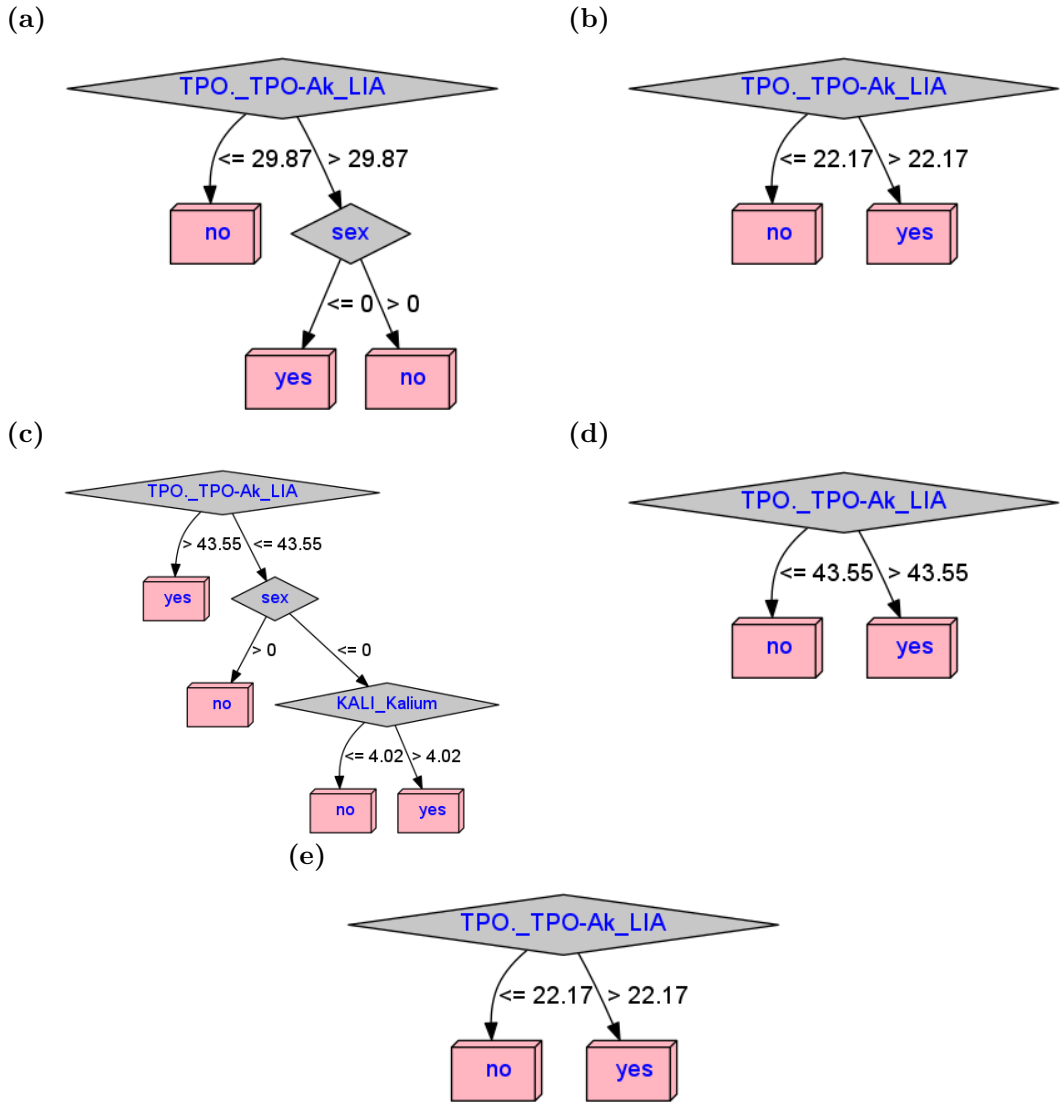

Figure 5.2. Decision tree ensemble predicting Hashimoto's disease. The single tree referring to each of the five trials is shown in **(a)**-**(e)**.

5.5 Discusion

One has to keep in mind that all diagnoses contained by the database are collected since a patient's first medical consultation but, laboratory tests are recorded not before the fourth quarter of 2008. Prediction is even more difficult because these diagnoses include also patients' medical history. So, it remains uncertain whether the first laboratory test represents a patient's diagnoses. In some cases a patient is healed or rather drug-treated before the first laboratory test contained in the database. These shortcomings can cause the moderate results of the models. In addition, the models are build without information about the age of the patients. This attribute was excluded because it covers laboratory values which are

the focal point of this analysis. This restriction affected especially the accuracy of the model generated to predict osteoporosis.

The higher variation in estimated accuracy of models predicting osteoporosis can be explained by the smaller number of diseases' occurrences. Small frequencies yield a greater variation in training data due to randomized selection of patients without disease of interest. Computing more resampling steps would reduce this variation.

Among thyroid diseases, the prediction of AITD and HT is more accurate than prediction of hypothyroidism. A reason could be that patients are already treated with thyroid hormones and selenium before their first laboratory test. Thus, thyroid hormone levels of patients suffering from hypothyroidism are similar to those from patients without hypothyroidism. Whereas, the concentration of TPO-Ab is moderately reduced over time [81, 82]. Hence, the prediction of AITD and HT is not affected as the prediction of hypothyroidism. AITD and HT are characterized by increased levels of TPO-Ab, TG, TSH-R and TSH. But only TPO-Ab and thyroid hormones are used in the ensembles predicting HT (figure 5.2) and AITD (figure C.1 in the appendix). Among these characteristic antibodies only TPO-Ab are measured often enough during first blood sample as demonstrated in table C.1 in the appendix. In this study the group of patients suffering from AITD mainly consists of HT and thus thyroid hormones and TPO-Ab are appropriate for prediction of AITD.

Excluding diabetes mellitus, models predicting MetS and its components show weak results for forecasting diagnoses referring to patients' second laboratory tests. This can have different causes. Firstly, models are not generated separately for women and men which is relevant in predicting diseases as MetS. They are correlated with gender specific values as e.g. body mass index (BMI), HDL-C or triglycerides. For instance, the wilcox test for HDL-C levels of patients suffering from MetS and patients that are not affected by MetS does not show significant differences in the mean referring to all patients. But tested on women and men separately, it yields significance. A similar effect happens in predicting chronic renal insufficiency using creatinine levels. Creatinine is a breakdown product of creatine phosphate in muscles [83]. Hence, the more muscular a person is, the more creatinine is present in the body [84]. Thus, standard values for creatinine are dependent on physical constitution as muscle mass. Increased creatinine levels indicate an impaired kidney function. Due to the lack of information concerning muscle mass in patients, creatinine levels are not scaled by that criteria. Sensitivity of 100 % and specificity of 62.9 % indicate that the threshold for creatinine tends to be underrated. Additionally, some characteristic parameters are measured rarely in the training set and are therefore excluded, as e.g. homeostatic model assessment index (HOMA-index) that is measured in 6 % concerning patients without insulin resistance.

The weak predictive power concerning mental disorders and cancer can be explained by the lack of laboratory values indicating such diseases as for example serotonin and tumor markers. In addition, in this study cancer is a generic term including different types as e.g.

breast cancer, uterus cancer, or prostate cancer which makes prediction even more difficult.

5.6 Conclusion

In this analysis, the power of C5.0 predicting diagnoses by means of everyday medical records could be partly demonstrated by predicting AITD, HT and diabetes mellitus. Additionally, the difficulties in predicting diseases resulting from sparse and heterogeneous data are demonstrated. The results would be even more powerful if one could obtain access to the database of the medical office software and therewith information about treatment and time of diagnoses. As a consequence of having this knowledge, one is able to exclude those patients from training data that are already treated. The scope of application includes prediction and thus prevention by reduced parameters as for instance in the discussed model for renal insufficiency. Furthermore, such models have the ability to generate patterns for diseases which can be used for composition of laboratory tests.

Besides the mentioned access on information about treatment, future work includes investigating already digitally provided laboratory values of another laboratory. After this improvement, referring to the quality of the data, one could try to evaluate different machine learning algorithms such as random forrest.

Chapter 6

Controlling the Course of Treatment

6.1 Introduction

This chapter introduces the STATIS approach into the analysis and interpretation of everyday medical records over five years. STATIS has already been successfully used in bioinformatics analyzing transcriptomic time-series data [85], dynamics of metabolomic processes [86], and biomedical data [87]. The idea of STATIS is to compare different data tables containing the same number of rows and/or columns on the basis of an optimum weighted consensus which captures what is common to all or a subset of analyzed tables [85, 88]. The compromise is obtained based on PCA of a specially constructed matrix. In addition, Euclidean distances are computed between laboratory values of a group of patients and an artificially modeled patient. The combination of these two approaches facilitates further investigations of individual patients, but also groups of patients.

One of the main goals of this part is to investigate whether one can find characteristic patterns for diseases. Additionally, it is analyzed if STATIS can be used as a detection tool for at-risk patients.

6.2 Theoretical Background

The following section introduces a method for simultaneously analysis of multiple data sets, called STATIS. Notations and explanations are mainly taken from works of Stanimirova et al. and Abdi et al. [88, 89].

6.2.1 STATIS

STATIS was developed by Hermier des Plantes and Thiébaut [90, 91] and is used for the analysis of a three-dimensional data set $\widetilde{X}$(figure 6.1a). $\widetilde{X}$ not necessarily consists of the

same sets of variables J collected on the same number of observations I measured at K different states, like times or locations. A single element of $\widetilde{X}$ is denoted by x_{ijk} where $i \in \{1, 2, \ldots, I\}$, $j \in \{1, 2, \ldots, J\}$ and $k \in \{1, 2, \ldots, K\}$. $\widetilde{X}$ can be divided in K data tables X_k, which are also called frontal slices, with dimensions $I \times J_k$ (figure 6.1b). For the sake of completeness, one can also decompose $\widetilde{X}$ in I horizontal slices with K observations. But the decomposition which is obtained by fixing J results in vertical slices, violates the required condition of dealing with identical observations in all tables. The concatenation of all frontal slices results in a data table with dimensions $I \times \sum_{k=1}^{K} J_k$ and is denoted as X.

STATIS can be divided in two main steps; (1) analysis of the inter-structure which provides an optimal weighting of the variance of individual tables and (2) analysis of the intra-structure which performs a generalized PCA of $\widetilde{X}$ that uses the calculated weights as constraints on the tables and their variables. A summary of the described procedure of STATIS is shown in figure 6.1c

Analysis of the Inter-Structure

The similarities between I observations of a frontal slice X_k is represented by the variance-covariance matrix S_k which is calculated as

$$S_k = X_k X_k^T, \tag{6.1}$$

where X_k is mean-adjusted. The linear closeness between two variance-covariance matrices S_k and $S_{k'}$ can be specified by the RV-coefficient [92], a commonly used measure that is obtained as

$$RV(S_k, S_{k'}) = \frac{\langle S_k, S_{k'} \rangle}{\sqrt{\langle S_k, S_k \rangle \langle S_{k'}, S_{k'} \rangle}}, \tag{6.2}$$

where $\langle S_k, S_{k'} \rangle$ denotes the inner product of S_k and $S_{k'}$ which can be rewritten as

$$\langle S_k, S_{k'} \rangle = \sum_{i=1}^{I} \sum_{j=1}^{I} s_{i,j,k} s_{i,j,k'}. \tag{6.3}$$

Thus, normalization of the matrices $S[k]$ and $S[k']$, such that the sum of squares of their elements is equal to one, leads to the following simplified formula for RV-coefficient

$$RV(S_k, S_{k'}) = \langle S_k, S_{k'} \rangle. \tag{6.4}$$

The RV-coefficients are non-negative, scaled between 0 and 1, and are stored in a square matrix matrix $C(K \times K)$. The greater the RV-coefficient, the more similar the two variance-covariance matrices $S[k]$ and $S[k']$ are. C is symmetric and positive semidefinite ([89] provide more details) which indicates that all eigenvalues are nonnegative and its eigenvectors are

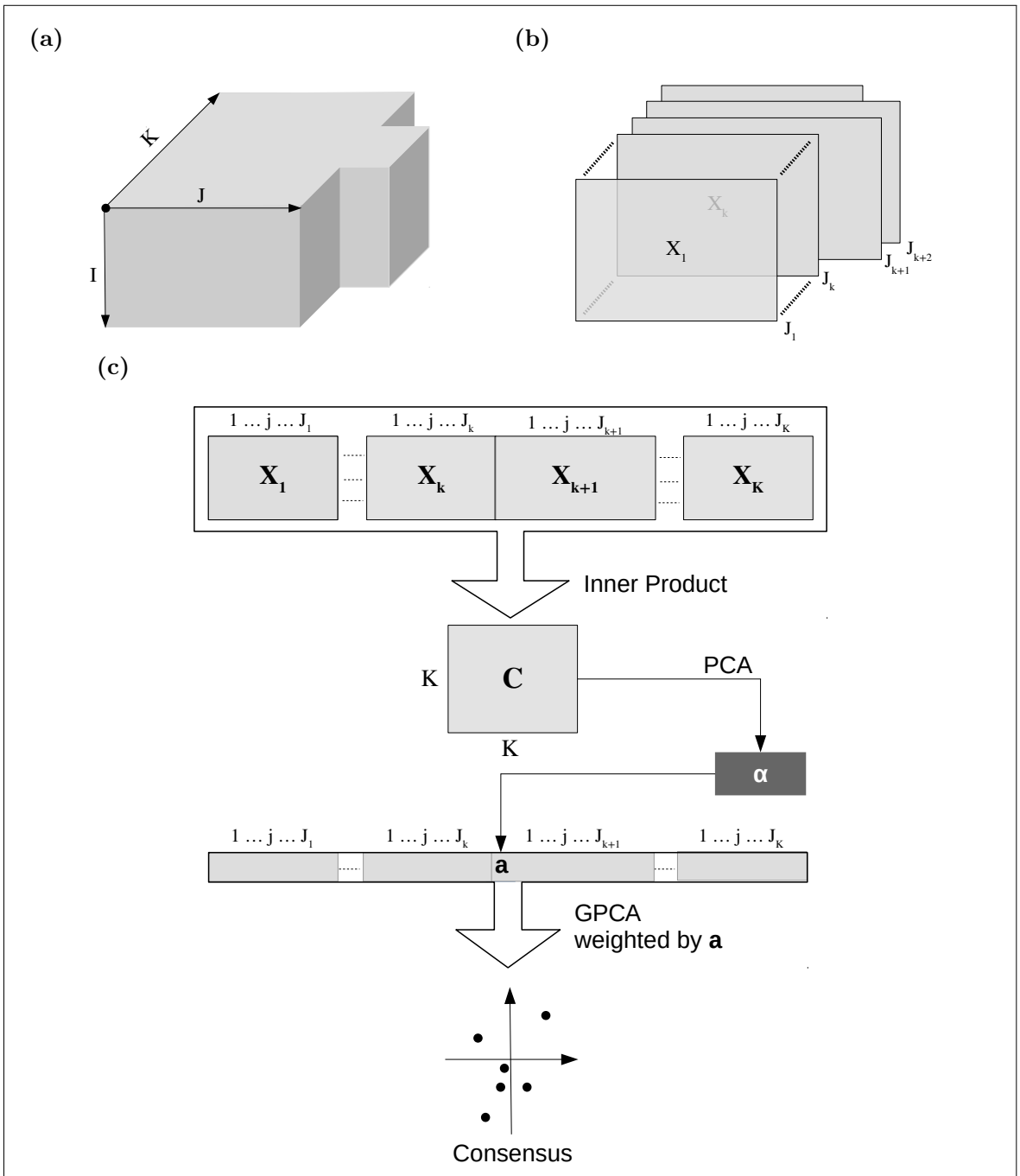

Figure 6.1. Sketch of input and steps of STATIS. **(a)** A three-way data array **(b)** Decomposition of the three-way data array into K frontal slices **(c)** Schematic overview about the procedure of STATIS

real and orthogonal to each other. Hence, C can be eigen-decomposed as

$$C = U\Theta U^T \quad \text{with} \quad U^T U = \mathbb{1}, \tag{6.5}$$

where $\mathbb{1}$ represents the identity matrix and Θ is the diagonal matrix whose diagonal comprises the PCA values. This yield the PCA of the similarity structure between the frontal slices X_k. The resulting first eigenvector denoted as u_1 represents the overall similarity to all the other tables. Thus, normalizing u_1, such that the sum of all elements is equal to one, lead to the optimal weighting for the tables. This optimal weighting is stored in a column vector of length K which is denoted as α and reduces the influence of atypical tables. For later purposes it is appropriate to gather the α weights in a column vector denoted as a by

$$a = \left[\alpha_1 \mathbf{1}_1^T, \alpha_2 \mathbf{1}_2^T, \ldots, \alpha_K \mathbf{1}_K^T\right], \tag{6.6}$$

where a is of length $\sum_{k=1}^{K} J_k$ and $\mathbf{1}$ is a column vector.

Analysis of the Intra-Structure

The obtained weights are used to compute a weighted version of GSVD of $\widetilde{X}$ under the constraints provided by the masses for the observations and the optimum weights for the K tables, which is expressed as

$$X = P\Delta Q^T \qquad \text{with} \qquad P^T M P = Q^T A Q = \mathbb{1}, \tag{6.7}$$

where the diagonal of Δ comprises the singular values of X, $A = diag\{a\}$ and M represents the masses which are often equal for all observations such that

$$M = diag\left\{\frac{1}{I} \times \mathbf{1}\right\}. \tag{6.8}$$

This GSVD corresponds to GPCA. Therefore, it provides factor scores F to describe the observations, and factor loadings Q to describe the variables. Because of the block structure of X one can express Q as a column block

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_K \end{bmatrix}, \tag{6.9}$$

where Q_k is of dimensions $J_k \times L$, where L denotes the rank of X and J_k reflects the number of variables (columns) of the frontal slice X_k. Q_k contains the right singular values

corresponding to the variables of matrix X_k. The factor scores reflect the best consensus for the set of the K slices and can be expressed as

$$F = P\Delta. \tag{6.10}$$

Combining equations 6.7 and 6.10 yield to

$$F = \overbrace{P\Delta \underbrace{Q^T AQ}_{=\mathbb{1}}}^{=X} = XAQ \tag{6.11}$$

which can be rewritten as

$$F = XAQ = [X_1|X_2|\cdots|X_K] \times A \times \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_K \end{bmatrix} \tag{6.12}$$

$$= \sum_{k=1}^{K} X_k \alpha_k Q_k \tag{6.13}$$

under consideration of the block structure of X,A and Q. Equation 6.12 indicates that partial factor scores F_k can be defined as the projection of this table onto its right singular vectors and computed as

$$F_k = X_k Q_k. \tag{6.14}$$

The consensus factor scores are the weighted average of the partial factor scores:

$$F = \sum_{k=1}^{K} \alpha_k F_k, \tag{6.15}$$

i.e. they represent the centroid.

6.3 Methodology

To draw conclusions about the course of treatment based on laboratory results requires the observation of a group of patients over time. Therefore, all patients that received laboratory tests in every consecutive year between 2010 and 2014 are analyzed. In best case, all patients can be described by the same set of variables, however the composition of laboratory tests varies heavily in reality. Hence, a set of 15 variables (see table 6.1) is selected which corresponds to the MetS and AITD. In case of multiple laboratory tests of one patient, the laboratory analysis including the most of the selected parameters is chosen. Consequently,

Table 6.1. Parameters of interest with ideal values accordingly to the physician.

Parameter	Value	Parameter	Value
TSH basal	0.8 μIU/mL	Vitamin D	60 ng/mL
free T3	3.5 pg/mL	HbA1c	5.2 %
free T4	1.4 ng/dL	Gamma-GT	14 U/L
Kreatinin	0.7 g/L	GPT (ALAT)	12 U/L
Potassium	4.3 mmol/L	Cholesterol	180 mg/dL
Calcium	2.42 mmol/L	LDL-C	100 mg/dL
total protein	7.4 g/dL	HDL-C	70 mg/dL
Homocysteine	5.0 μmol/L		

each patient can be described by five laboratory tests. These laboratory test results are stored in five data tables with dimensions $i \times 15$, where i denotes the number of patients and each table contains the laboratory results of one year. Missing values are reduced by iterative deletion (see section 3.1.1) of columns containing more than 45% missing data and rows including more than 50% missing data, respectively. Thus, it is possible that each of the five data tables contains a different number of variables. The remaining missing values are imputed by a modified KNN approach presented in section 3.1.2. This approach yields better estimates than the standard KNN algorithm in a testing procedure including nine parameters of 17 patients as shown in figure D.1 in the appendix. To enable comparison, the three-way data array is extended by an artificial patient with ideal laboratory values, as shown in table 6.1.

The differences and similarities of patients over time are investigated by STATIS. Conclusions about the course of treatment are allowed by computation of Euclidean distances (see, equation 3.1) between the Z-scores of laboratory results of the artificial modeled patient and each of the real patients for each year. Furthermore, changes in biomolecular status of the inspected patients are demonstrated by comparing first and last measurements of laboratory values affected by orthomolecular medicine such as vitamins, hormones, minerals, trace elements.

6.4 Results

This analysis is based on 20 patients who have visited the medical office at least one time a year between 2010 and 2014 for laboratory tests. These patients possibly suffer from multiple diseases. The frequency of diseases' occurrence is contained in table B.2 in the appendix. Originally, these patients' laboratory values contain on average 17.5 % missing values (table 6.2).

Table 6.2. Proportions of missing values for each data table.

Year	2010	2011	2012	2013	2014
Missing (%)	13.8	14.2	23.1	18.5	18.3

The bi-plot in figure 6.2a shows the factor scores (patients) and loadings (laboratory tests) for the first two components of the five tables' consensus. The loadings are averaged according to the weights for each table and re-scaled to have an equal variance to the singular values of the compromise analysis, as it is illustrated by Greenacre [93]. Note, the loadings include only laboratory tests that are present in each table. One can see that patients with and without hypertension are separated by the first component which explains 23 % of the inertia. The first component mainly includes vitamin D, creatinine, HbA1c and homocysteine.

Figure 6.2b demonstrates Euclidean distances between each patient and the artificially modeled patient per year. It points out that the Euclidean distance is reduced in 80 % of the observed patients after five years of treatment. In 2010 the Euclidean distance is on average 5.3 whereas it drops to 3.9 in 2014. Considering 2014, patients suffering from hypertension show an increased distance compared to patients without hypertension. These courses of treatment are obtained by applying orthomolecular medicine. Figure 6.3 depicts changes in the biomolecular status concerning the first and last measurements of laboratory values referring to the selected 20 patients. The investigated laboratory values are not imputed and thus, the number of measurements can vary between the categories *first* and *last*. Note, *first* and *last* include the first respectively last measurement of each considered laboratory value, not laboratory test.

The partial factor scores along with the re-scaled loadings are displayed in figure 6.4. Each figure represents the patient's condition each year. The number of laboratory tests differs from one patient to another, but eight of the contained laboratory tests are present each year. By means of these plots one can display the course of treatment of each patient. For instance, patient 155 shows varying distances to the ideal patient, i.e. in 2011 and 2013 the distance is larger than in 2010, 2012 and 2014. This observation is conformed by Euclidean distances in figure 6.2b.

6.5 Discussion

Hypertension is part of MetS but also induced by hypothyroidism and age. Patients with hypertension are, inter alia, characterized by increased homocysteine levels. This observation confirms recent studies which demonstrate the correlation of homocysteine with CVD [94]. Patients 18 and 101 who are spatially close to each other suffer from MetS. Both show increased levels in HbA1c, creatinine, homocysteine as well as increased liver values. Their

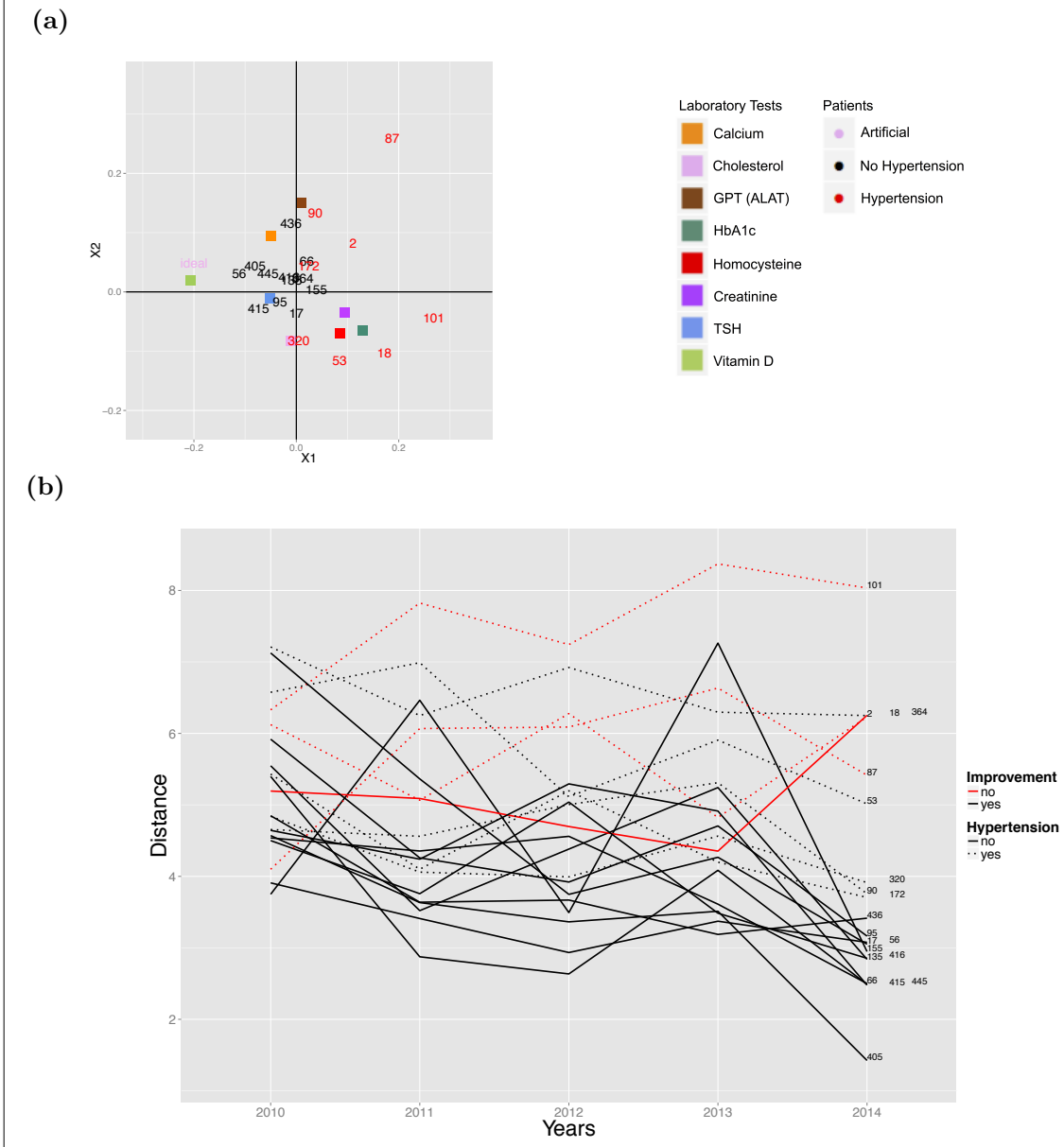

Figure 6.2. Multivariate analysis with STATIS and course of treatment. **(a)** Bi-plot depicting factor scores and loadings (only variables that are present at each time point are represented) of consensus representing 5 data tables between 2010 and 2014. Patients are represented by their colored identification number. ($\lambda_1 = 0.011, \tau_1 = 23\%, \lambda_2 = 0.007, \tau_2 = 14\%$) **(b)** Euclidean distance between each of the 20 patients and the artificially modeled patient which comprises ideal laboratory values. Patients are represented by their identification number.

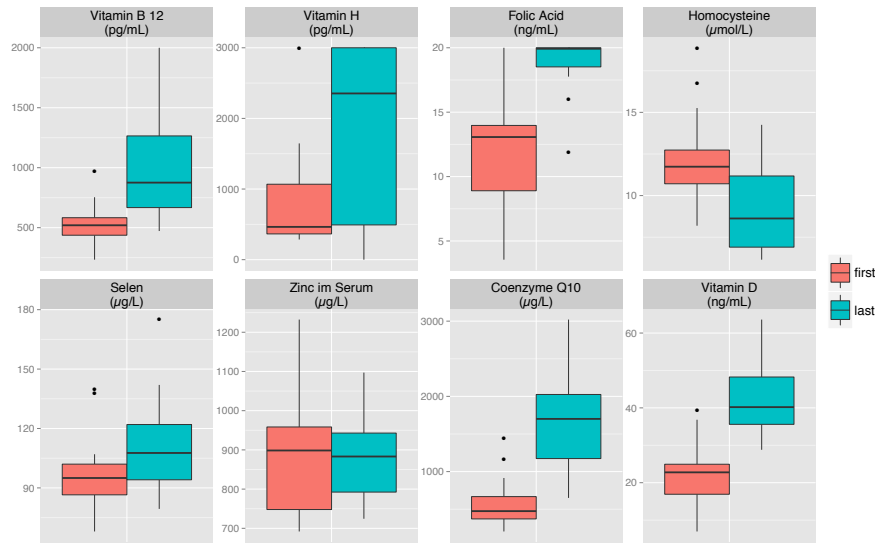

Figure 6.3. Comparison of first and last measurements of available laboratory tests. The boxplots represent the real data and thus, the number of measurements can vary between the categories *first* and *last*

closest neighbor is patient 53 who fulfills two criteria of MetS which is the most except for patients 18 and 101. Among the more distant neighbors 320, 87, 155 and 2 one can find one criterion of MetS whereas the remaining 13 patients fulfill at most one criterion. Nevertheless, the results of STATIS cannot be interpreted as a metric distance measure. For instance, patient 364 appears normal but STATIS points out that the patient's cholesterol level and liver values are highly increased. Due to opposing loadings the contributions of these values are nullified and patient 364 remains in the center.

However, STATIS' results yield a tendency. All patients who are located far away from the ideal patient in the consensus in figure 6.2a are seriously ill. Patients 18, 101, 87, 2 and 53 combine for 80 % chronic renal insufficiency, 50 % cancer and 100 % MetS, insulin resistance and diabetes mellitus. Patients in the near of the artificially modeled one are afflicted by hypothyroidism and psychosomatic disorders. Both, STATIS and the Euclidean distance demonstrate an improvement of patients laboratory values but even after five years non of the patients is located at the "ideal" point. This is a result from taking ambitious values into account, especially an homocysteine level of 5 μmol/L is hard to achieve. A more appropriate approach would include a collection of artificially modeled patients with satisfying values that span an "ideal" area in STATIS, e.g. include artificially modeled patients with homocysteine levels between 5 μmol/L and 7 μmol/L. Furthermore, patient 155 is a prime example of monitoring a patient's development with STATIS. In figure 6.4 the partial factor scores show an alternating distance to the ideal patient. The computed Euclidean distances referring

to patient 155 in figure 6.2b confirm this observation. After consultating the physician it turned out that patient 155 did not follow the therapy in 2011 and 2013.

The general improvement of laboratory values (figure 6.2b) is achieved by orthomolecular medicine. Laboratory values which are affected during therapy and contained in the database are depicted in figure 6.3. Except for homocysteine all of these values are given as nutritional supplements. Homocysteine is an indicator for CVD and is shown for demonstrating the success of this therapy. It is known that homocysteine levels are reduced by treating deficiencies in vitamin B12, vitamin H and folic acid [95]. The observation that laboratory values of patients suffering from hypertension have not improved as much as those of patients without hypertension, after being treated for five years can have several reasons. Firstly, as mentioned in chapter 4, it may occur because of the lack of laboratory values referring to psychosomatic disorders. Thus, the Euclidean distances of patients affected by psychosomatic disorders are underrated. Secondly, one could argue that it is more difficult to obtain an comparable improvement in these eight patients because they are on average 22.5 years older and more affected by chronic renal disease, cancer and MetS. In my opinion the second assumption is more convincing. But as long as there are not enough data available concerning mitochondrial function, neurotransmitters and hormones of adrenal gland the first one cannot be excluded.

A shortcoming of the applied methodology arises from creatinine levels as already mentioned on page 5.5. As a result creatinine levels of smaller people tend to be underrated. For instance, patient 364 who is female, 160 cm tall and weighs 56 kg suffering from chronic renal insufficiency. Neither the consensus nor the partial factor scores indicate any increased creatinine level although, her creatinine level is 1.05 g/L on average between 2010 and 2014. The differences in the measured creatinine levels of patient 364 and the mean values of creatinine in each table are relatively small. Hence, patient 364 is not located in the area of increased creatinine values. Analogously, the Euclidean distance is affected by unscaled creatinine levels.

6.6 Conclusion

The results presented in this chapter demonstrate that STATIS is able to monitor the course of treatment if one considers distance measures such as the Euclidean distance, as well. Furthermore, by means of STATIS one can obtain areas which are predestinated for specific diseases as MetS and hence, it can be used to identify potential at-risk patients. However, due to the discussed effect of nullification this scope of application has to be treated with caution. Additionally, based on the loadings one can detect disease characteristics (patterns).

Future work should deal with scaling of laboratory values as creatinine referring to height, weight and gender. This modification improves the reliability and interpretability of the

applied methods. Moreover, one could define the desired or ideal intervals of each laboratory value. Following a generation of n ideal patients in which each laboratory value is an element of the predefined interval one obtains an ideal space which is more realistic than only one point. To reduce the effect of these artificially modeled patients on the methodology of STATIS one could compute the compromise without them. Afterwards, the loadings can be used to transform the modeled patients into the two-dimensional space. To increase the explained inertia one could visualize STATIS' results by means of a three dimensional plot.

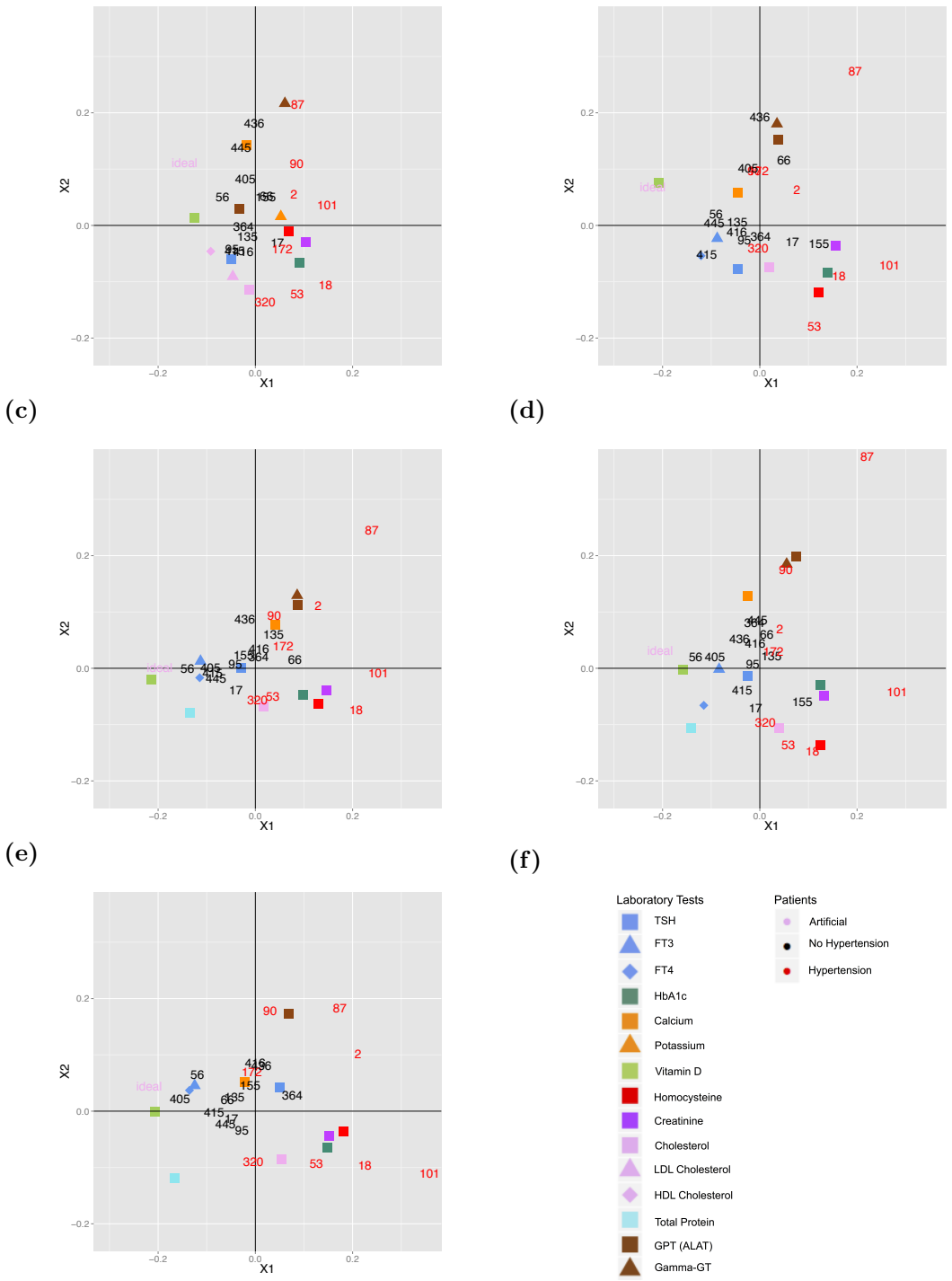

Figure 6.4. The partial factor scores and variable loadings for the first two dimensions of the compromise space. The loadings are rescaled to have the same variance as the singular values of the consensus analysis. Patients are represented by their identification number. **(a)-(e)** Contributions of the observations from 2010 to 2014. **(f)** The legend corresponding to subfigures (a)-(e).

Chapter 7

Conclusion and Future Work

In this work, data mining methods were applied on anonymized everyday data of a physician's office covering internal holistic medicine in order to enable research, prevention and monitoring of patients' course of treatment. Thereby, it was demonstrated that these often overlooked data are challenging due to the amount of missing data. At the same time analyzing these data is a great chance to do research on disease profiles of real patients. Thereby, one can hopefully improve healthcare. Additionally, turning more attention to everyday medical records can may be used to support physicians in their practical work.

Due to the lack of cooperation by the provider of the medical practice software a lot of actually existing records and information about the treatment are not available in this study. Nevertheless, the presented results considering comorbidity and the prediction of HT give an impression of the ability to find hidden relations and patterns. Furthermore, investigations in chapter 4 underpinned well-known criteria of MetS. This illustrates the power and validity of the underlying data. Assumptions for HPL are partially confirmed in this study. Moreover, a rather new method, called STATIS, was introduced and showed the ability to visualize dynamics in patients' biomolecular status and along with it the positive effect of combining internal and orthomolecular medicine.

The results presented here can be seen as a prototype that has to be further developed. Future work referring to this medical office is only promising if the provider of the medical office's software makes the database available for access. This yields information about treatment which is required to draw meaningful conclusions. Additionally, pre-processing has to be extended by scaling parameters which depend on other factors to enable more comparability. Moreover, one has to reduce the amount of missing data which can be achieved on two ways. Firstly, patients receive more laboratory tests leading to increased costs. Alternatively, one tries to improve imputing procedures in combination with previously and upcoming measurements of laboratory values to reduce sparseness of the data set. So, there is perhaps the possibility of virtually reducing the sparseness but preserving it in reality with

nearly no loss of information. Solving the problem of sparsity would improve results of this study and moreover, it permits the application of further powerful algorithms as biclustering or multiple regression.

Appendices

Appendix A

Metabolic Syndrom

Table A.1. Definition of the MetS for europeans.

Central obesity		
- Waist circumference		
	men	$\geq$ 94 cm
	women	$\geq$ 80 cm
Plus any two:		
- Raised triglyceride		$>$ 150 mg/dL
- Specific treatment for this lipid abnormality		
- Reduced HDL-C		
	men	$<$ 40 mg/dL
	women	$<$ 50 mg/dL
- Specific treatment for this lipid abnormality		
- Raised blood pressure		
	Systolic	$\geq$ 130 mm Hg
	Diastolic	$\geq$ 85 mm Hg
- Treatment of previously diagnosed hypertension		
- Raised fasting plasma glucose		$\geq 5.6 mmol/L$
- Previously diagnosed type 2 diabetes		

Appendix B

Disease Network

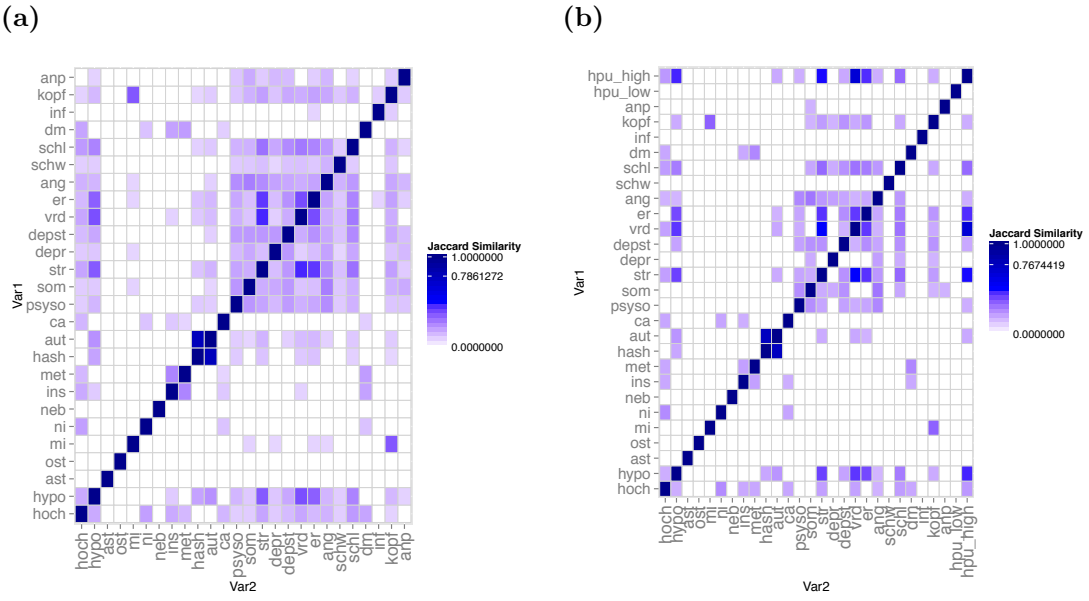

Figure B.1. Heatmaps according to disease networks. These maps show the Jaccard similarity index between diseases using (a) all patients or (b) only patients with measured HPL levels.

Table B.1. Absolute number of diseases' appearances overall patients and for women and men, separately.

Disease/Disorder	Shortcut	Total	Women	Men
Dyspepsia	ver	720	513	207
Stress	str	575	430	145
Exhaustion	er	463	377	86
Hypothyroidism	hypo	457	383	74
Insomnia	schl	346	266	80
Hypertension	hoch	283	156	127
Anxiety Disorder	ang	226	169	57
Depressive Disorder	depst	217	174	43
Headache	kopf	192	149	43
Psychosmomatic Disorder	psyso	178	133	45
Autoimmune Thyroid Disease	aut	173	158	15
Somatoform Disorder	som	147	118	29
Hashimoto's Disease	hash	136	125	11
Depression	depr	134	105	29
Vertigo	schw	131	99	32
Insulin Resistance	ins	106	57	49
Chronic Renal Insufficiency	ni	103	52	51
Cancer	ca	101	56	45
Adaptive Disorder	anp	100	80	20
Diabetes Mellitus	dm	80	42	38
Migraine	mi	70	59	11
Susceptibility to Infection	inf	65	53	12
Metabolic Syndrome	met	52	23	29
Adrenal Weakness	neb	49	44	5
Asthma	ast	42	27	15
Osteoporosis	ost	38	33	5
Chronic Obstructive Lung Disease	cold	26	13	13
Secondary Hyperparathyroidism	sek	24	12	12
Diabetes Mellitus(non-insulin dependent)	dm_ni	20	7	13
Psoriasis	pso	19	15	4
Rheumatoid Arthritis	ra	18	13	5
Hyperthyroidism	hyper	13	10	3
Graves' disease	base	8	8	0
Primary Hyperparathyroidism	pri	3	3	0
Polycystic Ovary Syndrome	pco	3	3	0

Table B.2. Absolute number of diseases' appearances considering the 20 patients analyzed with STATIS.

Disease/Disorder	Total
Dyspepsia	15
Stress	16
Exhaustion	13
Hypothyroidism	13
Insomnia	12
Hypertension	8
Anxiety Disorder	9
Depressive Disorder	7
Headache	9
Psychosmomatic Disorder	6
Autoimmune Thyroid Disease	4
Somatoform Disorder	9
Hashimoto's Disease	4
Depression	6
Vertigo	5
Insulin Resistance	1
Chronic Renal Insufficiency	5
Cancer	6
Adaptive Disorder	5
Diabetes Mellitus	2
Migraine	3
Susceptibility to Infection	0
Metabolic Syndrome	2
Adrenal Weakness	2
Asthma	1
Osteoporosis	2
Chronic Obstructive Lung Disease	1
Secondary Hyperparathyroidism	2
Diabetes Mellitus(non-insulin dependent)	1
Psoriasis	0
Rheumatoid Arthritis	2
Hyperthyroidism	1
Graves' disease	0
Primary Hyperparathyroidism	1
Polycystic Ovary Syndrome	0

Table B.3. Characteristics of the two groups `Low` (HPL < 1.0 nmol/L) and `High` (HPL ≥ 1.0 nmol/L) which differ significantly in their means. Numerical variables are described by *mean±standard deviation* whereas absolute numbers are used in the case of binary attributes. Percentages represent proportion of the attribute within each group. Statistical significance is indicated by an asterisk.

Characteristics	low (HPL < 1.0) $n = 70$	high (HPL ≥ 1.0 $n = 395$
Gender		
Men	20(28.6%)	104(26.3%)
Women	50(71.4%)	291(73.7%)
Age	47.7 ± 15.1(100%)	49.4 ± 13.0(100%)
Body Mass Index	23.7 ± 4.0	23.4 ± 4.0
Thyroid Diseases		
Hypothyroidism	33(47.1%)	198(50.1%)
AITD	11(15.7%)	75(19.0%)
Hashimoto's Disease	6(8.6%)	60(15.2%)
Thyroid Levels		
TSH (μIU/mL)	2.1 ± 1.6(75.7%)	2.3 ± 9.4(82.8%)
FT3 (pg/mL)	3.0 ± 0.5(60.0%)	3.0 ± 0.5(73.4%)
FT4 (ng/dL)	1.3 ± 0.2(61.4%)	1.2 ± 0.2(73.7%)
TPO-Ab* (U/mL)	19.2 ± 34.0(47.1%)	45.8 ± 107.1(60.3%)
Selen (μg/L)	89.0 ± 15.7(77.1%)	87.2 ± 15.8(81.8%)
Metabolic Syndrome	4(5.7%)	15(3.8%)
Hypertension	11(15.7%)	90(22.8%)
Diabetes Mellitus	4(5.7%)	21(5.3%)
Insulin Resistence	6(8.6%)	40(10.1%)
Levels		
HDL-C (mg/dL)	63.1 ± 16.6(28.6%)	67.7 ± 17.2(45.1%)
LDL-C (mg/dL)	130.7 ± 33.3(28.6%)	129.9 ± 34.2(45.8%)
ALAT GPT (U/L)	22.9 ± 21.4(55.7%)	23.0 ± 18.6(62.0%)
Gamma-GT (U/L)	22.0 ± 21.5(57.1%)	24.0 ± 24.7(62.5%)
HbA1c (%)	5.6 ± 0.3(45.7%)	5.6 ± 0.4(60.3%)
Other Diseases		
Psychosomatics	7(10.0%)	70(17.7%)
Stress	41(58.6%)	215(54.4%)
Exhaustion*	25(35.7%)	188(47.6%)
Insomnia	20(28.6%)	134(33.9%)
Dyspepsia *	43(41.4%)	296(74.9%)
Headache	13(18.6%)	72(18.2%)
Other Parameters		
Vitamin D* (ng/mL)	21.1 ± 11.4(72.9%)	24.4 ± 12.0(83.8%)
Apr/May/Jun/Jul/Aug/Sep	24	158
Oct/Nov/Dec/Jan/Feb/Mar	27	173
HPL (nmol/L)	0.8 ± 0.1(100%)	2.4 ± 1.7(100%)
Homocysteine* (μmol/L)	11.7 ± 3.1(74.3%)	11.2 ± 3.2(80.5%)
Kreatinin (g/L)	0.8 ± 0.2(51.4%)	0.8 ± 0.2(64.3%)

Appendix C

Predicting Diseases

Table C.1. Proportion of characteristic laboratory values for AITD and Hashimoto's disease measured in patients' first blood sample.

First Laboratory Test	AITD	Hashimoto's Disease	without AITD
TSH	82.7 %	80.9 %	66.5 %
FT3	65.6 %	65.4 %	50.4 %
FT4	67.6 %	67.7 %	50.7 %
TPO	59.0 %	59.6 %	37.0 %
!!SRE	7.5 %	5.1 %	2.2 %
TG	16.8 %	15.4 %	4.0 %

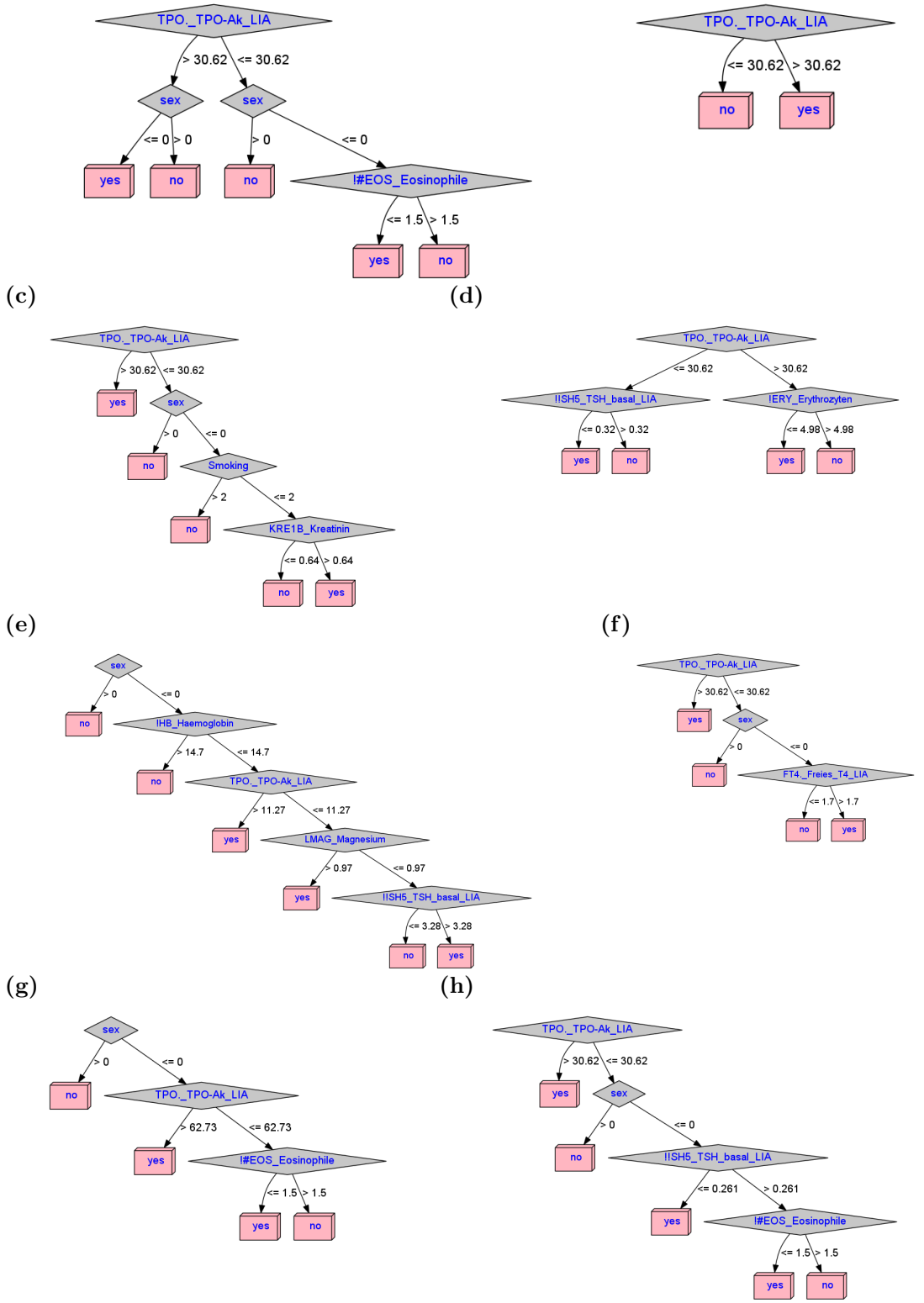

Figure C.1. Decision tree ensemble predicting AITD. The single tree referring to each of the eight trials is shown in **(a)**-**(h)**.

Appendix D

Controlling the Course of Treatment

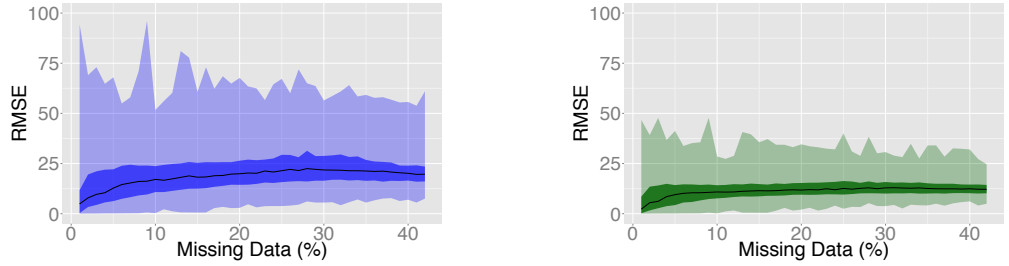

Figure D.1. Root-means-square errors in laboratory values due to imputation of inserted missing values. These test results are based on 17 patients and 9 laboratory vaues including TSH, FT3, FT4, HbA1c, calcium, vitamin D, creatinine, cholesterol and GTP (ALAT). **(a)** KNN **(b)** modified KNN

Bibliography

[1] Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic Health Records, Medical Research, and the Tower of Babel. New England Journal of Medicine. 2008;358(16):1738–1740. PMID: 18420507. Available from: http://dx.doi.org/10.1056/NEJMsb0800209.

[2] Goth G. Analyzing Medical Data. Commun ACM. 2012 Jun;55(6):13–15. Available from: http://doi.acm.org/10.1145/2184319.2184324.

[3] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 06;13(6):395–405. Available from: http://dx.doi.org/10.1038/nrg3208.

[4] Kierkegaard P. Electronic health record: Wiring Europe's healthcare. Computer Law & Security Review. 2011;27(5):503 – 515. Available from: http://www.sciencedirect.com/science/article/pii/S0267364911001257.

[5] Prousky J. Orthomolecular psychiatric treatments are preferable to mainstream psychiatric drugs: a rational analysis. J Orthomol Med. 2013;28:17–32.

[6] Janson M. Orthomolecular medicine: the therapeutic use of dietary supplements for anti-aging. Clinical interventions in aging. 2006;1(3):261.

[7] Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Affairs. 2012;31(12):2805–2816.

[8] Winters-Miner LA, Bolding PS, Hilbe JM, Goldstein M, Hill T, Nisbet R, et al. Chapter 5 - Electronic Medical Records: Analytics' Best Hope. In: Winters-Miner LA, Bolding PS, Hilbe JM, Goldstein M, Hill T, Nisbet R, et al., editors. Practical Predictive Analytics and Decisioning Systems for Medicine. Academic Press; 2015. p. 1019 – 1029. Available from: http://www.sciencedirect.com/science/article/pii/B978012411643600051X.

[9] Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81–106.

[10] Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.

[11] Zimmet P, Alberti KGMM, Shaw J. Global and societal implications of the diabetes epidemic. Nature. 2001 12;414(6865):782–787. Available from: http://dx.doi.org/10.1038/414782a.

[12] Reaven GM. Role of Insulin Resistance in Human Disease. Diabetes. 1988;37(12):1595–1607. Available from: http://diabetes.diabetesjournals.org/content/37/12/1595.abstract.

[13] DeFronzo RA. Insulin resistance: a multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidaemia and atherosclerosis. The Netherlands Journal of Medicine. 1997;50(5):191 – 197. Available from: http://www.sciencedirect.com/science/article/pii/S0300297797000120.

[14] Kaplan NM. The deadly quartet: Upper-body obesity, glucose intolerance, hypertriglyceridemia, and hypertension. Archives of Internal Medicine. 1989;149(7):1514–1520. Available from: +http://dx.doi.org/10.1001/archinte.1989.00390070054005.

[15] Eckel RH, Grundy SM, Zimmet PZ. The metabolic syndrome. The Lancet. 2005;365(9468):1415 – 1428. Available from: http://www.sciencedirect.com/science/article/pii/S0140673605663787.

[16] Alberti KGM, Zimmet P, Shaw J. The metabolic syndrome—a new worldwide definition. The Lancet. 2005;366(9491):1059 – 1062. Available from: http://www.sciencedirect.com/science/article/pii/S0140673605674028.

[17] Alberti KGMM, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation. Diabetic Medicine. 1998;15(7):539–553. Available from: http://dx.doi.org/10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.

[18] Balkau B, Charles MA. Comment on the provisional report from the WHO consultation. Diabetic medicine. 1999;16(5):442–443.

[19] on Detection EP, Evaluation, , of High Blood Cholesterol in Adults T. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). JAMA. 2001;285(19):2486–2497. Available from: +http://dx.doi.org/10.1001/jama.285.19.2486.

[20] Federation ID. The IDF consensus worldwide definition of the metabolic syndrome; 2005. Accessed March 3, 2016. http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf.

[21] Abe Y, Kikuchi T, Nagasaki K, Hiura M, Tanaka Y, Ogawa Y, et al. Usefulness of GPT for Diagnosis of Metabolic Syndrome in Obese Japanese Children. Journal of Atherosclerosis and Thrombosis. 2010;16(6):902–909.

[22] Yousefzadeh G, Shokoohi M, Yeganeh M, Najafipour H. Role of gamma-glutamyl transferase (GGT) in diagnosis of impaired glucose tolerance and metabolic syndrome: A prospective cohort research from the Kerman Coronary Artery Disease Risk Study (KERCADRS). Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2012;6(4):190 – 194. Available from: http://www.sciencedirect.com/science/article/pii/S1871402112001166.

[23] Nejatinamini S, Ataie-Jafari A, Qorbani M, Nikoohemat S, Kelishadi R, Asayesh H, et al. Association between serum uric acid level and metabolic syndrome components. Journal of Diabetes and Metabolic Disorders. 2015;14:70. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4570526/.

[24] Laurinavicius A, Conceicao R, Kashiwagi NM, Tabone VA, Carvalho JAM, Fogaca V, et al. GLYCATED HEMOGLOBIN: A NEW PARADIGM FOR THE METABOLIC SYNDROME? Journal of the American College of Cardiology. 2014;63(12_S). Available from: +http://dx.doi.org/10.1016/S0735-1097(14)61342-5.

[25] Selvin E BFCJ Crainiceanu CM. Short-term variability in measures of glycemia and implications for the classification of diabetes. Archives of Internal Medicine. 2007;167(14):1545–1551. Available from: +http://dx.doi.org/10.1001/archinte.167.14.1545.

[26] Young D, Bermes E. Preanalytical variables and biological variation. Tietz textbook of clinical chemistry and molecular diagnostics, 4th edition St Louis: Elsevier Saunders. 2006;p. 449–473.

[27] Franco RS. Measurement of Red Cell Lifespan and Aging. Transfusion Medicine and Hemotherapy. 2012 10;39(5):302–307. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678251/.

[28] Umar H, Muallima N, Adam J, Sanusi H. Hashimoto's thyroiditis following Graves' disease. Acta Med Indones. 2010;42(1):31–35.

[29] Majumder A, Sanyal D. A case of simultaneous occurrence of Graves' disease and Hashimoto's thyroiditis. Indian journal of endocrinology and metabolism. 2012;16(Suppl 2):S338.

[30] Iddah M, Macharia B. Autoimmune thyroid disorders. ISRN endocrinology. 2013;2013.

[31] Kamsteeg J. HPU und dann. . . .? Beschwerden und Erkrankungen infolge von „Pyrrolurie “ KEAC-Weert. 2005;195.

[32] Strienz J. Leben mit KPU-Kryptopyrrolurie: ein Ratgeber für Patienten. Zuckschwerdt Verlag; 2011.

[33] McGinnis WR, John McLaren-Howard DSc F, Lewis A, Lauda PH, Lietha R. Discerning the mauve factor, part 1. Alternative therapies in health and medicine. 2008;14(2):40.

[34] Mikirova N. Clinical Test of Pyrroles in Psychiatric Disorders: Association with Nutritional, Immunological and Metabolic Markers. Journal of Nutritional Therapeutics. 2015;4(1):4–11.

[35] DG I. Apparently non-indolic Ehrlich-positive substances related to mental illnesses. J Neuropsychiatr. 1961 August;(2):292–305.

[36] Irvine D, Bayne W, Miyashita H, Majer J. Identification of kryptopyrrole in human urine and its relation to psychosis. 1969;.

[37] Sohler A, Beck R, Noval JJ. Mauve factor re-identified as 2, 4-dimethyl-3-ethylpyrrole and its sedative effect on the CNS. Nature. 1970;228:1318–1320.

[38] Pfeiffer C, Sohler A, Jenney E, Iliev V. Treatment of pyroluric schizophrenia (malvaria) with large doses of pyridoxine and a dietary supplement of zinc. J Appl Nutr. 1974;26:21–28.

[39] Przyrembel H SM. Die (Krypto-) Pyrrolurie in der Umweltmedizin: eine valide Diagnose? Bundesgesundheitsbl - Gesundheitsforsch - Gesundheitsschutz. 2007;(50):1324–1330.

[40] Alpaydin E. Introduction to machine learning. MIT press; 2014.

[41] Han J, Kamber M, Pei J. Data mining: concepts and techniques. Elsevier; 2011.

[42] Matzk S. Systematic Management and Visualization of Complex and Heterogeneous Patient Data; 2015.

[43] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org/`.

[44] Wickham H, James DA, Falcon S. RSQLite: SQLite Interface for R; 2014. R package version 1.0.0. Available from: `http://CRAN.R-project.org/package=RSQLite`.

[45] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–525.

[46] Moni MA, Liò P. comoR: a software for disease comorbidity risk assessment. Journal of Clinical Bioinformatics. 2014;4:8–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4081507/.

[47] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proceedings of the National Academy of Sciences. 2007;104(21):8685–8690. Available from: http://www.pnas.org/content/104/21/8685.abstract.

[48] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005 10;437(7062):1173–1178. Available from: http://dx.doi.org/10.1038/nature04209.

[49] Kan WC, Wang JJ, Wang SY, Sun YM, Hung CY, Chu CC, et al. The New Comorbidity Index for Predicting Survival in Elderly Dialysis Patients: A Long-Term Population-Based Study. PLoS ONE. 2013 08;8(8):1–8. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0068748.

[50] Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol. 2009 04;5:1–11. Available from: http://dx.doi.org/10.1371%2Fjournal.pcbi.1000353.

[51] Jaccard P. Nouvelles recherches sur la distribution florale. Bulletin de la Sociète Vaudense des Sciences Naturelles. 1908;44:223–270.

[52] Florin E. Causality measures between neural signals from invasively and non-invasively obtained local field potentials in humans [Dr. (Univ.)]. Bergische Universitä. Wuppertal. Wuppertal; 2010. Record converted from VDB: 12.11.2012; Wuppertal, Bergische Universität, Fachbereich Mathematik und Naturwissenschaften, Diss., 2010. Available from: http://juser.fz-juelich.de/record/10390.

[53] Morganti S, Ceda G, Saccani M, Milli B, Ugolotti D, Prampolini R, et al. Thyroid disease in the elderly: sex-related differences in clinical expression. Journal of endocrinological investigation. 2004;28(11 Suppl Proceedings):101–104.

[54] Melton LJ, Chrischilles EA, Cooper C, Lane AW, Riggs BL. How many women have osteoporosis? Journal of bone and mineral research. 2005;20(5):886–892.

[55] Adami S, Marcocci C, Gatti D. Epidemiology of primary hyperparathyroidism in Europe. Journal of bone and mineral research: the official journal of the American Society for Bone and Mineral Research. 2002;17:N18–23.

[56] Organization WH. Gender and women's mental health;. http://www.who.int/mental_health/prevention/genderwomen/en/.

[57] Saito I, Ito K, Saruta T. Hypothyroidism as a cause of hypertension. Hypertension. 1983;5(1):112–5. Available from: http://hyper.ahajournals.org/content/5/1/112.abstract.

[58] Stabouli S, Papakatsika S, Kotsis V. Hypothyroidism and hypertension. Expert Review of Cardiovascular Therapy. 2010;8(11):1559–1565. Available from: http://dx.doi.org/10.1586/erc.10.141.

[59] Jackson D A HHH Somers K M. Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence? The American Naturalist. 1989;133:346–453.

[60] Kawada T, Otsuka T, Inagaki H, Wakayama Y, Li Q, Li YJ, et al. Increase in the prevalence of metabolic syndrome among workers according to age. The Aging Male. 2010;13(3):184–187.

[61] Botella-Carretero JI, Alvarez-Blasco F, Villafruela JJ, Balsa JA, Vázquez C, Escobar-Morreale HF. Vitamin D deficiency is associated with the metabolic syndrome in morbid obesity. Clinical Nutrition. 2007;26(5):573–580.

[62] Kivity S, Agmon-Levin N, Zisappl M, Shapira Y, Nagy EV, Dankó K, et al. Vitamin D and autoimmune thyroid diseases. Cellular & molecular immunology. 2011;8(3):243–247.

[63] Anglin RE, Samaan Z, Walter SD, McDonald SD. Vitamin D deficiency and depression in adults: systematic review and meta-analysis. The British journal of psychiatry. 2013;202(2):100–107.

[64] D'Aurizio F, Villalta D, Metus P, Doretto P, Tozzoli R. Is vitamin D a player or not in the pathophysiology of autoimmune thyroid diseases? Autoimmunity Reviews. 2015;14(5):363 – 369. Available from: http://www.sciencedirect.com/science/article/pii/S1568997214002201.

[65] Antunes CM, Oliveira AL. Temporal data mining: An overview. In: KDD workshop on temporal data mining. vol. 1; 2001. p. 13.

[66] Chen H, Fuller SS, Friedman C, Hersh W. Knowledge management, data mining, and text mining in medical informatics. In: Medical Informatics. Springer; 2005. p. 3–33.

[67] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences. 2013;29(2):93–99.

[68] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine. 2001;7(6):673–679.

[69] Shahhoseini R, Ghazvini A, Esmaeilpour M, Pourtaghi G, Tofighi S. Presentation of a model-based data mining to predict lung cancer. Journal of research in health sciences. 2015;15(3):189–195.

[70] Zorman M, Podgorelec V, Kokol P, Peterson M, Šprogar M, Ojsteršek M. Finding the right decision tree's induction strategy for a hard real world problem. International journal of medical informatics. 2001;63(1):109–121.

[71] Quinlan JR. C4. 5: programs for machine learning. Elsevier; 2014.

[72] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[73] Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied statistics. 1980;p. 119–127.

[74] Shahhoseini R, Ghazvini A, Esmaeilpour M, Pourtaghi G, Tofighi S. Presentation of a model-based data mining to predict lung cancer. Journal of research in health sciences. 2015;15(3):189–195.

[75] Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;5(1):3–55.

[76] Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005.

[77] Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999;14(771-780):1612.

[78] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21(15):3301–3307.

[79] Kuhn M, Weston S, Coulter N, code for C5 0 by R Quinlan MCC. C50: C5.0 Decision Trees and Rule-Based Models; 2015. R package version 0.1.0-24. Available from: `http://CRAN.R-project.org/package=C50`.

[80] Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al.. caret: Classification and Regression Training; 2015. R package version 6.0-62. Available from: `http://CRAN.R-project.org/package=caret`.

[81] Toulis KA, Anastasilakis AD, Tzellos TG, Goulis DG, Kouvelas D. Selenium supplementation in the treatment of Hashimoto's thyroiditis: a systematic review and a meta-analysis. Thyroid. 2010;20(10):1163–1173.

[82] Mazokopakis EE, Papadakis JA, Papadomanolaki MG, Batistakis AG, Giannakopoulos TG, Protopapadakis EE, et al. Effects of 12 months treatment with L-selenomethionine on serum anti-TPO Levels in Patients with Hashimoto's thyroiditis. Thyroid. 2007;17(7):609–612.

[83] Wyss M, Kaddurah-Daouk R. Creatine and creatinine metabolism. Physiological reviews. 2000;80(3):1107–1213.

[84] Kim Sw, Jung HW, Kim CH, Kim Ki, Chin HJ, Lee H. A New Equation to Estimate Muscle Mass from Creatinine and Cystatin C. PloS one. 2016;11(2):e0148495.

[85] Klie S, Caldana C, Nikoloski Z. Compromise of multiple time-resolved transcriptomics experiments identifies tightly regulated functions. Front Plant Sci. 2012;3(249):10–3389.

[86] Klie S, Osorio S, Tohge T, Drincovich MF, Fait A, Giovannoni JJ, et al. Conserved changes in the dynamics of metabolic processes during fruit development and ripening across species. Plant physiology. 2014;164(1):55–68.

[87] Bordag N, Klie S, Jürchott K, Vierheller J, Schiewe H, Albrecht V, et al. Glucocorticoid (dexamethasone)-induced metabolome changes in healthy males suggest prediction of response and side effects. Scientific Reports. 2015 11;5:15954 EP –. Available from: http://dx.doi.org/10.1038/srep15954.

[88] Stanimirova I, Walczak B, Massart DL, Simeonov V, Saby CA, Crescenzo ED. STATIS, a three-way method for data analysis. Application to environmental data. Chemometrics and Intelligent Laboratory Systems. 2004;73(2):219 – 233. Available from: http://www.sciencedirect.com/science/article/pii/S016974390400084X.

[89] Abdi H, Williams LJ, Valentin D, Bennani-Dosse M. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(2):124–167. Available from: http://dx.doi.org/10.1002/wics.198.

[90] l'Hermier des Plantes H, Thiébaut B. Étude de la pluviosité au moyen de la méthode S.T.A.T.I.S. Revue de Statistique Appliquée. 1977;25(2):57–81. Available from: http://eudml.org/doc/106041.

[91] Lavit C, Escoufier Y, Sabatier R, Traissac P. The {ACT} (STATIS method). Computational Statistics & Data Analysis. 1994;18(1):97 – 119. Available from: http://www.sciencedirect.com/science/article/pii/0167947394901341.

[92] P Robert YE. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. Journal of the Royal Statistical Society Series C (Applied Statistics). 1976;25(3):257–265. Available from: http://www.jstor.org/stable/2347233.

[93] Greenacre MJ. Biplots in practice. Fundacion BBVA; 2010.

[94] Kayacelebi AA, Willers J, Pham VV, Hahn A, Schneider JY, Rothmann S, et al. Plasma homoarginine, arginine, asymmetric dimethylarginine and total homocysteine interrelationships in rheumatoid arthritis, coronary artery disease and peripheral artery occlusion disease. Amino acids. 2015;47(9):1885–1891.

[95] den Heijer M, Brouwer IA, Bos GM, Blom HJ, van der Put NM, Spaans AP, et al. Vitamin supplementation reduces blood homocysteine levels A controlled trial in patients with venous thrombosis and healthy volunteers. Arteriosclerosis, thrombosis, and vascular biology. 1998;18(3):356–361.