DISSERTATION

The Composition of a statistical data-driven Workflow for untargeted Metabolomics Studies: Complexity and Applications

submitted by

Mag.rer.soc.oec. Mag.rer.nat. Sophie Helene NARATH

for the Academic Degree of

Doctor of Medical Science

(Dr. scient. med.)

at the

Medical University of Graz

Department of Internal Medicine

Division of Endocrinology and Diabetology

under the Supervision of

ASS. PROF. PRIV.-DOZ. DR. MED. UNIV. Harald SOURIJ

2016

Declaration

I hereby declare that this dissertation is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this dissertation. Due acknowledgement has been made in the text to all other material used. Throughout this dissertation and in all related publications I followed the guidelines of “Good Scientific Practice”.

Graz,

Foreword

This is a thesis written in the field of medical science, about a statistical data driven workflow in metabolomics. I work as a statistician at Joanneum Research. My academical-background is sport science, urbanism and sociology, which underlines, even intensifies the interdisciplinarity of this work: with a focus on methodology not excluding reflections about the actual position of “Metabolomics” in science including keywords in fashion like biomarker and precision medicine.

Acknowledgements

Supervisor

ASS.PROF.PRIV.-DOZ.DR.MED.UNIV.HARALD SOURIJ

Comité-Members

DR. CHRISTOPH MAGNES

UNIV.PROF. DR. THOMAS PIEBER

AO.UNIV.-PROF. DR. DR. MICHAEL G. SCHIMEK

Personal Support

Bettina, Markus, Joris Narath

Sportel-, Tratsch-Freunde

Beate Boulgaropoulos, JR Health-Kollegen

Table of Contents

Abbreviation and Definitions

AUC	Area Under the Curve of a peak; representing the intensities of metabolic features and metabolites
BCAA	Branched chain amino acids
BL	Solvent blank-sample for HPLC-HRMS quality control
Biomarker	NIH Definition 2001: „a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention“ (1)
ESRD	End Stage Renal Disease
GC-MS	Gas chromatography–mass spectrometry
HILIC	Hydrophilic Interaction Liquid Chromatography
HD	Hemodialysis
HDF	Hemodiafiltration
HMDB	The Human Metabolome Database (2)
HPLC-HRMS	High Performance Liquid Chromatography-High Resolution Mass Spectrometry
IPAH	Idiopathic Pulmonary Hypertension
OL-HDF	On-Line-Hemodiafiltration (HDF: Hemodiafiltration) used as synonymes
Metabolic feature	Metabolite or substance measured by defined by specific retention time and mass
Metabolomics	Metabolomics deals with low molecular weight metabolites (<1500 Dalton) that are ubiquitously present within organisms, cells or tissues.
MS-MS	Tandem mass spectrometry
Mz/Mzmed	Mass, median Mz-mass for a metabolic feature
PCA	Principal Component Analysis Statistical (visualisation) tool for multivariate data analysis

PRE-POST effects	Effects seen before and after intervention (bariatric surgery, dialysis)
Precision medicine	NIH 2016: “Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person” (3)
QC	Quality Control samples: pooled samples for HPLC-HRMS quality control
Quantile Regression	Method for drift correction on QCs for time-dependent measurement variation
R	Freely available matrix-based statistical software
RCT	Randomized Controlled Trial
Rt/RtMed	Retention time/Median retention time for a metabolic feature
T2DM	Type 2 diabetes mellitus
UHPLC-MS	Ultra-High Pressure Liquid Chromatography-Mass Spectrometry
XCMS	R-based software for peak-detection and -alignment

List of Figures

List of Tables

Zusammenfassung

Ziel der Dissertation war es, einen statistischen data-driven Workflow für untargeted Metabolomics Studien zu entwickeln und anzuwenden. Die vorliegende Arbeit befasst sich weiters mit Metabolomics im Allgemeinen und hier speziell mit der Problematik. Der Fokus liegt auf der Methodik. Die große Menge an Daten ist ein Merkmal für Metabolomics Studien und muss mittels statistischer Methoden bearbeitet und gefiltert werden: Ziel ist das Trennen wichtiger Information von unwichtiger Information bzw. Rauschen. Die Definition von „wichtiger“ Information ergibt sich aus der medizinischen oder biologischen Fragestellung. Der Workflow besteht aus insgesamt 15 Schritten, die sowohl Daten-Bearbeitungsschritte wie Filtern und Driftkorrektur als auch die statistische Analyse der End-Daten beinhalten. Der Workflow wurde anhand von präklinischen und klinischen Metabolomics Studien fortwährend weiterentwickelt und optimiert. Ein Auszug der wichtigsten und größten Studien wird hier als repräsentative Anwendungsbeispiele beschrieben. Die wichtigsten Ergebnisse der Dissertation sind:

- Implementierung des Workflows für klinische und präklinische untargeted LC-MS Metabolomics Studien, mit spezifischen Anpassungen ja nach wissenschaftlicher Fragestellung
- Erfolgreiche Anwendung der Driftkorrektur via Quantils-Regression an einem großen GC-MS Data-Set (> 1000 samples)
- Branched chain amino acids und armonatic amino acids gelten als Indikatoren für kardiovaskuläres Risiko, diese wurden als signifikante Metabolomics Marker bei Bariatric Surgery Patienten identifiziert.
- Bewusstsein, dass Reflexion und Diskussion in einem interdisziplinären Team ausschlaggebend für interpretierbare Metabolomics-Ergebnisse sind. Eine fortwährende Diskussion von Studiendesign, über Data-Processing bis hin zum statistischen Modellieren ist erforderlich.
- Jede Studie erfordert eine angepasste Muster-Erkennung. Obgleich die Methoden im Workflow standardisiert sein sollen, kann der gesamte Prozess nicht automatisiert werden.
- Fall-Kontroll-Studien eignen sich besser für untargeted Metabolomics Anwendung da hier die Variabilität besser kontrolliert werden kann als bei RCT.

Abstract

The aim of the thesis was to develop and apply a statistical data driven workflow for untargeted metabolomics studies. The present work further deals with metabolomics and challenges of metabolomics studies, with strong focus on methodology.

One challenge is the big amount of data deriving from the metabolomics studies; the statistical data driven workflow enables to handle this big amount of data by distinguishing between important and non-important information, depending on the particular scientific question (medical, biological). It consists of 15 steps in total including filtering-steps, drift correction via quantile regression and consecutively statistical analysis of the processed data.

The workflow[1] has been developed and constantly optimized with data from pre-clinical and clinical untargeted metabolomics studies. The largest and most important studies were used as representative examples and are described in the following. The main outcomes of the thesis are:

- Implementation of the statistical data driven workflow works for pre-clinical and clinical untargeted LC-MS Metabolomics studies, with several adaptions depending on the scientific question
- Successful application of the drift correction via quantile regression on a very large data set (<1000 samples) from GC-MS-Data
- The metabolomics identification of branched chain amino acids and aromatic amino acids which are indicators for cardiovascular disease and these could be identified to be significant in patients undergoing bariatric surgery.
- Awareness that reflections and discussion in an interdisciplinary team are crucial to get interpretable results. Ongoing discussion is required from the study-design throughout the data processing and statistical modelling
- Each study demands its proper pattern analysis. Although the tools of the workflow should be standardized, the whole proceeding cannot be automated
- Case-control-studies are better suited for untargeted metabolomics applications than RCTs as the variability can better be controlled

[1] For better reading „workflow“ consecutively means „statistical data-driven workflow“

1. Introduction

1.1. Metabolomics

Metabolomics deals with low molecular weight metabolites (<1500 Dalton) that are ubiquitously present within organisms, cells or tissues. Metabolomics is performed by using three techniques, namely NMR (Nuclear Magnetic Resonance Spectroscopy), LC-MS (Liquid Chromatography Mass Spectrometry) and GC-MS (Gas Chromatography Mass Spectrometry). In this thesis we will focus on MS techniques.

Metabolomics is applied in different fields of science (plants, food, biology, medicine…) - this thesis concentrates on the field of medical science, whereas the ambitious global aim of metabolomics in medical science has been formulated from Wishart in 2007:

> *"The aim is to be able to take urine, blood or some other body fluid, scan it in a machine and find a profile of tens or hundreds of chemicals that can predict whether an individual is on the road to a disease, say, or likely to experience side-effects from a particular drug"*(4).

Not surprisingly, mechanisms and humans are more complicated than that, but the citation reflects the overall-wish dedicated to metabolomics.

We distinguish between targeted and untargeted metabolomics. In clinical research targeted metabolomics is hypothesis driven (5), we are looking for levels of specific metabolites (6) to answer specific questions. Untargeted metabolomics serves to generate hypothesis and describes the global profile of a metabolome (6). Untargeted metabolomics is located in the setting of an exploratory search of putative biomarkers. In short, targeted metabolomics include studies to explore biological mechanistic processes and untargeted metabolomics include studies that search and assess **biomarkers** (7).

Putative biomarkers are defined as a marker or a set of identified metabolites that distinguish one group from the other. The way from a putative biomarker or potential biomarker candidate to biomarker tests in clinical routine is very long and not precisely defined, as the NIH-definition of 2001 points out: *„a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention“*(1).

Putative biomarkers can be categorized into diagnostic, prognostic and predictive biomarkers, they can be used in analytical validation, qualification and their utilization has to be discussed separately (8). Biomarker research has been increasingly funded during the last years but the terminology and validation is still incomplete and the discussion is ongoing (9,10). The selection, assessment or reporting of candidate biomarkers is not standardized (7). A clinical relevance based on the three crucial questions:

“1. Can the clinician measure them?

2. Do they add new information?

3. Do they help the clinician to manage patients? “

are still far away to be properly answered in the metabolomics field as recently mentioned by Mamas M, Dunn WB, Neyses L and Goodacre R. (11).

The biomarker-discussion can be seen from different perspectives; the overall-medical perspective for clinical improvement (1,9,10), the overall metabolomics-perspective (7,12–14), metabolomics disease-relevant perspective (15–20), the data-driven perspective (7,21,22), the economical perspective and many others. Biomarker discussion is not the focus of this thesis per se, but it plays a large role in communication and commercialization of metabolomics. The complexity and the vast meaning in terminology of biomarker should be kept in mind: biomarkers are hard to find, even harder to validate and no metabolomics biomarker has found its way to a clinical application in routine up to now.

To summarize the key message: that is why we talk here about distinctive metabolic features, putative markers or selective metabolic features and not about biomarkers per se.

This thesis deals with the implementation of a workflow for statistical data processing for untargeted metabolomics to select important information out of the huge amount of metabolic features. In the following the term “metabolomics” refers to “untargeted metabolomics”. The thesis also deals with the appropriate study design for metabolomics studies and reflections about the use of metabolomics in a broader scientific view.

One challenge of untargeted metabolomics is that around 10000 peaks are measured by LC-MS in one study, but very few of them refer to known metabolites. A representative peak is called “**metabolic feature**” and is defined by a specific retention time (Rt) and a specific mass (Mz). Rt refers to the chemical properties of a metabolite (the earlier the more lipophilic, the later the more hydrophilic) and Mz is a parameter for the size of a metabolite.

1.2. The statistical data-driven Workflow for untargeted Metabolomics Studies

A typical research question for untargeted metabolomics is to find differences in the metabolic profile between healthy subjects and controls, or between “before” and “after” samples of a clinical intervention. The aim is to find discriminatory features within two groups and thousands of metabolic features and to find patterns that separate one group from the other.

In untargeted metabolomics relevant information has to be distinguished from non-information. Therefore statistics are needed. Due to the huge amount of data, statistical analysis of metabolomics data is not directly comparable with the statistics used for conventional clinical studies. The number of variables is many times higher than the number of observations. This situation is called p<<N and described as High-Dimensional Problems (23).

To extract the relevant information and to get interpretable and reasonable results, advanced univariate and multivariate statistical methods are applied. The application of these methods bears risks of over-interpretation and incorrect handling: False discovery of discriminatory features and potential “biomarkers” is a known problem in the scientific community (24,25).

Many metabolomics tools to process data exist, due to the lack of harmonization of analytical methods and automatisation of spectral data processing (26). An overview of the state-of-the-art tools is listed in Misra and van der Hooft 2016 (see Table 16).

However, a reproducible and consistent way of data processing for application in several different metabolomics studies is still missing (25,27–29). Therefore, we propose a standardized workflow for metabolomics data processing and an ensuing statistical analysis, tylored to the JR-metabolomics platform, which is sketched in the following figure (Figure 1). Beside three other freely available tools namely MetAlgn, MZmine and SpectConnect, XCMS was used as a standard data preprocessing tool (26).

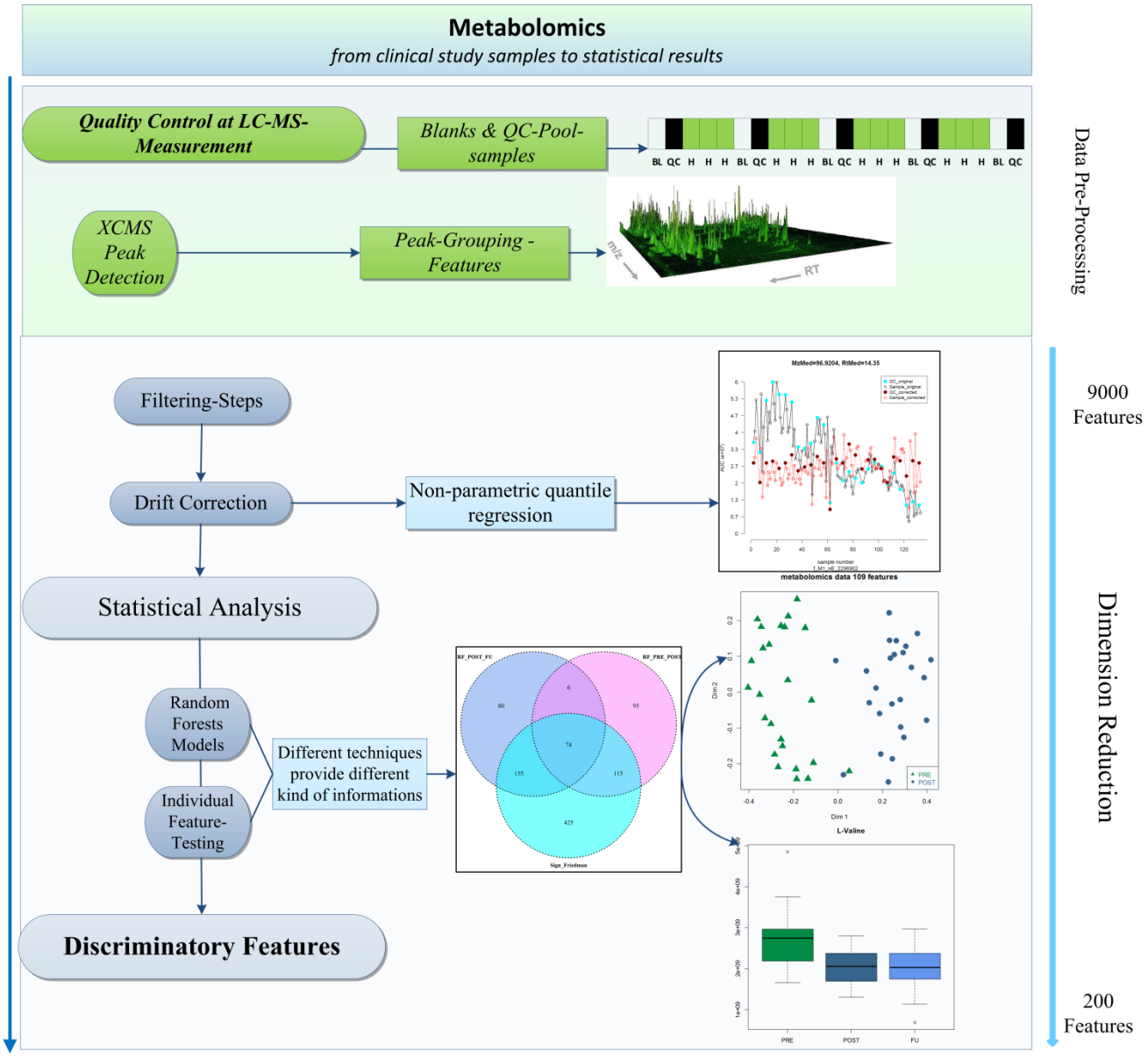


Figure 1: Overview of the standardized metabolomics data processing workflow

Metabolomics deals with low molecular weight metabolites (<1500 Dalton) that are ubiquitously presented within organisms, cells or tissue. It produces an enormous quantity of data, which needs to undergo several pre-processing steps, like quality control steps, peak detection, peak grouping, etc., before being further processed. This further data processing comprises data filtering and drift correction.

For drift correction, non-parametric Quantile Regression models were built and batch-to-batch variation was evaluated by using Random Forests Models. The processed data were the basis for further statistical analysis.

The resulting worflow includes validation steps to detect relevant and discriminatory features. The workflow has been optimized on various data sets of metabolomics studies. Our main examples of use are the following four clinical studies:

1. **Cardionor** – a clinical study investigating treatment responses from T2DM patients associated with cardiovascular risk over 2 years with the carotis intimia media thickness (CIMT) as primary response. Samples from 81 patients were collected to perform metabolomics.

2. **Bariatric Surgery** – a clinical study where responses of patients undergoing bariatric surgery over 1 year were investigated. Serum samples were collected from 44 obese patients at three visits to do metabolomics. The outcome of the clinical study “Bariatric surgery” has been published PLOS ONE[2] and takes an important part of this thesis

3. **Metaprol** – a clinical study where two dialysis modalities were compared in a cross-over design. Serum samples form 19 patients were collected at 4 different visits to do metabolomics. The manuscript for publication is in preparation for “Kidney International”.

4. **Nutritech** – an EU-wide project, its aim is to quantify the effect of diet on “phenotypic flexibility”. The study comprises 72 volunteers with several blood samples over 4 days. Drift correction was performed on metabolomics data from GC-MS of one day.

Additionally the workflow has been applied to preclinical studies, which are described briefly in the discussion-part.

[2] Accepted in August 2016

1.3. Realistic Expectations & Study Design

Knowing the challenges of metabolomics, realistic expectations and communication are essential. Sample size is a critical point in metabolomics studies because the number of variables is always a multiple amount of the number of samples. Another critical part of untargeted metabolomics studies is the description of distinctive groups within the study design.

Due to the preexisting high variability in the data thorough characterization of the clinical phenotype is essential to get valuable information from the metabolomics measurements. Potential bias can occur through various sources, such as sample collection and preparation, HPLC-HRMS. Analysis which, in some cases, can be corrected for, whereas in other cases, need to be included and described as restriction in the data interpretation.

A checklist will be described to place the data driven workflow as its best and to set the expectations to an appropriate and satisfying level.

2. Material and Methods

In the materials and methods section, materials, techniques and tools that are used for the composition of the statistical data driven workflow are described. The functionality and application of the workflow is presented in the results-section.

This chapter describes first sample preparation and UHPLC-MS Analysis for untargeted metabolomics, second the statistical methods used for data processing and data analysis and third the proof of concept studies used for the application of the data driven workflow.

2.1. Analytical Methods

The analytical methods are the most costly and a major part of metabolomics. But as they are not the main focus of my thesis, important cornerstones for the overall-understanding will briefly be sketched in the following. Sample preparation and HPLC-HRMS analysis were the same for all blood-samples.

Sample Preparation & HPLC-HRMS Analysis[3]

Serum samples were processed as described by Yuan et al (37). Briefly, 200 µl of serum were transferred to 1.5 ml tubes and centrifuged at 4°C for 10 min at 13,000 g. 800 µl methanol (cooled to -80°C) were added and mixed with the samples. Samples were incubated overnight at -80°C then centrifuged at 13,000 for 10 min and the supernatant was transferred to 1.5 ml tubes. Samples were evaporated to dryness by using nitrogen and reconstituted in 200 µl 30% methanol. Peak splitting was avoided by using small injection volumes for LC-HRMS analysis.

LC-HRMS analyses were performed with an Ultimate 3000 UHPLC system (Thermo Fisher Scientific, San Jose, CA, USA) coupled to a high resolution mass spectrometer Q-Exactive (Thermo Fisher Scientific, Bremen, Germany). The chromatographic separation was done by HILIC (hydrophilic interaction liquid chromatography) on a Luna NH2 column (2×150 mm; 3 µm; Phenomenex, Torrance, USA) following the procedure published by Bajad et al (38).

[3] Published in PLOS ONE (accepted in August 2016)

HILIC retains hydrophilic compounds and is ideal for polar low molecular weight compounds in contrast to the more commonly used reversed phase chromatography (37).

Separation was performed using eluent A: 20 mM ammonium acetate + 20 mM ammonium hydroxide in 95:5 water: acetonitrile, pH 9.45; Eluent B: acetonitrile. The gradient was as follows: t = 0 min, 85% B; t = 15 min 0% B; t = 20 min 0%B; t = 22 min 85% B; t = 37 min 85%B. Flow rate was set to 150µl/min. Full scan spectra were recorded in positive and in negative electrospray from m/z 70 – 1050 with a resolution of 140,000 (at m/z 200).

Quality Control: 10 µl of each sample were mixed together to generate a pooled quality control sample (QCs). QCs and solvent blank samples (BLs) were injected sequentially in-between the human serum samples. BLs, each followed by a QC, were measured after every third serum sample.

Identification and annotation of metabolites: Metabolites were identified according to Sumner et al. (39) *(1) Compounds* were identified by accurate mass and retention time in comparison to reference standards. (*2) Putatively annotated compounds* were annotated by accurate mass comparison using freely available metabolite databases (HMDB, KEGG, Metlin) (40,2,41–44).

Annotation of metabolic features

Metabolic features were annotated via mz-mass in databases such as `The Human Metabolome Database` (HMDB)[4] and the Joanneum Research database, which is in a permanent growing process. Annotation is the lowest level of identification of the features (27,48). A discussion about the level of identification is currently ongoing and a new guideline hasn't been published yet. 2014[5].

[4] http://www.hmdb.ca/
[5] http://interest-groups.metabolomicssociety.org/viewforum.php?id=13

Identification of discriminatory metabolic Features via MS-MS

The statistical workflow delivered features that discriminate between the time-points, groups or interventions. Annotation is the lowest level of identification of the features (27). For biological interpretations a higher level of identification is needed. Therefore, the discriminatory features were verified in MS-MS analysed human serum samples. MSMS-analysis delivered full scan of masses. As the feature identification is still a manual process more than 150 Mass-Identifications are not feasible in a reasonable amount of time and manpower. Therefore, the here presented approach is based on the following argumentation:

- Up to150 masses are reasonable to be identified via MS-MS
- Selection of metabolic profiles that are of special interest from a medical point of view

Most of the features were usually found in the MS-MS Spectrum. Over one third of the identified features were lipids, such as phosphatidylcholine (PC), phosphatidylinositol (PI) and triglycerides (TG). Also amino-acids, peptides and carnitines are among the identified features.

2.2. Statistical Methods for the data-driven Workflow

The steps in the data processing workflow need to be coordinated with the requirements of the measurement system to identify and compensate for bias from various sources such as sample collection and preparation, HPLC-HRMS Analysis.

The data-driven workflow includes filtering-steps and one drift correction step and statistical analysis. The filtering steps cope with artefacts, system impurities, bad signal-to-noise ratios and instable features and the drift correction is based on the Quantile Regression Models, established by selecting appropriate parameters. These parameters were found by following a selection procedure, based on the Quantile Regression Theory.

The whole data-driven workflow was programmed in the freely available statistics software R (R version 3.0.1). The measurement settings and the structure of the pre-processed data, as well as the subsequent steps of the data processing workflow will be described in the following. Filtering and drift correction steps were performed by means of the QCs and BLs and the order of the serum samples was randomized to avoid time-dependent bias. (Figure 2)

Figure 2: Sample sequence (BL: Blank samples, QC: quality control = pooled samples, H: human serum samples)

Measurement criteria is the CV (coefficient of variation) measured with the same sample (pooled QC) over the whole measurement period. The critical value of a variation of the QCs is CV > 0.3 (49).

Pre-processed data, which have already undergone peak detection, -matching, –alignment by using the R-package XCMS (50,51) and parameter-optimization IPO (52), are obtained in matrix form with certain intensities (representing the area under the curve = AUC) for each feature at a specific mass and a specific retention time. The matrix contains information on the metabolic features (columns) and on the samples (rows) (example is shown in Table 1). This matrix is the working-basis for all further data-processing steps.

Table 1: Segment of a feature-sample- matrix, number in the cells are AUC of the MS peak

Sample	MF1	MF2	MF3	MF4	MF5	…1000 MF
BL	6476309	17000303	743447389	15040872	1251612	
QC	9815286	12034871	1657779884	10148120	1799753	
Human Sample	6730046	12407047	790080306	9918036	275537	
Human Sample	6015698	9002199	665440607	7425585	694401	
Human Sample	7460523	14097114	914826248	12105009	287492	
…xxx Samples						

2.2.1. Filtering Steps

QC zero: Zero-intensities stem from artefacts and were therefore excluded. Because pooled samples are mixtures of all biological samples, they must show all possible feature signals. A signal, which does not appear in the QC-spectrum but in the spectrum of the biological sample is consequently considered as an artefact.

System-Peak-Filter: The measurement system produces so-called system peaks that derive from artefacts or system impurities. To detect these system peaks, BLs were induced after every fourth sample (Figure 2). Features were defined as system peaks, when BLs and QCs were significantly correlated (corr.test: p.value < 0.05) and the level difference was not significantly high (paired t-test: p.value >0.05).

Blank-Filter: BLs were used to identify system peaks and they served as a threshold for the signal-to-noise ratio. A feature was excluded as noise when it contained a Blank-signal higher than 10% of the class-mean signal in more than two sample-classes (e.g. patients and QCs).

Final Filter step: CV>30%: The final filter step was applied after the drift correction to exclude instable measured features. Those features with a high variation of QC-intensities (a coefficient of variation >0.3) were excluded. The coefficient of variation is based on the assumption of a parametric distribution, and therefore a non-parametric criterion (median/Inter Quartiles Range) was calculated additionally. Application of the filter steps is presented in chapter 3.1 *Modules programmed in R (Toolbox)*.

2.2.2. Drift Correction

Drift correction by Quantile Regression

The LC-MS measurements were performed with a very sensitive HILIC-FTMS device, which produces intensity fluctuations over time. These so called drifts need to be corrected, which can be done by using QCs (53–56). Different drift correction methods, exploiting the presence of QCs have been used (54,57,58). The choice of method depends on the data structure, the size of the study, the number of QC-samples and technical specificities of the measured features. Based on the QC-intensity-fluctuations (Figure 3) we built a regression model that works with the assumption that the intensity of a QC depends on its sample number. Several regression models were tested to fit the variation of the QC intensities. The various Quantile Regression models provide much better fits to the QCs than the linear regression model (Figure 3, left plot).

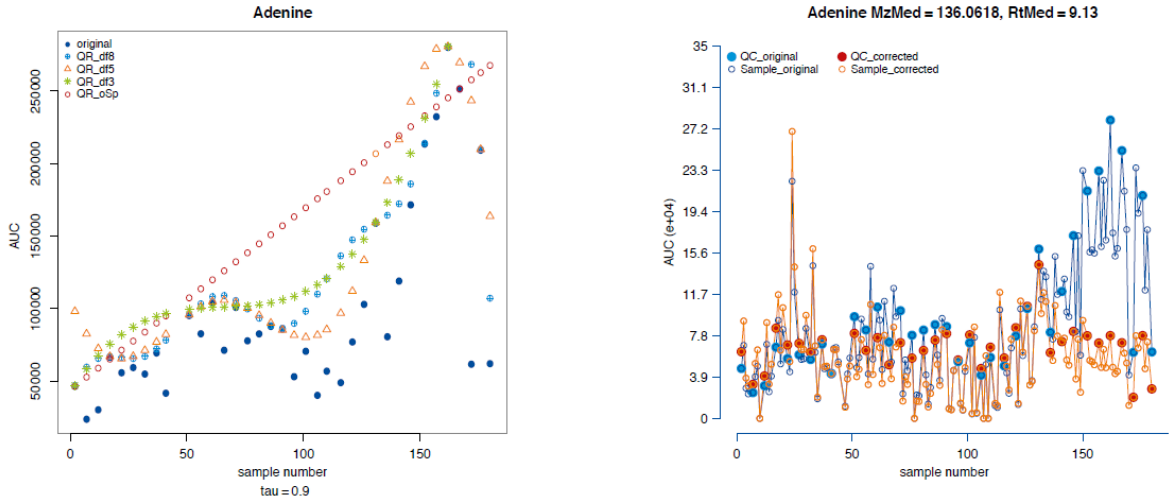


Figure 3: Drift correction using a Quantile Regression approach. Left: Model fits for the QCs. Right: Final correction using df = 5 and tau = 0.9

As the variability of features was high, smoothing by a locally adaptive regression technique was required. Quantile Regression is highly suited for data modeling with heterogeneous conditional distributions.

Theoretical Aspects: Quantile Regression

The Quantile Regression models the quantiles of the data distribution separately, in case, that dependencies are not the same in different quantiles. We used a nonparametric Quantile Regression to fit conditional quantile functions. The procedure fits a piecewise cubic polynomial with the number of one third of available data-points knots (breakpoints) arranged at the quantiles of the QCs-signals (59).

The R-function ´quant.reg´ [6](60) was applied in two steps to perform a drift correction. A model was built to configure the variation in time, following the QC-intensities in the sample order. The 90 % quantile of the QCs(y's) was estimated via nonparametric Quantile Regression, using regression splines depending on the sample number (x's). This procedure fits a piecewise cubic polynomial with 5 knots (df) (breakpoints in the third derivative) arranged at the 90 % quantile of the x's: rq (y ~ bs (x, df = 5), tau = 0.9). Through a multiplicative correction factor based on the median of the original QC-values, a further Quantile Regression model was estimated to attune all samples (Figure 1).

Two main criteria for the choice of a suitable regression method exist:

- The regression method must be adaptable for different numbers of QCs (minimum 10, dependent on the study-design)
- The regression method must be suitable for different kinds of drifts (linear, non-linear, jumping). The variations of the QCs differ, depending on the features. Therefore each feature requires an appropriate model.

A representative set of features was selected for Quantile Regression parameters to observe and choose suitable values for the parameters tau and df (tau means the quantile to model and df the number of knots, meaning inflexion points) for the final application of Quantile Regression in the data processing workflow.

[6] Roger Koenker (2013). quantreg: Quantile Regression. R package version 5.05. http://CRAN.R-project.org/package=quantreg

2.2.3. Multivariate Analysis (RF & PCA)

For the multivariate perspective, mostly Random Forests Models (RF) were applied, in various ways: unsupervised RFs to investigate potential clusters, supervised RFs to select most important features. Random Forests Models are already successful used in processing of metabolomics data (61–65). In general, RFs have a wide range of applications: they are used as regression and classification methods. Here, RFs were used for the analysis of metabolomics data as unsupervised and supervised classification methods and for feature selection. We combined their supervised and unsupervised classification abilities[7]- to show clusters and to detect discriminatory features.

Random Forests, used in the R-package randomForest (66), were based on the work of Adele Cutler and Leo Breiman (67,68). The method relies on decision trees, which are independently randomly distributed. These decision trees function as predictors. Each tree counts equally for the classification and at the end the votes are counted for one class in the whole forest. The trees are created by a bootstrap sample of size N as a training data set with replacement, from the original data, though. According to the number of independent variables (M), a number m << M is specified such that at each node m variables were selected at random. The variable that discriminated the best was selected to split the node. The number m was held constant during the 'Growing Process'.

The better classified each tree is, the lower is the error rate (ER) of the model. The greater the correlation between the trees, the lower is the ER of the model. The representation of the number of trees led to the number of trees needed to be created to minimize the errors in the model.

A special feature of Random Forests is the internal model validation using 'out of bag' (OOB) samples: Each tree is created using different bootstrap samples from the original data.

[7] A "supervised" method refers to a model that has already handled the class membership as information. "Unsupervised" models have no class information hereby classes and clusters are only visible, when clearly presented in the data. Therefore, as a first step an "unsupervised" model is built and, if trends are apparent, as the second step creates a supervised model to assess the strength classification, prognosis strength or variable-importance to which make up classes.

One-third of all samples is omitted from the sample and is not used to create the tree but only for validation of the classifications. Therefore, an intern cross-validation is executed. Hastie & Tibshirani stated S.593 (23):

'*For each observation $z_i = (x_i, y_i)$, construct its random forest predictor by averaging only those trees corresponding to boot-strap samples in which z_i did not appear. An oob error estimate is almost identical to that obtained by N-fold cross-validation"*.

The influence of variables on the model is indicated by Gini Importance and Mean-Decrease-Accuracy, calculated on the number of correct votes per variable and per node. The higher this value, the greater is the influence of the feature to the classification. This can also be represented graphically via a Variable-Importance-Plot (Figure 4).

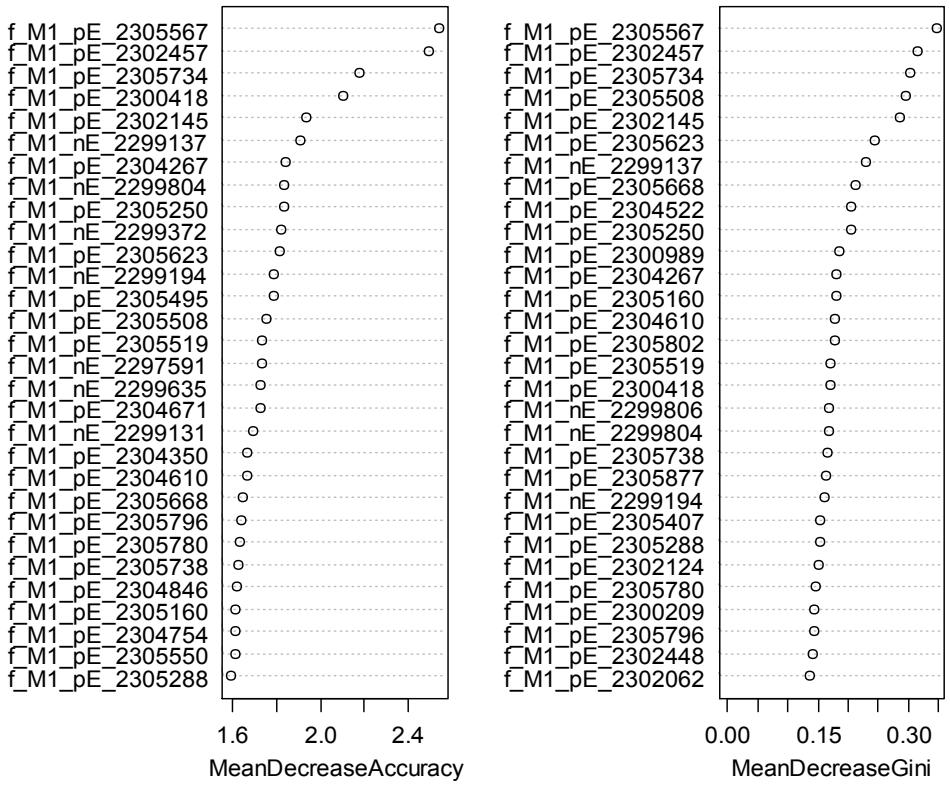


Figure 4: Variable- Importance-Plot shows the 30 most important features.

The presentation of the models with respect to potential sample clustering is done by means of MDS plots (Multi-dimensional Scaling Plot of Proximity matrix from Random Forests object). It represents the scaling coordinates of the proximity matrix of the Random Forests object (Figure 5):

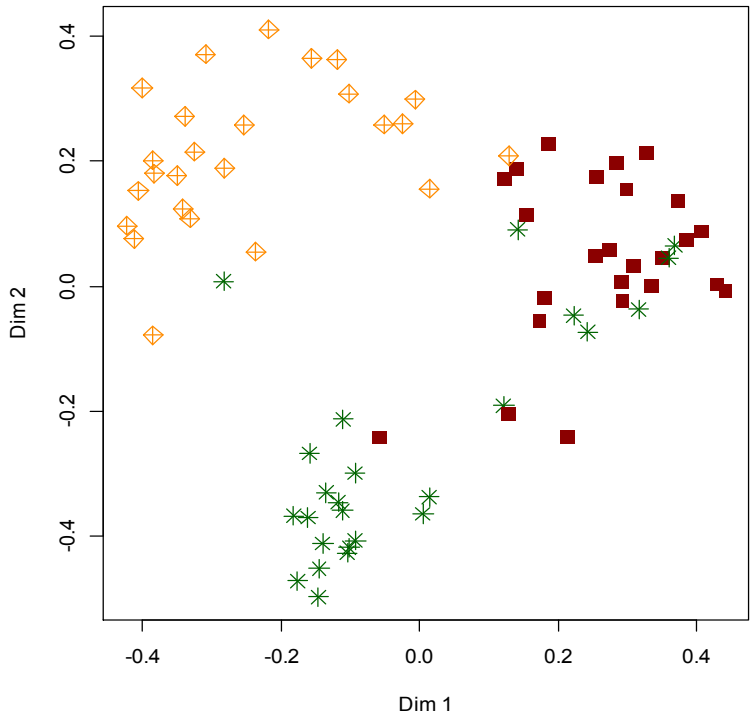


Figure 5: MDS plots of a supervised Random Forest object.

Random Forests Models have advantages over other classification models such as poor chance of overfitting and proper model prediction (62,64,69,70). As the method covers different steps of the statistical analysis process (supervised, unsupervised modelling and variable selection) the resulting outcomes can be directly combined to form a comprehensive statement.

Random Forests Models have already been applied in metabolomics for several years (62–64). They have been used as regression as well as classification methods. In this thesis, we utilized Random Forests Models as an unsupervised classification method to detect potential clusters by applying the R-package ´randomForest´ (66–68).

2.2.4. Univariate Testing

Applying univariate and multivariate technique yield different results, as has been reported recently (71). In this study a data-driven approach allows to show different perspectives on the data to get most valuable results. Statistical testing was done for features selection and not for hypothesis-testing in its classical sense of meaning.

To test features on its individual level student t-test followed by p-adjustment for multiple testing via false discovery rate[8] (Benjamini & Hochberg 1995) were applied. Frequency analysis of up- and downregulated metabolic features were done and tested via chi².Ratios between times were built to get a quantitative order. Boxplots and line-plots are the most widely used graphs to represent differences. Metabolic features with a tendency are named tendency features; meaning that p-values were < 0.05 in univariate testing but > 0.05 after p-value adjustment.

2.2.5. Evaluation of the data processing workflow

The quality of the data processing workflow was evaluated concerning batch-to-batch variations and time dependent drifts with two different methodological approaches. The first approach was a multivariate approach and consisted of unsupervised Random Forests Models used to detect potential clusters in the data. The second approach was an univariate one, based on the CV of QCs of representative metabolic features on the one hand and an overall CV of QCs over all metabolic features. The process with best amelioration was taken to be the optimal one.

[8] *"The "BH" (aka "fdr") and "BY" method of Benjamini, Hochberg, and Yekutieli control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error rate, so these methods are more powerful than the others."* R Core Team (2013)

2.3. Studies: Design, Data-sets & Background

This chapter characterizes the data-set and frame conditions of each study.

2.3.1. Cardionor

Cardionor was a clinical prospective, open, 2-years study, investigating treatment responses from T2DM patients associated with cardiovascular risk over 2 years. Samples from 81 patients were collected to perform metabolomics. The clinical study was published few times later (72) and presented clinical parameters with the change in carotis intimia media thickness (CIMT) after 2 years as primary outcome. For the clinical study 97 patients with type 2 diabetes and at least two insufficiently treated cardiovascular risk factors, i.e. HbA1c > 7.5% (58 mmol/mol); LDL-cholesterol >3.1 mmol/l or blood pressure >140/90 mmHg were included.

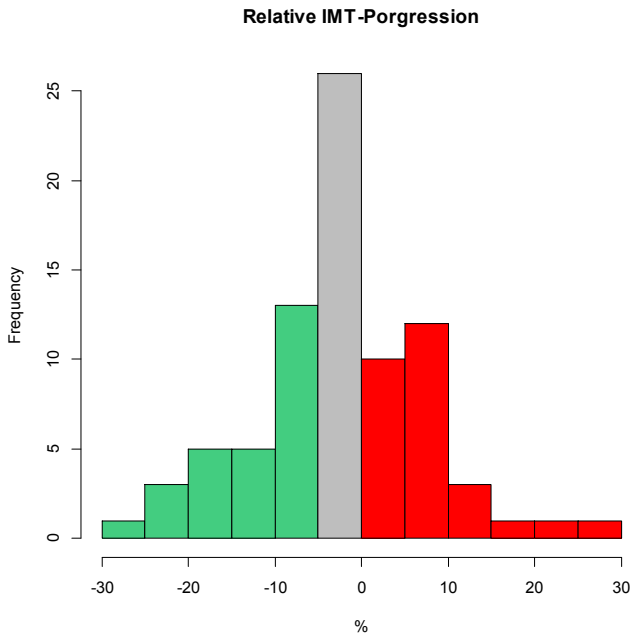


Figure 6 Distribution of relative IMT-Progression, colors are representing the tertile apporach

The patients were split with a „ tertile-approach“, a statistical quantitative method to build three groups with similar size, based on the relative IMT-progression after two years of therapy. The three groups were built without clinical differentiation. For statistical metabolomics analysis the intermedian group had been left out-to suggest clear separation between responders and non-responders.

In “doing metabolomics” we wish to differentiate metabolite profiles between one group and the other and to characterize these metabolite profiles for further research. All patients suffered all from diabetes. Diabetes type 2 (DMT2) is a large research field in metabolomics (73–78) - as it is in medicine as well. In this study appropriate measurements and filter modes were selected and time-dependent drifts were corrected. No separation between the responder-groups could be found. This study was financed by BMVIT project Met4CAD.

2.3.2. Bariatric Surgery[9]

Bariatric surgery is a clinical intervention study on patients who underwent bariatric surgery. Bariatric surgery is an important intervention for severe obese patients. The surgery is recommended as a very effective way of reducing weight (30). Not only the weight reducing effect is far-ranging recognized but also a short-term effect of insulin-sensitivity is discussed (31). The use of a metabolomics approach in this field of research is not new and rather well established (32–36).

Our aim is to detect differences in the metabolic profile before and after the intervention, if there are any, to describe them and associate metabolomics results with diabetes relevant outcomes like insulin resistance. Through a data driven approach, an unbiased way for exploratory investigations, hypotheses generating results for further research are delivered.

Serum samples were collected from 44 obese patients who underwent gastric bypass surgery: 29 females (BMI: 44.2 ±4.9, Age: 44±11) and 15 males (BMI: 45.2 ±7.1, Age: 48 ±15). Samples were collected at two study centers: Medical University of Graz (AUT) and Interdisciplinary Obesity Center in St. Gallen (CH) (Clinicaltrials.gov: NCT01271062). Serum samples were taken at three different time points: two to four weeks before the surgery (PRE), one to three weeks after the surgery (POST) and one year follow up after surgery (FU). Sample preparation was done according to standardized procedures. The study was approved by the respective local ethics committees and conducted in accordance with the principles of the Declaration of Helsinki, GCP-ICH and the requirements of the appropriate regulatory authorities. This study was financed by EAFSD, BMVIT.

[9] Published in PLOS ONE, accepted in August 2016

2.3.3. Metaprol

METAPROL is a pilot, cross-over study to generate a hypothesis why studies have shown an improved clinical outcome in end stage renal disease (ESRD) patients treated with post-dilution On-Line-Haemodiafiltration (OL-HDF).

The primary objective was to evaluate the influence of post-dilution OL-HDF versus haemodialysis (HD) on the metabolomic and proteomic profiles in patients with ESRD after a period of 4 weeks (short-term effect). The secondary objectives were to investigate metabolomic and proteomic profiles in pre- and postdialysis (PRE-POST effect) plasma samples and to investigate the plasma metabolomic and proteomic profile after 12 weeks (follow-up) as well as to investigate the effect of post-dilution OL-HDF versus HD on the specific chemistry parameters.

23 patients were included in the study; complete data from 18 patients were available and used for metabolomics analysis, split into two randomization groups; R1 resieving first 4 weeks of HD and 12 weeks of OL-HDF, R2 the other way round (Figure 7). The study was coordinated by Prof. Dr. Alexander Rosenkranz and supported by Fresenius Medical Care.

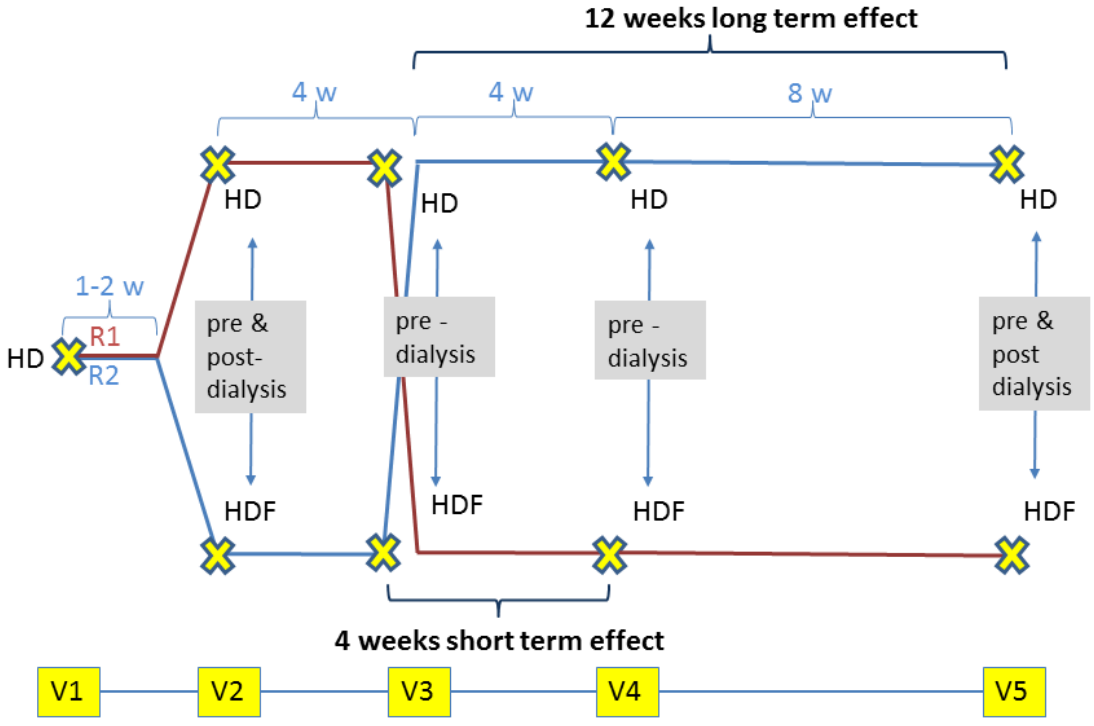


Figure 7: Study Design of metaprol study (HDF=OL-HDF)

2.3.4. Nutritech

"Nutritech" is an EU project[10] that aims to quantify the effect of diet on "phenotypic flexibility". This includes *"all underlying mechanisms and physiological processes of adaptation when homeostasis is challenged. Methods will in the first instance be evaluated within a human intervention study, and the resulting optimal methods will be validated in a number of existing cohorts against established endpoints"* (79).

The study comprises 72 volunteers including 38 women and 34 men with an average age: 59.7 ± 3.8 years and an average BMI of 29.6 ± 2.9.

All volunteers were asked to provide a fasting blood sampled (day 1) and went through an oral glucose tolerance test (day 2) when 7 blood samples were collected over 4 hours; a mixed meal tolerance test (day 3) when 7 blood samples were collected over 8 hours and a mixed meal tolerance test + physical activity (day 4), with the collection of 6 blood samples over the course of 6 hours.

OGTT consisted of 75 g of glucose; mixed meal consisted of 75 g glucose + 25 g protein + 70 g of palm oil. The physical activity consisted of a 40% VO2max during 30 minutes starting just after the drinking of the mixed meal. These 3 tests were performed before and after a lifestyle intervention for 13 weeks. During this 13 weeks period 40 volunteers went through a 20% energy restriction, while the other 32 had no energy restriction but were advised to eat more saturated fat in order to "mimic" the European diet. In total, 3024 plasma samples were supposed to be collected. The actual sample number is slightly smaller because some tests were not completed.

The work presented at the Metabolomics Conference in Washington (Poster, Figure 36) was the impulse for a cooperation with Prof. Hannelore Daniel, Professor for Nutrition Physiology, School of Life Sciences Weihenstephan, TUM - Technische Universität München.

[10] Nutritech - a Framework 7 project of the EU commission

3. Results

Result of the thesis is a data driven workflow for untargeted metabolomics. As it consists of a tool box, its application and adaption is described based on examples from clinical studies. The aim of the workflow can easily be modulated and applied on diverse metabolomics studies, as it consists of a tool-box.

The data driven workflow has been developed for feature selection, these features describe the difference in the metabolic profile between groups and/or time points. This data driven workflow was applied and optimized in preclinical and clinical studies (JR-Metabolomics-Projects from 2011-2016). Results from three selected clinical studies are presented in the following section as a proof of concept. One part of the data driven workflow (namely data processing) was also applied to GC-MS data of an EU project (Nutritech) to exhibit the versatile range of application.

A schematic representation of the workflow is given in Figure 8: XCMS data comprise more than 2000 metabolic features, these metabolic features contain still various potential bias that is corrected via filtering steps and drift correction. This results in a matrix of around 1000 “proper” metabolic features. This data-set is the basis for the feature selection process. Depending on the group distinction, a set of metabolic features that describe the differences in metabolite profiles is derived through random forests and univariate testing. At large, round 10% of metabolites can be identified, resulting eventually in putative markers.

The modules of the workflow will be repeated and redone; depending on the study design, interdisciplinary communication and data availability. The data driven approach is best applied in appropriate framing conditions. The whole metabolomics workflow consists of analytical and technical parts and reflection parts concerning study design literature research and data interpretation followed by additional data analysis. The specific challenge is the combination and timing of both parts with the persons involved.

Research question and study design determine analytical methods as well as data processing and statistical analysis. Therefore an interdisciplinary communication right from the project beginning is indispensable for meaningful results.

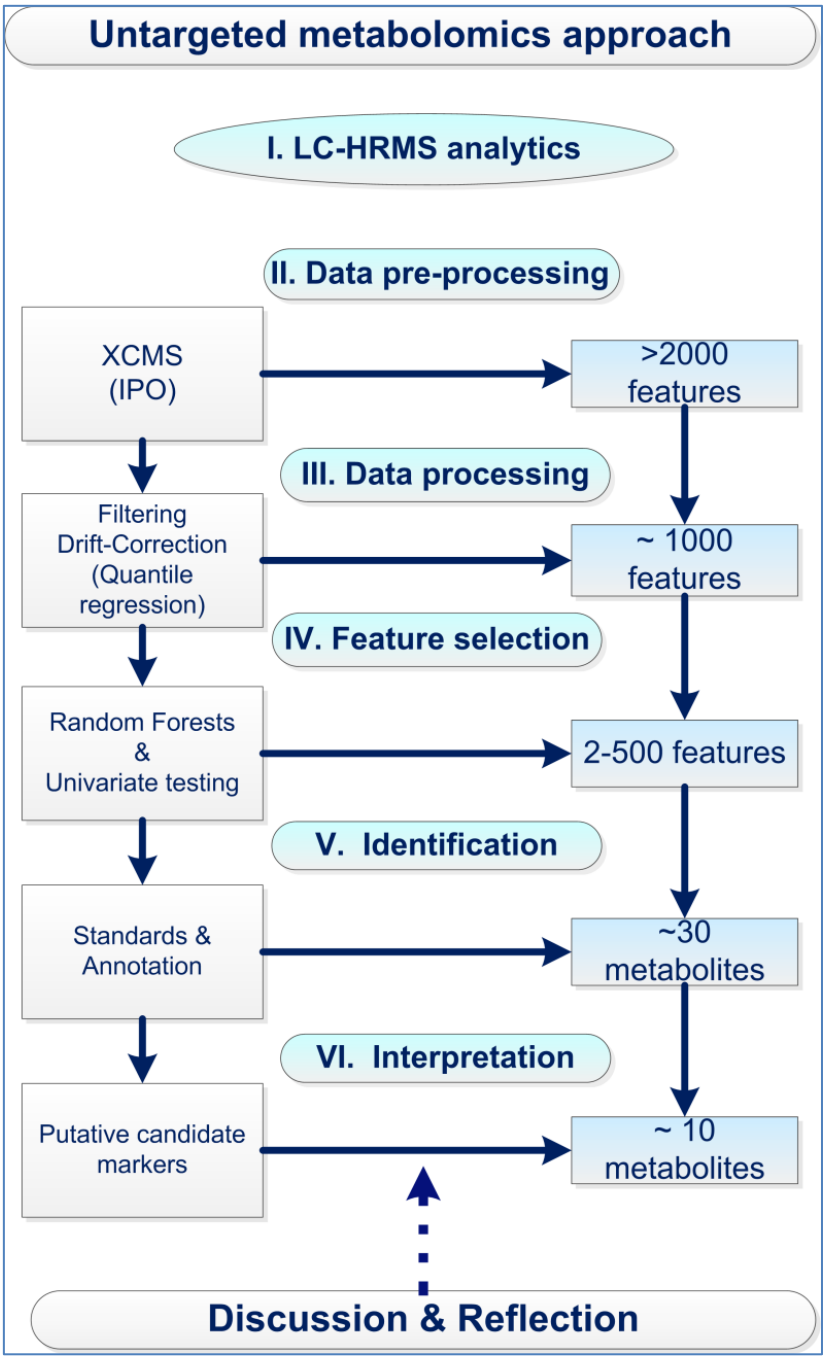


Figure 8: Data driven metabolomics approach

3.1. Statistical data-driven Workflow

The data-driven workflow is programmed in R and is presented as a toolbox, applied in XV steps (Figure 9). Each step requires an active check and decisions for the further process. In the following a short overview of the developing process of the applied workflow is described.

The **data import** had to be prepared, because two sorts of files are imported: the annotated file and the intensity file. For the annotated file all ‘‘ and ‘ (like “*HMDB29185 5-(3',4'-Dihydroxyphenyl)-gamma-valerolactone*” or “*HMDB37436 Kaempferol 3-(2'',6''-di-(E)-p-coumarylglucoside)*”) have to be removed as R will not read a “.tsv” file in the correct way. **Data check** for plausibility after data import was an important task; during data pre-processing errors can occur like an implausible number of features or lack of samples, inappropriate class-order. The sample names differ from file to file. Therefore scripts have to be checked for such differences.

For clinical data, for example, possibly randomization number might differ from subject-Id which refers to the intervention or treatment – which further is crucial for statistical analysis. **QCs** and Blanks (**BLs**) must be in the same structure to perform filtering and drift correction. If QCs were often stated as 0 – there might have been a problem with the measurement itself - the QC was excluded which has consequences for drift correction.

Drift correction performs best when the distance between the QCs are equal and do not “jump”. The drift correction should fit most of the metabolic features and its success was measured by the **CV** of the QCs – the number of improved metabolic feature and the overall improvement served as indicator for the most suitable drift correction. If a selection of metabolic features seemed very important, drift correction should be adapted individually decided from case to case of specific metabolic features for this selection. The selection could be based on a specific range of Mz (mass) and/or Rt (retention time) or already identified metabolites. A scrolling through the **drift correction-pdfs** gave a visual control of potential patterns that indicate artefacts. The final data-set “dat” was checked for **potential bias** that is known from patients’ characteristics.

This occured by visual control of multivariate methods like **PCA** – the dots on the multidimensional scaling plots are colored for study-centers, age and gender-groups or sample sequence. If a bias was clearly visible, a feature-selection process was needed to correct the bias (additional variable in modelling for interesting metabolic features) or remove “bias-features” as performed by Luchinat et al. (80).

After again **filtering** the data of highly correlated metabolic features and bias-features, the **feature-selection process** starts. This was done via multivariate supervised **random forests** models, unsupervised random forests models, PCA and univariate testing via t-test or ANOVA-models followed by p-value adjustment via **false-discovery-rate**: Benjamini Hochberg procedure. The intersection of RF-hits and univariate hits on different significance levels was saved in “hits-RF_univariate” and exported in Pdfs of boxplots and Lineplots for each hit-feature.

PCAs and RFs were built based on these hits to compare all filtered feature and hit-features. **Annotation** of significant features is based on several metabolite databases like KEGG and HMDB. The final approval of the **annotation** and the **identification of metabolites** had to be done by a chemist. Assuming an important number of metabolic features and metabolites were selected to be **important** for the scientific question; **pattern analysis** would start.

The pattern analysis evolves on the basis of questions like: How many metabolic features behave in the same way, where are differences, are there any frequencies significant. **Ratios** of changes were calculated and compared. Building **subgroups** provided better understanding of the metabolic profiles.

The **statistical modeling** was completed when a group of interesting metabolic feature is selected. Here, the integration of **clinical data** was meaningful to answer more specific questions. The **feature selection process** including clinical parameters might be redone, based on hypotheses that came up during the discussion process. For **publication** data needed to be exported and formatted to fit the publications requirements. The formation of data and the descriptions of methods were further adapted to the requirements of the metabolomics platform “**Metabolights**”, as this platform is state-of-the-art for data-publishing of metabolomics studies.

Statistical data-driven workflow for untargeted metabolomics:

Modules programmed in R (Toolbox)

I. Data Import: read the data from JR-interne database into R
II. Data Processing: bringing data in the following structure:
  a. Data-file. Rows: samples, columns: features (number + Ionization Info)
  b. Mass-file: containing information about retention time, Mzmass, variation and ionization and possible annotations
  c. Class: variable containing information about class-dependencies of each sample
III. Data-Check
  a. number of samples
  b. number samples per class
  c. number of patients/ time-points
IV. QC and Sample-Sequence
  a. Pdf-file containing plots of each feature to get a first impression of a possible drift or pattern (for example one zero-sample)
  b. Excel-file containing number of zeros per sample and class
  c. Calculation of QC- CV per feature
  d. Check for patterns
V. Feature-filtering (see methods-section)
  a. QC zero
  b. System-Peak-Filter
  c. Blank-Filter
VI. Sample-filtering
  a. Looking for QC-outliers (> 50% Zeros)
  b. Deep sensitivity loss (jump)
VII. Drift Correction via quantile regression (> 10 QCs)
  a. Taus: 0.3, 0.5, 0.8
  b. Comparison between drift-correction quality on CV-distribution and number of approved features based on CV
  c. Pdf with scatter plots to compare drift-corrected and original data for each feature
  d. Final Filter step: CV>30%
VIII. Bias check on filtered data
  a. PCA coloured for classes, study-centres, sample-sequence, gender, others depending on study design and research question
  b. Histogram over tested normality per feature (>80% features with p-value)
  c. Check for number of redundant features (corr>0.95)
IX. Feature-selection (based on research question: discriminatory for specific group or time)
  a. Multivariate unsupervised methods: PCA, RF
  b. Multivariate supervised methods: RF
  c. Univariate testing (ANOVA, t-test) followed by p-value adjustment via false discovery rate (Benjamini Hochberg)
X. Analysis of annotated features and known metabolites
  a. Annotated data file ready to merge with significant hits
XI. Pattern analysis
  a. Frequency Analysis
  b. Selection of features following unidirectional trends over time
  c. Ratios and mean-values (Box-plot pdf)
XII. Statistical Modelling
  a. Subgroup-analysis (MzMed-groups, RtMed groups)
  b. Random mixed effect models
  c. Integration of clinical parameters
  d. Specific analysis of annotated and identified metabolites
  e. Correlation and Boxplots
XIII. Explicitly search for known candidate markers
XIV. Statistical Modelling for medical, biological questions – rose up during discussion
XV. Export essential data files for publication (for example Metabolights)

Figure 9 Statistical data-driven workflow: Modules programmed in R

3.2. Study Results

In the second part of the results section the main outcomes of the chosen studies is represented.

The following describes results from metabolomics studies in a chronological order. The **Cardionor** study served as basis for analytical and statistical-methodical development. **Bariatric surgery** has been published with a strong medical emphasis. A lot of effort went into this study as more than one year of publishing was needed and study protocol dated from 2009. Therefore results from this study will be described in more detail. **Metaprol** was a cross-over intervention study to detect changes in metabolite profiles depending on the dialysis modalities. For the statistical analyis of Metaprol frequency analysis enlarged the toolbox of the statistical data driven workflow. **Nutritech** serves as an example for successful drift correction on a large data sets with few QCs.

3.2.1. Cardionor

For the Cardionor-study samples from 81 patients were used for metabolomics analysis. The data set served to establish drift correction based on quantile regression to minimize batch-dependencies and time-dependent drift:

The first application of the data processing workflow resulted in a feature reduction of more than 50% (initially detected features: 12000). Time dependent variations over the QC pool samples (>0.4 to <0.25 CV) were reduced (Figure 10). Batch dependencies were also reduced as shown in unsupervised Random Forests before and after the drift correction (Figure 11)

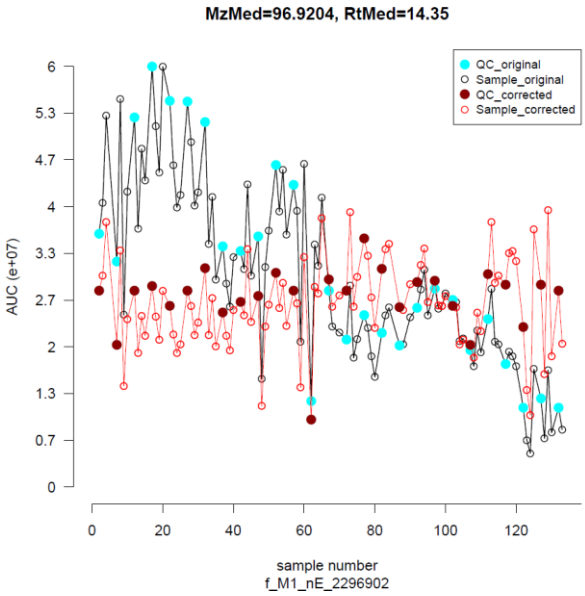


Figure 10: Feature intensities shown as peak area versus sample run order.

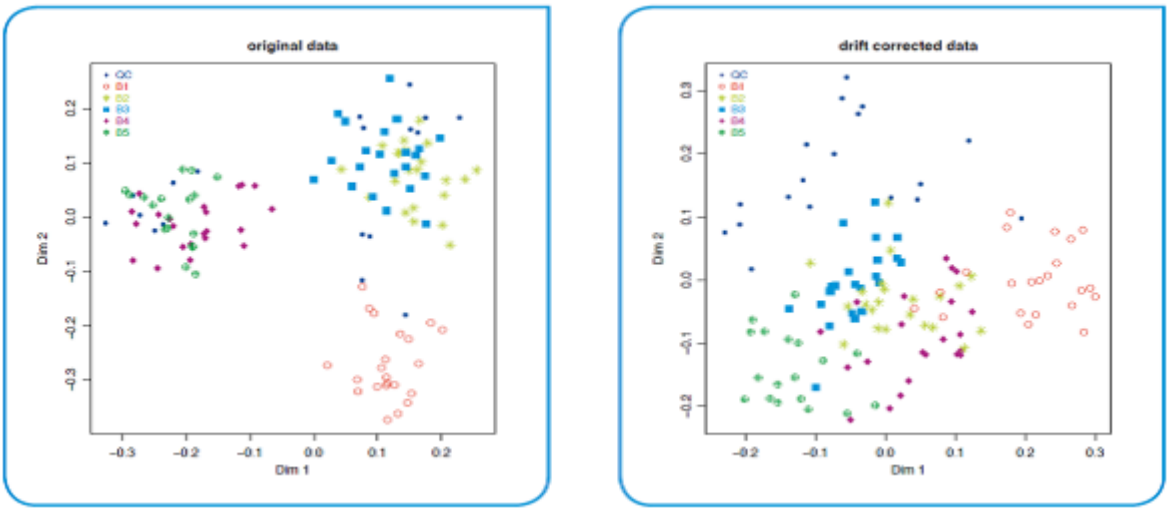


Figure 11: Unsupervised Random Forests Model calculated from the original data (left), and from drift corrected data (right).

We used 500 trees as default parameter to construct the unsupervised Random Forests models (Figure 11). The data-processing steps were programmed in a modular structure, which allowed the look into different perspectives concerning batches and overall study considerations. Depending on the study design, the workflow is adaptable in a semi-automated way.

The initial clinical hypothesis to find metabolic marker that distinguish responder from non-responder could not been answered (Figure 12).

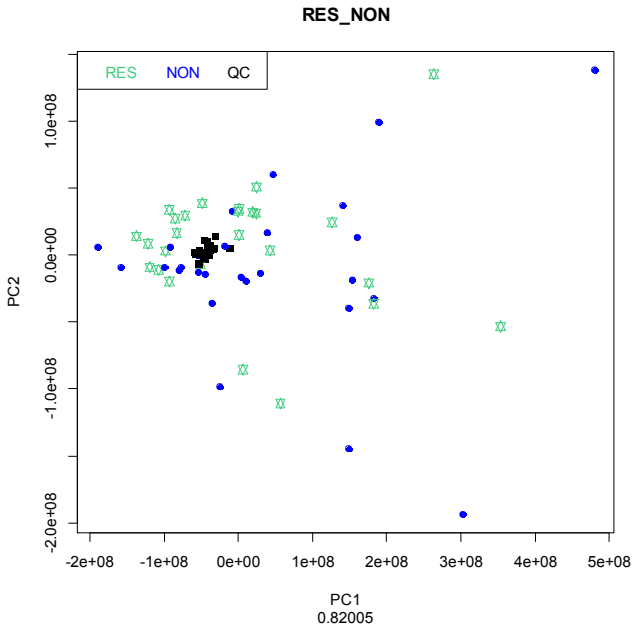


Figure 12: PCA showing no clustering between responder and non-responder.

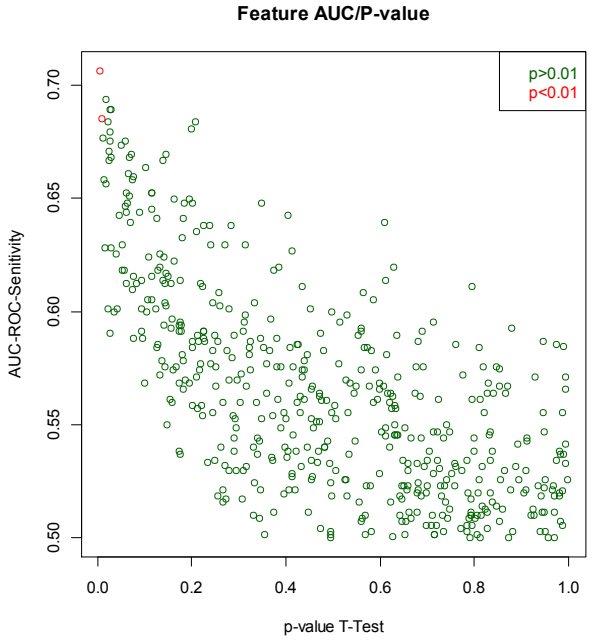


Figure 13: Scatter-Plot of metabolic features with AUC-Roc Sensitivity and adjusted p-values of t-test

3.2.2. Bariatric Surgery

All results as well discussion have been published in PLOS ONE.

Serum samples from 44 obese patients (25 patients from the center in Austria and 19 patients from the Swiss center) included in the study were analyzed (for characteristics of patients see Table 2).

Table 2: Patients characteristics

	PRE	POST	FU	paired-t-test PRE-POST	paired-t-test PRE-FU
Gender (male/ female)	15/29	-	-	-	-
Age (years)	46.8 (11.3)	-	-	-	-
Weight (kg)	126.4 (19.5)	117.2 (18)	86.3 (13.4)	<0.001	<0.001
BMI (kg/m^2)	43.9 (5.4)	40.8 (5.2)	30 (4.4)	<0.001	<0.001
HbA1c (%)	6.5 (1.3)	6.1 (1)	5.6 (0.8)	<0.001	<0.001
Sys	132.6 (15)	123.7 (14.1)	126.4 (17.1)	<0.001	0.029
Dias	83.7 (10.9)	77.3 (9.1)	77.8 (11.1)	<0.001	0.003
Chol	181.5 (39.7)	-	146 (27.9)	-	<0.001
HDL	50.2 (16.7)	-	49.8 (14.5)	-	0.806
LDL	47.7 (53.2)	-	36.8 (39.4)	-	<0.001
TG	159.2 (101.3)	-	88.8 (32)	-	<0.001

Untargeted metabolic feature selection resulted in 177 relevant metabolic features that represent short-term and long-term changes, out of which 32 metabolites were successfully identified and putatively annotated. A simultaneous explicitly search further identified 4 additional metabolites (BCAAs) affected by bariatric surgery. Supervised RFs showed a clear separation between the samples taken before and after surgery, with a class error of only 6% and 11%, respectively (Figure 14).

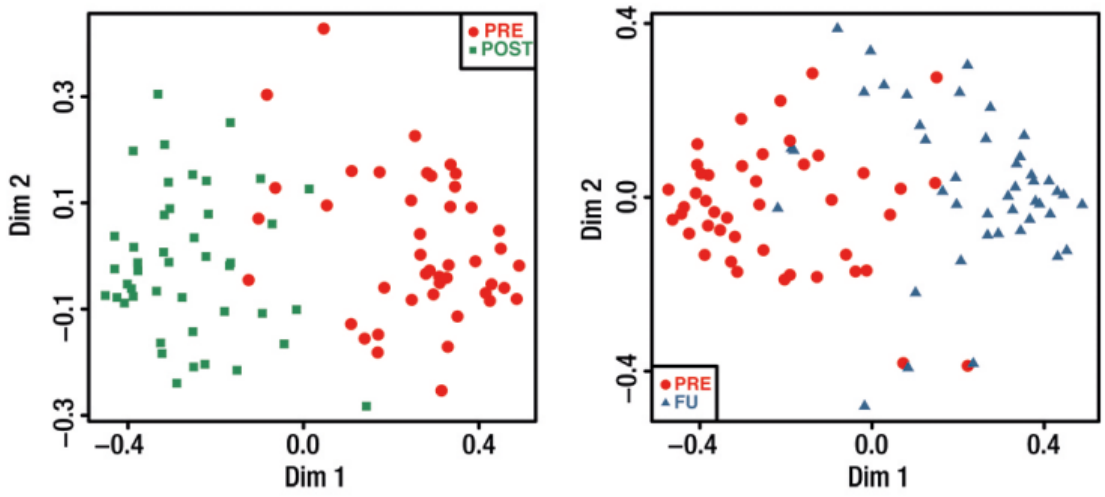


Figure 14: MDS-Plots from supervised random forests, showing clustering between before and after the surgery.

As a visual control, unsupervised RFs were built based on the selected 177 metabolic features. Samples taken before the surgery (PRE) clustered closer together than samples taken after the surgery (POST and FU) (Figure 15).

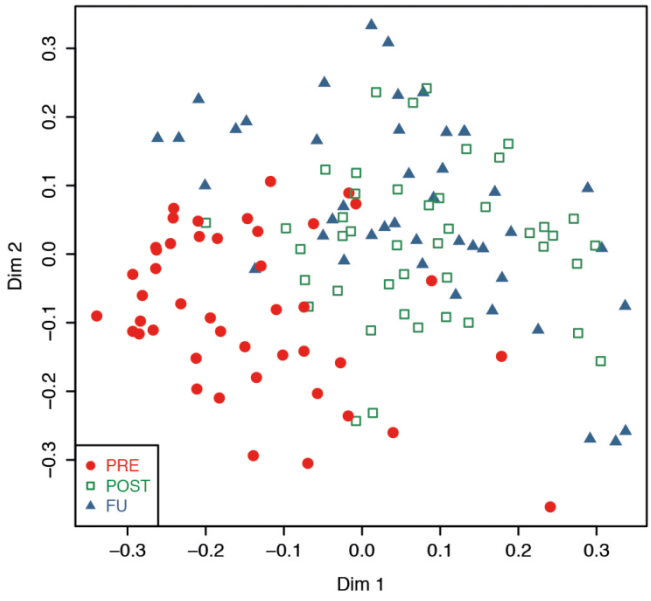


Figure 15: MDS-Plot of unsupervised Random Forests using 177 selected metabolic features from all three sampling points.

8 identified metabolites were linked to CVR factors (81,82): TMAO, indoxyl sulphate (increasing trend), choline, alanine, phenylalanine, tyrosine, valine, leucine/isoleucine (decreasing trend) (Fig. 4, Table 1).

The interpretation of the identified metabolites was based on trend-patterns which were assigned to one of four different pattern groups. 9 out of the 36 metabolites showed unidirectional trends in intensities (either increasing or decreasing): trimethylamine-*N*-oxide (TMAO) and indoxyl-sulfate, glycine and PC C40:7 (phosphatidylcholine) (increased after surgery), or branched chain amino acids (BCAA) choline, tyrosine, alanine and phenylalanine (decreased after surgery) (Table 2).

Examples for “V-pattern” were shown be the following metabolites: creatine, ornithine, tryptophan and LysoPC C16:1, LysoPC C18:2. “Λ-pattern” were shown in hydroxyisobutyric acid and acetylglycine (Table 3). All significantly changed metabolites are summarized in Table 2 and Table 3.

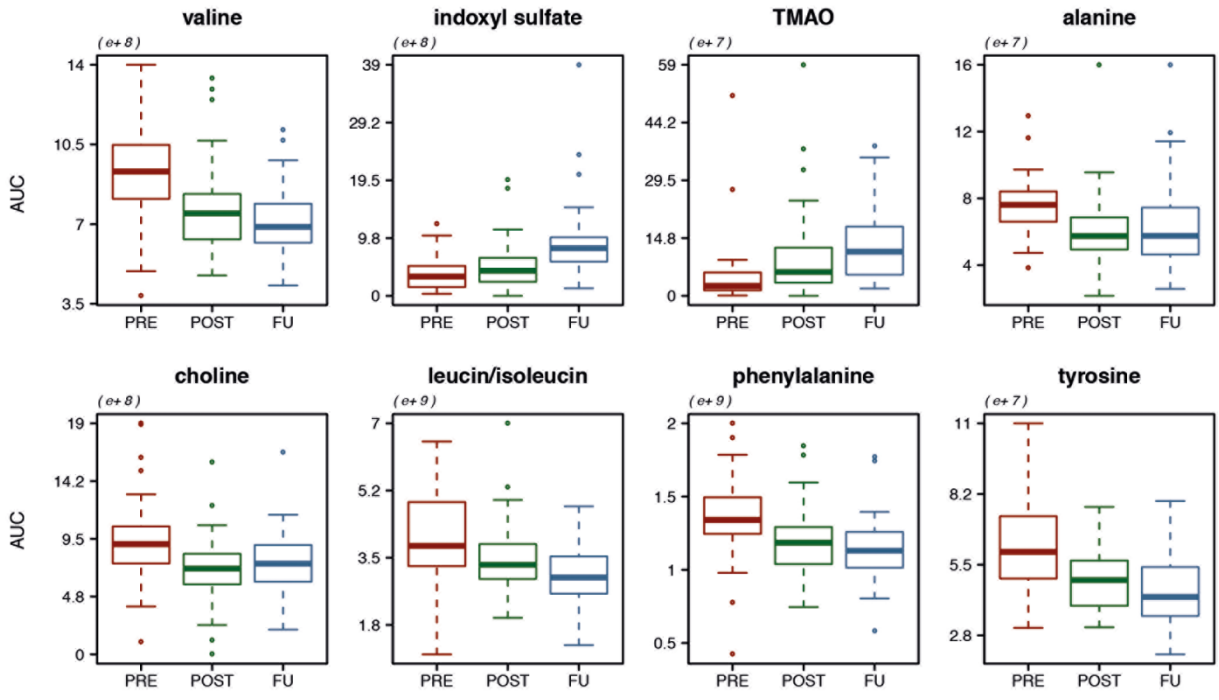


Figure 16: Boxplots of peak-AUC metabolites related to CVR for three different sampling points.

Table 3: Unidirectional trends of changes in the intensities (peak-AUC) of identified metabolites before and after bariatric surgery. Metabolites in bold have previously have been associated with CVR.

Metabolite* (Ionization-mode)	MzMed	RtMed	p-value** PRE-POST	p-value** PRE-FU	Ratio*** PRE,POST	Ratio POST, FU	Ratio PRE,FU
decreasing trend							
Alanine (+)	90.0556	12.20	<0.001	0.019	0.8	1	0.85
Choline (+)	104.1076	10.06	<0.001	0.003	0.74	1	0.79
Leucine/Isoleucine° (+)	132.1022	9.45	0.003	<0.001	0.87	0.89	0.77
Lysine (-)	145.0968	13.02	0.036	<0.001	0.91	0.97	0.88
Oxovaleric acid (-)	115.0384	9.90	<0.001	<0.001	0.81	0.91	0.74
Pentoses (-)	149.0441	9.96	0.127	<0.001	0.93	0.84	0.78
Phenylalanine (+)	166.0865	9.74	0.003	<0.001	0.88	0.95	0.83
Tyrosine (+)	182.0815	11.62	<0.001	<0.001	0.79	0.92	0.73
Uridine (-)	243.0617	7.20	0.004	0.006	0.82	0.99	0.81
Valine° (-)	116.0700	10.19	<0.001	<0.001	0.82	0.92	0.75
increasing trend							
Glutamine° (+)	147.0767	13.63	<0.001	0.003	1.17	1	1.13
Glycine° (+)	76.0400	14.2	<0.001	<0.001	1.89	1	1.85
Hydroxydecanoic acid (-)	187.1329	9.45	<0.001	<0.001	1.59	1.68	2.68
Indoxyl sulphate (-)	212.0013	9.13	0.067	<0.001	1.35	1.76	2.38
PC C40:7 (+)	832.5865	5.14	0.641	<0.001	1.04	1.34	1.4
Trimethylamine-*N*-oxid (+)	76.0764	11.88	0.022	<0.001	1.99	1.3	2.59

*details about category of identification according to Sumner et al. (39) are provided in the appendix Table 18

** unadjusted p-values from paired t-test, ° metabolites identified with explicitly search ***ratio based on mean-values

Table 4: Bidirectional trends of changes in the intensities (peak-AUC) of identified metabolites before and after bariatric surgery. Metabolites in bold have previously been associated with CVR.

Metabolite* (Ionization-mode)	MzMed	RtMed	p-value** PRE-POST	p-value** PRE-FU	Ratio *** PRE,POST	Ratio POST, FU	Ratio PRE,FU
		V-pattern					
Creatine (+)	132.0771	12.19	<0.001	<0.001	0.66	1.09	0.72
LysoPC C16:1 (+)	494.3249	4.83	0.077	0.288	0.85	1.29	1.09
LysoPC C18:2 (+)	520.3407	5.13	<0.001	0.863	0.68	1.48	1.01
Ornithine (-)	131.0815	12.54	0.004	0.019	0.83	1.34	1.11
PC C34:3 (+)	756.5550	5.27	<0.001	0.370	0.66	1.44	0.95
PC C36:5 (+)	780.5550	5.01	<0.001	0.006	0.67	1.21	0.81
PC C36:6 (+)	778.5389	4.61	<0.001	0.809	0.48	2.06	0.98
Sarcosine (-)	88.0386	11.27	<0.001	<0.001	0.78	1.1	0.86
Tryptophan (+)	205.0973	9.83	<0.001	<0.001	0.74	1.1	0.81
Uracil (+)	113.0351	6.98	<0.001	<0.001	0.75	1.04	0.78
		Λ-pattern					
Acetylglycine (-)	116.0337	12.97	<0.001	<0.001	2.78	0.74	2.05
Arginine (+)	175.1193	12.15	0.620	0.233	0.97	1.08	1.05
Carnitine (+)	162.1127	10.88	0.004	0.515	1.19	0.86	1.03
Hydroxyisobutyric acid (-)	103.0387	12.10	<0.001	<0.001	3.3	0.21	0.71
Leu Pro (+)	229.1548	9.19	<0.001	0.180	1.64	0.55	0.9
LysoPE C20:4 (+)	502.2936	8.27	0.022	0.227	1.16	0.94	1.09
Pantothenic acid (-)	218.1025	12.69	0.001	0.270	1.52	0.75	1.14
PC C38:6 (+)	806.5705	5.21	<0.001	0.020	1.31	0.88	1.15
Pyroglutamic acid (-)	128.0337	12.89	0.002	0.038	1.18	0.93	1.1
Threonine (+)	120.0660	13.11	0.602	<0.001	1.04	0.75	0.79

*details about category of identification according to Sumner et al. (39) are provided in the appendix Table 18
** unadjusted p-values from paired t-test, ° metabolites identified with explicitly search ***ratio based on mean-values

Metabolites linked to clinical outcomes

The median weight reduction at follow-up (FU) after one year was 37.7 kg (iQR: 16.25 kg). For relating weight-loss with metabolomics data, patients were allocated to a high weight loss (HWL) and low weight loss (LWL) group. The weight-loss ratio was calculated as: weight one year post surgery/weight at baseline=FU/PRE with a weight-loss median of 0.7. HWL was below the weight loss median and LWL was above the weight-loss median. From the originally identified metabolites, creatinine, ornithine, arginine and valine were significantly lower in the HWL group compared to the LWL group (Figure 17).

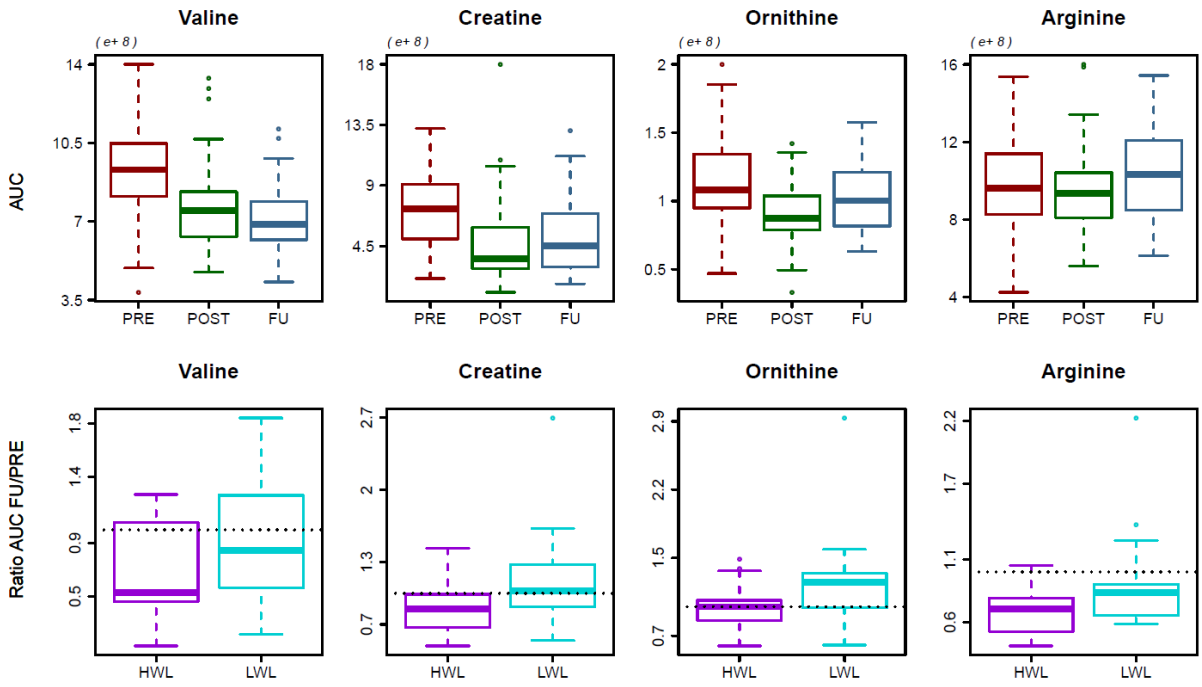


Figure 17: Metabolites with significant changes between high and low-weight loss patients.

Out of the 24 patients having T2DM at baseline, 9 patients were having a complete diabetes remission after one year. Diabetes remission was defined as an HbA1c below 48 mmol/mol (6.5%) without pharmacological treatment. Patients with a complete diabetes remission were younger (42 +/- 8 years vs. 55+/-10 years), had shorter diabetes duration (6+/- 7 years vs. 11+/- 7 years) but higher weights pre-surgery (138 +/-19 kg vs 124 +/- 22 kg) compared to non-remission patients. However, patients with diabetes remission also had a significantly larger weight reduction in the first year after bariatric surgery

From the originally identified metabolites, sarcosine, pyroglutamic acid, alanine and leucyl-proline showed a significantly larger decline in patients with complete diabetes remission compared to patients without diabetes remission (Figure 18).

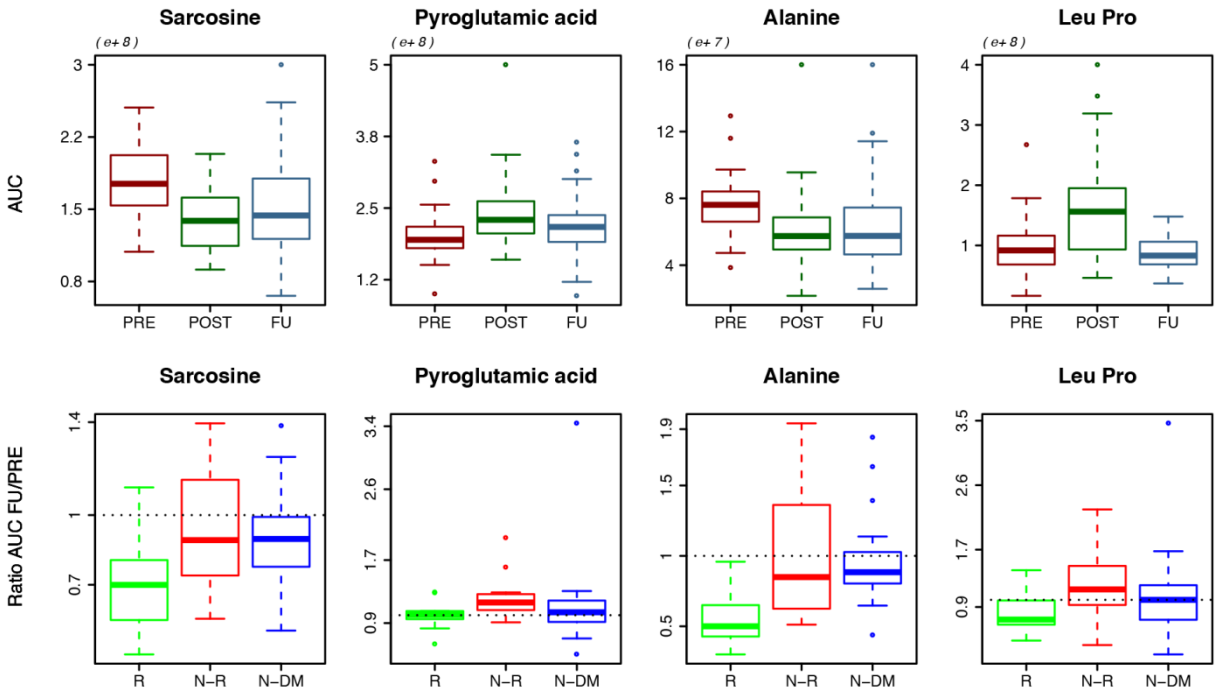


Figure 18: Metabolites showing a significant decline (FU/PRE) in diabetes remission (R) patients compared to non-remission (N-R).

3.2.3. Metaprol

23 patients were included in the study that compares one dialysis method (HD) with another (OL-HDF) with a cross-over design (see chapter 2 material and methods, Figure 7). Complete data from 18 patients were available for statistical analysis. Randomisation group 1 (R1) defines patients receiving first HD and at visit 3 OL-HDF and randomization group 2 (R2) defines patients receiving first OL-HDF and at visit 3 HD. The analytical methods and data pre-processing followed the same routine as described for the study “Bariatric Surgery” (chapter 2 material and methods). The study design and research questions were further complex with **pre-post** effect, **4-weeks short-term** and **12-weeks long-term** dialysis effects. For each comparison metabolomics analysis was done. The major outcomes are sketched in the following part.

PRE-POST dialysis shows pronounced effects, because the dialysis per se is a strong intervention and, therefore, the metabolic profile changes between before and after the dialysis in both treatments Figure 19.

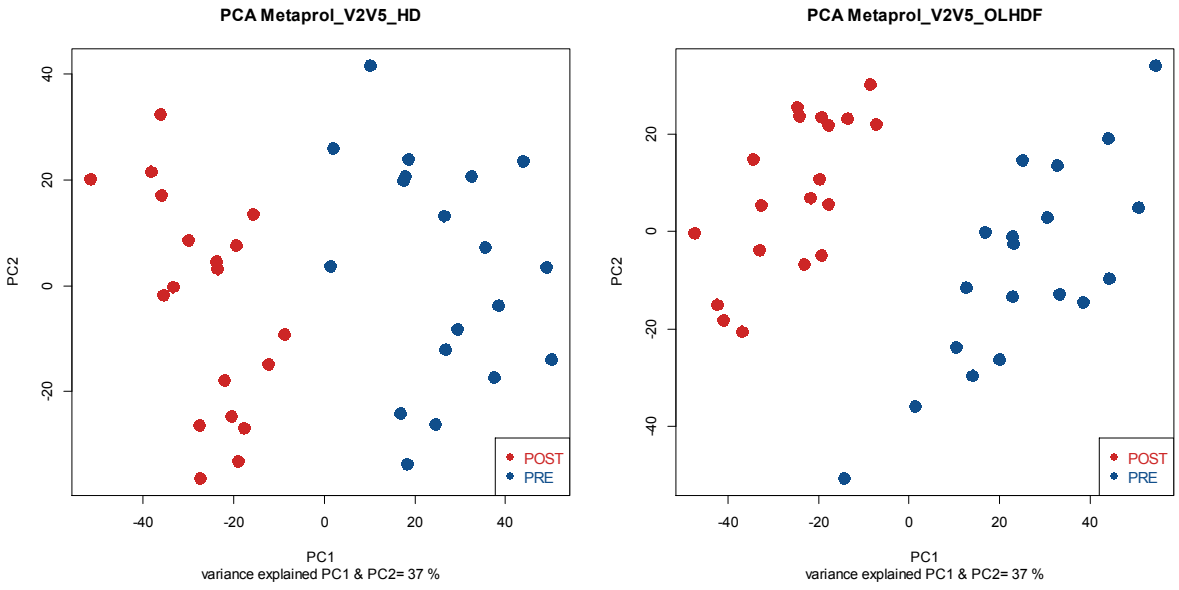


Figure 19: PCA for both treatments showing clear clustering between before and after dialysis.

1699 metabolic features were detected that differ significantly before and after dialysis in OL-HDF, compared to 1861 metabolic features in HD-treatment, 1408 in both treatments. The different metabolic features did not only show depletion from dialysis (Table 5) but also enrichment for annotated metabolic features (Table 6).

Uremic toxins have been searched for explicitly with a targeted approach - these are highly relevant metabolites from a nephrologist point of view.

Table 5: Metabolites showing a significant decrease after dialysis (depletion) in both treatments

PRE > POST			HD	OL-HDF
			mean ratio	mean ratio
Metabolite	Mzmed	Rtmed	POST/PRE	POST/PRE
D-Glyceric acid	105.018208	12.90835	0.45	0.47
Creatinine	112.050117	5.9854	0.37	0.37
Proline	114.05459	10.425292	0.70	0.72
L-Valine	116.070226	9.765358	0.72	0.72
L-Threonine	118.049543	11.219183	0.74	0.82
L-Lysine	145.096922	12.579692	0.74	0.80
L-Histidine	154.061005	11.759725	0.74	0.80
Allantoin	157.035511	8.458366	0.25	0.24
Uric acid	167.019884	15.086492	0.25	0.26
L-Arginine	173.10341	11.890783	0.73	0.79
Citrulline	174.087443	11.472333	0.41	0.41
Pantothenic acid	218.103121	12.174983	0.51	0.66
L-Cystathionine	221.059761	13.790258	0.19	0.07
Uridine	243.062135	9.256409	0.37	0.37
Deoxyguanosine	266.086196	9.085492	0.24	0.15
Inosine	267.072509	14.009084	0.11	0.11
Trehalose	341.109302	10.6890335	0.19	0.24
Trimethylamine N-oxide	76.0762881	11.846592	0.19	0.19
L-Alanine_Sarcosine	90.0555207	11.056392	0.69	0.71
L-Threonine_L-Homoserine	120.065807	12.832533	0.29	0.27
Ornithine	133.097361	12.681675	0.69	0.76
L-Glutamine	147.076499	15.955783	0.86	0.90
L-Lysine	147.112863	12.506017	0.83	0.87
L-Carnitine	162.112488	11.702367	0.33	0.34
Uric acid	169.035664	15.043325	0.21	0.20
Citrulline	176.103085	11.741208	0.48	0.47
L-Tyrosine	182.081247	13.520233	0.45	0.44
Cytidine	244.092859	8.980891	0.24	0.25
Uridine	245.076626	9.303875	0.33	0.32

Table 6: Metabolites showing a significant increase after dialysis (enrichment) in both treatments

POST > PRE			HD	OL-HDF
Metabolite	Mzmed	Rtmed	mean ratio POST/PRE	mean ratio POST/PRE
PI C34:2	833.520716	29.4395	1.20	1.30
PI C38:3	887.557975	32.8935	1.26	1.25
PC C32:0	734.569764	7.740875	1.36	1.27
PE C36:2	744.554161	5.2298	1.41	1.29
PC C34:3	756.553775	4.9602833	1.23	1.15
PC C34:1	760.575686	5.6940165	1.28	1.23
PC C36:5	780.553703	4.8050666	1.23	1.18
PC C36:4	782.569362	5.2141914	1.28	1.19
PC C36:3	784.584936	5.7889166	1.32	1.24
PC C36:2	786.600565	7.413025	1.45	1.30
PC C38:6	806.569228	4.914225	1.23	1.13
PC C38:5	808.584911	5.2556415	1.27	1.19
PC C38:4	810.60065	6.59825	1.44	1.31
PC C40:5	836.616813	6.6767335	1.40	1.21
PC C40:6	834.600738	6.050717	1.26	1.12

Table 7: Uremic Toxins showing a decrease (depletion) in both treatments

	Uremic Toxins	p_value	p_adjust	mean Ratio POST/PRE
HD	Indoxyl.sulphate.NEG	0.1213	0.3638	0.62
	p.Cresyl.glucuronide.NEG	<0.001	<0.001	0.16
	p.Cresyl.sulphate.NEG	0.0493	0.1479	0.74
OL-HDF	Indoxyl.sulphate.NEG	0.1184	0.3551	0.84
	p.Cresyl.glucuronide.NEG	0.0022	0.0067	0.14
	p.Cresyl.sulphate.NEG	0.0230	0.0691	0.78

A metabolic feature is defined by retention time (Rt) and mass (Mz), whereas Rt is a parameter that refers to chemical properties (lipophilic-hydrophilic), Mz describes the size of metabolites. Therefore, Rt- and Mz-groups were built to make distinctive statements about the influence of dialysis. A histogram (Figure 20) shows the number of metabolic features by retention time in minutes. Most of the metabolic features appear between 10 and 12 minutes. For further statistical frequency analysis quantitatively comparable groups were built (Table 8).

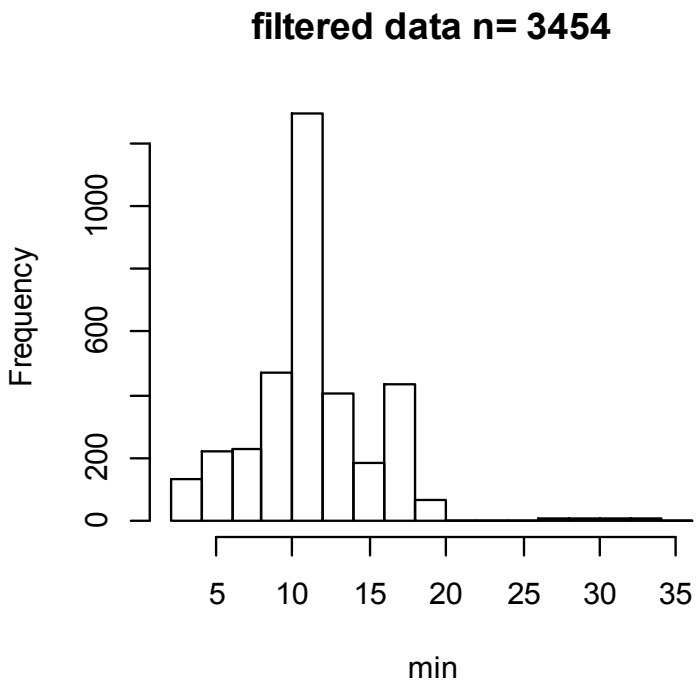


Figure 20: Number of metabolic features per retention time

Table 8: Number of metabolic features per retention time group

Rt	N-Features
Min. 3-8	726
Min. 9-10	843
Min. 11	784
Min. 12-16	840
Min. 17-35	261

To evaluate the amount of enrichment and depletion in each group, chi² tests were applied on each retention time group. The overall view represents more depletion than enrichment; as has been expected from dialysis. Having a more detailed look, metabolic features that differ just in one treatment (HD, OL-HDF specific), show higher values after the dialysis (POST > PRE).

Table 9: Number of metabolic features per retention time group with enrichment (yellow) and depletion (green) per treatment, p-values from pearson chi²-test

	HD				OL-HDF			
	PRE>POST	POST>PRE	all	p-value	PRE>POST	POST>PRE	all	p-value
Min3_8	241	189	430	0.0122	219	65	284	<0.001
Min9_10	188	207	395	0.3391	177	213	390	0.0683
Min11	147	163	310	0.3635	110	196	306	<0.001
Min12_16	398	181	579	<0.001	378	200	578	<0.001
Min17_35	127	54	181	<0.001	119	55	174	<0.001
			HD-specific				**OL-HDF-specific**	
	PRE>POST	POST>PRE	all	p-value	PRE>POST	POST>PRE	all	p-value
Min3_8	29	129	158	<0.001	7	5	12	0.5637
Min9_10	25	76	101	<0.001	14	82	96	<0.001
Min11	45	66	111	0.0462	7	100	107	<0.001
Min12_16	26	36	62	0.2041	7	54	61	<0.001
Min17_35	16	13	29	0.5775	8	14	22	0.2008

To estimate the overall amount of enrichment and depletion, median ratios for each group and treatment were built. The median ratios do not differ significantly between the treatments (Table 10).

Table 10: Median ratios of enrichment and depletion per treatment per retention time

		OL-HDF	HD
Min 3-8	median enrichment	1.43	1.49
	median depletion	0.29	0.32
Min 9-10	median enrichment	1.16	1.20
	median depletion	0.21	0.28
Min 11	median enrichment	1.18	1.16
	median depletion	0.23	0.47
Min 12-16	median enrichment	1.23	1.23
	median depletion	0.15	0.16
Min 17-35	median enrichment	1.24	1.24
	median depletion	0.25	0.27

A similar analysis is done for Mz groups. The histogram of Mzmed-values shows higher frequencies in metabolic features of about 300 DA (Figure 21).

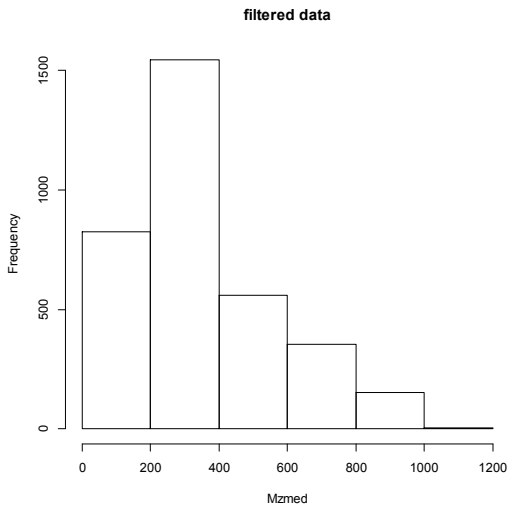


Figure 21: Histogram of Mz frequencies.

Table 11: Classification of Mz-groups Classification of Mz-groups

Mz	N-Features
0-200	832
200-400	1543
400-600	565
600-800	355
800-1200	159

Smaller metabolic features (Mz 200-400) were removed in both dialysis modalities (Table 12).

Table 12: Number of metabolic features per Mz-group with enrichment (yellow) and depletion (green) per treatment, p-values from pearson chi²-test

	HD				**OL-HDF**			
	PRE>POST	POST>PRE	all	p-value	PRE>POST	POST>PRE	all	p-value
M0_200	409	76	485	<0.001	366	109	475	<0.001
M200_400	573	246	819	<0.001	520	272	792	<0.001
M400_600	78	194	271	<0.001	77	186	263	<0.001
M600_800	30	199	229	<0.001	26	155	181	<0.001
M800_1200	13	106	119	<0.001	15	67	82	<0.001
	HD-specific				**OL-HDF-specific**			
	PRE>POST	POST>PRE	all	p-value	PRE>POST	POST>PRE	all	p-value
M0_200	56	18	74	<0.001	13	51	64	<0.001
M200_400	69	90	159	0.0958	16	116	132	<0.001
M400_600	9	67	76	<0.001	8	60	68	<0.001
M600_800	7	82	89	<0.001	2	39	41	<0.001
M800_1200	1	46	47	<0.001	1	9	10	0.0114

No significant differences between the treatments were detected for the 4 weeks short-term effect,.

Clearance could also be shown in clinical parameters like Beta2 Microglobulin showing OL-HDF with a ratio of 4.4 compared to 3.3 in HD a significant treatment effect in linear mixed model (Table 13, Figure 32).

Table 13: p-values from linear mixed model with patient as random effect and treatment as covariate

	p-value PRE-POST	p-value Treatment	Ratio: PRE/POST HD	Ratio: PRE/POST OL-HDF
Beta2-Microglobulin (mg/l)	<0.0001	0.0002	3.254	4.374
free light chain Lamda (mg/l)	<0.0001	0.0008	1.279	1.850
Free light chain Kappa (mg/l)	<0.0001	0.0034	2.078	3.396

4-weeks short-term effect did not show any effects between the two treatments. To efficiently exploit the data, annotated hits were searched explicitly. Tyrosine is the example for an identified metabolite in both data-sets, PRE-POST (Figure 22) and 4-weeks short-term effect (Figure 23), showing the expected result:

Tyrosine is removed by both dialysis modalities in all patients within 4 weeks, the amount of Tyrosine is less in OL-HDF-treated patients.

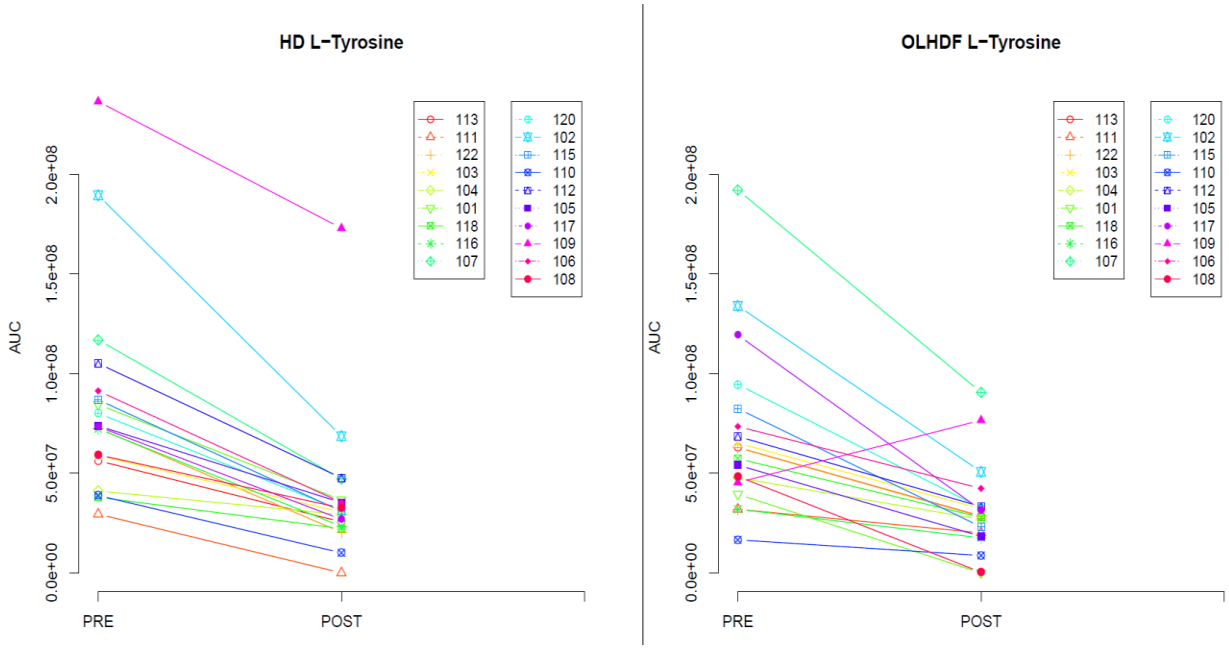


Figure 22: Tyrosine as an annotated example for a PRE-POST effect

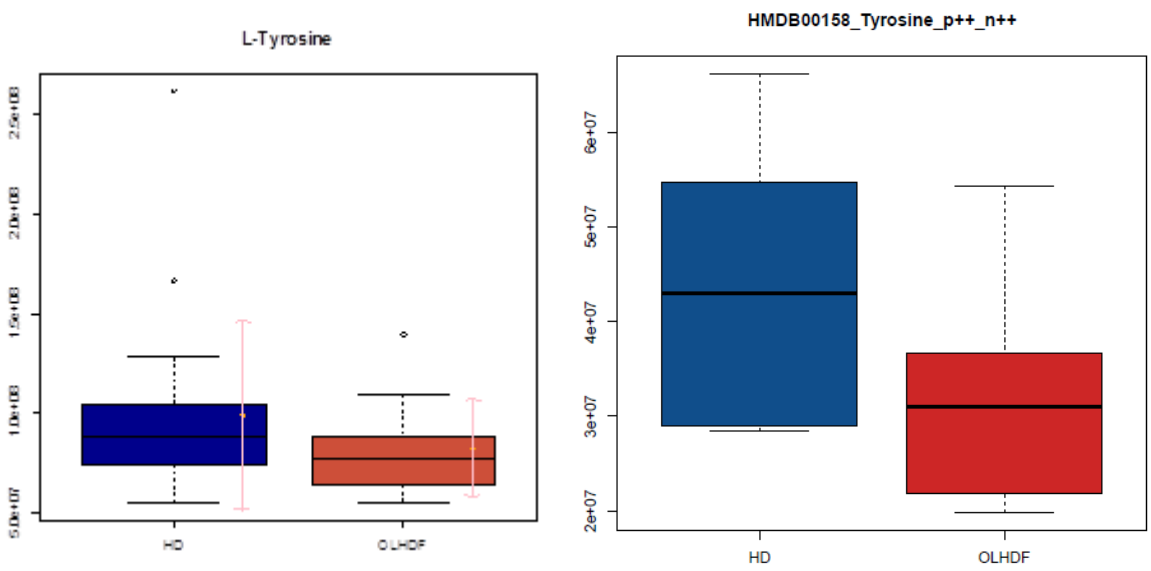


Figure 23: Tyrosine as an annotated example for 4-weeks short-term effect and 12 weeks long-term effect

In **12-weeks long-term effect** Paired t-tests were performed to identify significant differences between HD and OL-HDF. 373 features showed tendencies (unadjusted p-value of paired t-test<0.05) in R1, >98% of the features in the same direction, meaning that OL-HDF showed smaller signals than HD. For R2, 205 metabolic features showed tendencies, but the trend is rather in the other direction (>81% of the features show higher levels in OL-HDF) (Table 14).

Table 14: Number of features with treatment differences per group

	OL-HDF>HD	OL-HDF<HD
R1	6	367
R2	168	37

Comparing the median of MzMass, 367 specific OL-HDF metabolic features were significantly larger with mean of 292.6759 than 168 HD-specific features with a MzMass of 247.5724 (p-value of t-test: <0.001). OL-HDF specific features also had higher retention time compared to HD-specific features (10.67 min compared to 8.89 min, p-value<0.001).

Taking the intersection of 27 metabolic features (Figure 24) that show a certain tendency(unadjusted p-value of paired t-test < 0.05), each of them showed the same

pattern: first higher values in OL-HDF switch to significantly lower values, shown in one representative example (Figure 25), irrespective of total signal-amount. This could be the indication of a long-term effect of OL-HDF.

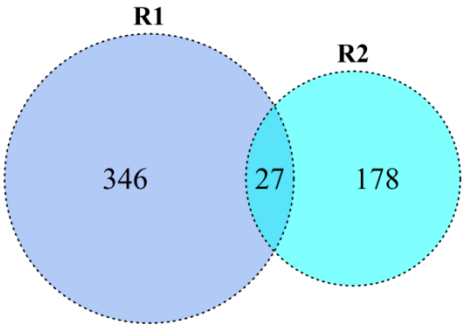


Figure 24: Venn-Diagram of tendency- metabolic features that differ between HD and OL-HDF

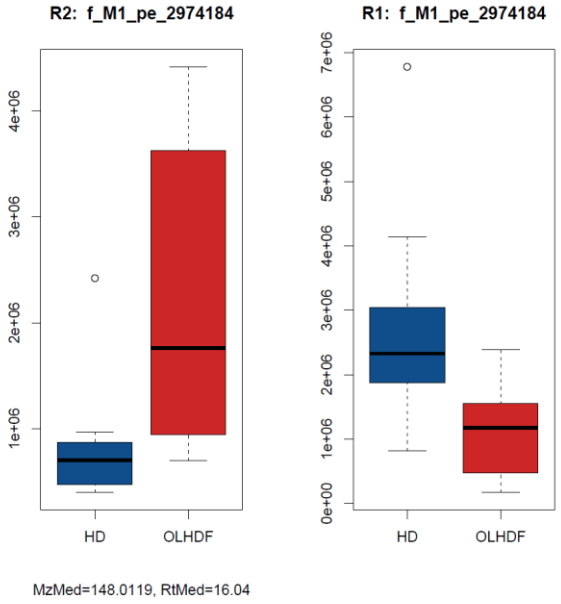


Figure 25: Example for inversion of intensities for specific changing metabolic features.

3.2.4. Nutritech: Drift Correction of a large Data Set

Samples of 72 volunteers were measured. GC-MS data were measured at the TU München (Prof. Hannelore Daniel) and processed with in-house software from Prof. Karsten Hiller Luxemburg.

Filter steps were performed as in the formal correction, resulting in 117 from initially 493 metabolic features. The drift correction was performed with batch-adjustment and quantile regression. The batch-adjustment was performed via a mean additive correction-factor from the QCs for each batch to adjust the level of the 16 batches (Figure 26). Drift correction resulted best with quantile-regression tau = 0.5 and df = 68. Improved metabolic features with a relative standard-deviation of <0.5 were included.

The median relative standard-deviation (RSD) of the QCs declined from 1.05 median RSD from the initial data set to 0.53 in the data-set to 0.28 in the batch-adjusted and to 0.23 in the corrected data-set.

The QC-samples D2B7_B7_QuK_Pl_16.cmp, D2B8_B8_QuK_Pl_5.cmp, D2B8_B8_QuK_Pl_7.cmp, D2B8_B8_QuK_Pl_15.cmp have low intensities, samples between these QCs were considered as outliers in the multivariate analysis of the PCA plot and sequence scatter-plot (Figure 27 shows PCAs from original and corrected data; the 16 different colours represent 16 different measurement batches; the batch-dependencies diminish after filtering and drift correction (right), outlier samples are visible on the bottom left part of the plot., Figure 28 shows Leucine as an example of well measured and corrected metabolite (red points represent QCs and blue dotted lines the different batches).

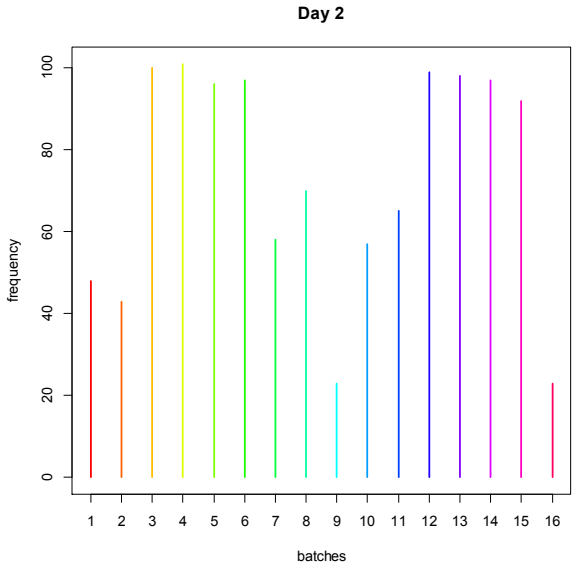


Figure 26: Nutritech day 2 samples were measured in 16 analytical measurement batches.

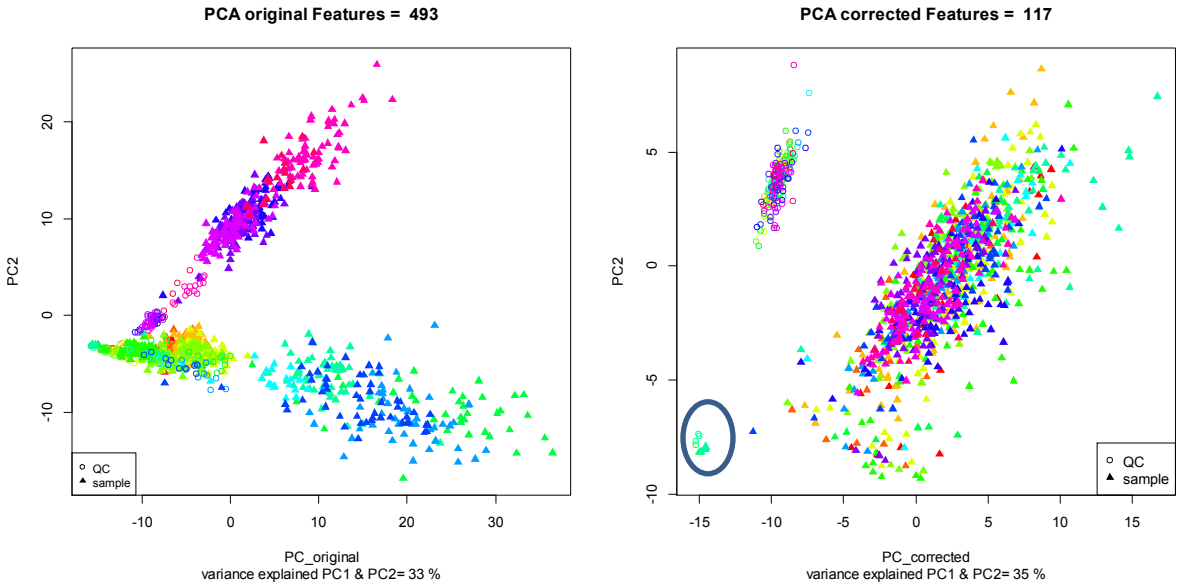


Figure 27: PCA-plot of original metabolic features (left) and drift-corrected metabolites (right).

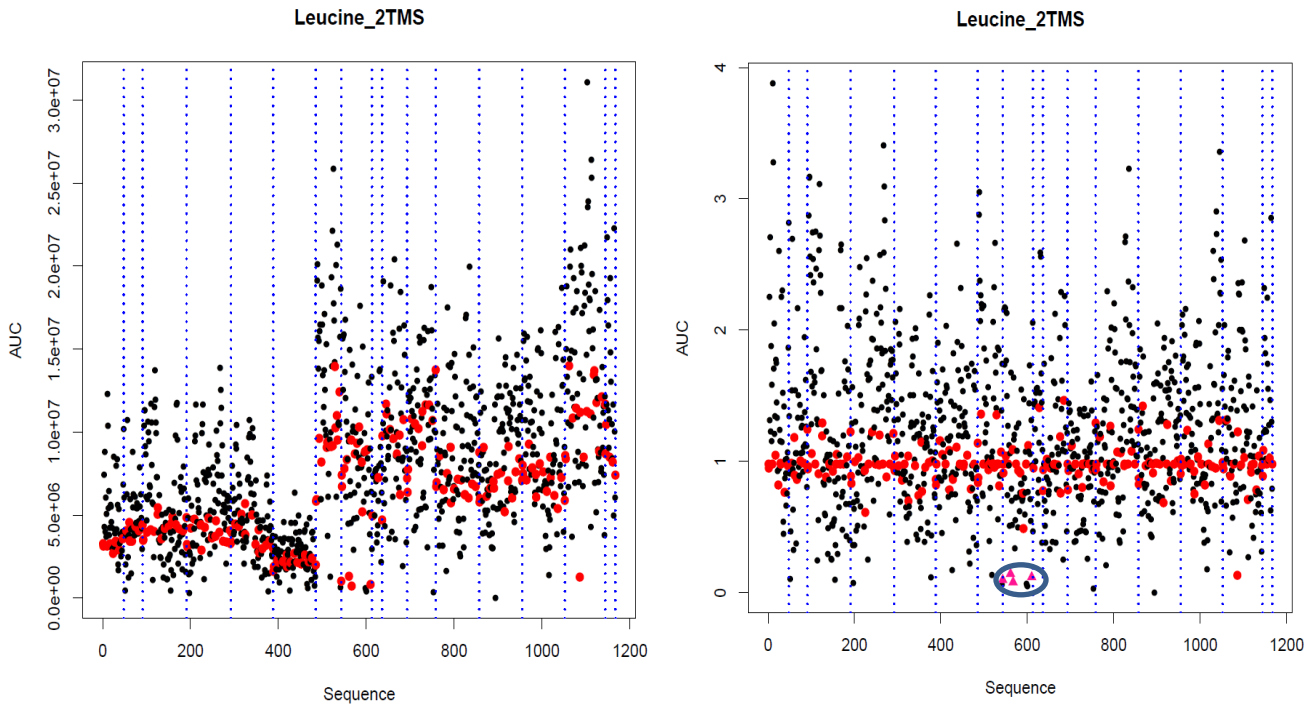


Figure 28: Drift correction of Leucine. Red points represent QCs and blue dotted lines the different batches

3.3. Metabolomics-Communication Check List

Metabolomics is an interdisciplinary field and metabolomics studies are quite complex as illustrated in the previous examples. Therefore regular communication between the persons involved is essential. The following checklist provides a set of questions that simplifies the process of a metabolomics project. The checklist is written from a statistician‘s point of view describing the optimal framing conditions to implement the data driven workflow in the best way.

The interdisciplinary team consists of

- Investigating people (physician, biologist)
- Analytical people (chemist, laboratory staff)
- Data people (statistician, informatician)

This team (or at least the group-representatives) should have an initial meeting to understand the study design and the research question and the following questions should be discussed: What is the research question?

1. Can we operationalize the research question in metabolomics parameters?
2. Can we formulate the questions that we want to have to answer?
3. Does a reference point in the literature exist?
4. How many groups do we compare and why?
5. How many time points do we consider and why?
6. Which sample size will be appropriate and why?
7. Which sample size will be affordable and why?
8. Is there any probable bias to consider?
    - Are several study centres involved?
    - Is there a age, gender, type-dependent bias to be expexted?
9. What can we expect from our study (e.g.):
    - Novelty?
    - Tendencies?
    - Confirmations of previous research?

The following technical choices have to be made

1. Choice of the appropriate statistical methods
    a. Combination with clinical data
    b. Planning explicitly search for markers known from literature and previous research
2. Choice of appropriate analytical methods
    a. Targeted versus untargeted, standards….
    b. Randomisation and measurement sequence depending on the research question
    c. Number of QCs (Important for drift correction but also important for measuring-time)
    d. Number of measurement batches
    e. Targeted explicit search (e.g. uremic toxins, fatty acids)
3. Choice of data pre-processing (depending on statistical methods)
    a. Grouping
    b. Number of data-sets
4. Proper labelling of
    a. groups
    b. samples

By answering these questions a discussion regarding challenges and requirements will arise:

5. Realistic timelines
6. Definition of tasks and assignment of persons
7. Definition of a common language
    a. Scientific glossary
    b. Clear terms and constant abbreviations
8. Planning of regular project meetings including
    a. Agenda
    b. Protocol

4. Discussion

The discussion is divided in four parts:

- Chapter 4.1 deals with the technical aspects of the statistical data driven workflow with focus on drift correction and statistical methodology.
- In chapter 4.2, the study results are discussed with focus on a rather medical point of view.
- Chapter 4.3 summarizes the diverse successful applications of the workflow and points out the diversity of application-fields.
- Chapter 4.4 gives a critical reflection of keywords in context with metabolomics, such as biomarker, big data and precision medicine.

Chapter 4 also comprises an outlook is the discussion where an appropriate study design for untargeted metabolomics studies is discussed for future applications.

4.1. Statistical data-driven Workflow

Our data-driven workflow for untargeted metabolomics is tailored for the JR-metabolomics platform. Its successful application on clinical data could was shown in the presented showcases. As already mentioned many tools available for metabolomics data processing (26) exist and therefore, the question comes up, why not using the actual and newest stuff freely available on the internet?

Time and resources are probably the most important reasons. The double challenge of producing results and developing new processes at the same time can hardly be achieved. Tool switching is not very common and practical for “data people” who are constantly working on data that need to be published or interpreted. Learning of new material is time and energy-costly. Therefore once a method or a workflow produces reliable and reasonable results, optimization the workflow is more feasible including completely new tools that require other formats and specifications.

Developing and producing results on the same time might not be the best scientific strategy but is favourable in terms of cost efficiency. However, comparing the presented workflow to new R-packages like MSPrep (83), a similar selection of statistical and mathematical tools are included to process metabolomics data. That was also the goal for the here presented workflow: the implication of a combination of best available methods, manageable by one person.

In the following section, the two main parts of the statistical data processing workflow-namely drift correction and filtering will be discussed separately.

Drift Correction: Drift correction is part of data normalization (84–87).The measurement system is highly sensitive and therefore time dependent drifts can occur. These drifts were corrected individually with adaptable regression models. As variability of features is high, smoothing by a locally adaptive regression technique, like Quantile Regression was necessary. Quantile Regression is highly flexible and well suited for the modelling of data with heterogeneous conditional distributions. Drift correction is used to correct batch to batch variations.

These variations appear because the sample sets were divided into batches of not more than 100 samples each, and the single batches were measured separately. Batch overlapping indicates high quality of the data processing workflow. The quality degree was depicted graphically by using unsupervised Random Forests Models.

Filtering Steps are applied to remove artefacts and serve to cope with outliers and missing values (57,84,87–93). Drift correction procedures are well described in the literature and drift correction is commonly done by using consecutive QCs (57,88–90) or internal standards (84,91). The regression models LR, LoReg, GLM and LOESS (locally weighted polynomial regression) are widely used as statistical correction methods, (84,87–90,94). LOESS represents the standard method for drift correction (84,95,96).

Metabolomics data processing has a huge influence on the results (97,98), but the data to be processed vary a lot between the different metabolomics studies. Reasons are the different instrumentation platforms used, eg. C-MS, GC-MS, UPLC-MS, LC-TOF, DIMS, NMR, etc. and the varying origins of the data, eg- from different research areas, such as microbe studies, plant studies, from mammalian and environmental systems studies, to name but a few.

Commercial and non-commercial tools for data processing are readily available, e.g. DanteR, XCMS-Online and MetaboAnalyst (87,92,93). These tools demand specific data formats that might not fit to every measurement system. Further, these tools provide consecutive statistical analysis that might not be suitable for all research issues.

Metabolomics data processing is still reported to be a manual process (6), because a lot of different data processing protocols are used, depending on the measurement abilities and kind of studies performed in the different laboratories, and no standardized workflow has been implemented yet. We successfully introduced a data processing workflow for untargeted metabolomics data. The single processing steps, which we applied, are filtering steps, drift correction and normalisation.

Filtering steps remove artefacts, outliers and deal with missing values (57,84,87–93). Drift correction is performed to reduce batch-to-batch Variations and measurement variabilities (88,90,91). In this study, we started with LOESS, which is the standard method for drift correction (84,95,96) and then switched to a Quantile Regression approach, because this approach is adaptable for different kinds of data distributions and it is implemented as an easily manageable function in R. The parameter selection for the Quantile Regression was based on only one set of data.

For further kinds of studies these selection should be revised. But as this parameter evaluation for Quantile Regression has already been developed in the data processing workflow, the parameters can be easily adopted individually for each metabolomics study. Routinely, drift correction procedures based on the 0.5 quantile and on the 0.8 quantile are compared for each upcoming data set. Normalization was done to reduce systematic bias and to reduce the impact of very large values. Normalization to reduce systematic bias was performed by drift correction. Normalization to reduce the impact of very large values is depending on the specific investigation problem. The kind of normalization depends on the type of variation to be corrected (99) and is also the basis for the following statistical analysis that will be part of the second progress report.

The data processing workflow was evaluated on an individual feature level, taking the CV as one target value and assessing the drift correction based on qualitative graphical representations. This evaluation procedure has been reported in several studies (57,87,89–91). Also, graphical representations of overall multivariate analyses, like PCA-Plots, are commonly presented to evaluate batch-to-batch variation (57,87,89–91). We favour unsupervised random Forests models over PCA-plots, because we further use them for the statistical analysis. To optimise the data processing workflow, a representative set of features was chosen. This selection depends on an arbitrary choice and as a consequence potentially important features may get lost.

The applicability of the workflow is limited by the number of samples: For a number of QCs below 10, drift correction with QCs cannot be performed properly. Drift correction is reported to be used in large-scale studies (89,96). For small studies, only filtering steps can be performed. An expansion of the data processing workflow will be a module for the use of internal standards additionally to the QCs for drift correction.

To conclude, we successfully developed a workflow on data processing and applied it as proof-of-concept on study samples from four clinical investigation study. The final data set is a mixture of not annotated, known and identified features. To obtain relevant information a pure statistics approach is not sufficient. On contrary, a tight interdisciplinary cooperation during the data processing between the chemist and the statistician is essentially needed.

The success of an untargeted metabolomics approach is largely determined by the applied data processing workflow. Our untargeted metabolomics approach included data processing and statistical selection of metabolic feature and was successfully used in the presented showcase-clinical studies. Data processing included filtering and time dependent drift correction on QC-intensities, measured by LC-HRMS. The use of the same pooled QC sample to observe the LC-HRMS measurement stability over time is a commonly used technique (96,100).

For QC based drift correction, we used quantile regression which is often used for data modelling with heterogeneous conditional distributions (59). Quantile regression has already been successfully used in metabolomics for drift correction and baseline alignment (101–104). A smoothing step by a locally adaptive regression technique was applied because of the high variability of metabolic features. By using quantile regression techniques, the distributions of data quantiles were modelled separately, when dependencies were not equally distributed in different quantiles. We used a nonparametric quantile regression to suit the conditional quantile functions, which fits a piecewise cubic polynomial with the number of one third of available data-points knots (breakpoints) arranged at the quantiles of the QCs-signals (59). An amelioration of QC-variance was achieved by drift correction: the median CV of QC-intensities from all 924 metabolic features was reduced from 0.2 to 0.1.

The selection of metabolic features was based on significant p-values of univariate paired t-tests and metabolic feature importance of Random Forest (RF), a well-established methodology, metabolomics data processing (61–64,100,105). RFs are well suited for metabolomics because of reduced overfitting and improved model prediction (62,64,69,70) compared to other supervised classification methods such as PLS-DA. In our study, RFs clearly indicated clusters of samples taken before and after bariatric surgery (MDS-Plot unsupervised RF). Remaining overlaps of clusters might be caused by individual responses to bariatric surgery.

We successfully applied a novel data-driven approach by combining quantile regression and RFs to clinical metabolomics data.

4.2. Discussion of Study-Results

4.2.1. Cardionor

The clinical results of this study have been published in Cardiovascular Diabetology 2014 (106). However, the metabolomics analysis of Cardionor could not distinguish a specific metabolic profile for responder based on the relative change of the CIMT. The Cardionor study suited well to develop the basic structure of the data-driven metabolomics workflow, as the sample size of 44 patients produced largely enough data. The normalization of batches as well as the pre-requirements of a successful drift correction was also developed within this study. The

4.2.2. Bariatric Surgery

Results and discussion of this study are published in PLOS ONE. We successfully used an untargeted metabolomics approach to identify a set of metabolites which characterizes short- as well as long-term changes after bariatric surgery. In addition, we also investigated known metabolites which are relevant in the pathogenesis of cardiovascular diseases or which have previously been shown to be associated with cardiovascular outcome. In total, we identifed 36 metabolites. Their changes over time are best described as trend patterns. Metabolites which display unidirectional trends of increasing or decreasing intensities are more likely to be of interest for future studies.

The considerable difference in patterns between short- and long-term changes highlights the importance of repeated measurements of metabolic patterns after an intervention. Focusing metabolic analysis on a single point in time can easily lead to false conclusions on the relevance of metabolites as potential biomarkers.

Our analyses underline the need for a comprehensive analysis of short-term as well as long-term changes in order to gain a more complete picture of the metabolic changes induced by bariatric surgery. We put a special focus on metabolites that are known to be associated with cardiovascular disease and CVR factors (including diabetes) such as amino-acids (including BCAA), phospholipids (PCs), phenylalanine, Trimethylaminoxid (TMAO) and indoxyl sulfate.

BCAA intensities showed a decreasing trend after bariatric surgery similar to most other amino acids. Previous studies have described lower levels of BCAA to be associated with improved glucose metabolism and insulin sensitivity (34,35,107–113). Data from the Framingham (Heart) Offspring Study demonstrated an association of high levels of BCAA with an increased risk for cardiovascular disease (81,82). Our results also showed an increasing trend in glycine levels which has previously been described as a short-term effect of bariatric surgery (114) and inversely related to type 2 diabetes (115,116). In general our results for BCAA and other amino acids support the central role of these metabolites in the improved metabolism after bariatric surgery.

In our analysis all identified PCs were strongly influenced by bariatric surgery but pattern differed considerably over time. For example, PC38:6 has been associated with an increased risk of T2DM (117). Our data show a short-term increase of PC38:6 but the increase was lower at the long-term visit relative to baseline, resulting in a Λ-pattern. A possible cause for the short-term increase might be the dietary pattern shortly after bariatric surgery or the surgical procedure itself (118,119). Also, our short-term results are in line with a recently published study which showed an increase of PC38:6 in the first 42 days after bariatric surgery (120). These Λ-patterns clearly demonstrate the relevance of long-term metabolic monitoring after bariatric surgery.

For other PCs such as PC36:5 we found a V-pattern, for PC40:7a steady increase. Both metabolites have been shown to be inversely associated with coronary artery disease and mortality (121).

Phenylalanine, choline and tyrosine levels decreased after surgery, an effect that was sustained after one year at follow-up, similar to previous findings that have described increased phenylalanine and tyrosine as biomarkers for CVR (82,122).

TMAO and indoxyl-sulfate were also among the identified metabolites and are known cardiovascular markers (123). TMAO is derived from the gut bacterial metabolism of choline, it has been shown to be directly associated with cardiovascular outcome and was suggested to be a potential metabolic link between gut microbiome and cardiovascular diseases (81,123,124). In contrast to previous studies, our results showed a significant increase of TMAO and indoxyl-sulphate similar to results of a bariatric study in rats (123). One possible explanation for the increase of TMAO in patients could be a surgery-induced change in the gut microbiome composition. Alternatively, carnitine can induce formation of TMAO (125) and carnitine is often promoted as a weight loss inducing supplement.

Although the supplements recommended in this study did not contain carnitine we cannot exclude the possibility that patients would take carnitine by their own. This interpretation is limited since we neither assessed dietary details to adjust the TMAO analysis to dietary composition nor did we collect stool samples to analyze gut microbiome composition. However, all patients underwent standardized nutritional counseling and received the same supplementation recommendations following international guidelines (118,119). To determine TMAO is a useful marker for CVR factors in patients undergoing bariatric surgery further studies are needed.

The in-depth analyses of patients with established diabetes mellitus confirm previously described predictors of diabetes remission after bariatric surgery such as age and diabetes duration at the time of the surgery (126). Furthermore, metabolomics analysis demonstrates that patients with a complete diabetes remission after one year showed larger declines in amino acids levels of alanine, proline and leucine as well as in the glycine metabolite sarcosine and the glutaminic acid derivate pyroglutamic acid.

Previous studies have demonstrated an association of increased levels of BCAA and aromatic amino acids with diabetes incidence (127). Our study adds valuable information to the few available human data on diabetes and associated levels of sarcosine, leucyl-proline and pyroglutamic acids.

Conclusion Bariatric Surgery

In summary, the data revealed both short-term and long-term metabolic effects of bariatric surgery in humans. The different identified patterns highlight the importance of repeated measurements over longer time periods in order to obtain a comprehensive understanding of the metabolic effects of bariatric surgery. Our study provides a better insight to changes in previously discussed metabolic CVR factors and to potential metabolic markers for diabetes remission. Our results also indicate that some metabolites might behave differently in patients with bariatric surgery than in other risk cohorts. For a future more précised medicine, more detailed understanding of the metabolic effects of such clinical interventions is needed.

4.2.3. Metaprol[11]

Metabolomics also has its application in the field of nephrology (128–130). Phospholipids are a group of annotated metabolic features that show enrichment after dialysis in both treatments, namely HD and OL-HDF (Table 6). There is no explanation for that from the medical point of view. Therefore this result could be basis for further targeted research. Beta2-Microglobulin, free light chain Lamda and free light chain Kappa showed better clearance in OL-HDF (Table 13) which is in line with literature from previous OL-HDF studies (131,132).

This result was a positive control of the planning and enrolment of the study. 12 weeks long-term effect showed significantly more OL-HDF specific features, with a decrease. For 27 metabolic features that have an effect in HD and OL-HDF, clearance was better after a long-term application which is an indication for a study effect-patients were treated particular attentive during the study.

As we found the trend of smaller values in OL-HDF than in HD, disregarding the individual p-values, a further long-term study would be beneficial. This study will have to concentrate on the top candidates (that separate HD and OL-HDF best) and targeted metabolomics analysis will have to be performed.

The top candidates were selected by taking the intersection of the 100 lowest p-values and the 100 highest ratios (HD/OL-HDF), resulting in 35 metabolic features. Assuming a power of 80% and an adjusted p-value of a paired t-test of 0.001 for each of the candidate-features, a sample-size calculation was performed. Assuming same sample stability and measurement quality a sample size of 35 patients would show a significant difference between HD and OL-HDF in all of the candidate features.

[11] Manuscript submission to Kideny International is planned with October 2016

Concentrating on identified metabolites, a higher reduction of tyrosine, alanine and valine by OL-HDF can be discussed in a cardiovascular risk context. Comparison with results from the bariatric surgery study showed that the reduction of BCAAs and aromatic amino acids is associated with a reduced cardiovascular risk.

Contradictory the idea was brought up that essential amino acids should not be removed by dialysis. Another idea, which came up in the discussion with Prof. Bernard Canaud from Fresenius Medical Care Bad Homburg was to concentrate on metabolites that have a known functionality and are easily interpretable regarding mechanistic aspects like metabolites from the muscle metabolism such as creatine and adenine.

4.3. Successful Applications of the Workflow

In summary, all here presented studies are similar regarding the overall data-processing and the aim to select relevant metabolic features. The feature selection process differs in its application on subgroups (like PRE-POST groups or Mz/Rt-groups). The methods like Quantile Regression, RF and PCA stay the same, as illustrated in Figure 9. The modelling on individual features related to patients' characteristics differs of course.

As these metabolomics studies are further pilot studies and not conformational studies the statistical analysis demands a certain creativity which can only be obtained in the discussion with an interdisciplinary team. The identification of relevant metabolic features with a significant influence is important for biological and mechanistic interpretation that will further probably bring up new questions that demand other statistical analysis.

The workflow has been also successfully applied in preclinical studies:

- EAE: Animal experiment: comparing blood plasma and cOFM of rats (Figure 33)
- Animal experiment: Comparing different feeding on mouse-cohorts, measuring serum samples and tissue samples to describe a rejuvenation effect of spermidine (Figure 34). This study was done in cooperation with Prof. Dr. Frank Madeo and Dr. Tobias Eisenberg from the Karl-Franzens University Graz. The work is submitted to nature and currently under review.
- Sample stability: Investigating EDTA sample stability under different storage temperatures (Figure 35)
- Difference between IPAH patients and healthy persons (Figure 41)

The Metabolomics communication checklist is based in study planning principles described in “good clinical practice” ICH guidelines and in the literature (133–135). Planning timelines, meetings, defining tasks and roles are known points from project management in scientific settings (136) and are also essential parts to be considered for a best performance of a data-drivenworkflow.

A clear structure of tasks and associated persons facilitate the communication and coordination as well as the identification with the project. A structured and transparent workflow with important methodological influence enables better and reproducible results, as has been recently described in a special critical series of Lancet 2014 (137,138).

4.4. Critical Reflection & Outlook

The following part deals with terms that are often used as keywords in combination with the application and future perspectives of metabolomics: **Biomarker**, **Big data** and **Precision Medicine**. Further, an appropriate study design combining clinical properties and metabolomics techniques is discussed as an outlook perspective.

4.4.1. Keywords in Fashion

As our study results reveal, the detection of a real biomarker still poses challanges.

This definition of biomarker has been given by the NIH in 2001 (1): „A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention“.

Other keywords in fashion that are often used in the context of metabolomics are “**Big Data**” and “Personalized-” respectively “**Precision- Medicine**”.

Big data is a field of research that touches different research areas. The subject related journal: “Big data and society” brings the idea out, that big data is messy (139,140), „big data” is also used for a new marketing field for the academic-industry(141) speaking of economic and commercial perspectives as it is the case for Joanneum Research.

Big data works in “translation” meaning the combination of all sort of data from clinical laboratory data to various omics data, only works for limited purposes (142,143). The discussion of big data opportunities and constraints is split into skepticism(142,144) and enthusiasm(145). The meaning of “Big” in “Big data” for personalized medicine is meant in both directions- first, a lot of variables and second, large sample size that might facilitate the identification of small populations of patients that might benefit from specific treatment or drug (146). Taking the metabolomics studies as examples, sample sizes are way too small to count for “big data”.

Personalized medicine, now **precision medicine** (147,148) as the correct term, has the goal to identify subgroups of patients that differ from the majority of patients and to subsequently develop new targeted treatment for this specific subgroup of patients. The term changed from personalized to precision because 'personalized' implies the prospect of devising a different treatment for each individual patient; what is "simply" the daily work of a physician. (149). The following paragraph gives the definition of "precision medicine" from The Economist in 2009 published in "The National Academic Press in 2011 (150):

> *""precision medicine" refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not. Although the term "personalized medicine" is also used to convey this meaning, that term is sometimes misinterpreted as implying that unique treatments can be designed for each individual. For this reason, the Committee thinks that the term "precision medicine" is preferable to "personalized medicine" to convey the meaning intended in this report. It should be emphasized that in "precision medicine" the word "precision" is being used in a colloquial sense, to mean both "accurate" and "precise" (in the scientific method, the* ***accuracy*** *of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value whereas the* ***precision*** *of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results). Accuracy and precision. the point where pharmacogenetics and personalised medicine meet (The Economist 2009)."*

4.4.2. Study Design & untargeted Metabolomics

Metabolomics has been applied to clinical and preclinical studies. The following questions serve to measure the success of the metabolomics application:

1. Were the analytical measurements successful, in terms of sensitivity and stability?
2. Did the data processing identify enough requested metabolic features from reliable peaks (5000-10000)?
3. Could a big part of the measured metabolic features be kept by filtering and drift correction steps?
4. Were biased avoided by appropriate randomisation and QC-distribution?
5. Could metabolites or new substances be identified?
6. Could any tendencies be observed?
7. Could any statistically significant results be observed by answering the research question?
8. Could any clinically relevant results be observed?

With only 10 samples from weakly characterized subjects, we can answer question 1 to 5 but any statistically significant results cannot be expected. We can expect to find tendencies that can be represented graphically. Following the technical progress, the success of being able to measure different materials from various matrices is crucial!

A reflection of metabolomics studies will help to clarify expectations:

- Was the study meant to be designed to do metabolomics or primary for other purposes and metabolomics was an “add-on”?

What did we learn during the past five years of performing untargeted metabolomics studies?
First, the expectations have to be kept realistic and second, not every study design seems to be suitable for metabolomics studies!

Experience showed that the clinical phenotype has to be well characterized in clear defined groups to get clear results from metabolomics. Clinical intervention studies

designed as randomized controlled trials are eventually not the best model for metabolomics studies, as seen from the examples of Cardionor and Metaprol. Variability in metabolomics is already high due to the chaining of complex technologies and adding another source of variability through weakly characterized patients group, does not make any results more clear.

How can we adapt a study design when metabolomics is planned? Clinical based research Case-Control studies with sharp characterized groups seem better suitable than randomized control trials for a given metabolomics setting. Table 15 gives a comparision of Case-Control-Studies versus RCT-Design (136).

Table 15: Comparison between Case Control and RCT (Cavalieri et al. 2014 S. 8-10 (136)**)**

Case Control Studies	Randomized, Controlled Trials
Case control studies are conducted to determine whether there is an association between a risk factor and an outcome. A characteristic feature is that subjects are enrolled as cases or controls based on the presence or absence of a specific outcome. The two groups are compared to assess the presence of a proposed risk factor. To optimize the ability to examine specific risk factors, the cases and controls should be matched as evenly as possible. The advantage of this type of study is that the investigator can enrol all defined cases, making it an attractive option when the outcome is rare. Multiple risk factors can be evaluated, provided that the investigator thinks of the various factors. The disadvantages of this type of study are that they are retrospective and subject to various types of bias. The investigator must be very careful when defining cases and controls that are representative of the population.	Randomized, controlled trials (RCTs) are prospective, and the randomization, if done effectively, limits bias. Ideally, subjects are assigned to intervention or control groups in a blinded fashion. RCTs are experimental rather than observational and are designed to test the effect of a planned intervention. Determination of the size of the trial (number of subjects in each group) is critical and depends on the incidence of the outcome measure in the control population as well as the effectiveness of the intervention. This type of study design is the most convincing demonstration of causality. The disadvantage to RCTs is usually cost and logistical considerations. At times, ethical considerations are quite thorny, such as whether a placebo should be utilized. Also, RCTs can be subject to confounding.
Variability is better described and controlled	More random-effects that represent population but cause more variability

From a medical point of view new insight, even mechanistic explanation of defined characteristics should deliver novelties, that require a new methodology and cannot be explained by conventional (analytical, laboratory) methods. As an example, only part of the patients of the bariatric surgery study was T2DM patients. Some of them could achieve complete diabetes remission. A further metabolomics bariatric surgery study should include a homogenous group of T2DM patients to observe changes in the metabolite profile after the surgery:

Can the remission-patients be separated from the non-remission patients due to the changing metabolite profile?

5. Conclusion

Is Metabolomics as a precise and exact science?

Outcome of Metabolomics studies (as it is the case for every other scientific studies) depends on the question we want to answer. Not every question can be answered by metabolomics. Realistic expectations and suitable study design including meaningful sample size will help to get most valuable results.

6. Bibliography

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. [Internet]. Clinical pharmacology and therapeutics. 2001 Mar [cited 2014 Jul 9]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11240971

2. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res [Internet]. 2009 Jan [cited 2015 Jan 26];37(Database issue):D603–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18953024

3. Precision Medicine Initiative | National Institutes of Health (NIH) [Internet]. [cited 2016 Mar 30]. Available from: https://www.nih.gov/precision-medicine-initiative-cohort-program

4. Wishart DS, Lewis MJ, Morrissey J a, Flegel MD, Jeroncic K, Xiong Y, et al. Meet the human metabolome. J Chromatogr B Anal Technol Biomed Life Sci [Internet]. 2007;446(2):8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18502700

5. Sas KM, Karnovsky A, Michailidis G, Pennathur S. Metabolomics and Diabetes: Analytical and Computational Approaches. Diabetes [Internet]. 2015 Mar 24 [cited 2016 Jan 21];64(3):718–32. Available from: http://diabetes.diabetesjournals.org/content/64/3/718.short

6. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. Nat Rev. 2012;13(Molecular Cell Biology):7.

7. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics [Internet]. 2013 Apr [cited 2013 Aug 14];9(2):280–99. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3608878&tool=pmcentrez&rendertype=abstract

8. Ludwig Boltzmann Institute, Zechmeister-Koss I, Kisser A. Procedural guidance for the systematic evaluation of biomarker tests. 2014.

9. Ptolemy AS, Rifai N. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. Scand J Clin Lab Invest Suppl. 2010;242(Suppl 242):6–14.

10. Puntmann VO. How-to guide on biomarkers: biomarker definitions, validation and applications with examples from cardiovascular disease. Postgrad Med J [Internet]. 2009 Oct 1 [cited 2016 Jan 27];85(1008):538–45. Available from: http://pmj.bmj.com/content/85/1008/538.short

11. Mamas M, Dunn WB, Neyses L, Goodacre R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Arch Toxicol [Internet]. 2011 Jan [cited 2016 Feb 8];85(1):5–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20953584

12. Blanchet L, Smolinska A, Attali A, Stoop MP, Ampt KA, van Aken H, et al. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2011 Jan 22 [cited 2011 Sep 1];12(1):254. Available from: http://www.biomedcentral.com/1471-2105/12/254/abstract

13. Emwas A-H, Roy R, McKay RT, Ryan D, Brennan L, Tenori L, et al. Recommendations and Standardization of Biomarker Quantification Using NMR-based Metabolomics with Particular Focus on Urinary Analysis. J Proteome Res [Internet]. American Chemical Society; 2016 Jan 8 [cited 2016 Jan 12];15(2):360–73. Available from: http://dx.doi.org/10.1021/acs.jproteome.5b00885

14. Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for Biomarker Discovery: Moving to the Clinic. Biomed Res Int [Internet]. 2015;2015:1–6. Available from: http://www.hindawi.com/journals/bmri/2015/354671/

15. Rhee EP, Gerszten RE. Metabolomics and cardiovascular biomarker discovery. Clin Chem [Internet]. 2012 Jan 1 [cited 2014 Feb 20];58(1):139–47. Available from: http://www.clinchem.org/content/58/1/139.short

16. Kang J, Zhu L, Lu J, Zhang X. Application of metabolomics in autoimmune diseases: Insight into biomarkers and pathology. J Neuroimmunol [Internet]. 2015 Feb [cited 2015 Jan 13];279:25–32. Available from: http://www.sciencedirect.com/science/article/pii/S016557281500003X

17. Menni C, Fauman E, Erte I, Perry JRB, Kastenmüller G, Shin S-Y, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. Diabetes [Internet]. 2013 Dec [cited 2014 Sep 3];62(12):4270–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23884885

18. Roberts LD, Koulman A, Griffin JL. Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome. Lancet Diabetes Endocrinol [Internet]. 2014 Jan [cited 2014 Feb 3];2(1):65–75. Available from: http://www.sciencedirect.com/science/article/pii/S2213858713701438

19. Den Ouden H, Pellis L, Rutten GEHM, Geerars-van Vonderen IK, Rubingh CM, van Ommen B, et al. Metabolomic biomarkers for personalised glucose lowering drugs treatment in type 2 diabetes. Metabolomics [Internet]. Jan [cited 2016 Feb 8];12:27. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4703625&tool=pmcentrez&rendertype=abstract

20. Urpi-Sarda M, Almanza-Aguilera E, Tulipani S, Tinahones FJ, Salas-Salvadó J, Andres-Lacueva C. Metabolomics for Biomarkers of Type 2 Diabetes Mellitus: Advances and Nutritional Intervention Trends. Curr Cardiovasc Risk Rep [Internet].

2015 Feb 17 [cited 2016 Jan 5];9(3):12. Available from: http://link.springer.com/10.1007/s12170-015-0440-y

21. Zhang A, Sun H, Yan G, Wang P, Wang X. Mass spectrometry-based metabolomics: Applications to biomarker and metabolic pathway research. Biomed Chromatogr. 2016;30(1):7–12.

22. Yun Y-H, Deng B-C, Cao D-S, Wang W-T, Liang Y-Z. Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery. Anal Chim Acta [Internet]. 2016 Jan [cited 2016 Jan 13]; Available from: http://www.sciencedirect.com/science/article/pii/S0003267016300216

23. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) [Internet]. Springer; 2009 [cited 2013 Jan 21]. 768 p. Available from: http://www.amazon.com/The-Elements-Statistical-Learning-Prediction/dp/0387848576

24. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics [Internet]. 2006 Nov [cited 2011 Jun 23];2(4):171–96. Available from: http://www.springerlink.com/index/10.1007/s11306-006-0037-z

25. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, et al. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics. Springer; 2007 Aug;3(3):231–41.

26. Misra BB, van der Hooft JJJ. Updates in metabolomics tools and resources: 2014-2015. Electrophoresis [Internet]. 2016 Jan [cited 2016 Feb 2];37(1):86–110. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26464019

27. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics [Internet]. 2007 Sep [cited 2014 Jul 10];3(3):211–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3772505&tool=pmcentrez&rendertype=abstract

28. Werf MJ, Takors R, Smedsgaard J, Nielsen J, Ferenci T, Portais JC, et al. Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments. Metabolomics [Internet]. 2007 Aug 20 [cited 2013 Nov 25];3(3):189–94. Available from: http://link.springer.com/10.1007/s11306-007-0080-4

29. Fiehn O, Robertson D, Griffin J, Werf M, Nikolau B, Morrison N, et al. The metabolomics standards initiative (MSI). Metabolomics [Internet]. 2007;3(3):175–8. Available from: http://www.springerlink.com/index/10.1007/s11306-007-0070-6

30. The Lancet Diabetes. Bariatric surgery: why only a last resort? lancet Diabetes Endocrinol [Internet]. 2014 Feb [cited 2014 Jun 4];2(2):91. Available from: http://www.thelancet.com/journals/a/article/PIIS2213-8587(14)70020-8/fulltext

31. Kashyap S, Daud S, Kelly K, Schauer PR. Acute effects of gastric bypass versus gastric restrictive surgery on β-cell function and insulinotropic hormones in severely obese patients with type 2 diabetes. Int J Obes. 2010;34(3):462–71.

32. Dunn WB. Diabetes - the Role of Metabolomics in the Discovery of New Mechanisms and Novel Biomarkers. Curr Cardiovasc Risk Rep [Internet]. 2012 Dec 7 [cited 2013 Nov 11];7(1):25–32. Available from: http://link.springer.com/10.1007/s12170-012-0282-9

33. Magkos F, Bradley D, Schweitzer GG, Finck BN, Eagon JC, Ilkayeva O, et al. Effect of Roux-en-Y gastric bypass and laparoscopic adjustable gastric banding on branched-chain amino acid metabolism. Diabetes [Internet]. 2013 Aug 22 [cited 2014 Jan 17];62(8):2757–61. Available from: http://diabetes.diabetesjournals.org/content/early/2013/04/17/db13-0185.short

34. Newgard C. Interplay between lipids and branched-chain amino acids in development of insulin resistance. Cell Metab [Internet]. 2012 [cited 2013 Jul 5];15(5):606–14. Available from: http://www.sciencedirect.com/science/article/pii/S1550413112001039

35. Batch BC, Shah SH, Newgard CB, Turer CB, Haynes C, Bain JR, et al. Branched chain amino acids are novel biomarkers for discrimination of metabolic wellness. Metabolism [Internet]. 2013 Jul [cited 2014 Feb 11];62(7):961–9. Available from: http://www.sciencedirect.com/science/article/pii/S0026049513000085

36. Bain JR, Stevens RD, Wenner BR, Ilkayeva O, Muoio DM, Newgard CB. Metabolomics applied to diabetes research: moving from information to knowledge. Diabetes [Internet]. 2009 Nov [cited 2013 May 23];58(11):2429–43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2768174&tool=pmcentrez&rendertype=abstract

37. Yuan M, Breitkopf SB, Yang X, Asara JM. A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. Nat Protoc [Internet]. Nature Publishing Group; 2012 May [cited 2013 Jan 28];7(5):872–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22498707

38. Bajad SU, Lu W, Kimball EH, Yuan J, Peterson C, Rabinowitz JD. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. J Chromatogr A. 2006;1125:76–88.

39. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin C a, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics [Internet]. Springer; 2007;3(3):211–21. Available from: http://www.springerlink.com/index/10.1007/s11306-007-0082-2

40. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res [Internet]. 2007 Jan [cited 2014 Sep 1];35(Database issue):D521–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1899095&tool=pmcentrez&rendertype=abstract

41. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res [Internet]. 2013 Jan [cited 2015 Jan 5];41(Database issue):D801–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531200&tool=pmcentrez&rendertype=abstract

42. Smith C a, O'Maille G, Want EJ, Qin C, Trauger S a, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit [Internet]. 2005 Dec;27(6):747–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16404815

43. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2011 Aug 3];28(1):27–30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract

44. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: Back to metabolism in KEGG. Nucleic Acids Res. 2014;42(D1):199–205.

45. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res [Internet]. 2007 Jan [cited 2014 Sep 1];35(Database issue):D521–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1899095&tool=pmcentrez&rendertype=abstract

46. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res [Internet]. 2009 Jan [cited 2015 Jan 26];37(Database issue):D603–10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686599&tool=pmcentrez&rendertype=abstract

47. Smith C a, O'Maille G, Want EJ, Qin C, Trauger S a, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit [Internet]. 2005 Dec;27(6):747–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16404815

48. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, et al. Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics [Internet]. 2014 Apr 8 [cited 2014 Apr 9];10(3):350–3. Available from: http://link.springer.com/10.1007/s11306-014-0656-8

49. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc Rev [Internet]. 2011 Jan [cited 2011 Jun 11];40(1):387–426. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20717559

50. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem [Internet]. American Chemical Society; 2006 Feb 1 [cited 2015 Feb 2];78(3):779–87. Available from: http://dx.doi.org/10.1021/ac051437y

51. Smith ACA, Tauten- R, Neumann S, Ben- P, Conley C. Package " xcms ." 2014;

52. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics [Internet]. 2015 Apr 16 [cited 2015 Apr 17];16(1):118. Available from: http://www.biomedcentral.com/1471-2105/16/118

53. Kamleh MA, Ebbels TMD, Spagou K, Masson P, Want EJ. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. Anal Chem. 2012 Mar;84(6):2670–7.

54. Draisma HHM, Reijmers TH, van der Kloet F, Bobeldijk-Pastorova I, Spies-Faber E, Vogels JTWE, et al. Equating, or correction for between-block effects with application to body fluid LC-MS and NMR metabolomics data sets. Anal Chem [Internet]. 2010 Mar 1;82(3):1039–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20052990

55. Draisma H, Kloet F Van Der, Reijmers T, Meulman J, Burk FE Van, Bartels M, et al. Fusion of LC – MS data from different measurement sessions using pooled blood plasma as transfer sample. 2007;1:2007.

56. Want E, Masson P. Processing and analysis of GC/LC-MS-based metabolomics data. Methods Mol Biol [Internet]. 2011 Jan [cited 2013 Jan 21];708:277–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21207297

57. Kamleh MA, Ebbels TMD, Spagou K, Masson P, Want EJ. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. Anal Chem [Internet]. 2012 Mar 20;84(6):2670–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22264131

58. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011 Jul;6(7):1060–83.

59. Koenker R. Quantile Regression (Econometric Society Monographs) [Internet]. Cambridge University Press; 2005 [cited 2013 Jun 24]. 366 p. Available from: http://www.amazon.com/Quantile-Regression-Econometric-Society-Monographs/dp/0521608279

60. Koenker R. Package "quantreg" [Internet]. 2013 [cited 2014 Feb 19]. Available from: http://cran.r-project.org/web/packages/quantreg/quantreg.pdf

61. Ai F-F, Bin J, Zhang Z, Huang J, Wang J, Liang Y, et al. Application of random forests to select premium quality vegetable oils by their fatty acid composition. Food Chem [Internet]. 2014 Jan 15 [cited 2015 Dec 30];143:472–8. Available from: http://www.sciencedirect.com/science/article/pii/S0308814613010820

62. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based

Complement Alternat Med [Internet]. 2013 Jan;2013:298183. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3594909&tool=pmcentrez&rendertype=abstract

63. Li S, Harner EJ, Adjeroh D a. Random KNN feature selection - a fast and stable alternative to Random Forests. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2011 Jan [cited 2012 Nov 22];12(1):450. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3281073&tool=pmcentrez&rendertype=abstract

64. Touw Wouter, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M. Data mining in the Life Sciences with Random Forest : a walk in the park or lost in the jungle ? Brief Bioinform. 2012;

65. Huang J-H, Yan J, Wu Q-H, Duarte Ferro M, Yi L-Z, Lu H-M, et al. Selective of informative metabolites using random forests based on model population analysis. Talanta [Internet]. 2013 Dec 15 [cited 2016 Feb 3];117:549–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24209380

66. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(December):18–22.

67. Breiman L. Random Forests. Mach Learn [Internet]. Kluwer Academic Publishers; 2001 Oct 1 [cited 2014 Jan 19];45(1):5–32. Available from: http://link.springer.com/article/10.1023/A:1010933404324

68. Random forests - classification description [Internet]. Available from: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

69. Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, et al. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. Metabolomics [Internet]. 2011 Jan 20 [cited 2011 Oct 18];7(4):549–58. Available from: http://www.springerlink.com/index/10.1007/s11306-011-0274-7

70. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics [Internet]. 2008 Jan [cited 2013 Nov 7];9(1):307. Available from: http://www.biomedcentral.com/1471-2105/9/307

71. Saccenti E, Hoefsloot HCJ, Smilde AK, Westerhuis JA, Hendriks MMWB. Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics [Internet]. 2013 Oct 26 [cited 2013 Nov 11]; Available from: http://link.springer.com/10.1007/s11306-013-0598-6

72. Tripolt NJ, Narath SH, Eder M, Pieber TR, Wascher TC, Sourij H. Multiple risk factor intervention reduces carotid atherosclerosis in patients with type 2 diabetes. Cardiovasc Diabetol [Internet]. 2014 Jan [cited 2016 Jan 4];13:95. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4041351&tool=pmcentrez&rendertype=abstract

73. Friedrich N. Metabolomics in Diabetes Research. J Endocrinol. 2012 Jun;

74. Roberts LD, Koulman A, Griffin JL. Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome. Lancet Diabetes Endocrinol [Internet]. Home | Journals The Lancet The Lancet Diabetes & Endocrinology The Lancet Global Health The Lancet Haematology The Lancet HIV The Lancet Infectious Diseases The Lancet Neurology The Lancet Oncology The Lancet Psychiatry The Lancet Respiratory Medicine | C; 2014 Jan [cited 2014 Feb 3];2(1):65–75. Available from: http://www.sciencedirect.com/science/article/pii/S2213858713701438

75. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011 Apr;17(4):448–53.

76. Wallaschofski H. What will metabolomics studies mean to endocrinology? J Endocrinol [Internet]. 2012 Oct [cited 2013 Jan 23];215(1):1–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22761276

77. Bain JR, Stevens RD, Wenner BR, Ilkayeva O, Muoio DM, Newgard CB. Metabolomics applied to diabetes research: moving from information to knowledge. Diabetes. 2009 Nov;58(11):2429–43.

78. Bain JR. Targeted metabolomics finds its mark in diabetes research. Diabetes [Internet]. American Diabetes Association; 2013 Feb 1 [cited 2013 Apr 11];62(2):349–51. Available from: /han/4943_0/diabetes.diabetesjournals.org/content/62/2/349.full

79. NutriTech [Internet]. [cited 2016 Feb 9]. Available from: http://www.nutritech.nl/nutritech/42521/7/0/30

80. Gralka AE, Luchinat C, Tenori L, Ernst B. Title : The metabolomic fingerprint of severe obesity is dynamically affected by bariatric surgery in a procedure-dependent manner. (3):1–37.

81. Tang WHW, Wang Z, Levison BS, Koeth R a, Britt EB, Fu X, et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. N Engl J Med [Internet]. 2013 Apr 25 [cited 2014 Mar 23];368(17):1575–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3701945&tool=pmcentrez&rendertype=abstract

82. Wurtz P, Havulinna AS, Soininen P, Tynkkynen T, Prieto-Merino D, Tillin T, et al. Metabolite Profiling and Cardiovascular Event Risk: A Prospective Study of Three Population-Based Cohorts. Circulation [Internet]. 2015 Jan 8 [cited 2015 Jan 12];CIRCULATIONAHA.114.013116 – . Available from: http://circ.ahajournals.org/content/early/2015/01/08/CIRCULATIONAHA.114.013116.abstract

83. Hughes G, Cruickshank-Quinn C, Reisdorph R, Lutz S, Petrache I, Reisdorph N, et al. MSPrep--summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. Bioinformatics [Internet]. 2014 Jan 1 [cited 2016 Feb 3];30(1):133–4. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3866554&tool=pmcentrez&rendertype=abstract

84. Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, Gronwald W. State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics [Internet]. 2011 Aug 12 [cited 2011 Aug 22]; Available from: http://www.springerlink.com/index/10.1007/s11306-011-0350-z

85. Ejigu BA, Valkenborg D, Baggerman G, Vanaerschot M, Witters E, Dujardin J-C, et al. Evaluation of Normalization Methods to Pave the Way Towards Large-Scale LC-MS-Based Metabolomic Profiling Experiments. OMICS [Internet]. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2013 Jun 29 [cited 2013 Aug 26]; Available from: http://online.liebertpub.com/doi/abs/10.1089/omi.2013.0010

86. Ranjbar MRN, Zhao Y, Tadesse MG, Wang Y, Ressom HW. Evaluation of Normalization Methods for Analysis of LC-MS Data.

87. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat Protoc [Internet]. 2011 May [cited 2011 Jun 11];6(6):743–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21637195

88. Wang S-Y, Kuo C-H, Tseng YJ. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Met. Anal Chem [Internet]. American Chemical Society; 2013 Jan 15 [cited 2014 May 16];85(2):1037–46. Available from: http://dx.doi.org/10.1021/ac302877x

89. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc [Internet]. 2011 Jul [cited 2012 Jul 12];6(7):1060–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21720319

90. Kirwan J a, Broadhurst DI, Davidson RL, Viant MR. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Anal Bioanal Chem [Internet]. 2013 Jun [cited 2014 May 12];405(15):5147–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23455646

91. Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J. Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. J Proteomics [Internet]. Elsevier B.V.; 2012 Jul 16 [cited 2012 Jul 31];75(13):3938–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22588121

92. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem [Internet]. 2012 Jun 5;84(11):5035–9. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3703953&tool=pmcentrez&rendertype=abstract

93. Taverner T, Karpievitch Y V, Polpitiya AD, Brown JN, Dabney AR, Anderson GA, et al. DanteR: an extensible R-based tool for quantitative analysis of -omics data. Bioinformatics [Internet]. 2012 Sep 15 [cited 2014 Apr 28];28(18):2404–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3436848&tool=pmcentrez&rendertype=abstract

94. Berg M, Vanaerschot M, Jankevics A, Cuypers B, Breitling R, Dujardin J-C. LC-MS metabolomics from study design to data-analysis – using a versatile pathogen as a test case. Comput Struct Biotechnol J [Internet]. 2013 Jan 1 [cited 2013 Mar 18];4(5). Available from: http://journals.sfu.ca/rncsb/index.php/csbj/article/view/csbj.201301002/201

95. Kirwan J a, Broadhurst DI, Davidson RL, Viant MR. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Anal Bioanal Chem. 2013 Mar;

96. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis [Internet]. Future Science Ltd London, UK; 2012 Sep 9 [cited 2013 Mar 18];4(18):2249–64. Available from: http://www.future-science.com/doi/abs/10.4155/bio.12.204

97. Gika HG, Theodoridis GA, Wingate JE, Wilson ID. Within-Day Reproducibility of an HPLC-MS-Based Method for Metabonomic Analysis : Application to Human Urine research articles. 2007;

98. Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S, et al. Bioinformatics tools for cancer metabolomics. Metabolomics [Internet]. 2011 Jan 12 [cited 2011 Aug 16];7(3):329–43. Available from: http://www.springerlink.com/index/10.1007/s11306-010-0270-3

99. Van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics [Internet]. 2006 Jan [cited 2012 Oct 28];7:142. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1534033&tool=pmcentrez&rendertype=abstract

100. Dunn WB, Lin W, Broadhurst D, Begley P, Brown M, Zelena E, et al. Molecular phenotyping of a UK population: defining the human serum metabolome. Metabolomics [Internet]. 2014 Jul 25 [cited 2014 Sep 9];11(1):9–26. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4289517&tool=pmcentrez&rendertype=abstract

101. Wang S-Y, Kuo C-H, Tseng YJ. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order

Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Met. Anal Chem [Internet]. American Chemical Society; 2013 Jan 15 [cited 2014 May 16];85(2):1037–46. Available from: http://dx.doi.org/10.1021/ac302877x

102. Lee J, Park J, Lim M, Seong SJ, Seo JJ, Park SM, et al. Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. Anal Sci [Internet]. 2012 Jan [cited 2015 Feb 16];28(8):801–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22878636

103. Liu X, Zhang Z, Sousa PFM, Chen C, Ouyang M, Wei Y, et al. Selective iteratively reweighted quantile regression for baseline correction. Anal Bioanal Chem [Internet]. 2014 Mar [cited 2015 Feb 16];406(7):1985–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24429977

104. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Curr Bioinform [Internet]. 2012 Mar;7(1):96–108. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299976&tool=pmcentrez&rendertype=abstract

105. Milburn M V, Lawton K a. Application of metabolomics to diagnosis of insulin resistance. Annu Rev Med [Internet]. 2013 Jan [cited 2014 Jul 14];64:291–305. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23327524

106. Tripolt NJ, Narath SH, Eder M, Pieber TR, Wascher TC, Sourij H. Multiple risk factor intervention reduces carotid atherosclerosis in patients with type 2 diabetes. Cardiovasc Diabetol [Internet]. 2014 Jan [cited 2016 Jan 4];13:95. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4041351&tool=pmcentrez&rendertype=abstract

107. Shah SH, Crosslin DR, Haynes CS, Nelson S, Turer CB, Stevens RD, et al. Branched-chain amino acid levels are associated with improvement in insulin resistance with weight loss. Diabetologia [Internet]. 2012 Feb [cited 2013 Mar 6];55(2):321–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22065088

108. Mutch DM, Fuhrmann JC, Rein D, Wiemer JC, Bouillot J-L, Poitou C, et al. Metabolite profiling identifies candidate markers reflecting the clinical adaptations associated with Roux-en-Y gastric bypass surgery. PLoS One [Internet]. 2009 Jan [cited 2013 Mar 14];4(11):e7905. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2775672&tool=pmcentrez&rendertype=abstract

109. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, et al. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. Cell Metab [Internet]. Elsevier Ltd; 2009 Apr [cited 2014 Jul 15];9(4):311–26. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3640280&tool=pmcentrez&rendertype=abstract

110. Walford G, Davis J, Warner S, Ackerman RJ, Billings LK, Chamarthi B, et al. Branched chain and aromatic amino acids change acutely following two medical therapies for type 2 diabetes mellitus. Metabolism [Internet]. Elsevier Inc.; 2013 Dec [cited 2014 Feb 12];62(12):1772–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23953891

111. Ferreira Nicoletti C, Morandi Junqueira-Franco MV, Dos Santos JE, Sergio Marchini J, Junior WS, Nonino CB. Protein and amino acid status before and after bariatric surgery: A 12-month follow-up study. Surg Obes Relat Dis [Internet]. Elsevier; 2013 Jan 1 [cited 2014 Jan 17];9(6):1008–12. Available from: http://www.soard.org/article/S1550-7289(13)00231-1/abstract

112. Hanzu FA, Vinaixa M, Papageorgiou A, Párrizas M, Correig X, Delgado S, et al. Obesity rather than regional fat depots marks the metabolomic pattern of adipose tissue: an untargeted metabolomic approach. Obesity (Silver Spring) [Internet]. 2014 Mar 26 [cited 2014 Feb 11];22(3):698–704. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23804579

113. Ho JE, Larson MG, Vasan RS, Ghorbani A, Cheng S, Rhee EP, et al. Metabolite profiles during oral glucose challenge. Diabetes [Internet]. 2013 Aug [cited 2013 Nov 12];62(8):2689–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23382451

114. Friedrich N, Budde K, Wolf T, Jungnickel A, Grotevendt A, Dressler M, et al. Short-term changes of the urine metabolome after bariatric surgery. OMICS [Internet]. 2012 Nov [cited 2014 Mar 8];16(11):612–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23095112

115. Wang-sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, et al. Novel biomarkers for pre-diabetes identified by metabolomics. 2012;(615).

116. Floegel A, Stefan N, Yu Z, Mühlenbruch K, Drogan D, Joost H-G, et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes [Internet]. 2013 Feb [cited 2015 Dec 14];62(2):639–48. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554384&tool=pmcentrez&rendertype=abstract

117. Walford G, Porneala BC, Dauriz M, Vassy JL, Cheng S, Rhee EP, et al. Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. Diabetes Care [Internet]. 2014 Sep [cited 2014 Sep 14];37(9):2508–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24947790

118. Ernährungsmedizin DA-G(D) DG für PM und PDG für. S3-Leitlinie : Chirurgie der Adipositas [Internet]. Evaluation. 2010. p. 1–59. Available from: http://www.adipositas-gesellschaft.de/fileadmin/PDF/Leitlinien/ADIP-6-2010.pdf

119. Heber D, Greenway FL, Kaplan LM, Livingston E, Salvador J, Still C. Endocrine and Nutritional Management of the Post-Bariatric Surgery Patient: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab [Internet]. 2010;95(11):4823–43. Available from: http://press.endocrine.org/doi/abs/10.1210/jc.2009-2128

120. Arora P. Metabolomics yield a novel biomarker for predicting diabetes mellitus risk in humans. Circ Cardiovasc Genet [Internet]. 2014 Feb 1 [cited 2014 Feb 22];7(1):95–6. Available from: http://circgenetics.ahajournals.org/content/7/1/95.short

121. Sigruener A, Kleber ME, Heimerl S, Liebisch G, Schmitz G, Maerz W. Glycerophospholipid and sphingolipid species and mortality: the Ludwigshafen Risk and Cardiovascular Health (LURIC) study. PLoS One [Internet]. 2014 Jan [cited 2015 Apr 1];9(1):e85724. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3895004&tool=pmcentrez&rendertype=abstract

122. Saccenti E, Suarez-Diez M, Luchinat C, Santucci C, Tenori L. Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. J Proteome Res [Internet]. 2015 Feb 6;14(2):1101–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25428344

123. Ashrafian H, Li J V, Spagou K, Harling L, Masson P, Darzi A, et al. Bariatric surgery modulates circulating and cardiac metabolites. J Proteome Res [Internet]. 2014 Feb 7;13(2):570–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24279706

124. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature [Internet]. Nature Publishing Group; 2011 Apr 7 [cited 2013 Feb 28];472(7341):57–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3086762&tool=pmcentrez&rendertype=abstract

125. Wang Z, Tang WHW, Buffa JA, Fu X, Britt EB, Koeth RA, et al. Prognostic value of choline and betaine depends on intestinal microbiota-generated metabolite trimethylamine-N-oxide. Eur Heart J [Internet]. 2014 Apr 3 [cited 2015 Jan 14];35(14):904–10. Available from: http://eurheartj.oxfordjournals.org/content/early/2014/02/03/eurheartj.ehu002.short

126. Panunzi S, Carlsson L, De Gaetano A, Peltonen M, Rice T, Sjöström L, et al. Determinants of Diabetes Remission and Glycemic Control After Bariatric Surgery. Diabetes Care [Internet]. 2016 Jan 1 [cited 2016 Jan 13];39(1):166–74. Available from: http://care.diabetesjournals.org/content/early/2015/11/29/dc15-0575

127. Klein MS, Shearer J. Metabolomics and Type 2 Diabetes : Translating Basic Research into Clinical Application. J Diabetes Res. 2016;2016.

128. Boelaert J, t'Kindt R, Schepers E, Jorge L, Glorieux G, Neirynck N, et al. State-of-the-art non-targeted metabolomics in the study of chronic kidney disease. Metabolomics [Internet]. 2013 Oct 26 [cited 2014 Apr 2]; Available from: http://link.springer.com/10.1007/s11306-013-0592-z

129. Glorieux G, Tattersall J. Uraemic toxins and new methods to control their accumulation: game changers for the concept of dialysis adequacy. Clin Kidney J [Internet]. 2015 Aug 1 [cited 2016 Feb 15];8(4):353–62. Available from: http://ckj.oxfordjournals.org/content/early/2015/06/01/ckj.sfv034.full

130. Mullen W, Saigusa D, Abe T, Adamski J, Mischak H. Proteomics and metabolomics as tools to unravel novel culprits and mechanisms of uremic toxicity: instrument or hype? Semin Nephrol [Internet]. 2014 Mar [cited 2016 Feb 23];34(2):180–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24780472

131. Lornoy W, Becaus I, Billiouw JM, Sierens L, Van Malderen P, D'Haenens P. On-line haemodiafiltration. Remarkable removal of beta2-microglobulin. Long-term clinical observations. Nephrol Dial Transplant [Internet]. 2000 Jan [cited 2016 Feb 23];15 Suppl 1:49–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10737167

132. Jean G, Hurot J-M, Deleaval P, Mayor B, Lorriaux C. Online-haemodiafiltration vs. conventional haemodialysis: a cross-over study. BMC Nephrol [Internet]. BioMed Central; 2015 Jan 9 [cited 2016 Feb 23];16(1):70. Available from: http://bmcnephrol.biomedcentral.com/articles/10.1186/s12882-015-0062-0

133. ICH Official web site : ICH [Internet]. [cited 2016 Feb 25]. Available from: http://www.ich.org/home.html

134. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin J a, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ [Internet]. 2013;346:e7586. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3541470&tool=pmcentrez&rendertype=abstract

135. König IR, Weitz G. Wie bereite ich mich auf eine biometrische Beratung vor? Dtsch Medizinische Wochenschrift. 2014;139(46):2354–6.

136. Cavalieri RJ, Rupp EM. Clinical Research Manual: Practical Tools and Templates for Managing Clinical Research. Sigma Theta Tau International; 2013.

137. Ioannidis JP a, Greenland S, Hlatky M a., Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet [Internet]. Elsevier Ltd; 2014;383(9912):166–75. Available from: http://dx.doi.org/10.1016/S0140-6736(13)62227-8

138. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP a, et al. Biomedical research: Increasing value, reducing waste. Lancet. 2014;383(9912):101–4.

139. Leonelli S. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. Big data Soc [Internet]. SAGE Publications; 2014 Jun 1 [cited 2015 Dec 15];1(1):2053951714534395. Available from: http://bds.sagepub.com/content/1/1/2053951714534395.abstract

140. Big Data: A Revolution That Will Transform How We Live, Work and Think: Amazon.co.uk: Viktor Mayer-Schonberger, Kenneth Cukier: 9781848547902: Books [Internet]. [cited 2016 Feb 3]. Available from: http://www.amazon.co.uk/dp/1848547900/ref=as_sl_pd_tf_lc?tag=finantimes-21&camp=1406&creative=6394&linkCode=as1&creativeASIN=1848547900&adid=0D1VHT6NNHB9R64PW6D9&&ref-

refURL=http%3A%2F%2Fwww.ft.com%2Fintl%2Fcms%2Fs%2F2%2Fafc1c178-8045-11e2-96ba-00144feabdc0.html

141. Jain S, Rosenblatt M, Duke J. Is Big Data the New Frontier for Academic-Industry Collaboration ? JAMA. 2014;311:2171–2.

142. Antes G. Eine neue Wissenschaft-(lichkeit)? Laborjournal online [Internet]. 2015 [cited 2016 Feb 2];10. Available from: http://www.laborjournal.de/editorials/981.lasso

143. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar D a., et al. Increasing value and reducing waste in biomedical research: Who's listening? Lancet. 2015;6736(15):1–15.

144. De Fortuny EJ, Martens D, Provost F. Predictive Modeling With Big Data: Is Bigger Really Better? Big Data [Internet]. 2013;1(4):215–26. Available from: http://online.liebertpub.com/doi/abs/10.1089/big.2013.0037

145. Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard D a, Najarian K. Big Data Analytics in Healthcare. Hindawi Publ Corp. 2015;2015:1–16.

146. Costa FF. Big data in biomedicine. Drug Discov Today [Internet]. Elsevier Ltd; 2014;19(4):433–40. Available from: http://dx.doi.org/10.1016/j.drudis.2013.10.012

147. Jameson JL, Ph D, Longo DL. Sounding Board Precision Medicine — Personalized , Problematic , and Promising. New Engl J Med Engl J Med. 2015;372(23):2229–34.

148. Peck RW. The right dose for every patient: a key step for precision medicine. Nat Rev Drug Discov [Internet]. Nature Publishing Group; 2015;1–2. Available from: http://www.nature.com/doifinder/10.1038/nrd.2015.22

149. Katsnelson A. Momentum grows to make " personalized " medicine more " precise ." Nat Publ Gr [Internet]. Nature Publishing Group; 2013;19(3):249. Available from: http://dx.doi.org/10.1038/nm0313-249

150. A C on AF for D, Disease NT of, Sciences B on L, Studies D on E and L. Toward Precision Medicine : Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease Committee on a Framework for Development a New Taxonomy of Disease ; THE NATIONAL ACADEMIES PRESS; 2011.

7. Publications

1 Narath SH, Mautner SI, Svehlikova E, Schultes B, Pieber TR, Sinner FM, Gander E, Libiseller G, Schimek MG, Sourij H, Magnes C. An untargeted metabolomics approach to characterize short-term and long-term metabolic changes after bariatric surgery. PLOS ONE accepted in August 2016

2 Tripolt NJ, Narath SH, Eder M, Pieber TR, Wascher TC, Sourij H. Multiple risk factor intervention reduces carotid atherosclerosis in patients with type 2 diabetes. Cardiovasc Diabetol. 2014;13(1):95.

3. Magnes C, Fauland A, Gander E, Narath S, Ratzer M, Eisenberg T, et al. Polyamines in biological samples: rapid and robust quantification by solid-phase extraction online-coupled to liquid chromatography-tandem mass spectrometry. J Chromatogr A. 2014 Feb 28;1331:44–51.

Conference Contributions

Narath SH, Svehlikova E, Schultes B, Pieber TR, Sinner FM, Gander E, et al. An untargeted metabolomics approach highlights short-term and long-term effects of bariatric surgery in humans. Metabolomics Conference 2015 San Francsico. 2015. p. 306.

Narath, Sophie; Magnes, C; Sourij, H; Pieber, T An Untargeted metabolomics approach to detect biomarkers for the effects of bariatric surgery in humans. Biomarkers and Diagnostics World Congress; MAY 5-7, 2015; Philadelphia, USA. 2015. [Poster]

Narath, SH; Libiseller, G; Fauland, A; Gander, E; Sourij, H; Kleb, U; Svehlikova, E; Pieber, TR; Sinner, FM; Magnes, C A data-drivenapproach to explore the link between metabolomics and diabetes relevant outcomes after bariatric surgery. Proceedings of the 4th European Lipidomic Meeting. 2014; 4th European Lipidomic Meeting; SEP 22-24, 2014; Graz, AUSTRIA. [Poster]

Narath, SH; Libiseller, G; Fauland, A; Gander, E; Sourij, H; Kleb, U; Svehlikova, E; Pieber, TR; Sinner, FM; Magnes, C Short-term effects of bariatric surgery: investigating the link of

metabolomics and insulin resistance with a data-drivenapproach. Metabomeeting; SEP 10-12, 2014; London, UK. 2014. [Poster]

Narath, S; Augustin, T; Sinner, F; Pieber, T; Tripolt, N; Libiseller, G; Sourij, H; Magnes, C Metabolite Biomarker to identify treatment non-responder in type 2 diabetes International Conference of the Metabolomics Society ; JUN 25-28, 2012; Washington, USA. 2012. [Poster]

Kirsch, AH; Narath, S; Krisper, P; Enzinger, G; Gießauf, W; Waller, I; Steppan, S; Tschulena, U; Magnes, C; Rosenkranz, AR; A Prospective, Randomized, Cross-Over, Open Pilot Study to Evaluate the In uence on Metabolome and Proteome in End Stage Renal Disease Patients with Post-Dilution On-Line-HDF versus Conventional Hemodialysis (METAPROL Study) Jahrestagung der österreichischen Gesellschaft für Nephrologie; OCT 1-3, 2015; Alpbach, Österreich. 2015. [Poster]

Zenz, S; Regittnig, W; Urschitz, M; Brunner, M; Korsatko, S; Raml, R; Narath, S; Magnes, C; Tschapeller, B; Augustin, T, Pieber, TR Endogenous Glucose Production during Hypoglycemia in Patients with Newly Diagnosed and Long-Standing Type 1 Diabetes. Diabetes. 2015; 64(S1):A506-A506.-75th Scientific Sessions of the American Diabetes Association (ADA); JUN 5-9, 2015; Boston, USA. [Poster]

Zenz, S; Regittnig, W; Urschitz, M; Brunner, M; Korsatko, S; Raml, R; Narath, S; Magnes, C; Tschapeller, B; Augustin, T; Pieber, TR Endogenous Glucose Production during an induced Hypoglycemia in newly diagnosed and long-term Type 1 Diabetes WIEN KLIN WOCHENSCHR. 2015; 127: S136-S136. [Oral Communication]

Tripolt, N; Narath, S; Eder, M; Pieber, T; Wascher, T; Sourij, H Multifactorial risk factor intervention in patients with type 2 diabetes significantly reduces carotid intima-media thickness. The Central European Journal of Medicine. 2013; 556-556.-44. Jahrestagung der Österreichischen Gesellschaft für Innere Medizin; SEP 26-28, 2013; Salzburg, AUSTRIA. [Poster]

Tripolt, N; Narath, S; Eder, M; Pieber, T; Wascher, T; Sourij, H; Multifactorial risk factor intervention in patients with type 2 diabetes significantly reduces carotid intima-media thickness. WIEN KLIN WOCHENSCHR. 2013; 125(17-18):556-556. [Oral Communication]

Tripolt, NJ; Narath, SH; Eder, M; Pieber, T; Wascher, T; Sourij, H Cardiovascular risk factor treatment in patients with type 2 diabetes significantly reduces carotid intima-media thickness Supplement 02/13 Wiener Klinische Wochenschrift. 2013; S7--41. Jahrestagung der Österreichischen Diabetes Gesellschaft; NOV 21-23, 2013; Salzburg, AUSTRIA. [Oral Communication]

Tripolt, NJ; Narath, SH; Eder, M; Pieber,T; Wascher, T; Sourij, H Effect of multiple risk factori intervention on carotid atherosclerosis in patients with type 2 diabetes Proceedings of International Congress on Coronary Artery Disease 2013. 2013; -ICCAD - International Congress on Coronary Artery Disease ; OCT 13 - 16, 2013; Florence, ITALY. [Poster]

Neubauer, K; Mader, J; Plank, J; Schaupp, L; Beck, P; Augustin, T; Narath, S; Pieske, B; Pieber, T Persistent Hyperglycemia in Hospitalized Patients with Diabetes Despite Considerable Operating Expense. Proceedings of the 72 Scientific Sessions of the American Diabetes Association. 2012; A628-A628.-72nd Scientific Sessions of the ADA (American Diabetes Association); JUN 8-12, 2012; Philadelphia, USA.

Konig, C; Kohler, G; Haas, W; Seereiner, S; Bruner, F; Schwarz, S; Haring, C; Gharibeh, A; Narath, S; Beck, P; Pieber, T Preliminary results of the Diabetic Foot Clinics (DFC) in Styria under the Disease Management Programme "Active therapy diabetes under control" WIEN KLIN WOCHENSCHR. . 2011; 123: S15-S15.-39. Jahrestagung der Österreichischen Diabetes Gesellschaft; NOV 17-19, 2011; Salzburg, AUSTRIA. [Oral Communication]

8. Appendix

8.1. Materials and Methods

Table 16: Tools for metabolomics data analysis printed in Misra & van der Hooft 2016 (26)

Table 1. Table enlisting the metabolomics tools and resources developed during the period 2014–15, consisting of the name, and displaying their platform dependencies in terms of analytical input and computational dependencies, web addresses (URL), and their reference if available

Name	Platform dependencies		URL	Reference
	Analytical Input	Computational		
Data handling and preprocessing tools				
MetMSLine	LC-MS	R	http://wmbedmands.github.io/MetMSLine/	[14]
MRM-DIFF	LC-MS	Windows	http://prime.psc.riken.jp/	[17]
MRMPROBS	LC-MS	Windows	http://prime.psc.riken.jp/	[18]
FragPred	MS	R	http://pattilab.wustl.edu/software/FragPred/index.php	[20]
MUSCLE	LC-MS	C++	http://www.muscleproject.org/	[21]
IsoMS	LC-MS	R, Windows	www.mycompoundid.org/IsoMS	[22]
MyCompoundID.org	MS	Internet	http://mycompoundid.org/	[23]
IsoMS Quant	LC-MS	R, Windows	www.mycompoundid.org/IsoMS	[24]
isoMETLIN	MS/MS	Internet	http://isometlin.scripps.edu/	[25]
IPO	LC-MS	R	https://github.com/glibiseller/IPO	[27]
Massifquant/ Optimize-it	MS	Bioconductor	https://github.com/topherconley/optimize-it	[28]
intCor	LC-MS	R	http://b2slab.upc.edu/software-and-downloads/intensity-drift-correction/	[33]
Peak-group-alignment	LC-MS	R	https://github.com/joewandy/peak-grouping-alignment	[38]
Parametric time warping	LC-MS, LC-UV-DAD	R	https://github.com/rwehrens/ptw	[37]
PeakANOVA	LC-MS	R	http://research.ics.aalto.fi/mi/software/peakANOVA/	[38]
MSPrep	MS	R	http://sourceforge.net/projects/msprep/	[48]
Statistical tools				
DeviumWeb	Any	R, Internet	https://github.com/dgrapov/DeviumWeb	[42]
EigenMS	MS	R, Matlab	http://sourceforge.net/projects/eigenms/	[45]
RepExplore	Any	Internet	http://www.repexplore.tk	[46]
Normalyzer	Any	Internet	http://quantitativeproteomics.org/normalyzer	[47]
BioStatFlow	Any	Internet	http://biostatflow.org/	-
Annotation tools				
POCHEMON	LC-MS	Matlab	http://www.ru.nl/science/analyticalchemistry/research/software/	[49]
BioSM	MS/MS	Java	http://metabolomics.pharm.uconn.edu	[53]
BioSMXpress	—	—	http://engr.uconn.edu/~rajasek/BioSMXpress.zip	[54]
CFM-ID	ESI-MS/MS	Internet	http://cfmid.wishartlab.com/	[56]
GS-align	—	C++	http://www.glycanstructure.org/gsalign	[58]
Cosmiq	LC-MS, GC-MS	R	http://www.bioconductor.org/packages/release/bioc/html/cosmiq.html	[59]
Feature Credential	MS/MS	R	http://pattilab.wustl.edu/software/credential/	[60]
HAMMER	MS/MS	Java, Python	http://www.biosciences-labs.bham.ac.uk/viant/hammer/)	[61]
HR3	LC-MS, GC-MS	Windows	www.metalign.nl	[62]
LipidPro	MS/MS	Windows	http://www.neurogenetics.biozentrum.uni-wuerzburg.de/en/project/services/lipidpro/	[63]
MAGMa	MS/MS	Java	https://www.emetabolomics.org/	[65]
MAIT	MS	R	http://b2slab.upc.edu/software-and-downloads/metabolite-automatic-identification-toolkit/	[67]
Metabolome searcher	—	Internet	http://procyc.westcent.usu.edu/cgi-bin/MetaboSearcher.cgi	[68]
MET-COFEA	LC-MS	Windows	http://bioinfo.noble.org/manuscript-support/met-cofea/	[69]
MetAssign	LC-MS	Java	http://mzmatch.sourceforge.net/MetAssign.php	[70]
MS2Analyzer	MS/MS	Java	http://fiehnlab.ucdavis.edu/projects/MS2Analyzer/	[71]
MIDAS	MS/MS	Internet	http://midas.omicsbio.org	[72]
mzCloud	MS/MS	Internet	https://www.mzcloud.org/	[82]
MS-DIAL	LC-MS	Windows	http://prime.psc.riken.jp/	[78]
mzGroupAnalyzer	LC-MS	Matlab	http://www.univie.ac.at/mosys/software.html	[80]

continued

Table 1. Continued

Name	Platform dependencies		URL	Reference
	Analytical Input	Computational		
ProbMetab	LC-MS	R	http://labpib.fmrp.usp.br/methods/probmetab/	[81]
RAMClustR	MS/MS	R	https://github.com/cbroeckl/RAMClustR	[82]
Pathway and networks analysis, and biological interpretation tools				
PathCaseMAW	Any	Internet	http://nashua.case.edu/PathwaysMAW/Web	[50]
MINE	—	Java, Perl, Python	http://minedatabase.mcs.anl.gov/#/home	[73]
PathWhiz	—	Internet	http://smpdb.ca/pathwhiz	[85]
TrackSM	MS/MS	Matlab	http://metabolomics.pharm.uconn.edu/?q=Software.html	[86]
MarVis-Suite	Any	Java	http://marvis.gobics.de	[87]
InCroMAP	Any	Java	http://www.cogsys.cs.uni-tuebingen.de/software/InCroMAP	[88]
iPEAP	Any	Windows, Java, R	http://www.tongji.edu.cn/~qiliu/ipeap.html	[89]
kpath	Any	Internet	http://browser.kpath.khaos.uma.es/	[90]
Pathomx	Any	Python, Matlab, R	http://pathomx.org/	[91]
MetaMapR	Any	R, Internet	http://dgrapov.github.io/MetaMapR/	[95]
Metabnet	LC-MS	R	https://sourceforge.net/projects/metabnet/	[100]
MetaNET	Any	Galaxy	http://metanet.osdd.net	[105]
Funrich	—	Internet	http://www.funrich.org/	[106]
Network portal	Any	Internet	http://networks.systemsbiology.net	[107]
SimIndex	Any	Python	http://tyolab.northwestern.edu/tools/	[109]
BiNChE	—	Internet	http://www.ebi.ac.uk/chebi/tools/binche/	[110]
PlantSEED	—	Internet	http://bioseed.mcs.anl.gov/~seaver/FIG/seedviewer.cgi?page=PlantSEED	[111]
SPICA	LC-MS	—	http://cmcr.columbia.edu/metabolomics/informaticstools.html	[112]
KiMoSys	Any	Internet	http://kimosys.or	[113]
Integrated interactome system	Any	Internet	http://www.lge.ibi.unicamp.br/lnbio/IIS/	[114]
MetaDB	LC-MS, GC-MS	R	https://github.com/rmylonas/MetaDB	[144]
GC-MS-based tools				
MetaMS	GC-MS	R	http://www.bioconductor.org/packages/release/bioc/html/metaMS.html	[143]
Maui-VIA	GC-MS	Java	http://bimsbstatic.mdc-berlin.de/kempa/software/kempaSoftware.html	[158]
PScore	GC-MS	R	http://raphaelaggio.github.io/	[161]
BiPACE 2D	GC-MS, 2D GC-MS	Java	http://maltcms.sf.net/	[162]
NMR-based Tools				
Focus	NMR	MATLAB	http://www.urr.cat/FOCUS	[164]
Metabnorm	NMR	R	http://sourceforge.net/projects/metabnorm/	[166]
BATMAN	NMR	R	http://batman.r-forge.r-project.org/	[165]
Bayesil	NMR	Internet	http://bayesil.ca/	[167]
SENECA	NMR	Java	http://sourceforge.net/projects/seneca/	[168]
COLMAR	NMR	Internet	http://spin.ccic.ohio-state.edu/index.php/colmar	[169]
$^{1}H(^{13}C)$-TOCCATA	NMR	Internet	http://spin.ccic.ohio-state.edu/index.php/toccata2/index	[170]
mQTL.NMR	NMR	R	http://www.ican-institute.org/tools/	[171]
MVAPACK	NMR	Matlab	http://bionmr.unl.edu/mvapack.php	[172]
ChemoSpec	NMR	R	http://cran.r-project.org/web/packages/ChemoSpec/	[173]
Library, databases, and others				
TAPIR	MS/MS	Python	https://github.com/msproteomicstools/msproteomicstools	[115]
TMDB	—	Internet	http://pcsb.ahau.edu.cn:8080/TCDB/	[121]
BioPhytMol database	—	Internet	http://ab-openlab.csir.res.in/biophytmol/	[122]
EssOilDB	—	Internet	http://nipgr.res.in/Essoildb/	[123]
mVOCs	—	Internet	http://bioinformatics.charite.de/mvoc	[124]
domdb	—	SQL	https://github.com/joefutrelle/domdb	[126]

continued

Table 1. Continued

Name	Platform dependencies		URL	Reference
	Analytical Input	Computational		
PhenoMeter	—	Internet	https://www.metabolome-express.org/phenometer.php	[127]
T3DB	—	Internet	www.t3db.ca	[128]
SwissLipids	MS	Internet	http://www.swisslipids.org/	[129]
Metabolonote	—	Internet	http://metabolonote.kazusa.or.jp/	[141]
KOMICS	—	Internet	http://www.kazusa.or.jp/komics/	[142]
QTREDS	—	Ruby, Rails	http://qtreds.crs4.it	[145]
MASTR-MS	—	Internet	https://mastr-ms.readthedocs.org/en/latest/	[146]
Yabi	—	Internet	http://ccg.murdoch.edu.au/yabi/	[147]
jmzTab	—	—	http://mztab.googlecode.com	[150]
PolySearch2	—	Internet	http://polysearch.ca	[152]
SpeckTackle	MS/MS, NMR	Java	https://bitbucket.org/sbeisken/specktackle	[153]
BioMet Toolbox 2.0	—	Internet	http://biomet-toolbox.org/	[154]
Metabolite Imager	—	Internet	www.metaboliteimager.com	[156]
EXIMS	MALDI-MS	Matlab	http://exims.sourceforge.net/	[157]
Multifunctional tools				
XCMS Online	LC-MS, GC-MS	Internet	https://xcmsonline.scripps.edu/	[177]
Mass++	LC-MS, GC-MS	Java	http://www.first-ms3d.jp/english/	[179]
MASSyPup	LC-MS, GC-MS	Linux	http://www.bioprocess.org/massypup	[180]
MetaboNexus	Any	Windows, Java	http://www.sph.nus.edu.sg/index.php/research-services/research-centres/ceohr/metabonexus	[181]
MetaboLyzer	LC-MS	Python	https://sites.google.com/a/georgetown.edu/fornace-lab-informatics/home/metabolyzer	[182]
Workflow4Metabolomics	Any	Galaxy	http://workflow4metabolomics.org	[183]
Haystack	LC-MS	Internet	http://binf-app.host.ualr.edu/haystack/	[184]
ALLocator	LC-MS	Internet	https://allocator.cebitec.uni-bielefeld.de	[185]
MeKO	GC-MS	Internet	http://prime.psc.riken.jp/meko/	[186]
MassCascade	LC-MS^n	KNIME	https://bitbucket.org/sbeisken/masscascadeknime/wiki/Home	[187]

Terms used in Table: Any, input data from different kinds of analytical platforms; GC, gas chromatography; Internet, tool running on server and assessable by online access; LC, liquid chromatography; MS, mass spectrometry; MS/MS, mass spectrometry fragmentation; NMR, nuclear magnetic resonance spectroscopy; *R, R* package; UV-DAD, ultraviolet diode array detection; —, not applicable/available. Abbreviations used in Table: BATMAN, Bayesian automated metabolite analyzer for NMR; COLMAR, complex mixture analysis by NMR; KNIME, Konstanz Information Miner; MAIT, metabolite automatic identification toolkit; MET-COFEA, metabolite compound feature extraction and annotation; MIDAS, metabolite identification via database searching; iPEAP, integrative Pathway Enrichment Analysis Platform; SPICA, selective paired ion contrast.

8.2. Bariatric Surgery

Table 17: T2DM patients in St. Gallen and Graz

DMR	Graz	St.Gallen	Sum
Non-DM	14	6	20
Complete remission	4	5	9
Non-or partial remission	7	8	15

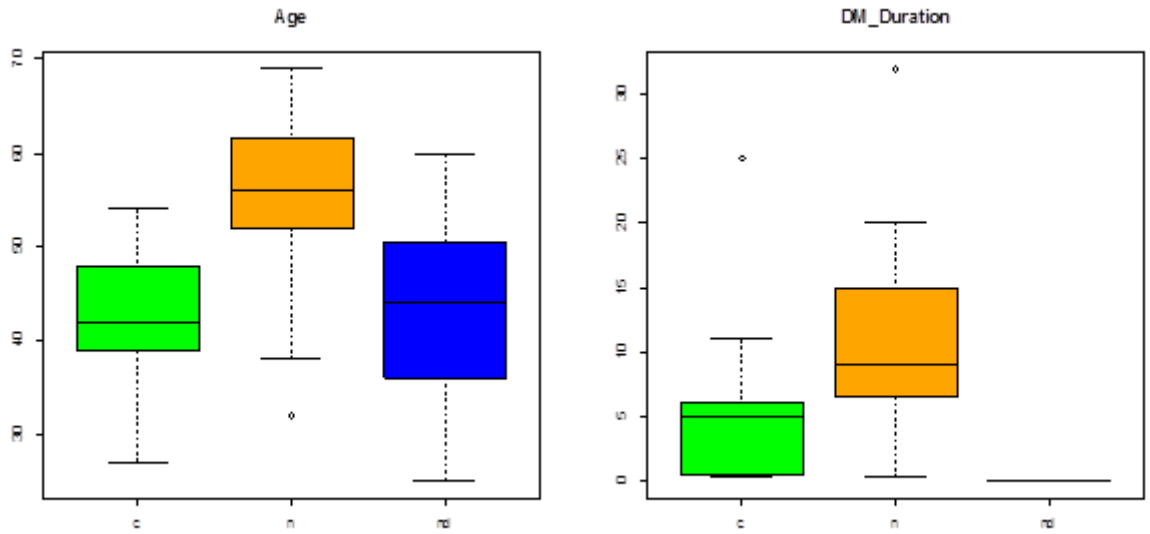


Figure 29: Non-remission patients (n) are significantly older than patients with complete remission (c) (42(9) years vs 55(9)). Nd=non-diabetes

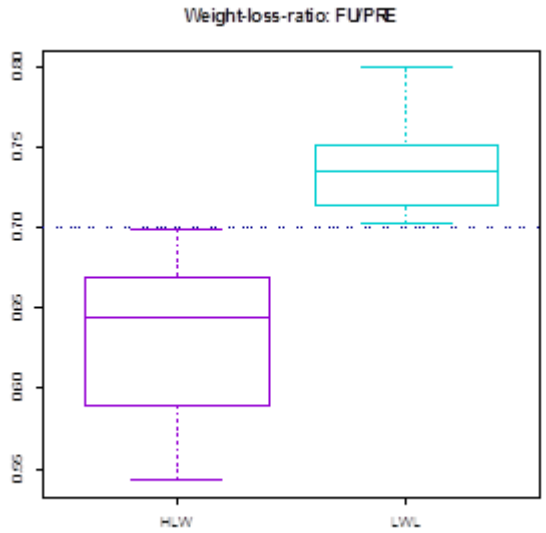


Figure 30: Distribution of high weight loss (HWL) and low weight loss (LWL) group

The median weight reduction follow up was 37.7 kg (iQR: 16.25 kg). We calculated a weight-loss ratio (weight 1 year post surgery-weight at baseline: FU/ POST) and allocated subjects into a high weight loss (HWL) and low weight loss (LWL) group, if they were below or above the median of the weight loss median of 0.7.

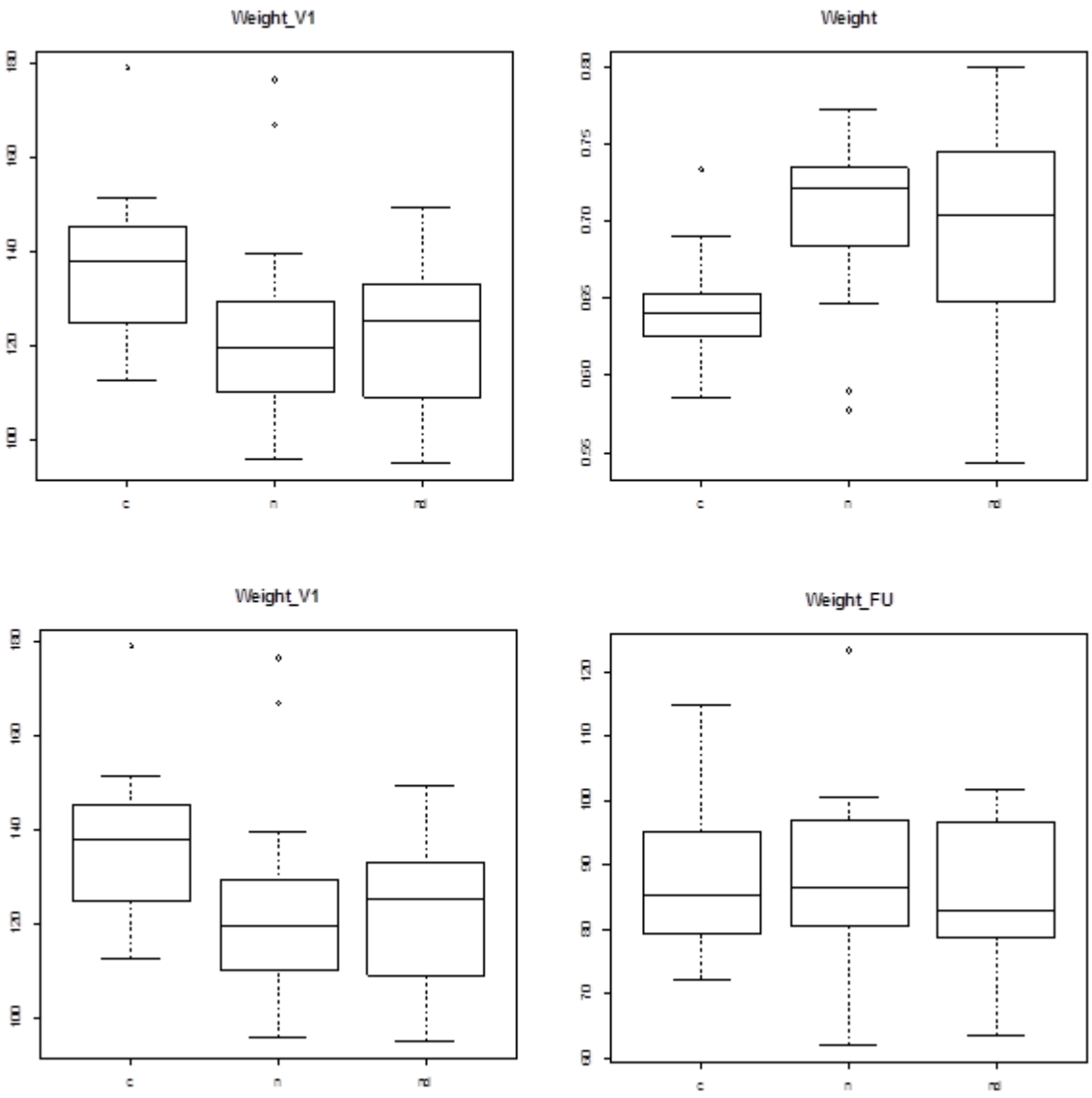


Figure 31: Weight per time and weight reduction for patients with complete and non-remission and non-diabetes patients

Table 18: Identified Metabolites

Feature-ID	Metabolite	HMDB-Number	Identification	Mzmed	Rtmed	ppm (Mz	delta-Rt	Comment
f_M1_nE_2656919	Sarcosine	HMDB00271_Sarcosine	Identified (RT, AM)	88,0386247	11,273666	20	0,57	
f_M1_nE_2656964	Hydroxyisobutyric acid	HMDB00729_Alpha-Hydroxyisobutyric acid	Identified (RT, AM)	103,038651	12,104067	14	0,30	
f_M1_nE_2657012	Oxovaleric acid	HMDB01865_2-Oxovaleric acid	Identified (RT, AM)	115,038377	9,90085	15	1,50	
f_M1_nE_2657015	Acetylglycine	HMDB00532_Acetylglycine	Identified (RT, AM)	116,033688	12,965983	14	0,63	There is a higher systematic ppm deviation for very small
f_M1_nE_2657054	Pyroglutamic acid	HMDB00267_Pyroglutamic acid	Identified (RT, AM)	128,033729	12,8914	12	0,91	molecules in nESI on the Q Exactive System. Values are
f_M1_nE_2657069	Ornithine	HMDB00214_Ornithine	Identified (RT, AM)	131,08152	12,5446415	8	1,16	confirmed by testing the deviation with standards.
f_M1_nE_2657135	Lysine	HMDB00182_Lysine	Identified (RT, AM)	145,096783	13,017792	10	0,78	
f_M1_nE_2657157	Pentoses	HMDB00283_Ribose	Identified (RT, AM)	149,044059	9,959267	10	0,96	
f_M1_nE_2657303	Hydroxydecanoic acid	HMDB_10725_(R)-3-Hydroxydecanoic acid	Putatively annotated	187,132944	9,448833	5		
f_M1_nE_2657381	Indoxyl sulfate	HMDB00682_Indoxyl sulphate	Identified (RT, AM)	212,001331	9,129958	7	1,77	
f_M1_nE_2657402	Pantothenic acid	HMDB00210_Pantothenic acid	Identified (RT, AM)	218,102471	12,69265	4	0,81	
f_M1_nE_2657469	Uridine	HMDB00296_Uridine	Identified (RT, AM)	243,061688	7,20455	2	2,60	
f_M1_pE_2657978	TMAO	HMDB00925_Trimethylamin-n-oxid	Identified (RT, AM)	76,0764079	11,877075	9	0,88	
f_M1_pE_2657993	Alanine	HMDB00161_Alanine	Identified (RT, AM)	90,0556094	12,197708	7	0,60	
f_M1_pE_2658002	Choline	HMDB00097_Choline	Identified (RT, AM)	104,107581	10,061075	0	2,44	
f_M1_pE_2658015	Uracil	HMDB00300_Uracil	Identified (RT, AM)	113,035089	6,981975	5	0,72	
f_M1_pE_2658030	Threonine	HMDB00167_Threonine	Identified (RT, AM)	120,065977	13,10835	4	0,61	
f_M1_pE_2658057	Creatine	HMDB00064_Creatine	Identified (RT, AM)	132,077105	12,1899	3	0,81	
f_M1_pE_2658151	Carnitine	HMDB00062_Carnitine	Identified (RT, AM)	162,112715	10,877383	2	0,62	
f_M1_pE_2658165	Phenylalanine	HMDB00159_Phenylalanin	Identified (RT, AM)	166,086499	9,742742	1	1,06	
f_M1_pE_2658184	Arginine	HMDB00517_Arginine	Identified (RT, AM)	175,119263	12,152066	2	0,95	
f_M1_pE_2658210	Tyrosine	HMDB00158_Tyrosine	Identified (RT, AM)	182,081532	11,621533	2	1,58	
f_M1_pE_2658275	Tryptophan	HMDB00929_Tryptophan	Identified (RT, AM)	205,097272	9,830733	1	1,97	
f_M1_pE_2658333	Leu Pro		Metlin, annotated	229,154834	9,188733	1		
f_M1_pE_2658750	LysoPC C16:1		Putatively annotated, retention time according to compound class	494,324878	4,831617	2	0,17	
f_M1_pE_2658758	LysoPE C20:4		Putatively annotated, retention time according to compound class	502,293597	8,268617	2	0,03	
f_M1_pE_2658777	LysoPC C18:2		Putatively annotated, retention time according to compound class	520,340656	5,1253085	2	0,13	
f_M1_pE_2658959	PC C34:3		Putatively annotated, retention time according to compound class	756,554998	5,274375	2	1,17	
f_M1_pE_2658999	PC C36:6		Putatively annotated, retention time according to compound class	778,53895	4,611767	1	0,51	
f_M1_pE_2659000	PC C36:5		Putatively annotated, retention time according to compound class	780,554963	5,009742	2	0,91	
f_M1_pE_2659031	PC C38:6		Putatively annotated, retention time according to compound class	806,57048	5,205025	1	1,11	
f_M1_pE_2659048	PC C40:7		Putatively annotated, retention time according to compound class	832,586531	5,144725	2	1,04	
f_M1_pE_2658059	Leucin/Isoleucin		Identified (RT, AM)	132,102226	9,449425	6	0,86	metabolites identified with explicitly search
f_M1_pE_2658106	Glutamine		Identified (RT, AM)	147,076703	13,634358	2	0,23	metabolites identified with explicitly search
f_M1_nE_2657016	Valine		Identified (RT, AM)	116,070004	10,1924	15	1,28	metabolites identified with explicitly search
Glycine	Glycine		Identified (RT, AM)	76,0400322	12,2655992	10	1,93	metabolites identified with explicitly search

Table 19: Nutritional Information: Supplements after bariatric surgery.

Supplement	Amount	Frequency
Calcium-Carbonate (-Citrate)	1.5 g	daily
Iron III iv. *or*	200 mg	every 3-6 months
Iron II p.o.	100-200 mg	
Vitamine D3 p.o. *(OleovitD3) or*	1200 IU	
Vitamine D3 i.m.	300 000 IU	every 3-6 months
Vitamine B12 i.m. *(Erycytol)*	1000 µg	every 3-6 months
Vitamine B combination		twice a week
Multivitamin micronutrient supplement		daily

All subjects underwent standardized nutritional counseling and received the same supplementation recommendations according to the guidelines (German S3 guideline on obesity and surgery)[12]

[12] *http://www.adipositas-gesellschaft.de/fileadmin/PDF/Leitlinien/ADIP-6-2010.pdf*

8.3. Metaprol

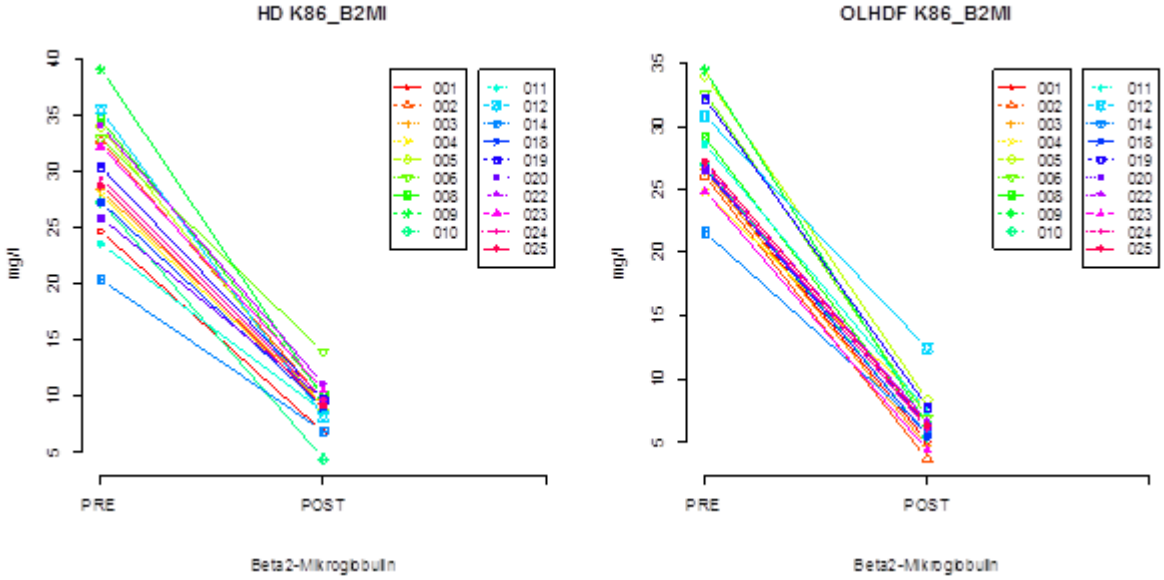


Figure 32: Clearance of Beta2 Microglobulin (mg/l) in HD (left) and OL-HDF (right)

8.4. Pre-clincial Studies: Application of data-driven Workflow

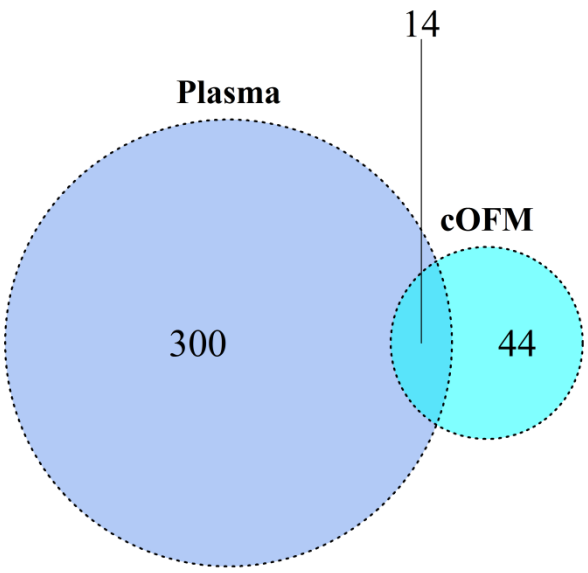


Figure 33: EAE discriminatory features in blood and cOFM samples, 14 metabolic features are in common.

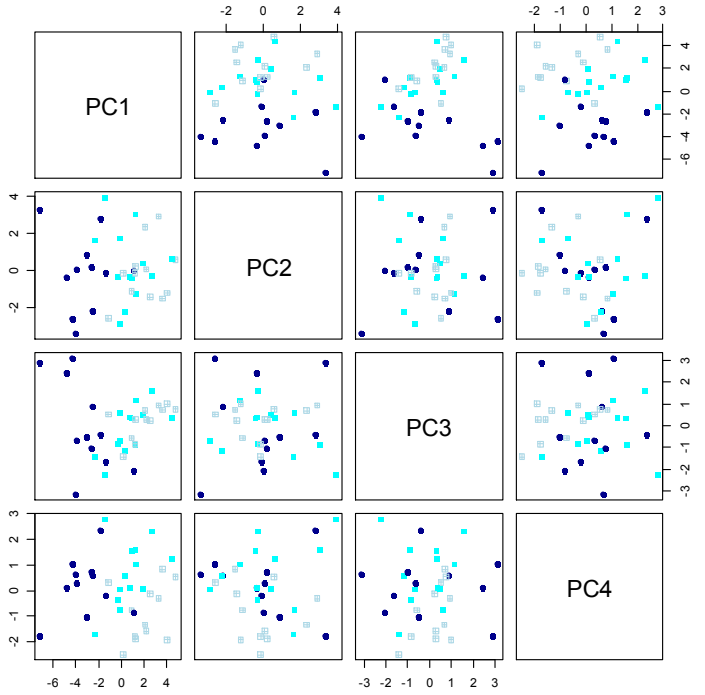


Figure 34: Mouse samples describing clustering for age, despite feeding conditions.

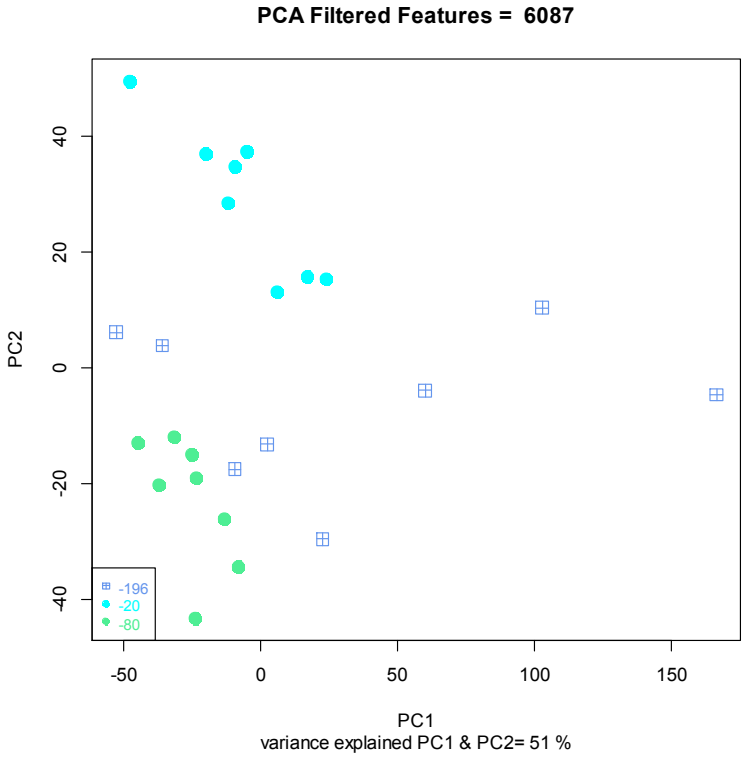


Figure 35: EDTA sample stability: samples frozen at -20°C show distinctive clustering to other termperatures

8.5. Posters

JOANNEUM RESEARCH HEALTH

Metabolite Biomarker to identify treatment non-responder in type 2 diabetes

Narath Sophie[1], *Augustin Thomas*[1], *Sinner Frank*[1,2], *Pieber Thomas*[1,2], *Tripolt Norbert*[2], *Libiseller Gunnar*[1], *Sourij Harald*[2], *Magnes Christoph*[1]

CONTACT

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
Christoph Magnes
Elisabethstraße 5
8010 Graz, Austria
Phone +43 316 876-4000
Fax +43 316 8769-4000
christoph.magnes@joanneum.at
health@joanneum.at
www.joanneum.at/health

Medical University of Graz

2
Medical University of Graz
University Clinic of Internal Medicine
Division of Endocrinology and Metabolism

Objective

Even if known cardiovascular risk factors, such as blood glucose, hypertension or lipids are adequately treated, cardiovascular disease (CVD) risk remains doubled in patients with type 2 diabetes mellitus (T2DM) compared to their non-diabetic counterparts.

The aim of this study is the detection of candidate biomarkers in low molecular weight compounds allowing identifying patients with progressive atherosclerosis despite attempts to intensify risk factor treatment. First results are presented here, using a holistic approach involving a subgroup of a larger trial for developing statistical algorithms.

Study Population

100 T2DM patients not reaching current glucose, blood pressure and blood lipid treatment targets were included in a prospective trial, where risk factor therapy was intensified within 3 months. To develop algorithms we used serum samples of a subgroup of 18 patients. We selected 9 patients, who responded well to cardiovascular risk factor management while the other 9 patients basically did not respond (Figure 1). Serum samples were collected before therapy start.

Methods

LC/FTMS-Metabolite Fingerprints (HILIC and IP/RP-LC coupled to LTQ Orbitrap XL) of deproteinized serum samples, multivariate statistical approach for dimension reduction and data overview (statistical analysis and data preparation were done with R-2.13.1).

Results

1286 Features (filters based on QC-sample analyses were applied, blank and high correlating features-intensities were excluded) from LC/FTMS-Metabolite Fingerprints did not show any clustering between responders and non-responders (PCA, Figure 2). Clinical parameters known to be associated with cardiovascular outcome were chosen to reduce the dimension of available data. Linear models with HbA1c, Carotis Intima-media Thickness (IMT), systolic blood pressure and LDL cholesterol as predictor variables were applied to select highly explanatory features. The selected features are explained by the relevant clinical parameters with a correlation coefficient of more than 0.5. A data overview with PCA of 32 Features shows on the first two components a tendency towards a clustering between responders and non-responders (Figure 3).

The small sample size has to be considered when interpreting the results. However, these tendencies are hypothesis generating and suggest to look more closely and targeted into metabolomics for the upcoming blood samples. We identified 32 features, which represent the basis for further research on the whole study group (100 subjects), where stronger tendencies are expected.

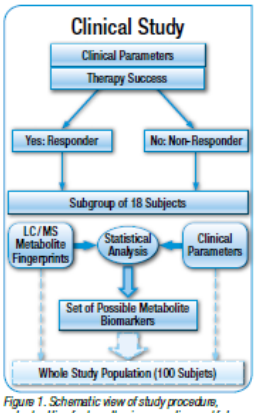


Figure 1. Schematic view of study procedure, actual setting for hypothesis generating and future prospects for the whole study population.

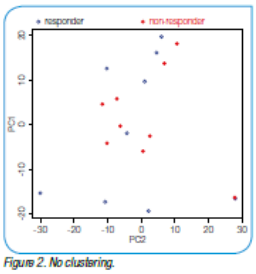


Figure 2. No clustering.

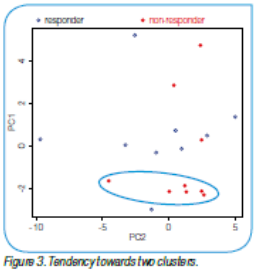


Figure 3. Tendency towards two clusters.

Acknowledgements
This work was supported financially by the Austrian Federal Ministry of Transportation, Innovation and Technology (bmvit), Project MetKAD and the Jubiläumsfond of the Austrian Nationalbank (Proj. 13699 to H.S.).

THE INNOVATION COMPANY

Figure 36: Conference of the Metabolomics Society ; JUN 25-28, 2012; Washington, USA

JOANNEUM RESEARCH HEALTH

Quantile Regression-Based Drift Correction

Applied to Metabolomics Data from Human Serum Samples

Sophie Narath[1,2], Michael G. Schimek[2], Gunnar Libiseller[1], Edgar Gander[1], Harald Sourij[3], Frank M. Sinner[1,3], Thomas R. Pieber[1,3], Christoph Magnes[1]

JOANNEUM RESEARCH HEALTH

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
Sophie Narath
Elisabethstrasse 5
8010 Graz, Austria
Phone +43 316 876-4000
Fax +43 316 8769-4000
health@joanneum.at
www.joanneum.at/health

Medical University of Graz

2
Medical University of Graz
Institute for Medical Informatics, Statistics and Documentation
Graz, Austria

Medical University of Graz

3
Medical University of Graz
Clinic of Internal Medicine
Division of Endocrinology and Metabolism
Graz, Austria

Objective

We aim to compose a workflow to identify and compensate analytical data for bias from various sources (sample collection and preparation, HILIC-FTMS analysis). The workflow comprises data filtering and drift correction. The statistical methods to be applied for signal drift correction depend primarily on data structure, study size, the number of QC samples, and technical specificities of the analytical method. QC samples are generally used to correct metabolomics data[1]. We present here our quantile regression approach.

Methods

R[2] was used for peak detection, peak grouping and the complete workflow for drift correction.

Filtering

Filtering steps are based on excluding technical artefacts, redundant information, batch differentiation, and highly spread features.

Quantile Regression

The main advantage of quantile regression over other regression techniques is its flexibility for modelling data with heterogeneous conditional distributions. Since the variability of features was high, smoothing by a locally adaptive regression technique was required to retrieve the maximum valuable information.

In this case, the 90 % quantile of the QCs/y's) was estimated via a nonparametric quantile regression using regression splines depending on the sample number (x's). This procedure fits a piecewise cubic polynomial with 5 knots (df) (breakpoints in the third derivative) arranged at the 90 % quantile of the x's: rq(y ~ bs(x, df = 5), tau = 0.9).

Through a multiplicative correction factor based on the median of the original QC-values, a further quantile regression model is estimated to correct all samples (Figure 1).

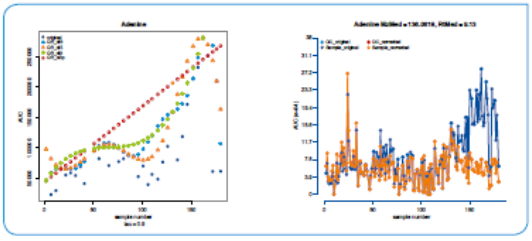


Figure 1: Drift correction using a quantile regression approach. Left: model fits for the QCs. Right: final correction using df = 5 and tau = 0.9

Evaluation of the drift correction

Criteria to evaluate the success of the workflow were overall lower variation in QC samples, graphical representations of drift features and multivariate modelling based on batches as class variables to determine the degree of batch overlap. Batch separation before and after data preparation is compared through multivariate modelling approaches (PCA & Random Forest). We have chosen Random Forest because of its adaptability for nonlinear properties of the data (Figure 2). In the present case 500 trees were used as a default parameter.

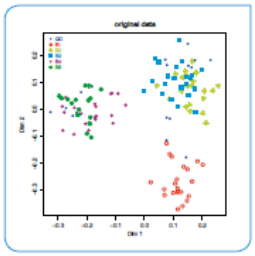


Figure 2: Unsupervised Random Forest calculation from original data, plotting batches and QCs as class variable

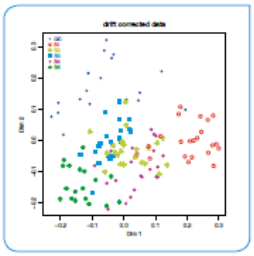


Figure 3: Unsupervised Random Forest calculation from drift corrected data, plotting batches and QCs as class variable

Results

In the present project, metabolite fingerprints of deproteinized serum samples from 106 type 1 diabetes patients were generated using HILIC-FTMS (LTQ Orbitrap XL). Study samples were divided into 5 different analysis batches.

The workflow application resulted in a feature reduction of more than 50 % (from 12,000 initially detected features), lower variation over the QC pool samples (from a RSD of 0.35 to 0.26), and less visible batch clustering (Figure 3).

Discussion

To consider various feasible drift patterns, a systematic analysis of the behaviour of the quantile regression approach for such data is currently under investigation. The workflow will be optimized with data from upcoming clinical studies.

References

[1]
Dunn, W.B., Broadhurst, D., et al. (2011). Nature protocols, 6(7).
Kamleh, M.A., Ebbels, T.M.D., et al. (2012). Analytical chemistry, 84(6).
Kirwan, J.A, Broadhurst, D.I., et al. (2013). Analytical and bioanalytical chemistry.

[2]
R Packages: "xcms", C.A. Smith et al. (2006), "quantreg", R. Koenker (2013), "randomForest", A. Liaw and M. Wiener (2002)

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), project MetZNet.

THE INNOVATION COMPANY

Figure 37: Conference of the Metabolomics Society ; JUL 01-04, 2013; Glasgow, UK

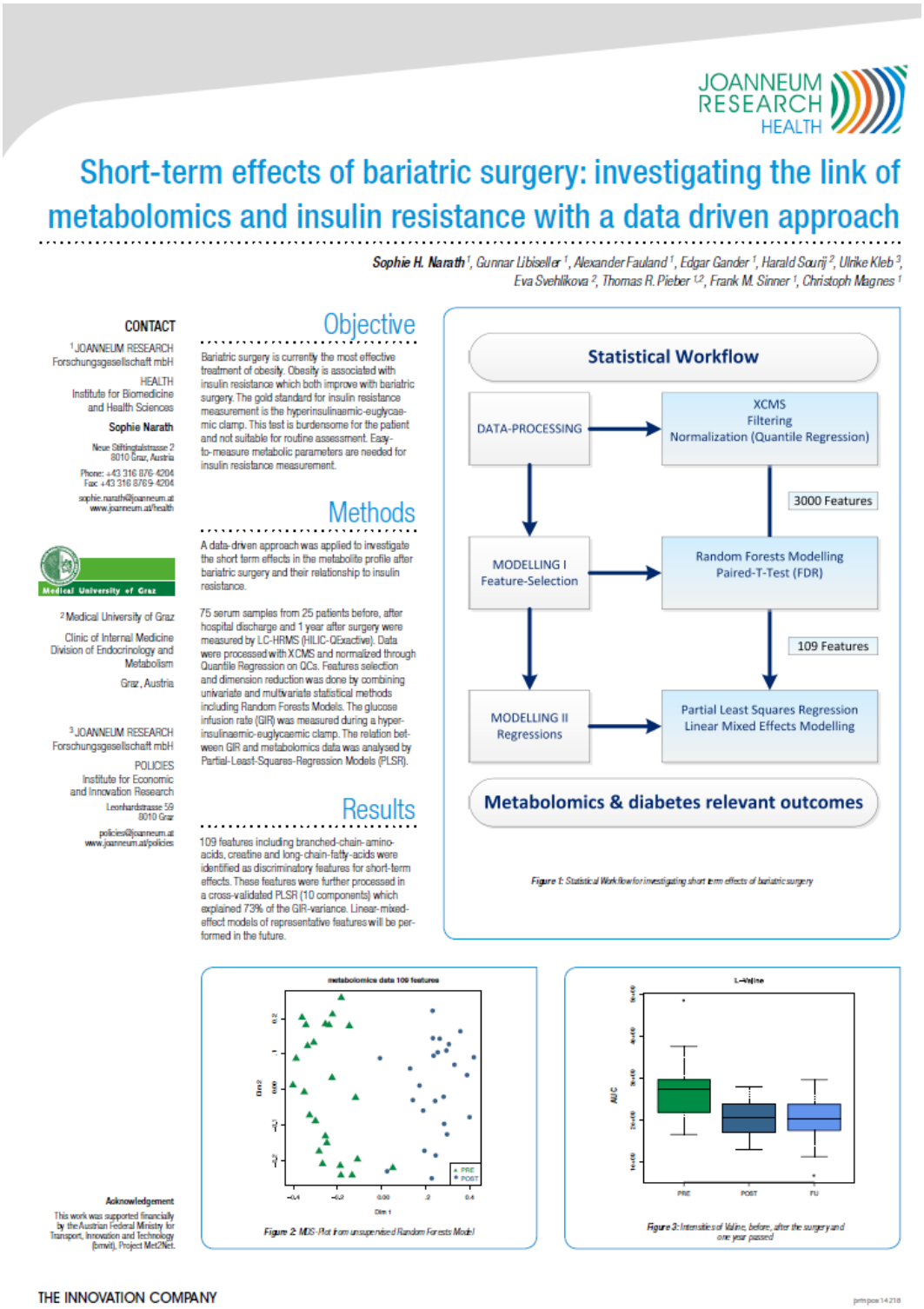


Figure 38: Metabomeeting; SEP 10-12, 2014; London, UK. 2014.

JOANNEUM RESEARCH HEALTH

An untargeted metabolomics approach highlights short-term and long-term effects of bariatric surgery in humans

Sophie H. Narath[1], Eva Svehlikova[2], Bernd Schultes[4], Thomas R. Pieber[1,2,3], Frank M. Sinner[1,2], Edgar Gander[1], Harald Sourij[2,3], Christoph Magnes[1]

CONTACT

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
Sophie H. Narath
Neue Stiftingtalstrasse 2
8010 Graz, Austria
Phone +43 316 876-4100
Fax +43 316 8769-4100
health@joanneum.at
www.joanneum.at/health

2
Medical University of Graz
Division of Endocrinology & Metabolism
Auenbruggerplatz 15
8036 Graz, Austria

3
CBmed
CENTER FOR BIOMARKER RESEARCH IN MEDICINE
Stiftingtalstrasse 5
8010 Graz, Austria

4
eSwiss
Medical and Surgical Center
9016 St. Gallen, Switzerland

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project MezNet, the Österreichische Forschungsförderungsgesellschaft FFG, Project 3014Brain and EFSD/MSD Clinical Research Programme 2009.

References

[1] Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. 2006, 78, 779–787.

[2] Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., Neumann, S., Trausinger, G., Sinner, F., Pieber, T., et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 2015, 16, 1–10.

Introduction

Bariatric surgery is currently considered one of the most effective treatments of obesity, leading to a recovery of the patient's metabolism. Besides the highly individual metabolic effects of bariatric surgery, the reduction of cardiovascular risks is an important aspect to be assessed. The aim of this study was to identify and quantify relevant metabolic changes which occur shortly after the surgery and after one year in a long term follow-up.

Methods

LC-HRMS (HILIC-QExactive) was used to analyze 132 serum samples from 44 patients before surgery (PRE), after hospital discharge (POST) and at a 1 year follow-up (FU). The raw data was processed with open source R package XCMS, optimized by IPO(1,2) and normalized through quantile regression based on quality controls.

Combining Random Forests and paired-t-tests defined a process to select metabolic features that describe the changes between two time points. The same metabolic feature selection process was applied to compare time-point 1 and 2 (PRE-POST) as well as time-point 1 and 3 (PRE-FU). The intersection of the selected characteristic metabolic features frowm both comparisons represents the combined information of short- and long-term effects of bariatric surgery.

Results

Metabolic feature selection resulted in 177 metabolic features that describe short-term and long-term effects and subsequently 36 metabolites were identified.

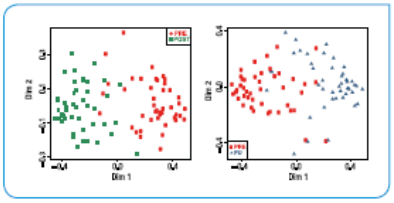


Figure 1: Multi-Dimensional-Scaling showing distinct clustering of plots from supervised Random Forests, showing clustering for samples taken before and after the surgery for both time-points

8 metabolites were linked to cardiovascular risk: TMAO, indoxyl sulphate (increasing trend), choline, alanine, phenylalanin, tyrosine, valine, leucine/ isoleucine (decreasing trend).

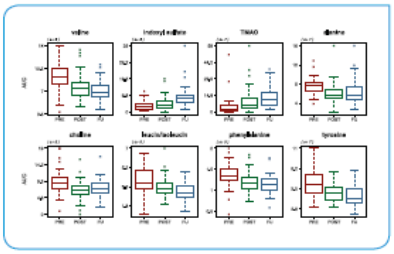


Figure 2: Changes in the intensities (peak-AUC) of identified metabolites before and after bariatric surgery

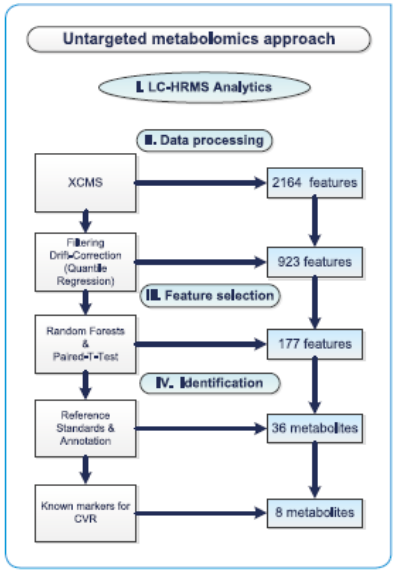


Figure 3: Scheme of untargeted metabolomics approach (CVR=cardiovascular risk)

Conclusion

The identified metabolites showed different trend-patterns over time, highlighting the importance of short and long-term metabolomics assessment after bariatric surgery. For the development of potential prognostic markers, metabolites with a negative or a positive median trend are better suited for further investigations rather than metabolites with an alternating zigzag course. This study links short-term and long-term metabolic changes after bariatric surgery by using an untargeted metabolomics approach.

THE INNOVATION COMPANY

Figure 39: Metabolomics Conference 2015 San Francsico. 2015. p. 306, USA

An untargeted metabolomics approach to detect biomarkers for the effects of bariatric surgery in humans

Sophie H. Narath[1], Christoph Magnes[1], Harald Sourij[2,3], Thomas R. Pieber[1,2,3]

[1] JOANNEUM RESEARCH, HEALTH – Institute for Biomedicine and Health Sciences, Graz, Austria
[2] Medical University of Graz, Division of Endocrinology and Metabolism, Graz, Austria
[3] CBmed – Center for Biomarker Research in Medicine, Stiftingtalstrasse 5, 8010 Graz, Austria

Background

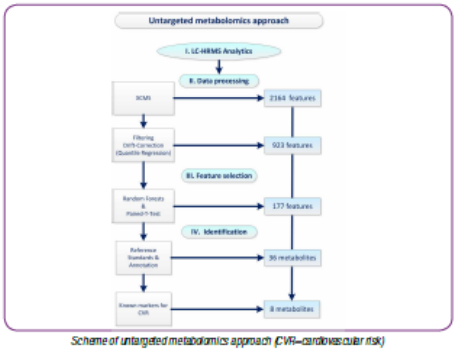
- Bariatric surgery is currently considered one of the most effective treatments of obesity, leading to a recovery of the patient's metabolism.
- An untargeted metabolomic profile can be used as a first step in the development of biomarkers to predict and survey the effects of bariatric surgery.
- Besides the highly individual metabolic effects of bariatric surgery, the reduction of cardiovascular risks is an important aspect to be assessed.

Aims

- The aim of this study was to identify relevant metabolic changes not only shortly after bariatric surgery but also up to one year after surgery in a long-term follow-up.

Methods

- LC-HRMS (HILIC-QExactive) was used to analyze 132 serum samples from 44 patients before surgery (PRE), after hospital discharge (POST) and at a 1 year follow-up (FU). The raw data was processed with open source R package XCMS, optimized by IPO (1,2) and normalized through quantile regression based on quality controls.

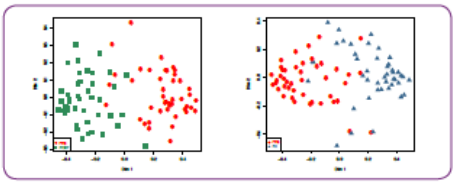


Scheme of untargeted metabolomics approach (CVR–cardiovascular risk)

- Combining Random Forests and paired-t-tests defined a process to select metabolic features that describe the changes between two time points.
- The same metabolic feature selection process was applied to compare time-point 1 and 2 (PRE-POST) as well as time-point 1 and 3 (PRE-FU).
- The intersection of the selected characteristic metabolic features from both comparisons represents the combined information of short- and long-term effects of bariatric surgery.

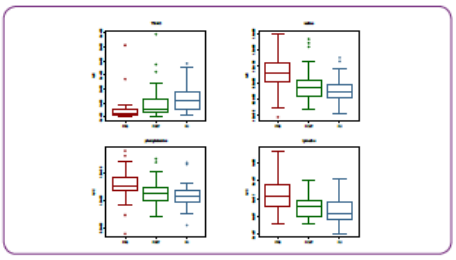
Results

- Metabolic feature selection resulted in 177 metabolic features that describe short-term and long-term effects and subsequently 36 metabolites were identified



Multi-Dimensional-Scaling showing distinct clustering of plots from supervised Random Forests, showing clustering for samples taken before and after the surgery for both time-points.

- 8 metabolites were linked to cardiovascular risk: TMAO, indoxyl sulphate (increasing trend), choline, alanine, phenylalanin, tyrosine, valine, leucine/isoleucine (decreasing trend).



Changes in the intensities (peak-AUC) of identified metabolites before and after bariatric surgery

Conclusions

- We identified short-term and long-term metabolic effects of bariatric surgery in humans through an untargeted metabolomics approach.
- The different identified metabolite-trends highlight the importance of repeated measurements in order to obtain a comprehensive understanding of the mechanisms and potential indications for treatment response in the field of bariatric surgery.
- Our study is also able to give a better insight into changes of previously identified cardiovascular risk factors and potential biomarkers.

(1) Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. 2006, 78, 779–787.
(2) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 2015, 16, 1–10.

funded by:

CONTACT
CBmed
CENTER FOR BIOMARKER RESEARCH IN MEDICINE
Stiftingtalstrasse 5
8010 Graz, Austria
Phone: +43 316 385 28801
office@cbmed.at

CBmed is an internationally recognized biomarker research center with focus on cancer, metabolism and inflammation

Figure 40: Biomarkers and Diagnostics World Congress; MAY 5-7, 2015; Philadelphia, USA.

Metabolomics indicates potential biomarkers for IPAH

Bence M. Nagy[1], Natalie Bordag[2], Sophie H. Narath[3], Vasile Foris[1, 4], Katharina Leithner[4], Gabor Kovacs[1, 4], Andrea Olschewski[1, 5], Horst Olschewski[1, 4], Cristoph Magnes[3]

[1]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, [2]CBmed–Center for Biomarker Research in Medicine, Graz, [3]JOANNEUM RESEARCH, HEALTH–Institute for Biomedicine and Health Sciences, Graz, [4]Department of Internal Medicine, Division of Pulmonology, Medical University of Graz, [5]Department of Anaesthesiology and Intensive Care Medicine, Experimental Anaesthesiology, Medical University of Graz

Introduction

The terminal stage of pulmonary arterial hypertension (PAH) is characterized by a significant reduction in the pulmonary vascular lumen, subsequently leading to right ventricular failure and premature death. In animal models with chronic hypoxia-induced PAH, vascular changes that are characteristic for the disease have been directly linked to an imbalance between glycolysis, glucose oxidation, and fatty acid oxidation, suggesting that there are similarities with cellular mechanisms detected in cancer. Data from both *in vitro* and animal models suggests that metabolic alterations play an important role in the molecular pathogenesis of the early or developing stage of pulmonary hypertension. Therefore detection of metabolic biosignatures may help us reveal potential biomarker candidates in idiopathic PAH (IPAH).

Materials and methods

LC-HRMS (HILIC-QExactive) was used to analyze 16 serum samples from 8 healthy (CRTL) and 8 IPAH patients. The raw data was processed with open source R package XCMS, optimized by IPO, and filtered for analytical stability yielding 1404 untargeted metabolic features. For targeted analysis metabolites were identified by mass and retention time and whenever possible against reference standard yielding 102 suitable for multi- and univariate analysis and additional 33 metabolites suitable for univariate analysis only.

Conclusion

These pilot findings draw our attention towards circulating fatty acids and lipid components in IPAH and provides a starting point for further detailed investigations.

Patient characteristics	IPAH	Ctrl
Number of patients	8	8
Age [years]	61 ± 11	61±11
mPAP [mmHg]	47 ± 9	-
CI [l/min/m²]	2,6 ± 0,8	-
PVRI [WU m²]	15,3 ± 7,6	-

Table 1. Patient characteristics

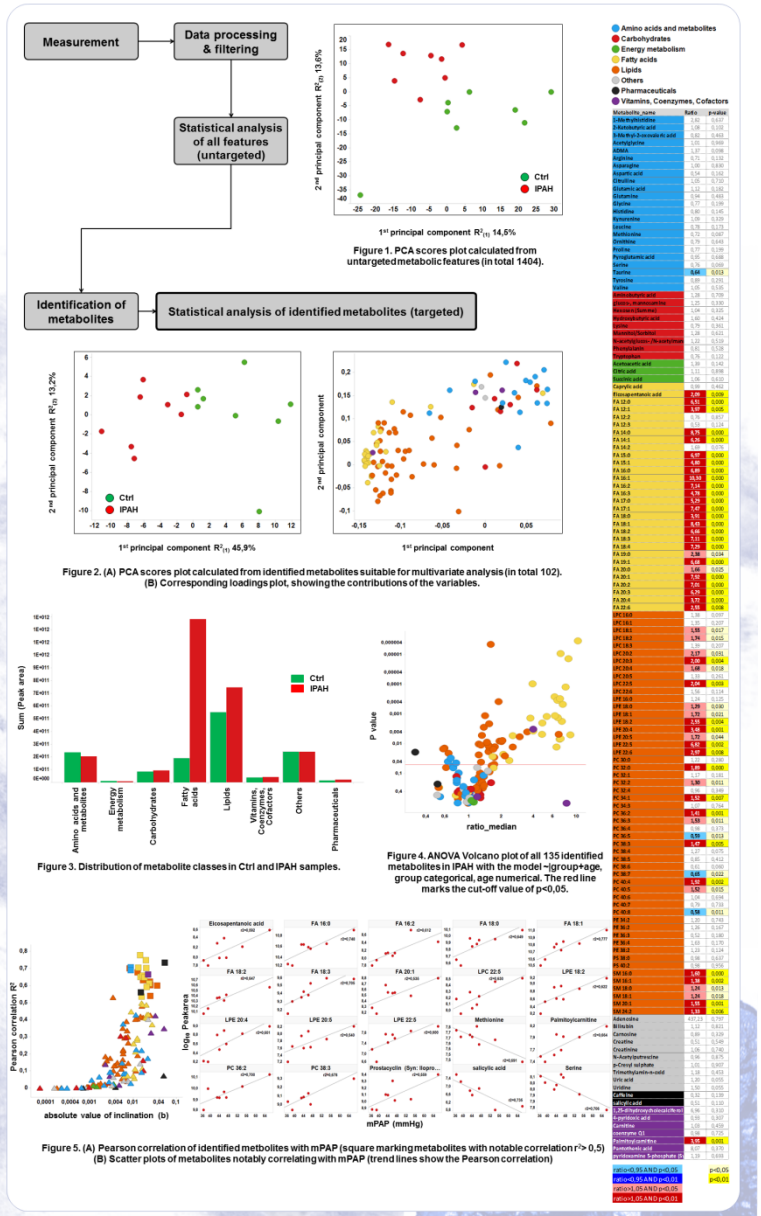


Figure 1. PCA scores plot calculated from untargeted metabolic features (in total 1404).

Figure 2. (A) PCA scores plot calculated from identified metabolites suitable for multivariate analysis (in total 102). (B) Corresponding loadings plot, showing the contributions of the variables.

Figure 3. Distribution of metabolite classes in Ctrl and IPAH samples.

Figure 4. ANOVA Volcano plot of all 135 identified metabolites in IPAH with the model ~group+age, group categorical, age numerical. The red line marks the cut-off value of p<0,05.

Figure 5. (A) Pearson correlation of identified metbolites with mPAP (square marking metabolites with notable correlation r²> 0,5) (B) Scatter plots of metabolites notably correlating with mPAP (trend lines show the Pearson correlation)

Ludwig Boltzmann Institute Lung Vascular Research
Medizinische Universität Graz
CBmed BIOMARKER RESEARCH
JOANNEUM RESEARCH

Acknowledgements
This work has been carried out with the K1 COMET Competence Center CBmed, which is funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG.

Figure 41: Poster presented at Scientific Advisory Board Ludwig Boltzmann Institute 9th and 10th July, Graz, 2015

A Prospective, Randomized, Cross-Over, Open Pilot Study to Evaluate the Influence on Metabolome and Proteome in End Stage Renal Disease Patients with Post-Dilution On-Line-HDF versus Conventional Hemodialysis (METAPROL Study)

19

Alexander H. Kirsch (1), Sophie Narath (2), Peter Krisper (1), Günter Enzinger (3), Werner Gießauf (3), Ingmar Waller (4), Sonja Steppan (5), Ulrich Tschulena (5), Christoph Magnes (2), and Alexander R. Rosenkranz (1).

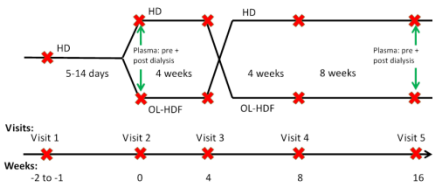
1. Clinical Division of Nephrology, Internal Medicine, Medical University of Graz,
2. Joanneum Research, Graz
3. Dialysis Institute Gießauf, Graz
4. Dialysis Institute Feldbach, Feldbach, Austria
5. Fresenius Medical Care, Bad Homburg, Germany

Medical University of Graz

Introduction

Renal replacement therapy has become a standard of care for patients suffering from end-stage renal disease. Despite numerous technical advances, patients treated with the standard of care i.e. hemodialysis (HD) still experience high annual morbidity and mortality. Evidence of better clearance of middle molecular weight uremic toxins has led to the introduction of hemodiafiltration (HDF), which combines convective solute removal with traditional dialysis. Some studies have shown a survival benefit for patients treated with HDF compared to conventional HD. To date, there is no comprehensive metabolomic and proteomic characterization of the different clearance profiles of these two methods. Such an analysis could serve as a basis for generating hypotheses to explain the potential survival benefit in HDF-treated patients.

Study design



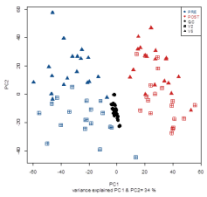
Prevalent dialysis patients were randomly assigned to receive 4 weeks of high-flux HD using FX80 membranes or post-dilution online-HDF using FX800 (both Fresenius Medical Care, Bad Homburg, Germany) membranes. Thereafter, the patients were crossed-over to HDF or HD treatment respectively for 12 weeks. Blood samples for routine clinical parameters as well as for metabolomics and proteomic assessment were drawn at week -2, 0, 4, 8, and 16 before the first randomized treatment. The primary endpoint was the short-term changes in the metabolome and proteome after 4 weeks of HDF versus HD. The secondary endpoints included short-term changes in the metabolome and proteome in HDF vs HD before and after treatments as well as long-term changes after 12 weeks of HDF vs HD.

These preliminary data show that metabolomics are a feasible analytical tool in ESRD (HD) patients. Analysis of proteomic as well as clinical data is still ongoing.

Patient characteristics

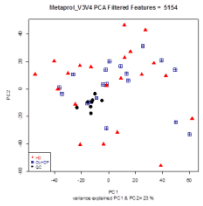
Patient Demography			
Group	HD → OL-HDF	OL-HDF → HD	Total
# patients	11	12	23
Sex (f/m)	3/8	3/9	6/17
Age	66.40 ±9.65	69.13±9.17	67.8
BMI	28.18±7.46	28.05±5.38	28.12
Comorbidities			
Arterial hypertension	8	8	16
Diabetic nephropathy	5	2	7
Diabetic retinopathy	4	1	5
Peripheral artery disease	3	4	7
Renal anemia	10	9	19
sHPT	9	9	18

Pre- vs. Post-Dialysis



Using HR-LCMS we identified > 3400 metabolic features. Differences between pre- and postdialytic samples were clear as 2155 out of a total of 3443 metabolites showed significant changes.

Treatment effect



196 out of 5154 features show a tendency between HD and OLHDF (unadjusted paired t-test < 0.05).

POSTERPRÄSENTATION ÖGN

Österreichische Gesellschaft für Nephrologie

Figure 42: Poster showing results from METAPROL-Study, presented at the ÖNG 1.-3. October 2015