Multivariate statistical methods to analyse multidimensional data in applied life science

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der

Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Frau Trutschel (geb. Boronczyk), Diana
Geb. am 18.02.1979 in Hohenmölsen

Gutachter:

1. Prof. Dr. Ivo Grosse, MLU Halle/Saale
2. Dr. Steffen Neumann, IPB alle/Saale
3. Prof. Dr. André Scherag, FSU Jena

Datum der Verteidigung: 18.04.2019

Eidesstattliche Erklärung / *Declaration under Oath*

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.
I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

____________________________ ____

Datum / Date Unterschrift des Antragstellers / *Signature of the applicant*

This thesis is a cumulative thesis, including five research articles that have previously been published in peer-reviewed international journals. In the following these publications are listed, whereby the first authors are underlined and my name (Trutschel) is marked in bold.

1. **Trutschel, Diana** and Schmidt, Stephan and Grosse, Ivo and Neumann, Steffen, ”Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data”, *Metabolomics*, 2015, available at: `https://link.springer.com/content/pdf/10.1007/s11306-014-0742-y.pdf`

2. **Trutschel, Diana** and Schmidt, Stephan and Grosse, Ivo and Neumann, Steffen, ”Joint analysis of dependent features within compound spectra can improve detection of differential features”, *Frontiers in Bioengineering and Biotechnology*, 2015, available at: `https://www.frontiersin.org/articles/10.3389/fbioe.2015.00129/full`

3. Mönchgesang, Susann and Strehmel, Nadine and **Trutschel, Diana** and Westphal, Lore and Neumann, Steffen and Scheel, Dierk, ”Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent”, *International Journal of Molecular Sciences*, 2016, available at: `http://www.mdpi.com/1422-0067/17/9/1565`

4. **Trutschel, Diana** and Palm, Rebecca and Holle, Bernhard and Simon, Michael, ”Methodological approaches in analysing observational data: a practical example on how to address clustering and selection bias”, *International Journal of Nursing Studies*, 2017, available at: `http://www.sciencedirect.com/science/article/pii/S0020748917301426?via%3Dihub`

5. Palm, Rebecca and **Trutschel, Diana** and Simon, Michael and Bartholomeyczik, Sabine and Holle, Bernhard, ”Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study”, *Journal of the American Medical Directors Association*, 2016, available at: `https://www.jamda.com/article/S1525-8610(15)00557-5/fulltext`

I hereby declare that the copyright of the content of the articles Trutschel et al., 2015b (2), Mönchgesang et al., 2016 (3) and Trutschel et al., 2017 (4) is by the authors (under Creative Commons License).

I hereby declare that the copyright of the content of the article Trutschel et al., 2015a (1) is by ©Springer Science+Business Media New York 2015.

I hereby declare that the copyright of the content of the article Palm et al., 2016 (5) is by ©2016 AMDA – The Society for Post-Acute and Long-Term Care Medicine.

Zusammenfassung

Angwandte Lebenswissenschaften sind interdisziplinäre Forschungsbereiche, die umfangreiche statistische und computergestützte Methoden benötigen um gesammelte Daten zu organisieren, visualisieren und analysieren, insbesondere seitdem die Komplexizität der Daten in diesen Forschungsfeldern selbst immer mehr zunimmt.

Für die Datenanalyse ist es wiederum wichtig, Studien durchdacht zu konzipieren und geeignete Methoden für die Analyse zu wählen, um valide Ergebnisse für wissenschaftliche Entscheidungen zu erhalten. Um dieses Ziel zu erreichen, wird Wissen über die Datenstruktur und -eigenschaften benötigt, unabhängig in welchem wissenschaftlichen Bereich gearbeitet wird. Benutzerfreundliche Programme, die komplizierte mathematische und computergestützte Methoden aufbereiten und für den praktischen Anwender zugänglich machen, sind dabei ebenfalls unverzichtbar geworden.

Der Fokus dieser Dissertation liegt auf der methodologischen Erarbeitung solcher Verfahren ebenso wie auf deren Anwendung bei der Analyse von Daten in realen Studien. Die Herausforderungen bei der Datenanalyse kommen durch die verschiedenen Dateneigenschaften zustande und werden hier am Beispiel von zwei Lebenswissenschaften aufgezeigt: Metabolomik und Gesundheitsversorgung. Während meiner Arbeit in beiden Bereichen hat sich gezeigt, dass obwohl beide Wissenschaften verschiedene Fragestellungen zu beantworten versuchen, die methodische Vorgehensweise ebenso wie die mathematischen Lösungsansätze ähnlich sind.

Metabolomik ist eine Schlüsseldisziplin in der Systembiologie. Das komplette Set an kleinen Molekülen in einem Organismus, das Metabolom, wird hier untersucht. Zur Identifizierung und Quantifizierung dieser kleinen Moleküle (Metabolite) in solchen komplexen Gemischen werden oft Methoden der Massenspektrometrie genutzt. Die Metabolomforschung beschäftigt sich mit metabolischen und regulatorischen Mechanismen, die das Wachstum, die Entwicklung und die Stressantwort von Organismen beeinflussen. Einen großen Teil nehmen dabei Experimente mit analytischem Character ein um diese Informationen zu erhalten. Die Daten aus solchen Experimenten müssen jedoch mit geeigneten Methoden ausgewertet werden können.

Ein Teilbereich der Gesundheitsversorgung ist die Pflegewissenschaft. Sie hat unter anderem zum Ziel, die Pflegepraxis anzuleiten und die Pflege und Lebensqualität der Patienten zu verbessern. Durch die Überalterung der Gesellschaft liegt heutzutage ein vermehrtes Interesse auf der Reduktion der Gesundheitsversorgungskosten und der Erhaltung der Lebensqualität von Erkrankten mit neurokognitive Störungen (Demenz), ebenso wie auf der Erleichterung der Pflege und dem Schutz vor extremer Arbeitsbelastung der Pflegenden. Um sich dieser Fragen anzunehmen werden zum Teil sehr komplexe Systeme untersucht, so dass der Vorgang für das Sammeln und die Analyse der Daten nach keinen festen Muster ablaufen kann, sondern eher, je nach Fragestellung in Bezug auf die Pflege von Personen mit Demenz, flexible Antworten benötigt werden.

In beiden wissenschaftlichen Gebieten werden spezifische wissenschaftliche Fragen gestellt. Die Eigenschaften der Daten, die zur Beantwortung der Fragen gewonnen werden, können sich zwischen den zwei Gebieten sehr unterscheiden oder auch ähneln. Aber unabhängig davon, wie sehr sich beide Wissenschaftsgebiete unterscheiden, in den meisten Fällen findet man in den Daten Abhängigkeiten und Korrelationen zwischen verschiedenen Variablen, die mit multivariate Methoden analysiert werden. Diese Arbeit beinhaltet drei methodische Artikel, die verschiedene multivariate Methoden untersuchen, und zwei Artikel, die die Analyse einer realen Studie unter Anwendung dieser Methoden präsentieren.

Summary

Applied life sciences are interdisciplinary fields, which require profound statistical and computational methods to organize, visualize and analyse the obtained data, especially, since the complexity of data has grown.

For data analysis carefully designed studies and appropriate methods are important to make conclusions on the basis of valid results. This needs knowledge about data structure and characteristics, whatever in which scientific field. Furthermore, user-friendly applications to make difficult mathematical and computational methods available for practitioners are essential for these applied sciences.

In this thesis the focus is on a methodological point of view as well as showing the application of provided methods in analysing data of real studies. The challenges on data analysis depending on several data characteristics are shown within two applied life sciences: metabolomics and health care. During my work in both fields, it has been shown, that the possible solutions and mathematical approaches are similar although both sciences have different issues.

Metabolomics - a key discipline in system biology - investigates the metabolome, which is a complete set of small molecules in an organism. For the identification and quantification of such molecules, the metabolites, in complex mixtures mass spectrometry based methods are often used. Metabolomic research helps to get insights into the metabolic and molecular regulatory mechanisms contoling the growth, development and stress responses of organism. A major part therefore takes the conduction of experiments with analytical character, which have to be analysed with appropriate methods to receive these insights.

Nursing science is one part of health care, where clinical nursing service research has the aim to guide nursing practice and to improve care and quality of life of patients. Related to the population ageing, nowadays, there is a special interest within nursing service research on neurocognitive disorders, popularly known as dementia, to reduce health care costs and maintain life quality of affected people, both patients and their caregivers. Complex systems are under investigation and thus, the proceeding how to obtain and analyse data is not a restrictive approach, but rather there is a need of flexible answers according to several scientific questions in terms of care of persons with dementia.

Both scientific fields have their own research questions with different aims. The characteristics of the data, obtained to answer these questions, between the two fields have differences as well as similarities. But no matter how different the research question and apparent data characteristics are, patterns reoccur. For example in most data dependencies and correlations between several variables are present and hence requires multivariate methods for the data analysis. Within this thesis three methodological articles, which investigate several multivariate methods, and two articles presenting real life study analysis, which shows the usage of such methods, are included.

Contents

Introduction 1

1.1 Background . . . 1

1.2 Research design to study causal effects . . . 2

1.2.1 Data analysis in studies . . . 4

1.2.2 Methodology to explore data characteristics . . . 6

1.3 Multivariate statistics . . . 7

Research objectives 10

Metabolomics - Analyse mass spectrometry data of plants 12

3.1 Metabolomics research . . . 12

3.2 Mass-spectrometry method . . . 12

3.3 Methodological challenges of mass spectrometry data analysis in metabolomics studies . . . 13

3.4 Publications . . . 14

3.4.1 Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. . . . 14

3.4.2 Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent. . . . 17

3.4.3 Joint analysis of dependent features within compound spectra can improve detection of differential features. . . . 19

Health care - Analyse data of people with dementia in clinical nursing services research 22

4.1 Nursing services research . . . 22

4.2 Dementia research . . . 22

4.3 Dementia research with the focus on the quality of care . . . 22

4.4 Methodological challenges of analysing data of people with dementia within nursing services research studies . . . 23

4.5 Publications . . . 24

4.5.1 Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing . . . 24

4.5.2 Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study 27

Conclusions and outlook 29

Bibliography 31

Publications 36

6.1 Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data . . . 37

6.2 Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent . . . 48

6.3 Joint analysis of dependent features within compound spectra can improve detection of differential features 58
6.4 Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing . . 68
6.5 Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study 78

Appendix - Methodological details 85
A.1 Hypothesis testing theory, test quality and performance of simulation studies . 85
A.2 Hierarchical experiments: distribution of observations and statistical test 90
A.2.1 Assumption 90
A.2.2 Traditional frequentist approach using the linear mixed model notation 91
A.2.3 Likelihood-ratio approach 93
A.2.4 Distribution of an observation with K replicates is K-dimensional . . . 103

Appendix - Supplemental material to the publications 109
B.1 Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data 109
B.1.1 Variance estimation 109
B.1.2 Test statistics and their distribution of hierarchical and non-hierarchical models 110
B.2 Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent 111
B.3 Joint analysis of dependent features within compound spectra can improve detection of differential features 113
B.4 Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing . . 116

List of Figures

1.1 Data analysis 2
1.2 Study design model 3
1.3 Type of study designs 4

3.1 Hierarchical experiment design 15
3.2 The distribution of estimated variances 16
3.3 Nested experimental design with three levels 17
3.4 Biological variability of secondary metabolites 18
3.5 Different decisions from univariate and multivariate test 20
3.6 Results of univariate and multivariate methods in feature detection 21

4.1 Interaction between study and the tasks of researchers 25

A.1 Probabilities of errors 86
A.2 Concepts of a hierarchical model within a biological experiment 90

B.1 Linear hierarchical model 109
B.2 Biological variability of primary metabolites 111
B.3 Accession-specific ICCs of secondary metabolites 112
B.4 Distribution of size of compound spectra identified with CAMERA 113
B.5 Balance check on propensity score 117

List of Tables

1.1 Study design types 5
1.2 Data matrix 8
1.3 Summary statistic 8

2.1 Articles included within this thesis. 10

4.1 A contingency table and the estimated parameters 26

A.1 Probabilities of errors within testing decisions 86
A.2 Confusion matrix of binary classification problem 87
A.3 Connection of measurement and estimate of probability due to the descision problems 88
A.4 Performance measurements of simulation studies 89

B.1 Test statistic for hypotheses tests and their distributions 110
B.2 Formula of the univariate Student's t test 115

1. Introduction

1.1. Background

Applied life sciences cover interdisciplinary research fields on living organisms. This comprises, among others, metabolomics (a part of biochemistry) and nursing service research (a special topic in health care).

Mass spectrometry-based methods play an important role in the metabolomics field (Gowda and Djukovic, 2014), studying the complete set of small molecules in organisms, so-called metabolites. Because mass spectrometry is a method, which is able to measure the molecule masses very precise, it is used to identify as well as to quantify the amount of metabolites occurring in organisms.

In contrast, nursing research generates knowledge with an impact for nurses. This could be for example knowledge that affects the delivery of nursing care (Burns and Grove, 2009) or that have an impact on patient care decisions.

Both, mass spectrometry methods and nursing research, are part of different areas of life sciences, but usually deal with multidimensional and complex data sets . There is a demand for suitable methods to obtain valid information from these complex data sets (Belle et al., 2015; Tugizimana et al., 2016). Hence, despite different research questions asked, different data at first glance, but repeating characteristics (bias, dependency etc.) require profound statistical and bioinformatical methods (Boccard and Rudaz, 2014; Beisken, Eiden, and Salek, 2015).

Bioinformatics is a scientific field, where methods for storing, retrieving, organizing, visualizing and analysing biological data are developed (Chicurel, 2002). The advances in computational power allow to use complex statistical methods (Díaz-Emparanza, 2000; Scott, 2015; Gentle, Härdle, and Mori, 2012). The rapid introduction of new biological measurement technologies and the increasing relation to many disciplines raise the need for carefully designed, conducted and analysed studies. This helps to make research reproducible and with regard to assessing research critically relevant to their clinical practice (Ioannidis et al., 2014). Here, the aim of computer scientists is to provide methods for user friendly application (Chicurel, 2002) of difficult mathematical as well as computational algorithms to practitioners.

The aim of this work is to make statistical methods available for researchers in two scientific fields with large datasets: metabolomics and health care research. On the one hand there is a need to specify the problems/issues of the data within each field and on the other to give recommendations to acquire data with most powerful experiment designs and analyse the data with appropriate methods. Thereby, the overall aim remains analysing data to answer a particular scientific question.

Figure 1.1 illustrates the scope of this thesis, that although the two scientific fields have different questions and also data with different characteristics, the overall analytical reflections remain the same. It describes the different aspects which have to be considered when conducting a real study (green box) and is explained with more detail in 1.2.1. This thesis has the focus to elaborate on these aspects from a methodological point of view (yellow box), which are discussed more comprehensive in 1.2.2. Some advice which have to be considered before a study is conducted and during data analysis are also given. Figure 1.1 is an overview of the context of this methodological consideration and the articles in this thesis are related to a specific position in this figure.

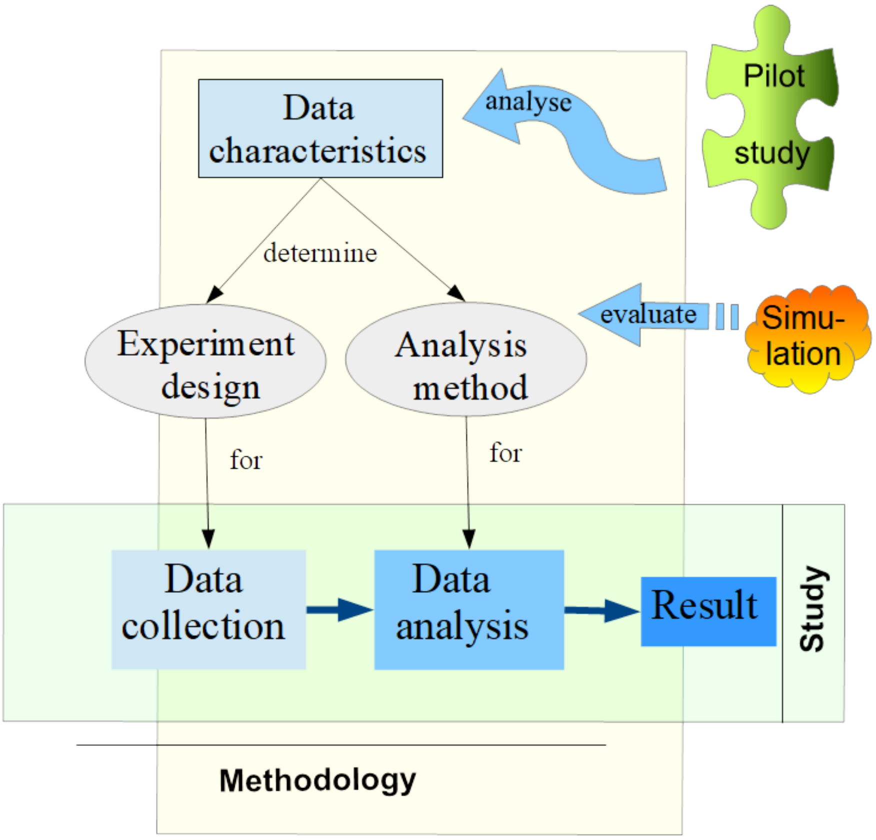


Figure 1.1.: To answer a scientific question appropriately, the data characteristics have to be known. The characteristics, e.g. the type of measurement method or the type of the outcome variable, which is strongly related to the scientific question and context, directly determines the experimental design and the statistical method to analyse the collected data. Methodological investigations about these data characteristics can help to find the best suitable experiment design for data collection or analysis method within a real study. A pilot study is one possibility to obtain insights into data characteristics.

1.2. Research design to study causal effects

Often, the primary aim of applied life science is to detect relationships or even causal effects between independent and dependent variables. A cause can be defined as any condition tending to increase the probability of the effect (Glymour, 2012). Figure 1.2 illustrates the simplest model of a study design type analysing the relationship between variables. The causal variable, often called exposition, is the independent variable and influences the objective criterion as dependent variable. Other associated causal variables, which are related to the exposure (direct causal effects) as well as to the objective criterion, are possible. They are known as confounder and responsible for bias because the two effects cannot be separated. Some examples for such relationships of exposition and objective can be given: in metabolic studies the genetic manipulation to change the metabolic state of organisms or in health care studies the use of an intervention to change the quality of life of people with dementia.

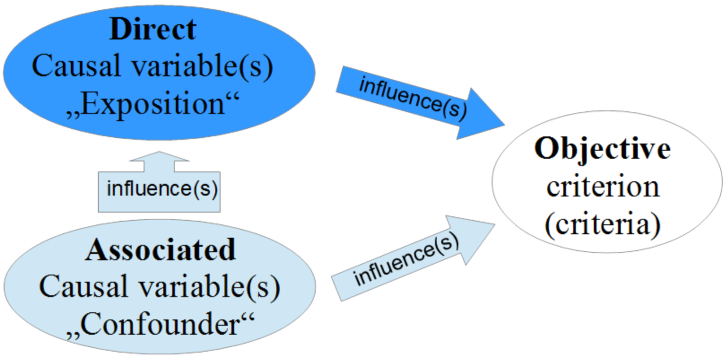


Figure 1.2.: The model of a study design illustrates that the direct causal variable(s) and possible associated causal variable(s) can influence the objective criterion (criteria).

To address a particular a problem and find a solution with confidence, careful consideration how to conduct the study is required. It is called the study design, also known as experimental or research design (Polit and Beck, 2004), and plays an important role in quality and interpretation of results related to a scientific question (Thiese, 2014; Tugizimana et al., 2016). This includes the design type of the study as well as data collection, statistical analysis and interpretation of the results (Knight, 2010). However, poor design choices can reduce the generalization of the study results (Ioannidis et al., 2014) and thus, should be avoided.

Types of studies designs. Figure 1.3 shows that different study design types are possible for translating the conceptual hypothesis into an operational one. They can be defined by different criteria: 1) the temporal nature (time), 2) the role of the investigator (objective) or 3) the investigated purpose (manipulation) of the study (Thiese, 2014).

Categorised by time, prospective versus retrospective study designs are possible. In a prospective study, at the beginning of the study the exposition (dark blue ellipse in Figure 1.2) is identified and thus, the defined population is followed for effect detection (forward-oriented). A retrospective study begins with the outcome (white ellipse in Figure 1.2) for a defined population (dependent variable) and looks back in time to identify the exposure factors or the cause (backward-oriented) (Polit, Beck, and Hungler, 2004).

If the study types differ by the objective of the study, descriptive or analytic study designs were distinguished. A descriptive study gathers ideas about relationships and identifies questions, so that hypotheses can be generated. In contrast, an analytic study attempts to validate hypotheses established by descriptive studies. The effect of a cause has to be identified as well as the effect size estimated.

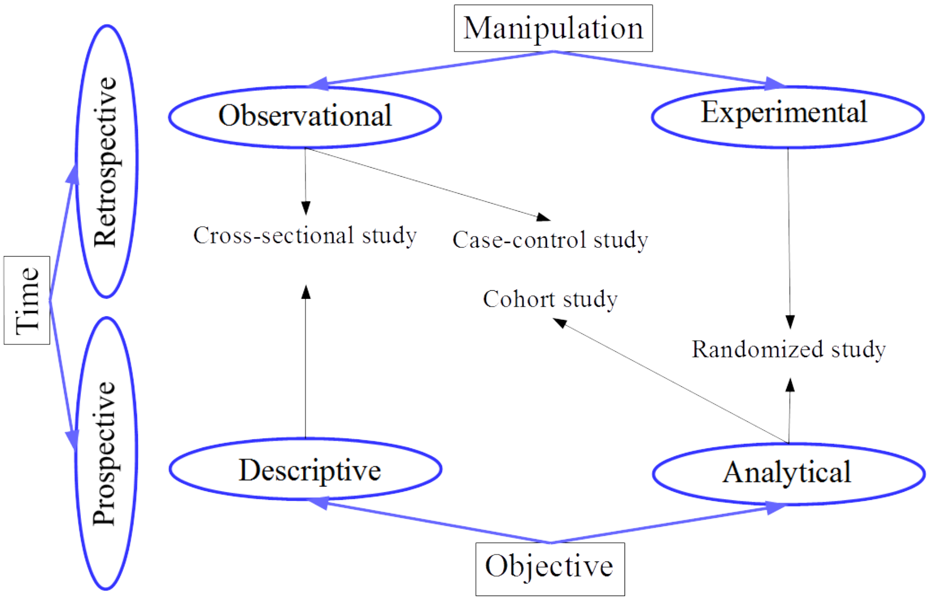


Figure 1.3.: Different design types of studies for translating the conceptual hypothesis into an operational can be categorised by different criteria, here shown for type of objective, time and manipulation. Classified by the objective, hypothesis-generating (descriptive) or -verifying (analytic) studies are distinguished. Depending on the degree of manipulation, study types are possible from observational to experimental (with increasing manipulation of the independent variable). Time separate studies in prospective (forward-oriented) and retrospective (backward-oriented) ones.

Observational on the one hand and experimental study types on the other hand are possible, if a researcher makes decisions about the manipulation of the independent variable. In the first case no manipulation is needed, whereas in the second manipulation is done. In observational studies the cause-effect relationship has to be found, described and assessed or summarized the status of a phenomena. Hence, observational studies can provide insights into how an intervention works in a representative population (Ho, Peterson, and Masoudi, 2008). In contrast, in experimental studies the effect of planned and actively applied exposition is analysed in a prior planned manner. Hence, experimental studies should be randomized.

In Table 1.1 the four special cases of study design types, shown in Figure 1.3, are described in detail, ordered from the lowest grade of evidence to the highest (top down), referring to the strength and weakness of the designs (Ho, Peterson, and Masoudi, 2008). Hence, each scientific question requires its own study design strongly related to the scientific context and aim of the research question. In metabolomics Case-Control studies can often be found to understand such relationships between genotypes and the presence of a metabolic substance. In contrast, several types of study designs like cross-sectional, cohort or randomized studies can be found in the field of nursing research. For example, in this thesis a cross-sectional (health care) and a case-control study (metabolomics) was conducted.

1.2.1. Data analysis in studies

Figure 1.1 illustrates that data analysis in studies (shown by the green box) implies three steps: 1) the data collection, 2) the analysis of the data and 3) the interpretation of the results. The first step is data generation, whereby data can be collected with or without prior knowledge about their characteristics. The kind of how the data are generated in turn depends

Type		Main characteristics
Cross-sec.	Def.	exposures and disease status at a single point in time
	Cat.	descriptive, observational
	App.	often findings are basis for other studies, can prove and/or disprove assumptions
	+	cost and time efficient
	−	no temporality demonstrated
Case-control	Def.	compare a sample group, where each member has an outcome of interest (condition), with another sample group, where each member has not determine relationship between outcome and interested risk factor (exposition)
	Cat.	retrospective, analytical, observational
	App.	instrumental to understand relationships (Ho, Peterson, and Masoudi, 2008)
	+	cost and time efficient
	−	no temporality demonstrated
Cohort	Def.	samples are separated by the exposition of interest involves data collection over two or more time points
	Cat.	retrospective and prospective possible, analytical, observational special cases: Follow-up, longitudinal and intervention studies (Hilgers, Bauer, and Scheiber, 2007)
	App.	examine long-term effects of a specific expositions
	+	prospective cohort studies best suited for suggesting causation (Thiese, 2014)
	+	temporality demonstrated
	−	in retrospective cohort studies randomized allocation not given and causal conclusions cannot be made
	−	expensive, time consuming
Randomized	Def.	all member are randomly allocated to receive one of the several interventions
	Cat.	prospective, analytical, experimental
	App.	determine any effects of the exposition
	+	provide most suitable equating groups on all possible characteristics (Polit and Beck, 2004)
	+	avoid bias
	−	expensive, time consuming, sometimes not practical
	−	an intervention may result in hidden events

Table 1.1.: Definition (Def.) of different study design types, ordered from the lowest grade of evidence to the highest (top down), are explained by their categories (Cat.), application (App.), advantages (+) and disadvantages (-).

on the context of the scientific question. However, an appropriate experimental study design (left grey ellipse Figure 1.1) is essential to obtain sufficient information. R.A. Fisher already stated that statistical procedure and experimental design are only two different aspects of the same issue (Fisher, 1966). Thus, a suitable statistical method (right gray ellipse Figure 1.1) using the correct assumptions related to the data characteristics is required and should be prespecified before data collection or at least before starting the analysis. Hence, analysing data of an applied life science study should answer the scientific question with a) valid methods and b) the most powerful experiment design. Finally, the goal of statistical data analysis through a study is to extract the maximum information from the data and results that are as accurate and as useful as possible (Scott, 2015; Boccard and Rudaz, 2014; Tugizimana et al., 2016) and avoid bias from all stages of research(Sackett, 1979). In other words, it reveals valid and reproducible knowledge about a practical issue by using models to make inference concerning the process (Gentle, Härdle, and Mori, 2012). Hence, the aim of successful experiments is make conclusions to causal relations, which could only be realised by randomised studies.

Data analysis, the second step, and causal effect estimation are closely connected to statistical interference. Statistical inference is a method to investigate the characteristics of causes and includes two related principles: estimation of population parameters and testing of hypotheses (Bortz, 2005). The best possible causal relation estimation is one of the principal aims of statistical analysis (Glymour, 2012), whereby causal relations describe how variables influence each other. Statistical hypothesis testing theory is a widely-used method for statistical inference. The statistical hypothesis is a statement about the characteristics of random variables, e.g. a parameter or a distribution of a population, and represents a scientific hypothesis. Data analysis uses the information gained from a sample of individuals in order to make inference about the relevant population (Ilakovac, 2009). Hence, statistical hypothesis testing is a procedure that is based on parameter estimation from a sample, a subset of the whole population, for which the statistical hypothesis should be evaluated.

The appropriate study design choice is related to statistical hypothesis testing as it is related to effect size, sample size and power (see for more detail Appendix A.1). Hence, before the study is conducted, considerations about the appropriate design should be made.

In this thesis, after methodological considerations of (Trutschel et al., 2015a, Section 6.1) and (Trutschel et al., 2017, Section 6.4), two examples of data analysis within a study and how to interpret the results are given: (Mönchgesang et al., 2016, Section 6.2) within the field of plant metabolomics and (Palm et al., 2016, Section 6.5) within nursing service research.

1.2.2. Methodology to explore data characteristics

Figure 1.1 also illustrates the methodological investigations before a study is done (yellow box): first, scientists have to gather prior knowledge about data characteristics, second, they have to find the best experiment design and the appropriate statistical method and finally, to analyse data of a real study.

It is important to understand the data characteristics before realising a possibly expensive study to answer a scientific question, because they limit the possibilities for experimental design and analysis method. Hence, pilot studies (green jigsaw in Figure 1.1) and computational simulations can help to get deeper insights on data characteristics and find a powerful experiment design before the procedure of a real study is defined, shown in (Trutschel et al., 2015a, section 6.1).

While knowing the data characteristics is essential, simulation studies can be performed to evaluate and compare experimental designs or analysis methods (Gentle, Härdle, and Mori, 2012) (see Figure 1.1). Hence, computational inference is a viable and useful alternative to traditional statistics inference. Simulation as a numerical technique, often use compute intensive procedures, can help to answer questions that could not have been answered on real data alone (Burton et al., 2006). Two types of simulations are classified by the type of input data: they are derived from 1) measured data of a real system or 2) via sampling from probability distributions using random numbers (Balci, 1990), known as Monte Carlo simulation. The aim is to test particular hypotheses and assess the performance of a variety of statistical methods in relation to a known truth. Thereby, several scenarios should reflect common circumstances. Then methods can be tested and compared referring to a) accuracy of estimation method or b) the quality of hypothesis testing approaches (For more details about simulation studies and their performance see Appendix A.1). In this dissertation a computational simulation study is used to find the appropriate statistical model to analyse data applied in (Trutschel et al., 2015b, section 6.3).

Many key issues are related to the data characteristics and each single choice is important for the analysis as it determines the experiment design and requires the appropriate analysis method (see section before, gray ellipses in Figure 1.1). For example according to the scientific field and question several type of measurement methods to obtain data are possible and determine the data characteristics. In metabolomics often mass spectrometry based methods are used, whereby in nursing research questionnaires are common instruments. Furthermore, different types of outcome variables can be found, e.g. nominal, ordinal or continuous outcomes. The type of study design additionally determines data characteristics. For example several number of groups can be compared or several replicates of a measurement unit can be obtained.

One key issue of data characteristics is that various types of dependencies have to be taken into consideration. In the past, statisticians like Laplace were already faced with dependent observations, for example calculating probabilities associated with the game of chance. In the present-day data analysis the consideration of *dependencies* still plays a central role. Dependencies are possible in many ways, for example due to a) *the study design*, b) *the type of manipulation* or c) *the measurement method*.

Dependencies due to a) study design occur, when a hierarchical structure is given. Here, repeated measurements of a unit, which are then dependent, are observed. Sometimes, this is called nested structure in time or space and is discussed in (Trutschel et al., 2015a, section 6.1) and (Mönchgesang et al., 2016, section 6.2).

When dependencies are present due to b) type of manipulation, this may be observed as selection bias. Then, observational independence of allocation to treatment and control is not guaranteed and so dependencies due to the lack of randomization are possible. In observational studies it is possible that covariates permit an assignment of observations to a specific group, where in the opposite case in randomized studies the assignment is independent of the covariates. An example is discussed in (Trutschel et al., 2017, section 6.4).

The kind of outcome is also influenced by c) the measurement method. Hence, dependencies or correlations between several outcomes may be occur corresponding to the method, e.g. using mass spectrometry methods it is possible to obtain a number of dependent signals for a single metabolite. The consequence of this kind of dependencies on model choice is discussed in (Trutschel et al., 2015b, section 6.3).

Dependencies within data of applied life science must be taken into account and often require, in addition to the complexity of the obtained data, *multivariate statistics* for data analysis (with more detail in next subsection). This is now widely performed using computational power. Disregarding dependencies may lead to statistical errors and false conclusions. Hence, available results of acquired data have to be interpreted in a correct manner.

1.3. Multivariate statistics

Analysing multi-dimensional data often requires multivariate approaches, because multivariate analysis takes all variables simultaneously into consideration (Beisken, Eiden, and Salek, 2015). Such data sets contain an amount of variables, generated by observations. A $n \times p$ data matrix contains measurements x_{ij} of p variables on n objects, shown in Table 1.2.

For example, in nursing science research on n individuals p variables, e.g. different characteristics like age and sex or interested outcomes like quality of life and challenging behaviour, can be measured, where the measurement method is often questionnaires. Another example is in the metabolomics field the measurement of p features resulting from q $(\leq p)$ metabolites from n plants using mass spectrometry methods.

In Table 1.2 each row corresponds to an object (e.g. individual) and each column to a variable (e.g. characteristic). This matrix can be analysed in two alternative ways: column-wise examining the relationship between different variables and row-wise between different

Object	Variable 1	Variable 2	$\cdots$	Variable p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Table 1.2.: Data matrix containing measurements x_{ij} of p variables on n objects.

objects (Mardia, Kent, and Bibby, 2003).

Formula	Summary statistic
$\overline{x_j} = \frac{1}{n} \sum\limits_{i=1}^{n} x_{ij}$	sample mean of variable j
$s_j = \frac{1}{n-1} \sum\limits_{i=1}^{n} (x_{ij} - \overline{x_j})^2$	sample variances of variable j
$s_{j\tilde{j}} = \frac{1}{n-1} \sum\limits_{i=1}^{n} (x_{ij} - \overline{x_j}) \left(x_{i\tilde{j}} - \overline{x_{\tilde{j}}}\right)$	sample covariances between two variables j and $\tilde{j}$

Table 1.3.: Summary statistics of multivariate data.

Summary statistics of this kind of data, listed in Table 1.3, are similar to univariate summaries. The sample mean vector $\overline{\mathbf{x}} = (\overline{x_1}, \ldots, \overline{x_p})^T$ (vectors are represented in bold letters) is an estimate of the true mean vector μ, where $\overline{x_j}$ is the sample mean of variable j.

The key characteristic using multivariate methods is the sample covariance matrix S, an estimate of the true covariance matrix Σ. It includes variances s_j of each variable j as diagonal matrix elements and sample covariances $s_{j\tilde{j}}$ between two variables j and $\tilde{j}$ as off-diagonal matrix elements (Table 1.3). Hence, the variance-covariance matrix has the following symmetric form:

$$\Sigma = \begin{pmatrix} s_1 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & s_{p-1p} \\ s_{1p} & \cdots & s_{p-1p} & s_p \end{pmatrix}. \tag{1.1}$$

For a special experiment design (nested structures) the form of the variance-covariance matrix of possible multivariate-normal distributions is derived in Appendix A.2.

Graphical visualisation of multivariate data is often used to get an impression of the data characteristics and discover the unexpected. It represents an explorative data analysis without an underlying parametric model. Although the non-parametric world is much more complex and more flexible than its counterpart, parametric methods, although they have key assumptions, are often used because they are the most powerful (Scott, 2015).

In this dissertation parametric tests are used and the distribution of the observations are assumed known. If it is assumed that the variables are conjointly distributed due to the dependencies within the data, it requires a multivariate parametric analysis. This is different to the univariate analysis, where each variable is analysed independently. While the distribution of one single random variable is univariate, the joint distribution of p variables is called a multivariate distribution (DeGroot, 2004). Inductive analysis that are interested on a multivariate distributed p-dimensional vector $\mathbf{x}$ often use the multivariate normal density (Scott, 2015) (Equation 1.2), which is a multidimensional generalisation of the normal distribution.

The multivariate normal density is defined by:

$$f(\mathbf{x}) \quad = \quad \frac{1}{\sqrt{2\pi\Sigma}} \exp -\frac{1}{2} (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu), \tag{1.2}$$

where $\mathbf{x}$ is a vector of observations on p variables (for example one row of Table 1.2), μ the p-dimensional vector of means and Σ^{-1} the inverse of the $p \times p$-dimensional variance-covariance matrix (Eq. 1.1). For illustration: different design types of a study (cross-over or case-control e.g.) determine different mean vectors and different structured data due to dependencies specify different variance-covariance matrices of the multivariate-normal distribution (examples how to sample such distributed data are given iby the provided *samplingDataCRT* [1]).

Hence, multivariate approaches are used to analyse more than one dependent or independent variable (Rasch et al., 2010). Techniques used on this kind of data are sometimes just a generalization of the univariate ones. For example the multivariate analysis of variance (MANOVA) is the extension of the univariate analysis of variance (ANOVA) for more than one dependent variable as the variables are analysed simultaneously. The different statistical inference methods that exist, for example: classical frequentist approach, likelihood-based or even Bayesian inference, can also be used in a multivariate manner.

The problems investigated in this dissertation focus on the multivariate representation of the data and their analysis by inductive multivariate approaches. Three main topics are considered within the thesis: 1) *multivariate hypothesis testing*, 2) *multiple regression analysis* and 3) *multilevel structures in data* and can be classified to multivariate analysis approaches (Mardia, Kent, and Bibby, 2003).

First, multivariate hypothesis testing means a joint hypothesis test on two or more parameters, which results in a multidimensional test statistic. This arises for example when more than one variable of interest (more than one dependent variable) are analysed simultaneously. The approximate confidence region for the parameter vector is a k-dimensional ellipsoid, where k is the number of tests (Millar, 2011). Using univariate hypothesis tests for each parameter individually instead causes the multiple testing problem of increasing Type I error. Hence, the adjustment by the correlation structure between test statistics within a joint analysis is then less conservative than ad hoc methods based on a Bonferroni adjustment of the Type I error rate (Stucke and Kieser, 2012). One example for a joint analysis is given in this thesis by (Trutschel et al., 2015b, section 6.3).

Second, if a dependent variable is affected by more than one variable, a set of variables can be used to predict another. This leads to multiple regression analysis, the extension of the univariate type with only one predictor variable. Observations on n objects are fitted by a linear combination of all independent variables, e.g. applied in (Trutschel et al., 2017, section 6.4) and (Palm et al., 2016, section 6.5).

Third, if multilevel structures in data are present, dependent observations due to multiple measurements on different levels of one object are obtained. Hence, a multivariate representation of the data with a covariance structure to adjust for the dependencies between observations is required. It is the base of the articles (Trutschel et al., 2015a, section 6.1), (Mönchgesang et al., 2016, section 6.2), (Trutschel et al., 2017, section 6.4) and (Palm et al., 2016, section 6.5). The R-package *samplingDataCRT* [1] provides as an easy tool sampling data sets of cluster randomized studies, which are characterized by such depended structures.

[1] `https://CRAN.R-project.org/package=samplingDataCRT`

2. Research objectives

Statistical inference formalizes the process of learning through observation, whereby the learning process includes two principal parts: design an experiment and analyse the experimental data (Berry, 1996). The aim of applied life science studies is with 1) the most powerful experiment design and 2) appropriate methods 3) analysing data to answer the scientific questions in a correct manner (Figure 1.1). Whereas researchers are confronted with large and complex data sets, they have to study the data characteristics as well as the meaningful analytical process (Belle et al., 2015). Thereby, evaluation within pilot studies or by simulations may help to find the best suitable experiment design and analysis method for analysing data, which is related to the data characteristics and according to the scientific question.

	Scientific field		
	Mass spectrometry		**Nursing services**
	Experiment design	Analysis method	Analysis method
Methodology	Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data (Trutschel et al., 2015a) Section 6.1 + Vignette	Joint Analysis of Dependent Features within Compound Spectra Can Improve Detection of Differential Features (Trutschel et al., 2015b) Section 6.3 + Vignette	Methodological approaches in analysing observational data: a practical example on how to address clustering and selection bias (Trutschel et al., 2017) Section 6.4 + Vignette
Study	Plant-to-plant variability in root metabolite profiles of 19 *Arabidopsis thaliana* accessions is substance-class-dependent (Mönchgesang et al., 2016) Section 6.2		Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study (Palm et al., 2016) Section 6.5

Table 2.1.: Articles included within this thesis.

In this dissertation these aims are addressed within two applied sciences: mass spectrometry as part of metabolomics and nursing service as part of health care. Table 2.1 gives an overview of the included articles and their focus on either experimental design and analysis method evaluation or practical application within a real study. For an easy orientation of the scope of each article the same color code as in Figure 1.1 is used in the table. Thus, the yellow shaded articles have the focus on methodology and the green on a real study analysis applying evaluated methods or designs to answer a relevant scientific question.

When writing the included articles an additional aim was to provide tools for real (other researcher's) problems within data analysis. For all three methodological articles within this thesis a vignette is provided. Thereby, an overall focus was to make mathematical or informatics methods available for practitioners. For all implementation the free statistical software R (R Core Team, 2016), which is especially conceptualized for statistical computing and graphics, was used.

In the following the thesis is partitioned for these two scientific fields, described in chapter 3 and 4. Each chapter includes its own introduction to the field with a special interest to the methodological challenges followed by a short explanation of the articles. This includes the topics of the articles and their context within this thesis, which show exemplary one of the aspects of Figure 1.1 within one of the research field. The Conclusion and Outlook chapter is then written for both in chapter 5, because this consideration can be made interdisciplinary as it is done in the Introduction chapter. The complete articles (as they are published) are given in chapter 6.

3. Metabolomics - Analyse mass spectrometry data of plants

3.1. Metabolomics research

Metabolomics is a discipline which globally studies metabolites - small molecules participating in metabolic reactions in a biological system - and their concentrations, interactions and dynamics within complex samples (Boccard, Veuthey, and Rudaz, 2010; Beisken, Eiden, and Salek, 2015). It is a key discipline in system biology (Fiehn, 2002; Weckwerth, 2003).

The complete set of small molecules in an organism is called metabolome. At the metabolite level the phenotype of cells is represented, influenced by perturbation of gene expression and the modulation of protein functions, which are caused by the environment or mutations (Saito and Matsuda, 2010; Beisken, Eiden, and Salek, 2015). The aim of metabolomics is to quantify all metabolites in order to find answers to biological questions (Koek et al., 2011). Besides that, untargeted metabolomics analysis starts with unknown structure and the objective is to measure many metabolites simultaneously and find potential biomarkers (Eliasson et al., 2012; Yi et al., 2016). Whereby the WHO describes biomarkers as any measurement reflecting an interaction between a biological system and a potential chemical, physical or biological risk (safety, 1993). Particularly, untargeted metabolomics aims at the simultaneous measurement of the full set of metabolites - not knowing the compound nature (structure or annotation).

The field of metabolomics has important applications in areas of life sciences and beyond (Gowda and Djukovic, 2014). Plant metabolomics has become a powerful tool to explore various aspects of plant physiology and biology. Insights into the metabolic and molecular regulatory mechanisms regulating plant growth, development and stress responses can be obtained with the help of metabolomic research (Hong et al., 2016). For example, the aim of case-control studies is to detect metabolites relevant to a specific genotype (Beisken, Eiden, and Salek, 2015), where the participants are identified based on their outcome (genotype/phenotype) and then the presence of the risk factor (exposition), here the metabolic substance, is compared. So the relationship between both is evaluated (see Section 1.1, Figure 1.2 and Table 1.1).

3.2. Mass-spectrometry method

Mass spectrometry-based methods are often used for comprehensive identification and accurate quantification of metabolites in complex mixtures of them (Fiehn, 2002; Weckwerth, 2003). Because of the sensitivity of the methods, which need less sample material, these methods play an increasingly dominant role in the metabolomics field (Gowda and Djukovic, 2014) and show their power in plant metabolomic studies in many common plant species (Hong et al., 2016).

Due to different analytical conditions (solvents, ionization etc.) different adducts and in-source fragments are measured by mass spectrometry. Additionally, different isotope patterns of one molecule can occur and hence, measured. They are the readout of the elemental composition and their natural distribution. Thus, this measurement method gives rise to a number of features (Brown et al., 2009), which are related to each other.

Liquid chromatography-mass spectrometry (LC-MS) or gas chromatography-mass spectrometry (GC-MS) combine chromatographic methods for separation (retention time) and subsequent mass spectrometry for detection of metabolites (Beisken, Eiden, and Salek, 2015). Hence, this method provides two-dimensional signals, called features (Tautenhahn, Böttcher, and Neumann, 2008), with information about retention times and mass-to-charge (Werner et al., 2008). Three-dimensional signals are obtained when the intensity of this features are also considered. GC-MS has been one of the most popular metabolomics techniques to determine the levels of primary metabolites (amino acids, organic acids, and sugars), while LC-MS is a method for the profiling of secondary metabolites (for example phenylpropanoids and alkaloids) of plants (Saito and Matsuda, 2010). Both are the scope of scientific investigations (Hong et al., 2016).

3.3. Methodological challenges of mass spectrometry data analysis in metabolomics studies

Measuring complex metabolomics samples containing hundreds to a few thousands metabolites using GC- or LC-MS leads to high dimensional data sets with many features. It results in a two-dimensional data matrix of $n \times m$ feature intensities, where n is the number of features (here variables correspond to the number of rows), including information such as retention times and mass-to-charges, and m the number of measurements within the experiment (here observations correspond to the number of columns). For the intensities or at least the logarithmic intensities of features obtained by GC/LC-MS a normal distribution can be assumed, so for data analysis all parametric tests with the assumption of normal-distributed observations are available. In the past, basic statistical tests like the univariate Student's t-test has found applications to identify metabolite differences between groups of, for example, different genotype.

Measuring many variables simultaneously requires sophisticated and powerful methods to analyse the data and turn it into biological knowledge (Steuer et al., 2007). However, researchers tend to use methods that are common and easy to apply (Moseley, 2013). Then, the challenge is to find the appropriate way for data analysis in metabolomics studies (Beisken, Eiden, and Salek, 2015; Yi et al., 2016). This depends, in mass spectrometry (as well as in other research fields) on data characteristics such as the type of study design or data processing method (instrument used and data collection) (Broadhurst and Kell, 2006).

Mass spectrometry data has some key characteristics, like dependencies between observations - between columns of the MS-data matrix - due to study design (e.g. technical replicates) or dependencies between different features - between rows of the MS-data matrix - (e.g. different species of one metabolite after ionisation). Thus, answering biological questions requires sophisticated statistical methods, which take such dependencies into account. Furthermore, a lack of statistical power due to a poor design is an example for obtaining interpretative bias and should be avoided (Moseley, 2013).

In this thesis both objectives, finding a suitable experiment design and evaluating statistical methods to answer biological questions are addressed for mass spectrometry data. (It should be noted that pre-processing of the data, such as treatment of missing values or normalisation techniques, will not be the objective.) Therefore, three articles are included: two methodological and one example for a real study analysis (See Table 2.1 and Figure 1.1).Their specific aims are explained below. Although the example data sets are obtained by measuring leaves, shoots or roots of plants, the methods are also usable for other tissues as well as other organisms and other measurement methods like GC-MS. R-codes are provided within each article for an easy use of the discussed methods or to adapt the analysis process in other contexts.

3.4. Publications

3.4.1. Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data.

The topic of the first article (section 6.1) is the identification and quantification of possible sources of variances in mass spectrometry experiments. Therefore, statistical methods are used to take dependencies between observations due to nested designs into consideration, that means dependencies between the columns of the two-dimensional data matrix of obtained mass spectrometry data sets. Additionally, it is focused on investigations of key data characteristics, possible dependencies due to the design, to find the appropriate and powerful experiment design (Figure 1.1) for metabolomics studies.

Introduction Depending on the experimental design, several sources of variance are present in metabolomics data and influences the type and result of hypothesis tests. This article presents a detailed analysis of known variance levels. Therefore, a pilot study with a hierarchical experiment design is performed. Due to this design, the different replicated observations on several levels are not independent any more. Such dependent observations follow a multivariate normal distribution (shown in appendix A.2). Thus, to obtain exact and unbiased estimates of individual variances at different levels, it requires the usage of a nested linear regression model using random effects for the different levels (also known as linear mixed models) (Davis, 2002).

In studies with the aim of detecting metabolite differences, technical replicates made on biological samples are often averaged to avoid the occurrence of dependent observations. The article describes how these dependencies can be handled even with the commonly used t-test statistics by a hierarchical t-test, and also for the more general case the (hierarchical) ANOVA, which correctly includes both biological and technical replicates without distorting the results. The derivation for estimates is given and shows that estimates obtained by ANOVA-based methods in special cases are equal to those obtained by likelihood based methods (appendix A.2).

These new insights into data characteristics can contribute to find cost effective experiment designs to answer relevant biological questions. Therefore, the impact of the respective number of replicates of each level on the statistical power of the test is considered. The aim is to find a compromise between expenses, associated with costs related to different levels of replication, and the quality of inference possible in a particular experiment.

Materials and Methods A pilot study for a typical *Arabidopsis thaliana* (*A. thaliana*) metabolomics experiment (Figure 3.1) was performed to the quantify variation at different levels of the experiment. Three sources of variation in MS experiments have been considered: (i) instrumental variation, (ii) preparation variation and (iii) biological variation The total variation is then the sum of all three variations. A hierarchical set of samples at different levels of variation was prepared.

Only the overall variance σ^2_{tot} - the sum of technical and biological variances - can be estimated directly from the dataset. To obtain an unbiased estimation at *individual* hierarchical levels (Figure 3.1), the instrumental σ^2_{instr}, preparation σ^2_{prep} and biological variances σ^2_{biol} were modelled as random effects with a three-level linear regression model for each detected feature:

$$Y_{nei} = \mu + \beta_n + \gamma_{ne} + \delta_{nei} \tag{3.1}$$

where Y_{nei} is the observed measurement of injection i of extraction e of plant n, μ the overall population mean, β_n the independent random biological effect on plant n, γ_{ne} the independent random preparation effect on preparation e in plant n and δ_{nei} the independent random instrumental effect on injection i in preparation e in plant n. The random effects β_n, γ_{ne}, δ_{nei} are mutually independent. The unbiased estimator can then be estimated (for formula see Section B.1, Figure B.1). The proportion of biological variance to total variance $\frac{\sigma^2_{biol}}{\sigma^2_{tot}}$ is known as intra-class correlation (ICC).

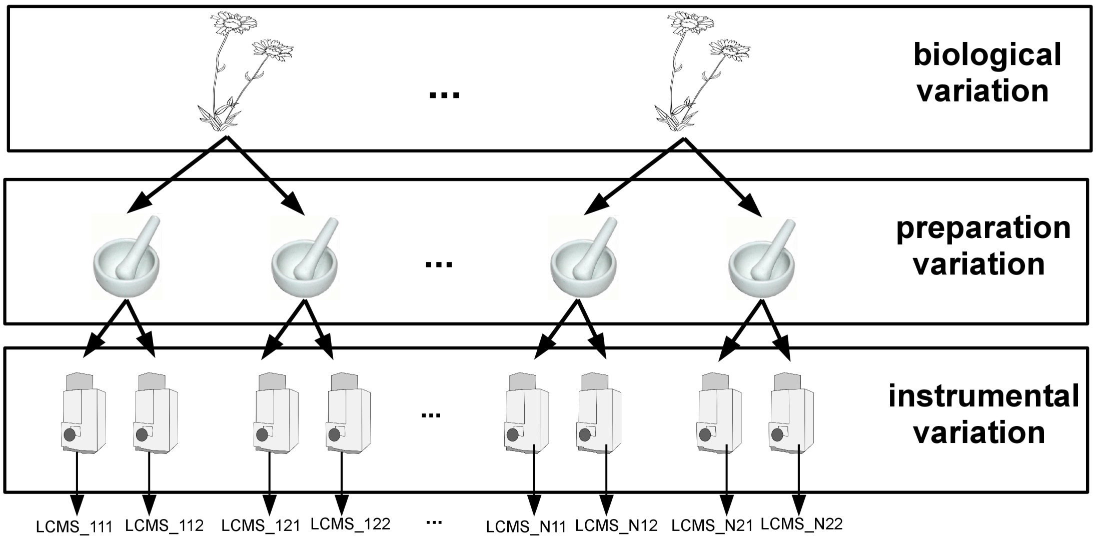


Figure 3.1.: Hierarchical experiment design. At all levels of variation replicates were prepared: To extract biological variation several plants were grown. From each plant, several extractions were performed, to assess the preparation variation. To identify the instrumental variation each extract was measured several times. The number of LC-MS datasets is the product of the number of plants N, extracts E per plant and injections I per extract.

For multilevel observations the hierarchical version of ANOVA and of the Student's t-test can also used (see subsection B.1.2) to find differences in means of observed intensities between groups. Then technical replicates are implicitly averaged and thus, multiple levels with biological and technical replicates within an experiment can be used. Both are special cases of linear mixed models (Raudenbush and Bryk, 2002) like the ANOVA is a special case of linear models.

If, though hypothesis testing, in non-hierarchical experiments four of the five parameters (i) power $1-\beta$, where β is the probability of error type II, (ii) number of samples N, (iii) effect δ between two groups, (iv) variance σ^2, and (v) α defined as the maximum allowed probability of Type I errors are given, the missing parameter can be calculated (Broadhurst and Kell, 2006). Since in multilevel models the interest is on the influence of different sources of variation, replication strategies and sample sizes on the statistical power (Snijders, 2005), here, the missing parameter can be calculated, if six of the seven parameters (i)$1-\beta$, (ii.a)number of biological replicates N and (ii.b) number of technical replicates M, (iii) δ, (iv.a) biological variance σ^2_{biol} and (iv.b) technical variance σ^2_{tech}, and (v) α are given.

Results and Discussions Figure 3.2 (left) shows the estimated variances for all obtained $S = 642$ features. The mean values of all feature variances are $\sigma^2_{instr} = 0.043$, $\sigma^2_{prep} = 0.076$, $\sigma^2_{biol} = 0.172$, hence they increases from technical to biological variation $\sigma^2_{instr} < \sigma^2_{prep} < \sigma^2_{biol}$ and the mean total variance $\sigma^2_{tot} = 0.291$ is the sum of these individual contributions. On average across all features the instrumental variance is 16.7%, the preparation variance is 29.1%, and the plant variance is 54.2% of the total. Figure 3.2 (right) shows the distribution of ICC_i, the proportion of each variance source on total variance (Sampson et al., 2013b), of the individual features and illustrates that half of the features have an ICC above 0.58.

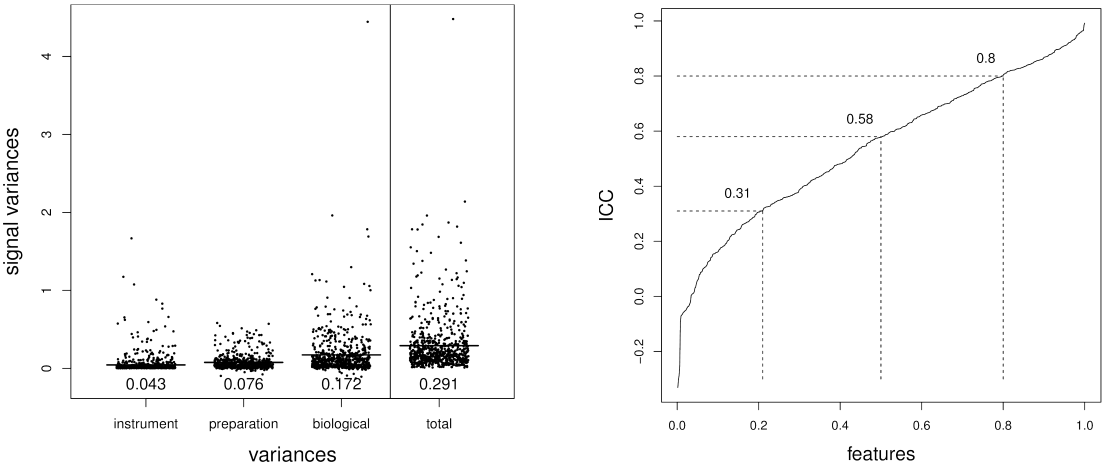


Figure 3.2.: The distribution of estimated variances of all measured features in leaf samples. Left: From left to right the estimated variances of all measured features $S = 642$ in leaf samples for σ^2_{instr}, σ^2_{prep}, σ^2_{biol}, and σ^2_{tot} are plotted. Each dot represents the estimated variance of one feature in the sample. The mean of all estimated feature variances for each variance level is given below and shown as black bar. Right: The cumulative distribution of ICC_i for all features i. E.g. 80% of the features have an ICC above 0.31, half of the features have an ICC above 0.58, and even 20% are above 0.8. The higher the proportion of features with a large ICC, the more important is a hierarchical experiment.

The experimentalists will have to decide whether the increased quality of the test justifies the added costs and the experimental effort when using more replicates. For this, a two step decision has been made: 1) find all possible combinations of biological and technical replicates in a two-level hierarchical experiment design using power calculation approach, 2) choose the combination which has the lowest costs, given a ratio of the costs between biological and technical replicates. This comparison of costs can help to choose an efficient experimental design. For example, given a fixed cost ratio of 9:1 (biological vs. technical), for a real effect of $\delta = 1.5$ or below and the given mean varaince estimates, technical replicates and the hierarchical t-test are superior (i.e. cheaper) than a normal t-test without technical replication.

Conclusion In mass spectrometry-based metabolomics there are several sources of variance. Based on a pilot study, it is shown that the hierarchical variance analysis is a method to quantify and separate these additive sources of variances. Such a pilot study is also a tool to determine the different sources of variance relative to the overall observed variance in a MS experiment and should be performed for each analytical setup and each organism or tissue type. In this setup it was found that the biological variance is larger than both the instrumental and preparation variance combined.

The statistical power depends on 1) the observed variance, and 2) the number of biological replicates and 3) the real effect that is relevant for the biological question and which is desired to be statistically significant. To decrease the influence of non-biological variance, technical replicates can be acquired and analysed with a hierarchical type of Student's t-test, or having more than two classes with nested ANOVA, or in general with multilevel mixed models.

For large studies following the prior conducted pilot experiment, an optimal experiment design is highly requested to save costs and effort, while maintaining statistical power.

3.4.2. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent.

The second article (section 6.2) benefits from the methodological investigations of the first article (Trutschel et al., 2015a, section 6.1) and used the proposed method to answer a metabolomic related question. Here, a real study (blue box Figure 1.1) with a hierarchical experimental design was conducted on the problem of plant-to-plant variability in root metabolite profiles of 19 *A. thaliana* accessions.

Introduction In plant science *A. thaliana* is a model species to investigate secondary metabolic pathways, whereby plant-to-plant variability has neither been investigated in root metabolism nor have previous studies incorporated more than two *A. thaliana* accessions into a comprehensive root metabolic profiling analysis.

In general, especially in roots, material of several plants is pooled before sample preparation, because of biomass is very little. The previous article (Trutschel et al., 2015a, chapter 6.1) shows a solution for how to incorporate different kinds of replicates into a powerful experimental design without the need for sample pooling. Instead, a hierarchical experiment design was used to be able to decompose the total observed variance of root metabolite profiles in the general physiological state into the components attributable to natural variation between accessions, experimental batch and individual variability between plants. Furthermore, the relative biological variability of three important substance classes was investigated: glucosinolates (GSLs), flavonoids, and phenylpropanoids including oligolignols, the latter playing a vital role in root metabolism.

Materials and Methods Seeds of 19 *A. thaliana* accessions were analysed by LC-MS and GC-MS. The used hierarchical experimental setup of the study is shown in Figure 3.3 (compare to the similar setup of the previous article, Figure 3.1).

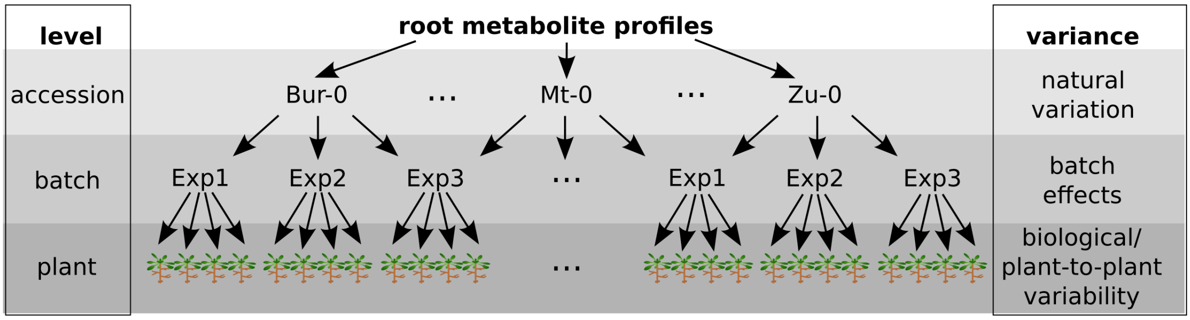


Figure 3.3.: Nested experimental design with three levels. Each variance level had multiple replicates to assess natural variation, 19 accessions of *A. thaliana* were grown. Three independent biological experiments were performed to estimate non-biological variance derived from the experimental batch. To assess individual variability, four plants were harvested in each biological experiment for each accession. Single-plant root extracts were subjected to LC-MS and GC-MS analysis.

The method of linear mixed models (lmm), which is more widely applicable compared to using the ANOVA-based variance estimation method (of the previous article), was used to dissect the total observed variance and quantify the amount of different sources of variation. With the obtained log-transformed metabolite abundances the variance contribution at each experimental level - accession, batch and plant - were estimated. Furthermore, lmms with only batch and plant as random effects were applied separately to each accession to examine accession-specific variances. Intraclass correlations (ICCs) were also calculated as the ratio of plant variance σ^2_{plant} and total variance σ^2_{total} according to (Sampson et al., 2013a) for each feature and additionally for known metabolites.

Results For non-targeted metabolic profiles of primary metabolites the estimated mean between-plant variance $\sigma^2_{plant} = 0.50$ is larger than the between-accession variance $\sigma^2_{accession} = 0.37$, whereby the estimated mean between-experiment variation $\sigma^2_{batch} = 0.19$ is less than σ^2_{plant}. It indicates, that for root metabolic natural variation, plant-to-plant variability seems to be larger than between-accession variance.

In addition, three sets of annotatable compounds were quantified (Figure 3.4): GSLs, flavonoids, and phenylpropanoids. Figure 3.4(a) separates the feature variance estimates according to the substance classes and Figure 3.4(b) interprets plant-to-plant variability in the context of total variance using ICC estimates. GSLs and phenylpropanoids show a large range of ICCs, where for flavonoid metabolites, the ICCs are rather high but similar for all analysed members of the substance class.

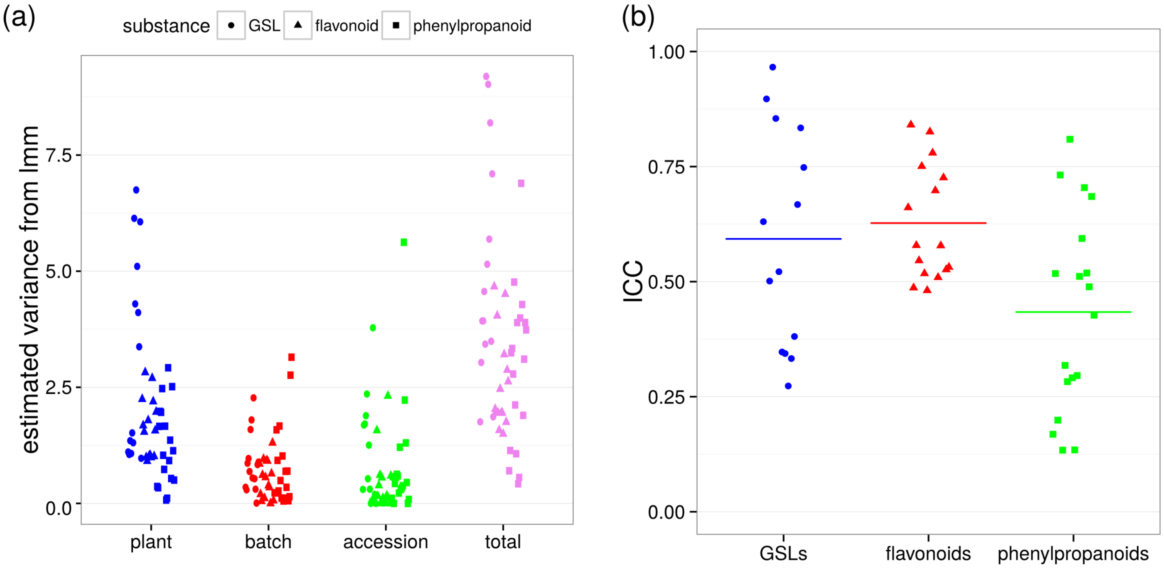


Figure 3.4.: Biological variability of annotated secondary metabolites. (a) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot: variance of one metabolite; (b) ICCs for glucosinolates (GSLs), flavonoids, and phenylpropanoids, dot: ICC of one metabolite, bar: the mean ICC for a substance class.

Discussion Measuring single plant extracts prevented the irreversible information loss resulting from pooling plant material and allows to distinguish between accessions and still analyse plant-to-plant variability. If a broad range of metabolites are of interest, it is important to know the biological variability that is exhibited by most metabolites. For example, calculations with the mean ICCs will provide sufficient power for analyses of flavonoids, but not for all metabolites of the classes GSLs and phenylpropanoids due to the high variability.

Conclusion The provided knowledge within this article about the variances of different substances classes can be exploited to appropriately design an experiment prior, because it may differ between a non-targeted screen and the analysis of specific substance classes. To exploit the full potential of a non-targeted metabolite profiling, single-plant measurements should be acquired and correctly integrated into the analysis. Hence, different substance classes of interest might require a specific experimental set-up guided by obtained variance values.

3.4.3. Joint analysis of dependent features within compound spectra can improve detection of differential features.

The third article (section 6.3) has the aim to find statistical methods to jointly analyse dependent features (adducts, fragments, isotopic peaks of one metabolite). Here, dependencies between rows of a two-dimensional data matrix obtained by mass spectrometry data sets are taken into consideration. The focus lies on data characteristics, which determine the analysis method (Figure 1.1). In addition, the problem of multiple hypothesis tests is addressed.

Introduction A typical research question in the field of metabolomics is biomarker discovery. Therefore, univariate hypothesis tests like Student's t-test (Student, 1908) and Analysis of Variances (ANOVA) can be used to detect differences between two or more sample classes, but one of the underlying assumptions is the independence between individual metabolic features. However, it is known that in mass spectrometry a single metabolite usually gives a rise of mass spectral features, e.g. isotopes, adducts or fragments (Brown et al., 2009), which are observed together and show a common behaviour across samples. Thus, methods for a joint analysis of such features are required instead of multiple univariate tests.

Multivariate methods like MANOVA are global approaches to analyse all features together taking correlations between all features into account. Nevertheless, in metabolomics the number of samples is usually much smaller than the number of features to be analysed. Therefore, correlation and covariance structures are difficult to estimate for all given features, and require an initial variable selection step. An alternative way to this joint analysis of all features is treating only related features together. This means to group those which originate from the same metabolite referred to as "compound spectra" in the following for a subsequent analysis. Then a multivariate analysis on the level of compound spectra instead of a global multivariate approach can be applied to determine differential metabolites.

Materials and Methods For the analysis, two metabolomics data sets from *A. thaliana* were used. The first is a dataset, where 26 independent plant profiles and a simulated effect were used to evaluate the proposed methods. The methods are then demonstrated on a second dataset including *A. thaliana* wildtype and a mutant line. Therefore, several multivariate methods to jointly analyse compound spectra representing metabolites are proposed.

The **univariate** Student's t-test (Student, 1908) assumes normal-distributed observations of independent features. The difference of the intensity mean between the two classes is estimated for each feature. The confidence interval (CI) of the estimated mean difference determines the accuracy of this estimation, and the CI size depends on the number of observations and the standard error (SE) of the estimated difference between means. Figure 3.5 shows that if independent univariate tests for two features in a compound spectrum are combined, the confidence interval becomes a rectangular confidence region or in general for groups with p features a p-dimensional hypercube. Even if multiple testing correction is done the confidence region holds a hypercube.

The **multivariate** extension, here Hotelling's T^2 (Hotellings, 1931), compare the difference of p-dimensional mean intensity vectors in relation to their $p \times p$ covariance matrices. Observations of features in a compound spectrum are then assumed to be multidimensional normal-distributed. Figure 3.5 illustrates for this multivariate analysis (of two depended features), the confidence region has an ellipsoid shape and hence, is not so conservative as usage multiple univariate tests. Using the multivariate tests, this statistic requires at least $\binom{p+1}{2}$ replicates (samples), where p is the number of features per metabolite group, to estimate the unknown entries of each covariance matrix. Additionally, a variant of the multivariate methods is proposed, named diagonal Hotelling's T^2. Therefore, only the diagonal entries of the covariance matrix are estimated, with the rest fixed to zero. This simplification ignores the correlation between features, but makes the covariance estimation more robust in the case where a compound spectrum consists of more features than samples are available to modify the idea of spectra-wise analysis on small data sets. In comparison to a full covariance matrix estimation, the main axes of the ellipsoid confidence region in Figure 3.5 are then parallel to the coordinate axes.

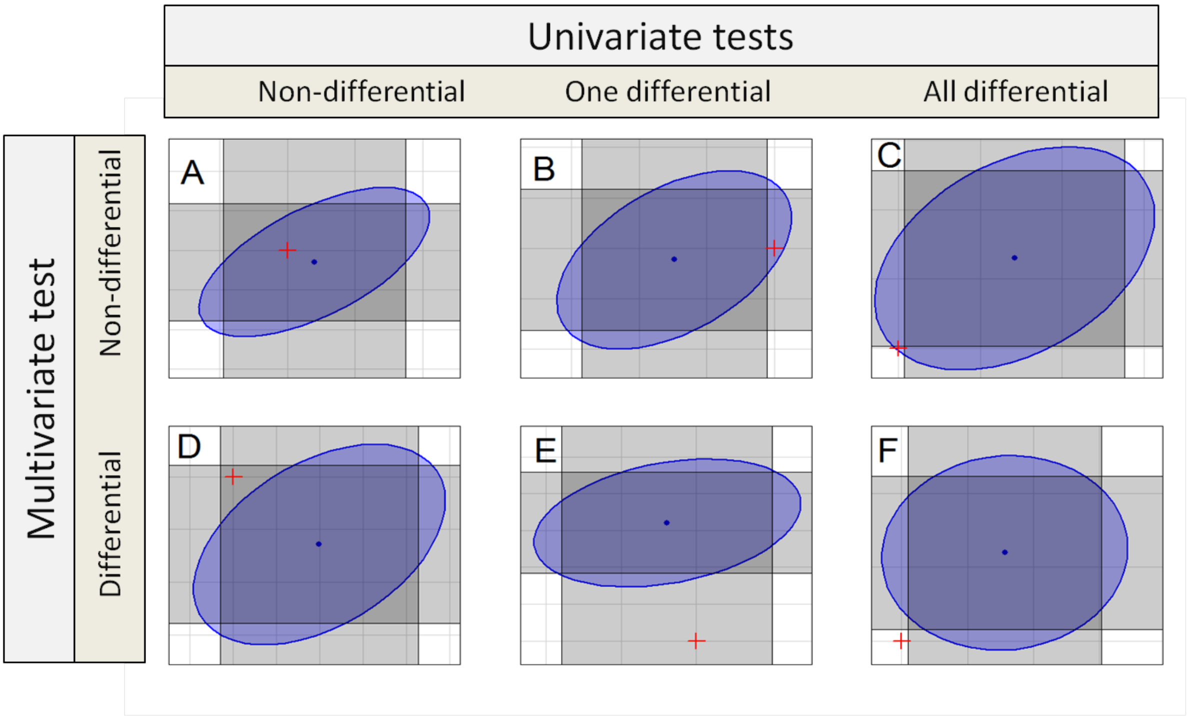


Figure 3.5.: Different decisions from univariate and multivariate test to detect differential features or compound spectra. Each gray rectangles marks the confidence interval of one test dimension, so the intersection of two rectangles marks the combined confidence region. The blue ellipse is the confidence region for a multivariate test. There are six different possibilities (six different coloured spaces) for the position of the origin corresponding to the null hypotheses marked by a red '+'.

Results and Discussions The performance of the three statistical analysis – univariate, and multivariate with both Hotellings-T^2 and the diagonal Hotellings-T^2 – was compared on two sample classes dataset of metabolite profiles from *A. thaliana*. Therefore, the negative set (effect 0.0) with 686 features in 153 compound spectra was combined with a positive set consisting of the same 686 features but with an added effect. For each effect between 0.0 and 1.4 the final ground truth dataset thus contained 306 compound spectra with a total of 1372 features.

Figure 3.6 demonstrates the behaviour of the proposed methods for all different effects. Therefore, the area under the curve (AUC) was used as a summary metric of the performance. For the feature detection (Figure 3.6, top) it is shown that the multivariate T^2 as well as the diagonal T^2 method has a better AUC compared to the univariate approach for all effects of $0.2, 0.3, \ldots, 1.4, 1.5$. Especially for smaller effects, the benefit of the multivariate approach is visible and also that the simplified diagonal T^2 approximates to the original T^2 for larger effects. Between the different compound spectra level (or grouping) approaches (Figure 3.6, bottom) no particular differences are shown. Thus the main benefit results from a joint analysis of compound spectra, while less differences are observed between the joint analysis methods.

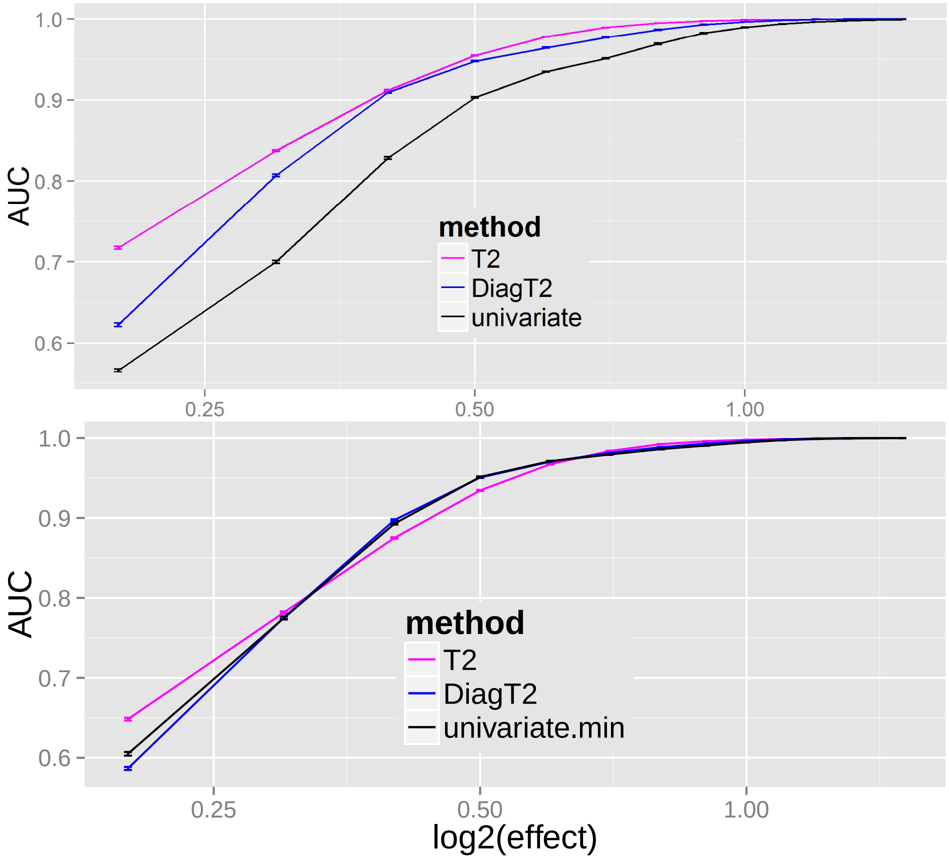


Figure 3.6.: Top: Results of univariate and multivariate methods in feature detection are compared at the feature level. Bottom: At the compound spectra level the results of different grouping analysis approaches are shown. For each simulation step, several added effects of $0.2, 0.3, ..., 1.4, 1.5$ on the 'mutant' class, the mean and SE of the evaluated AUCs (results from 100 repetitions) are plotted.

Conclusion In mass spectrometry-based metabolomics data will usually give a rise of multiple spectral features. In recent years, methods were developed to group these related features into compound spectra. However, the statistical analysis was still based in either individual univariate tests, or global multivariate analysis. Within this article the feature-wise univariate statistic tests to a compound spectra-wise analysis has been extended .

At the feature level the resulting AUCs for the multivariate analysis of compound spectra were better than in the univariate case. Hence, for biomarker discovery in mass spectrometry metabolomics data the analysis of data compound spectra-wise can now be recommended. At the compound spectra level the advantage of T^2 over the other spectra-wise approaches is most prominent for noisy data and/or if very small effects should be detectable.

The proposed joint analysis of features of a metabolite group as a spectra-wise analysis is the key idea and bridges an important gap between hypotheses tests on individual features on the one hand, and global multivariate methods which might be more difficult to interpret.

4. Health care - Analyse data of people with dementia in clinical nursing services research

4.1. Nursing services research

Nursing science develops systematically knowledge with impact for nurses. Clinical nursing science has the aim to guide nursing practice and to improve care and quality of life of patients (Polit, Beck, and Hungler, 2004). The aim is to validate, refine and generate knowledge, for example from empirical studies, that directly and indirectly affects the delivery of nursing care (Burns and Grove, 2009). Results of clinical research have also an impact on patient care decisions and recommendations for cost reduction of health care. This means an evaluation of the efficiency of cost reduced, but quality maintained health care services (Polit and Beck, 2008). The so called evidence-based practice is the use of the current best available evidence in patient care decision-making (Sackett et al., 1996; Murray et al., 2013; American Association of Colleges of Nursing, 2015). Hence, high reliability of results require research methods to derive unbiased effect estimates of an intervention in a certain population in real world settings and means in particular the need of randomization (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, 2017; Adams et al., 2018).

4.2. Dementia research

Nowadays, there is a special interest within nursing service research on care of people with neurocognitive disorders. Although these disorders are categorised by different subtypes (e.g. Alzheimer's disease) and different severe stages (mild, moderate and severe), they are popularly known as dementia. Neurocognitive disorders are defined by a decline of several cognitive functions, which is related to a decrease of independence in every day activity (APA, 2013). Related to the population ageing, there is a shift to disease burden at older ages worldwide. Dementia is one of the major neurological disorders, which is in the list of the top increases from 1990 to 2010 (Murray et al., 2013; Prince et al., 2015a). The number of people with dementia worldwide will increase further with an estimation of 131.5 million by 2050 (Prince et al., 2015b). Currently, this progressive disease cannot be cured, thus, the increase of people with dementia requires specialized care. Dementia is associated with increasing healthcare costs in European countries (Wimo et al., 2013) and therefore for society means an increase of costs. This will have implications to health-services planning, manpower and education, e.g. (Murray et al., 2013).

4.3. Dementia research with the focus on the quality of care

Although for a high proportion of people with dementia it is possible to live at home (alone) in Germany, UK and US (Eichler et al., 2016), the behavioural and psychological symptoms affect the quality of life of people with dementia and also of their carer. This circumstance and also the absence of cure are the reason why it is necessary to develop interventions to improve or maintain quality of life (Prince et al., 2015a; Klapwijk et al., 2016). As a consequence, studies are conducted to evaluate the quality of care and the effectiveness of interventions to improve or maintain quality of life. For example, one scientific question to answer is, whether residents who live in care units which are specialised in care of residents with dementia are more likely to receive an intervention that is increasing quality of life than residents of other care units.

4.4. Methodological challenges of analysing data of people with dementia within nursing services research studies

Data collection within nursing research studies are displayed as a two-dimensional data matrix of n rows × m columns, where n is the number measurements (e.g. one measurement for each resident/patient) and m the number of variables obtained from the used questionnaires.

Complex problems are present in this scientific field and the proceeding how to obtain and analyse data is not a restrictive approach, but rather there is a need of a flexible answer according to several scientific questions in terms of care of persons with dementia. Nevertheless, the specific characteristics of the collected data according to the type of i) study design ii) manipulation and iii) outcome variable or measurement method (see Figure 1.1) play a decisive role in choosing analysis methods within this scientific field. The characteristics and the resulting consequences for analysis of this specific data are described in the following.

In health care research the cohort study design type is often used to identify potential risk factors for outcomes and study changes or development over time. To follow patients over a period of time it is necessary to measure interested variables repeatedly during the study. One example is to decide, if two cohorts differ from frequencies of behaviour occurrence (incidence) over time. Furthermore, the collection of resident or patient measurements within clusters, for example nursing homes or hospitals, is common. The reason for that is often randomised studies could only be realised through cluster randomisation because in order to avoid contamination of the intervention (exposition). Both, repeated measurements over time and collection data within clusters, result in dependencies between observations (dependencies between rows of the data matrix). This violates the assumptions of independent measurements, which is fundamental to most hypothesis tests. This is one of the main data characteristics that need to be considered. Then, an adjustment for such dependent observations within estimation models to obtain unbiased intervention effects are necessary.

While randomized studies are the "gold standard" and this study design generates the highest level of evidence to answer questions about effectiveness of different interventions (treatments), in nursing science various circumstances hinder the conduction of a randomized controlled trial (not feasible or practical). Hence, other study types like observational ones must be considered (Ho, Peterson, and Masoudi, 2008). When data are collected in observational studies, dependencies due to allocation of individuals to the treatment classes is possible. Here, randomization is not part of the study design and assignment to treatments may be associated with the potential outcomes of the treatment (Ridder and Graeve, 2011). This is a source of selection bias and statistical methods are needed to reduce it.

Within nursing services research very complex systems are under investigation. Within this complex system very different interaction partner are involved, e.g. persons being cared for, care giver or social environment. Within a study setting often some variables cannot be fixed and thus, the existence of different sources of bias is usual. Due to the influence of several confounders on outcomes, the data are often acquired with a high variability. Within the statistical model the bias can only be adjusted for, if the confounder are known and measured, otherwise bias remains hidden.

Furthermore, health care research often deals with so-called complex interventions. Complex interventions are characterized by having multiple components, which interact with each other. Thus a single primary outcome may not make best use of the data; a range of measures will be needed. Hence, if multiple testing problem is present, then a correction of p-values has to be considered by researchers. Furthermore, different types of outcome distributions due to the used questionnaires are possible. Hence, different tests or histograms according to their respective assumptions (e.g. normality) are required and sometimes transformation of data (e.g. for binary data) could be necessary.

One of the objectives of this thesis, the usage of appropriate statistical methods to answer scientific question regarding the specific data characteristics, is addressed in two articles with the special interest on studies on people with dementia within nursing research. One article is a methodological one and the other a real study example, which profit from the methodological insights of the first. The used methods may also be interesting for other parts of health care. Therefore, R-code is provided within the methodological article for an easy use of the discussed methods or to adapt the analysis process in other contexts.

4.5. Publications

4.5.1. Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing

This article (section 6.4) focuses on considerations about data characteristics, for example a special type of outcome, the study type or a specific design, which determine the choice of methods for the data analysis (Figure 1.1). While data analysis generates results, the analysis method then directly influences the interpretation of the results. These considerations are shown on a practical example in health services research.

Introduction Motivated by an observational study in health services research, there was an interest on a special study type: observational studies in which three main data characteristics (*dichotomous* outcome, *clustered* data, *observational* study) need to be addressed to find a suitable analysis method, as illustrated in Figure 4.1.

First, the distribution of the outcome variable, influences the choice of the statistical method. Here, the use of case conferences as a binary outcome is analysed. Binary variables are summarised by probabilities, odds and odds ratios (OR) , whereby the interpretation of odds is more difficult for practitioners (Greenland, 1987; O'Connor, 2013).

Second, in observational studies, the possibility of controlling factors that may influence the study outcome is limited to observed variables because randomisation is not part of the study design. Therefore, other options must be applied to reduce selection bias (Ridder and Graeve, 2011), which can contribute to over-/underestimations of the intervention effect (Starks, Diehr, and Curtis, 2009).

Third, this study example is an evaluation of special care units. Studies about dementia special care units typically have a multi-stage clustered data structure: residents are clustered within units, units are clustered within nursing homes, and nursing homes are clustered in provider systems. If a clustered or nested data structure is most likely present, the error terms within a cluster are no longer independent (see also Trutschel et al., 2015a, section 6.1). Hence, this clustering must be considered when choosing the analysis method to avoid overestimating the significance of the effects.

Materials and Methods From an observational study (Palm et al., 2014) a dataset from 64 care units in 36 nursing homes ($n = 835$ participants) is provided. The primary question for this analysis was whether a dementia special care unit (treatment group) more frequently performs case conferences than traditional care units (control group).

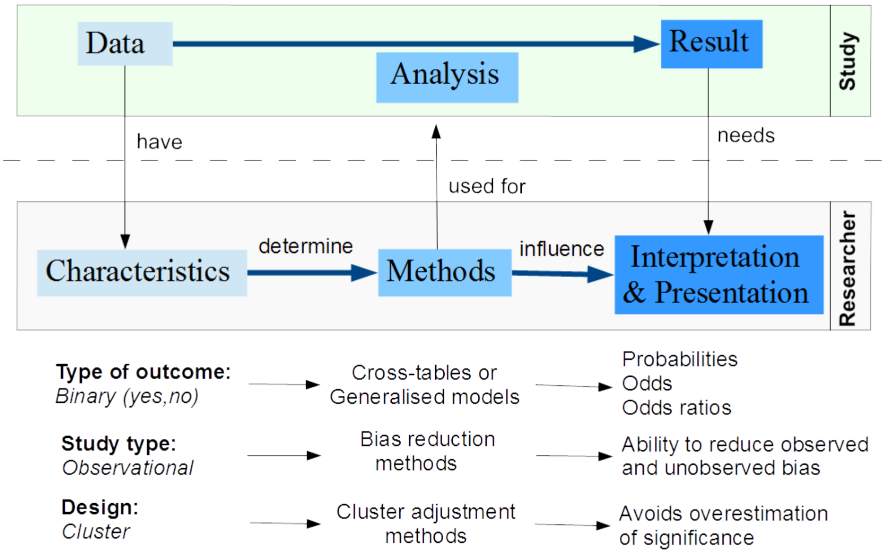

Data
Analysis
Result
Study
have
needs
used for
Characteristics
determine
Methods
influence
Interpretation & Presentation
Researcher
Type of outcome: Binary (yes,no)
Cross-tables or Generalised models
Probabilities Odds Odds ratios
Study type: Observational
Bias reduction methods
Ability to reduce observed and unobserved bias
Design: Cluster
Cluster adjustment methods
Avoids overestimation of significance

Figure 4.1.: In a study, the data analysis generates results. The data have their own characteristics, for example, a special outcome type, a unique study type or a specific design. These characteristics determine the choice between sophisticated methods for data analysis. Hence, the method directly influences the interpretation of the results and therefore must be carefully chosen using the skill of the researcher.

Here, two levels of analytical strategies are illustrated: i) different models with different abilities to adjust for dependencies (due to clustered data) to analyse binary data and ii) different methods to adjust for selection bias .

Different models and their ability to adjust for dependencies.

First, the crude model is a simple contingency table. Table 4.1 (upper part) provides an initial overview: the distribution of a binary outcome variable, here, the performance of case conferences. Table 4.1 (bottom part), illustrates also that the (estimated) probability of 'receive a case conference' p can be calculated as a proportion from the frequencies in each group, and differences in (estimated) probabilities between the two groups (treatment and control) can easily be calculated by $p_1 - p_2$. The odds of each group are then defined as the ratio of the probabilities p and $1-p$ ($\frac{p}{1-p}$). It compares how much larger one probability is relative to another in a specific group. The widely used odds ratio is thus the ratio of both odds, namely, the odds of the treatment group related to the odds of the control group. The crude model provides almost the same results as the logistic regression model with only one independent variable for group assignment (treatment versus control). The logistic regression belongs to the family of generalised linear models (GLMs), which can handle different distributions of outcome variables (Gelman and Hill, 2007). The generalised linear model adapts the linear relationship between the probability that an event occurs and the predictor variable (group) by using the logit function.

Secondly, a generalised linear mixed model is used when observations are not independent, because of clustering in different nursing homes for example. Generalised linear mixed models are an extension of the generalised linear models (Hardin and Hilbe, 2012; Stroup, 2012; Li et al., 2011). They combine two statistical concepts: using linear mixed models to include random effects and using generalised models to model non-normal distributed data.

Methods for selection bias adjustment.

Two methods for bias reduction that can be used for analysing data (by crude or advanced mixed model) to achieve balance are used: 1) genetic matching (Pimentel, Yoon, and Keele,

		Group		**M**arginal
		Treat (Special care)	Control (Traditional care)	
Outc.	no	n_{11} (22)	n_{12} (119)	$\mathrm{n}_{1.}=\mathrm{n}_{11}+\mathrm{n}_{12}$ (141)
	yes	n_{21} (224)	n_{22} (470)	$\mathrm{n}_{22}=\mathrm{n}_{21}+\mathrm{n}_{2.}$ (694)
	Marginal	$\mathrm{n}_{.1}=\mathrm{n}_{11}+\mathrm{n}_{21}$ (246)	$\mathrm{n}_{.2}=\mathrm{n}_{12}+\mathrm{n}_{22}$ (589)	$\mathrm{N}=\mathrm{n}_{.1}+\mathrm{n}_{.2}=\mathrm{n}_{1.}+\mathrm{n}_{2.}$ (835)
Interpr.	Probabilities	$p_{\text{Treat}} = \frac{n_{21}}{n_{.1}}$ (0.91)	$p_{\text{Control}} = \frac{n_{22}}{n_{.2}}$ (0.8)	Diff. $= p_{\text{Treat}} - p_{\text{Control}}$ (0.11)
	Odds	$\text{Odd}_{\text{Treat}} = \frac{p_{\text{Treat}}}{1-p_{\text{Treat}}}$ (10.18)	$\text{Odd}_{\text{Control}} = \frac{p_{\text{Control}}}{1-p_{\text{Control}}}$ (3.95)	$\text{OR} = \frac{\text{Odd}_{\text{Treat}}}{\text{Odd}_{\text{Control}}}$ (2.58)

Table 4.1.: Upper: A contingency table of a two-group comparison for a dichotomous outcome variable, where $\mathrm{n}_{i,j}$ is the absolute amount of outcome i in group j. Lower: Parameters, their estimates calculated from the contingency table and their interpretation. (p = probability, OR = Odds ratio)

2015; Rosenbaum, 2002; Rubin, 2006; Stuart, 2010) on samples and 2) adjustment via the regression model (Cepeda et al., 2003; Gelman and Hill, 2007). Balancing in this context means that the baseline characteristics in the treatment and control groups are the same (matching) or that balance differences are taken into account (regression).

The first method, balances the sample independent from the outcome, which means matching of similar individuals in the treatment group with individuals from the control group with the disadvantage of loosing information, but for balancing observed control variables in both groups (Baser, 2006). After that the matched sample can be further analysed, but needs additional adjusting for produced dependencies.

The other adjusts for selection bias by estimating the contribution of each variable to the outcome within a regression framework by inclusion of independent variables (covariates) into the model. Thus, the analysis and bias adjustment are not separated and provide a conditional estimate of the treatment effect (given levels of the covariates).

Results According to the crude model (Table 4.1), 91% of residents in dementia special care units received a case conference, whereas only 80% in traditional care units received a case conference. The substantive interpretation would be that a patient is more likely to receive a case conference in dementia special care units than in traditional care units. The table shows also an odds ratio of 2.58, which indicates that the odds of receiving a case conference is more than two and a half times higher in the group of special care units than in the group of traditional care units; in other words, being in the treatment group (relative to the control group) raises the odds of receiving a case conference.

In the opposite, through the generalised linear mixed model, the estimated odds ratio is more than three times higher than the odds ratio in the generalised linear model ignoring the clustered data (see article). This study example shows, that addressing the dependencies with a mixed model has an impact on the estimation of odds ratio. Here, this considerable difference can be explained by the strong clustering effect present in these data.

Discussion and Conclusion Although the different analysis methods present different results, they at least point in the same direction, indicating that the estimated probability of receiving a case conference might be higher in the treatment group than in the control group. However, in this study, when adjustment for bias and dependencies is performed, the null hypothesis of a difference in the use of condition between the two groups could not be rejected.

Before collecting data for an observational study, two major concerns should be taken into account: 1) covariates that may obtain selection bias and hence require measurement are determined and 2) a larger sample size is needed to ensure a sufficient sample size; although there is a loss due to adjustment methods. However, further investigations should be performed to allow drawing conclusions regarding the minimum required sample size within observational studies, which has to be adjusted for bias, or, if bias appears, how much of the sample is being lost via matching.

4.5.2. Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study

This article shows the application of the methodological problems due to analysing observational studies with clustered data, which is discussed in (Trutschel et al., 2017, section 6.4) on a real study example (see Figure 1.1). Here, a real cross-sectional study was performed to address the research questions: *Are residents who live in dementia special care units with additional funding more likely to receive case conferences than residents of traditional care units?* This dataset is faced with the problem of selection bias, as well as clustered data. Hence, this article shows the statistical methods to adjust for these issues to receive unbiased and reliable estimates.

Introduction A real study example was performed with the aim to evaluate, whether residents who live in dementia special care units are more likely to receive case conferences, a common tool used to find a solution to clinically relevant problems, than residents of traditional care units.

At present, dementia special care units (DSCUs) form one of the most frequently implemented structural care interventions. Although a definition of DSCUs does currently not exist, there is agreement regarding special structural and residential characteristics of DSCUs in comparison to traditional care units (TCUs). For example specially designed environments, higher staff-to-resident ratios, and dementia-oriented therapy programs are provided to residents with dementia, severe cognitive impairments, and severe challenging behaviours. However, that means higher costs that are attributed to more intensive resource use. From a health policy perspective, the extra costs raise questions about the benefits of DSCUs and the regulations for preventing misuse of the funds.

Case conferences are a common tool that is used in long-term care practice to find a solution to a clinically relevant problem. In Germany, case conferences are usually not part of the routine care process but are provided when specific problems occur that require an adaptation of the care plan (e.g.. enduring refusal of food or drinks). Thus, the provision of multidisciplinary case conferences is considered to be a feature of DSCUs in Germany.

Methods To address the research questions a cross-sectional study was performed and 1808 residents living in 109 care units in 51 German nursing homes were recruited. Data were collected at the levels of the nursing home, care units and residents. Due to exclusion criteria and missings, at the end, a data sets from 888 participants were used for the analysis. The provision of a case conference (dichotomous) was investigated as the dependent variable to answer the research question.

Based on the discussion of observational studies, it is assumed that the two samples (TCU, DSCU) were not equally distributed with regard to subject characteristics. This resulting selection bias can be adjusted (see investigations of Trutschel et al., 2017, section 6.4) by creating an new dataset (e.g. via matching) or by adaptation of the estimation model (e.g. covariate adjustment). For matching, we used criteria similar to those used for admission to DSCUs (care level, cognitive impairment, diagnosis of dementia, mobility) and relevant sociodemographic variables (age, sex, place of residence before moving into the nursing home). At least, the expected values of the outcomes in each group were estimated by model-based odds and odds ratios using logistic regression models.

Results From the sample of 888 residents, a matched sample with 264 residents in each group were drawn, which means a information loss through matching process. It was discovered that DSCU residents received case conferences more often than TCU residents. Using the generalized linear mixed model, the odds of receiving a case conference was estimated to be nearly 10 : 1 in TCUs and 40 : 1 in DSCUs. This resulted in an OR of 4 between the two groups and means that the chance of receiving a case conference was 4 times higher for residents in a DSCU than for residents in a TCU. However, this OR was not significant and had a large 95% confidence interval.

Additionally, the results indicates that it was more common for DSCUs to conduct case conferences as a matter of routine compared with TCUs, although the majority of the case conferences in both types of units were conducted for specific reasons. In DSCUs, residents, relatives, head nurses, and physicians took significantly smaller roles in case conferences compared within TCUs, whereas therapeutic, housekeeping, and other care staff took on significantly greater roles. The topics of nutrition, falls/risk of falling, cognition, and psychosocial situations were discussed more often in DSCUs than in TCUs. Challenging behaviours were discussed more often in TCUs than in DSCUs. Regarding the performance of the case conferences, the only difference between the groups that remained significant was the topic of challenging behaviours, which was addressed more often in TCUs.

Discussion In this study, after adjusting for differences in the resident sample and the clustered structure of the data, the hypothesis that DSCU residents were not more likely to receive a case conference could not be rejected. Data from both groups indicate that case conferences are a common intervention but that they do not occur more often in DSCUs. Only, the performance of these conferences differed in regard to the topic of challenging behaviours, which was discussed more often in TCUs than in DSCUs.

Case conferences on the management of residents' behaviours appear to be an important issue in TCUs. In DSCUs, the staff members are more likely to be faced with challenging behaviours and may use strategies to manage them more routinely than in TCUs. It is possible that TCU staff need more support for handling problematic situations and reducing behaviour-related distress. In addition, TCUs are often not designed with consideration for dementia-related problems (such as wandering behaviours and orientation problems); thus, the staff members must find alternative solutions to address these problems.

Conclusion Case conferences including a multidisciplinary exchange are a widespread care intervention in DSCUs and TCUs. The results of this study indicate that case conferences are a common care intervention but that they do not occur significantly more often in DSCUs, when adjustment for clustering and reducing selection bias was included into the model.

5. Conclusions and outlook

The aim of this thesis was to discuss different methodological challenges of conducting experimental studies in two scientific fields with large datasets: mass spectrometry and nursing services research. Within such scientific fields, recommendation to 1) acquire data with most powerful experiment designs and 2) analyse data with appropriate methods are given. Additionally, all used statistical methods are made available for practitioners for application. Thereby, the overall aim persists 3) analysing data of a study, appropriately, to answer a scientific question. Thus, real study analysis show the application of the methods discussed in the methodological articles at least. Although, both scientific fields are faced with the same type of data (two-dimensional large datasets) and also the same aims, the study designs and required methods differ between them.

Using more complex statistical methods on large datasets requires computational power and methods. Hence, a key skill lies on programming. In this thesis simulation studies are implemented in different articles (chapter 6.1 and 6.3). Additionally, for all methodological articles a vignette was given to make the implementation of these complex methods available for practitioners. An R-package for sampling multivariate normal-distributed data, the base of the underlying data within both fields, is also provided.

Similarities between both scientific fields. The primary aim of both fields, detection of relationships or even causal effects between independent and dependent variables, asks for valid and reproducible results. Hence, appropriate statistical methods for data analysis to obtain less biased effect estimates are required to make conclusions about causal relations. and was discussed in the articles of section 6.3 and 6.4. Different experiment designs, the basis of following sufficient data analysis, is discussed in the article of section 6.1.

The common ground of all articles is the consideration of dependencies within data. Such dependencies are the basis of multivariate statistics on two-dimensional data. Two kinds are discussed in this thesis: dependencies between columns and between rows of two-dimensional data. The type of a multi-dimensional normal distribution, the basis of the multivariate statistical method used in the articles within this thesis, is given by its mean vector and its variance-covariance matrix. This distribution is derived and explained with more detail in the section A.2 in this thesis.

Differences between both scientific fields. The suitable statistical method using the correct assumptions is related to the data characteristics and differs between between metabolomics and nursing science, because of the scientific context and research question. Hence, the different used study types and outcome distributions require different methods. For example, in mass spectrometry the distribution of logarithm intensities of obtained features are assumed normal, where in nursing services research a variety of distributions are possible because of nominal, ordinal or continuous outcomes due to used data acquisition method. Hence, in this area a generalization of the commonly used linear models are used. In the examples of real data analysis within the two fields (section 6.2, 6.5), different models are used. Also the study designs varies widely: case control studies are usually used for mass spectrometry experiments (section 6.1, 6.3), while randomized and observational studies are present in nursing services research (section 6.4). Thus, different analysis methods are required to adjust for possible sources of bias.

Possible extensions. With the five included articles the objectives of this thesis have been implemented. All frequentist approach methods given in this dissertation are transferable to data analysis of other scientific fields, as long as the assumptions that are made are equal. Nevertheless, some common statistical methods were not scope of this thesis, but could be valuable extensions in future.

For example, a Bayesian way of data analysis, which includes a-priori assumptions in the model to re-allocate the results with given knowledge. The problem of using Bayesian analysis in relative new sciences is that such a priori knowledge is not known or generated. Often in publications the necessary information (e.g. estimation of distribution parameters for prior) is not given due to the fact, that this importance is not known. The aim of future analysis is to collect knowledge from previously studies in order to weight the results of subsequently performed studies with them using Bayesian methods.

Another approach, which should be considered more often, is the use of meta-analysis for a number of comparable single studies. It helps to arrange the results of studies in order to previously published studies within the same context. Nevertheless, here it is the same problem, that within articles often not all required information are given or comparable studies has not yet been conducted.

A future task could also be providing the methods for applied researcher, which are not familiar with programming. For example, all R-codes given by Vignettes and the sampling package could be implemented through a web tool, which make the application of the methods easier, but still requires statistical knowledge and experience.

Conclude. This thesis shows, there is a need for methodological discussion before analysing data (Moseley, 2013). Therefore, a permanent statistician within research teams or institutes, who is integrated in each step of answering scientific questions with empirical studies, is preferable. At the best case, this expert is engaged with the field specific characteristics of the data and has knowledge about a variety of methods. Furthermore, the appropriate preparation for practitioners could also help that such suitable methods were used. Hence, a further recommendation is to train researchers in applied sciences for a good understanding in statistics and how to interpret the results. Furthermore, a good cooperation of scientists and statisticians within interdisciplinary fields still remains very necessary.

Bibliography

Adams, Yenupini Joyce et al. 'Revisiting the Quality of Reporting Randomized Controlled Trials in Nursing Literature'. In: *Journal of Nursing Scholarship* 50.2 (2018), pp. 200–209. ISSN: 1547-5069.

American Association of Colleges of Nursing. *Nursing research - position statement.* Web Page. 2015.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5.* 5th ed. Washington, DC: Autor, 2013.

Balci, Osman. 'Guidelines for Successful Simluation Studies (Tutorial Session)'. In: *Proceedings of the 22Nd Conference on Winter Simulation.* WSC' 90. New Orleans, Louisiana, USA: IEEE Press, 1990, pp. 25–32. ISBN: 0-911801-72-3.

Baser, Onur. 'Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching'. In: *Value in Health* 9.6 (2006), pp. 377 –385. ISSN: 1098-3015.

Beisken, Stephan, Michael Eiden, and M. Reza Salek. 'Getting the right answers: understanding metabolomics challenges'. In: *Expert Review of Molecular Diagnostics* 15.1 (2015), pp. 97–109. eprint: `http://dx.doi.org/10.1586/14737159.2015.974562`.

Belle, Ashwin et al. 'Big Data Analytics in Healthcare'. In: *BioMed Research International* 2015 (2015), p. 16.

Berry, Donald A. *Statistics: A Bayesian Perspective.* An Alexander Kugushev book. Duxbury Press, 1996. ISBN: 9780534234720.

Boccard, Julien and Serge Rudaz. 'Harnessing the complexity of metabolomic data with chemometrics'. In: *Journal of Chemometrics* 28.1 (2014). CEM-13-0118.R1, pp. 1–9. ISSN: 1099-128X.

Boccard, Julien, Jean-Luc Veuthey, and Serge Rudaz. 'Knowledge Discovery in metabolomics: An overview of MS data handling'. In: *Journal of Separation Science* 33 (2010), pp. 290–304.

Bortz, Jürgen. *Statistik: Für Human- und Sozialwissenschaftler.* 6th ed. Heidelberg: Springer Medizin Verlag, 2005. ISBN: 978-3-540-33305-0.

Broadhurst, David I. and Douglas B. Kell. 'Statistical strategies for avoiding false discoveries in metabolomics and related experiments'. In: *Metabolomics* 2.2 (Dec. 2006), pp. 171–196.

Brown, M. et al. 'Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics.' In: *Analyst* 134.7 (2009), pp. 1322–1332.

Burns, Nancy and Susan K. Grove. *The practice of nursing research. Appraisal, synthesis and geeration of evidence.* Vol. 6th edition. St. Louis: Saunders Elsevier, 2009.

Burton, Andrea et al. 'The design of simulation studies in medical statistics'. In: *Statistics in Medicine* 25.24 (2006), pp. 4279–4292. ISSN: 1097-0258.

Cepeda, M. Soledad et al. 'Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders'. In: *American Journal of Epidemiology* 158.3 (2003), pp. 280–287. eprint: `http://aje.oxfordjournals.org/content/158/3/280.full.pdf+html`.

Chicurel, Marina. 'Bioinformatics: Bringing it all together technology feature'. In: *Nature* 419.6908 (Oct. 2002), pp. 751–757. ISSN: 0028-0836.

Davis, Charles. *Statistical Methods for the Analysis of Repeated Measurements.* Springer, 2002.

Díaz-Emparanza, Ignacio. *Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test.* Econometrics. EconWPA, 2000.

DeGroot, Morris H. *Optimal Statistical Decisions.* Wiley Classics Library. Wiley, 2004. ISBN: 9780471680291.

Eichler, Tilly et al. 'Living Alone with Dementia: Prevalence, Correlates and the Utilization of Health and Nursing Care Services'. In: *Journal of Alzheimer's Disease* 52.2 (2016), pp. 619–629. ISSN: 1875-8908.

Eliasson, Mattias et al. 'Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach.' eng. In: *Analytical Chemistry* 84.15 (2012), pp. 6869–6876.

Fiehn, Oliver. 'Metabolomics – the link between genotypes and phenotypes'. In: *Plant Molecular Biology* 48.1 (2002), pp. 155–171. ISSN: 1573-5028.

Fisher, R. A. *The design of experiments.* Hafner Pub. Co., 1966.

Gelman, Andrew and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Analytical Methods for Social Research. New York: Cambridge University Press, 2007. ISBN: 978-0-521-68689-1.

Gentle, James E., Wolfgang Karl Härdle, and Yuichi Mori. 'Handbook of computational statistics: concepts and methods'. In: Springer Science & Business Media, 2012. Chap. 1: How to computational statistics became the backbone of modern data science.

Glymour, Clark. *Methods and Applications of Statistics in the Social and Behavioral Sciences: Causation II.* Ed. by N. Balakrishnan. Methods and Applications of Statistics. Wiley, 2012. ISBN: 9780470405079.

Gowda, G. A. Nagana and Danijel Djukovic. 'Overview of mass spectrometry-based metabolomics: opportunities and challenges'. In: *Mass Spectrometry in Metabolomics: Methods and Protocols* (2014), pp. 3–12.

Greenland, Sander. 'Interpretation and choice of effect measures in epidemiologic analyses'. In: *American Journal of Epidemiology* 125 (1987), pp. 761–768.

Hardin, J. W. and J. M. Hilbe. *Generalized Estimating Equations, Second Edition.* CRC Press, 2012. ISBN: 9781439881149.

Hilgers, Ralf-Dieter, Peter Bauer, and Viktor Scheiber. *Einführung in die Medizinische Statistik.* Springer, 2007.

Ho, P. M., P. N. Peterson, and F. A. Masoudi. 'Evaluating the Evidence: Is There a Rigid Hierarchy?' In: *Circulation* 118.16 (2008), pp. 1675–1684. ISSN: 1524-4539.

Hong, Jun et al. 'Plant Metabolomics: An Indispensable System Biology Tool for Plant Science'. In: *International Journal of Molecular Sciences* 17.6 (2016), p. 767.

Hotellings, Hrold. 'The generalization of Student's ratio'. In: *The Annals of Mathematical Statistics* 2 (1931), pp. 360–378.

Ilakovac, Vesna. 'Statistical hypothesis testing and some pitfalls'. In: *Biochemia Medica* (2009). ISSN: 1846-7482.

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, ed. *Allgemeine Methoden 5.0.* 2017.

Ioannidis, John P. A. et al. 'Increasing value and reducing waste in research design, conduct, and analysis'. In: *The Lancet* 383.9912 (2014), pp. 166–175.

James, G. S. 'Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown'. In: *Biometrika* 41(1/2) (1954), pp. 19–43.

Klapwijk, M. S. et al. 'Characteristics Associated with Quality of Life in Long-Term Care Residents with Dementia: A Cross-Sectional Study'. In: *Dementia and Geriatric Cognitive Disorders* 42.3-4 (2016), pp. 186–197. ISSN: 1420-8008.

Knight, Kenneth L. 'Study/Experimental/Research Design: Much More Than Statistics'. In: *Journal of Athletic Training* 45.1 (2010), pp. 98–100. ISSN: 1938-162X.

Koek, Maud M. et al. 'Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives'. In: *Metabolomics* 7.3 (2011), pp. 307–328. ISSN: 1573-3890.

Kuhl, Carsten et al. 'CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets'. In: *Analytical Chemistry* 84.1 (2012), pp. 283–9. ISSN: 1520-6882 (Electronic) 0003-2700 (Linking).

Li, Baoyue et al. 'Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes'. In: *BMC Medical Research Methodology* 11.1 (2011), p. 77. ISSN: 1471-2288.

Mardia, K. V., J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, 2003.

Millar, R. B. *Maximum likelihood estimation and inference: With examples in R, SAS and ADMB.* Statistics in practice. 2011, 1 online resource (xvi, 357. ISBN: 9780470094822.

Mönchgesang, Susann et al. 'Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent'. In: *International Journal of Molecular Sciences* 17.9 (2016), p. 1565. ISSN: 1422-0067.

Moseley, Hunter N. B. 'Error analysis and propagation in metabolomics data analysis'. In: *Computational and structural biotechnology journal* 4.5 (2013), pp. 1–12.

Murray, Christopher J. L. et al. 'Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010 '. In: *The Lancet* 380.9859 (2013), pp. 2197 –2223. ISSN: 0140-6736.

O'Connor, A. M. 'Interpretation of Odds and Risk Ratios'. In: *J. Vet. Intern. Med.* 27.3 (2013), pp. 600–603. ISSN: 1939-1676.

Palm, R. et al. 'Structural characteristics of specialised living units for people with dementia: a cross-sectional study in German nursing homes'. In: *International Journal of Mental Health Systems* 8.1 (2014), p. 39. ISSN: 1752-4458 (Electronic) 1752-4458 (Linking).

Palm, Rebecca et al. 'Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study'. In: *Journal of the American Medical Directors Association* 17.1 (2016), 91.e9–91.e13. ISSN: 1525-8610.

Pimentel, Samuel D., Frank Yoon, and Luke Keele. 'Variable-ratio matching with fine balance in a study of the Peer Health Exchange'. In: *Statistics in Medicine* 34.30 (2015), pp. 4070–4082. ISSN: 1097-0258.

Polit, D. F. and C. T. Beck. *Nursing Research: Generating and Assessing Evidence for Nursing Practice.* Nursing Research. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008. ISBN: 9780781794688.

— *Nursing Research: Principles and Methods.* Nursing Research: Princ & Practice. Lippincott Williams & Wilkins, 2004. ISBN: 9780781737333.

Polit, Denise F., Cheryl Tatano Beck, and Bernadette P. Hungler. *Lehrbuch Pflegeforschung: Methodik, Beurteilung und Anwendungen.* German. 1., Aufl. Huber, Bern, Apr. 2004. ISBN: 3456839375.

Prince, Martin J. et al. 'The burden of disease in older people and implications for health policy and practice'. In: *The Lancet* 385.9967 (2015), pp. 549–562. ISSN: 0140-6736.

Prince, Martin James et al. *World Alzheimer Report 2015 - The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends.* Alzheimer's Disease International, Aug. 2015.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2016.

Rasch et al. *Quantitative Methoden 1.* Springer, 2010.

Raudenbush, Stephen W. and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods.* SAGE, 2002.

Ridder, Annemieke De and Diana De Graeve. 'Can we account for selection bias? A comparison between bare metal and drug-eluting stents'. In: *Value in Health* 14.1 (2011), pp. 3 –14. ISSN: 1098-3015.

Rosenbaum, P. R. *Observational Studies.* Springer Series in Statistics. Springer, 2002. ISBN: 9780387989679.

Rubin, Donald B. *Matched sampling for causal effects.* English. Formerly CIP. Cambridge : Cambridge University Press, 2006. ISBN: 9780521857628 (hbk.)

Sackett, David L. 'Bias in analytic research'. In: *Journal of Cronic Disease* 32.1 (1979), pp. 51–63. ISSN: 0895-4356.

Sackett, David L. et al. 'Evidence based medicine: what it is and what it isn't'. In: *British Medical Journal* 312.7023 (1996), pp. 71–72. ISSN: 0959-8138. eprint: `http://www.bmj.com/content/312/7023/71`.

safety, WHO International Programm on chemical. *Biomarkers and Risk Assessement: Concepts and Principles.* WHO, 1993.

Saito, Kazuki and Fumio Matsuda. 'Metabolomics for functional genomics, systems biology, and biotechnology'. In: *Annual review of plant biology* 61 (2010), pp. 463–489.

Sampson, J. N. et al. 'Metabolomics in epidemiology: sources of variability in metabolite measurements and implications'. In: *Cancer Epidemiol Biomarkers Prev* 22.4 (2013), pp. 631–40. ISSN: 1538-7755 (Electronic) 1055-9965 (Linking).

Sampson, Joshua N. et al. 'Metabolomics in epidemiology: sources of variability in metabolite measurements and implications.' eng. In: *Cancer Epidemiol Biomarkers Prev* 22.4 (2013), pp. 631–640.

Scott, David W. *Multivariate density estimation: Theory, practice, and visualization.* Second edition. Wiley series in probability and statistics. 2015. ISBN: 9780471697558.

Snijders, Tom A. B. 'Power and Sample Size in Multilevel Linear Models'. In: *Encyclopedia of Statistics in Behavioral Science* 3 (2005), pp. 1570–1573.

Starks, Helene, Paula Diehr, and J. Randall Curtis. 'The Challenge of Selection Bias and Confounding in Palliative Care Research'. In: *Journal of Palliative Medicine* 12.2 (2009), pp. 181–187. ISSN: 1557-7740.

Steuer, Ralf et al. 'A Gentle Guide to the Analysis of Metabolomic Data'. English. In: *Metabolomics.* Ed. by Wolfram Weckwerth. Vol. 358. Humana Press, 2007, pp. 105–126. ISBN: 978-1-58829-561-3.

Stroup, W. W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012. ISBN: 9781439815120.

Stuart, Elizabeth A. 'Matching methods for causal inference: A review and a look forward'. In: *Stat. Sci. Rev. J. Inst. Math. Stat.* 25.1 (2010), pp. 1–21. ISSN: 0883-4237.

Stucke, Kathrin and Meinhard Kieser. 'A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design'. In: *Statistics in Medicine* 31.28 (2012), pp. 3579–3596. ISSN: 1097-0258.

Student. 'The probable error of a mean'. In: *Biometrika* 6 (1908), pp. 1–25.

Tautenhahn, Ralf, Christoph Böttcher, and Steffen Neumann. 'Highly sensitive feature detection for high resolution LC/MS'. In: *BMC Bioinformatics* 9.1 (2008), p. 504. ISSN: 1471-2105.

Thiese, Matthew S. 'Observational and interventional study design types; an overview'. In: *Biochemia Medica* 24.2 (2014), pp. 199–210. ISSN: 1846-7482.

Trutschel, D. et al. 'Methodological approaches in analysing observational data: a practical example on how to address clustering and selection bias'. In: *International Journal of Nursing Studies* (2017).

Trutschel, Diana et al. 'Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data'. In: *Metabolomics* 11 (2015), pp. 851–860.

— 'Joint analysis of dependent features within compound spectra can improve detection of differential features'. In: *Frontiers in Bioengineering and Biotechnology* 3.129 (2015). ISSN: 2296-4185.

Tugizimana, Fidele et al. 'A Conversation on Data Mining Strategies in LC-MS Untargeted Metabolomics: Pre-Processing and Pre-Treatment Steps'. In: (2016).

Weckwerth, Wolfram. 'Metabolomics in systems biology'. In: *Annual review of plant biology* 54.1 (2003), pp. 669–689.

Welch, B. L. 'The generalization of 'Student's' problem when several different population variances are involved'. In: *Biometrica* 34 (1947), pp. 28–35.

Werner, Erwan et al. 'Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends'. In: *Journal of Chromatography B* 871.2 (2008). Hyphenated Techniques for Global Metabolite Profiling, pp. 143 –163. ISSN: 1570-0232.

Wimo, Anders et al. 'The worldwide economic impact of dementia 2010'. In: *Alzheimer's & Dementia* 9.1 (2013), pp. 1–11.

Yi, Lunzhao et al. 'Chemometric methods in data processing of mass spectrometry-based metabolomics: A review'. In: *Analytica Chimica Acta* 914 (2016), pp. 17 –34. ISSN: 0003-2670.

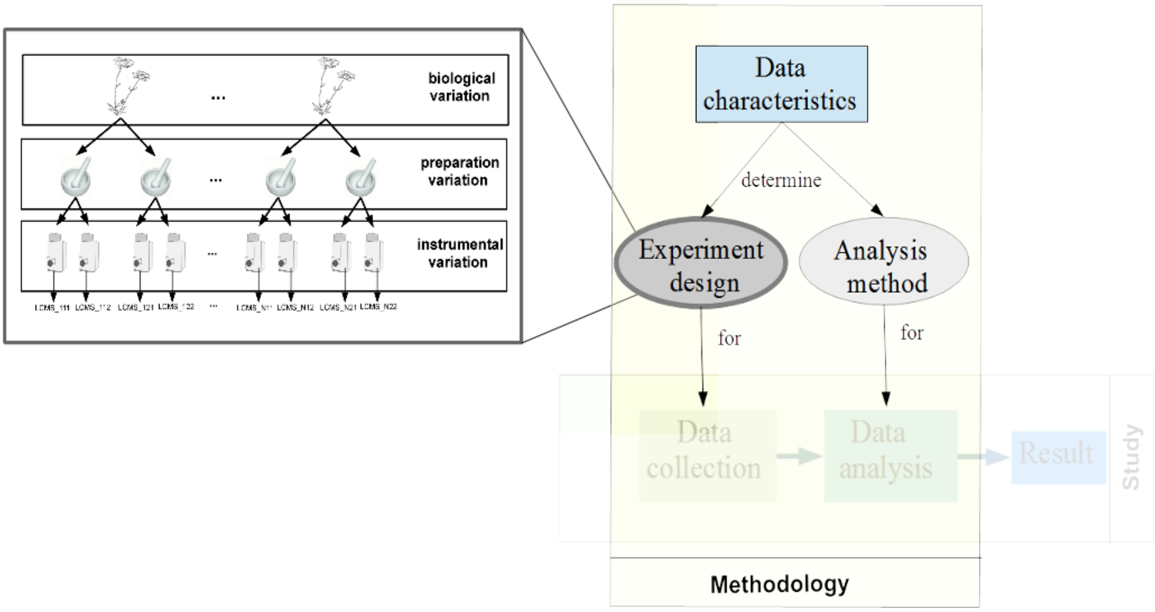
6. Publications

6.1. Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data

Metabolomics, August 2015, Volume 11, Issue 4, pp 851 - 860

Diana Trutschel[1,2],
Stephan Schmidt[1],
Ivo Grosse[2,3] and
Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
[2]Martin-Luther-University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany



Abstract: Univariate hypotheses tests such as Student's t-test or variance analysis (ANOVA) can help to answer a variety of questions in metabolomics data analysis. The statistical power of these tests depends on the setup of the experiment, the experimental design and the analytical variance of the actual observations.

In this paper, we demonstrate how a well-designed pilot study prior to an experiment with the aim to find differences between e.g. several genotypes, can help to determine the variance at multiple levels ranging from biological variance, sample preparation to instrumental variances. Next, we illustrate how these variances can be used to obtain several parameters (e.g. minimum statistically significant effect, number of required replicates and error probabilities) which influence the design of the actual study. In particular, we are going to sketch how technical replicates can improve the performance of a test, when they are correctly used in the statistical analysis, e.g. with a hierarchical model. Finally, we demonstrate the process of evaluating the trade-off between different experimental designs with different replication strategies. The choice of an experimental design beyond the gut feeling can be influenced by factors such as costs, sample availability and the accuracy of of the tests.

We use metabolite profiles of the model plant *Arabidopsis thaliana* measured on an UPLC-ESI/QqTOF-MS as real-world dataset, but the approach is equally applicable to other sample types and measurement methods like NMR based metabolomics.

Availability: The R code and vignette for the calculations presented in this article are available as supplementary material under the GPL license.

Keywords: Metabolomics, statistics, variances, hierarchical experiment design

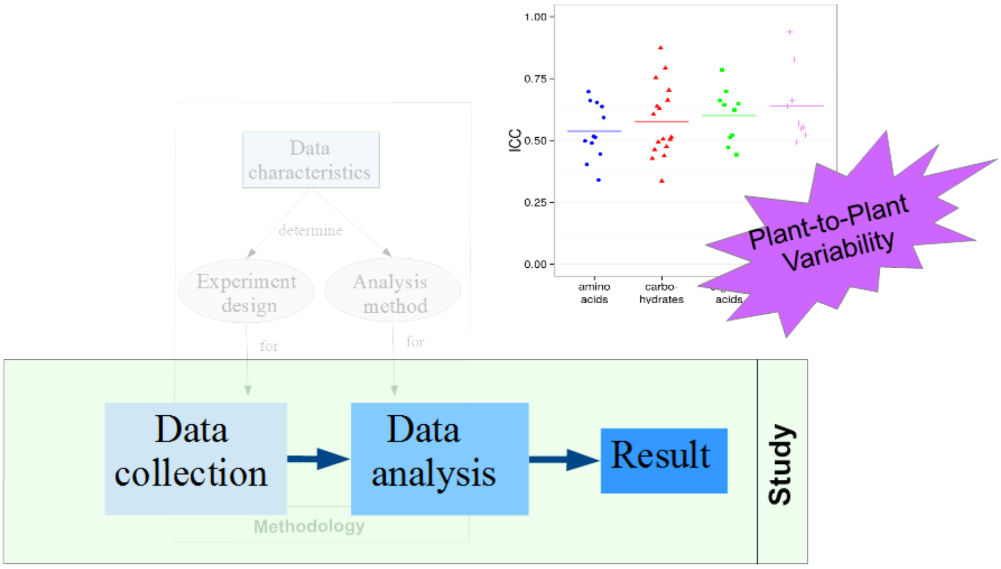
full article on pages 39-47 is available under
`https://link.springer.com/article/10.1007%2Fs11306-014-0742-y`

6.2. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent

International Journal of Molecular Sciences, 17(9), 2016

Susann Mönchgesang[1],
Nadine Strehmel[1],
Diana Trutschel[1,2,3],
Lore Westphal[1],
Steffen Neumann[1] and
Dierk Scheel[1]

[1]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Halle (Saale), Germany
[2]Martin-Luther-University Halle-Wittenberg, Institute of Computer Science, Halle (Saale), Germany
[3]German Center for Neurodegenerative Disesaes, Witten, Germany



International Journal of
Molecular Sciences

Communication

Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent

Susann Mönchgesang [1,*,†], Nadine Strehmel [1,†], Diana Trutschel [1,2,3], Lore Westphal [1], Steffen Neumann [1] and Dierk Scheel [1,*]

[1] Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany; nadine.strehmel@ipb-halle.de (N.S.); diana.trutschel@ipb-halle.de (D.T.); lore.westphal@ipb-halle.de (L.W.); steffen.neumann@ipb-halle.de (S.N.)

[2] Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle (Saale), Germany

[3] German Center for Neurodegenerative Disesaes, Stockumer Straße 12, 58453 Witten, Germany

[*] Correspondence: susann.moenchgesang@ipb-halle.de (S.M.); dierk.scheel@ipb-halle.de (D.S.); Tel.: +49-345-5582-1475 (S.M.); +49-345-5582-1400 (D.S.)

[†] These authors contributed equally to this work.

Academic Editor: Ute Roessner
Received: 30 June 2016; Accepted: 12 September 2016; Published: 16 September 2016

Abstract: Natural variation of secondary metabolism between different accessions of *Arabidopsis thaliana* (*A. thaliana*) has been studied extensively. In this study, we extended the natural variation approach by including biological variability (plant-to-plant variability) and analysed root metabolic patterns as well as their variability between plants and naturally occurring accessions. To screen 19 accessions of *A. thaliana*, comprehensive non-targeted metabolite profiling of single plant root extracts was performed using ultra performance liquid chromatography/electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QTOF-MS) and gas chromatography/electron ionization quadrupole mass spectrometry (GC/EI-QMS). Linear mixed models were applied to dissect the total observed variance. All metabolic profiles pointed towards a larger plant-to-plant variability than natural variation between accessions and variance of experimental batches. Ratios of plant-to-plant to total variability were high and distinct for certain secondary metabolites. None of the investigated accessions displayed a specifically high or low biological variability for these substance classes. This study provides recommendations for future natural variation analyses of glucosinolates, flavonoids, and phenylpropanoids and also reference data for additional substance classes.

Keywords: LC/MS; GC/MS; *Arabidopsis*; secondary metabolism; natural variation; individual variability; metabolite profiling

1. Introduction

Metabolomics is one of the "-omics" disciplines in plant science. With the help of hyphenated techniques such as gas chromatography coupled to mass spectrometry (GC/MS) or liquid chromatography-coupled mass spectrometry (LC/MS), a large spectrum of small molecules within a plant can be analysed. *Arabidopsis thaliana* (*A. thaliana*) is a model species to investigate secondary metabolic pathways. Naturally occurring accessions and their distinct phenotypes have evolved in different habitats and full genome sequencing revealed a substantial number of single nucleotide polymorphisms [1]. Compared to seeds and shoots, root metabolism is not as well investigated, but in plants it is crucial in order to provide the molecular building blocks for physical anchorage in the ground and to regulate all belowground processes. By root exudation, plants also communicate with their surrounding rhizosphere and soil microorganisms. In general, due to the relatively low biomass of *Arabidopsis*, especially in roots, material of several plants is pooled before sample preparation. With increasing sensitivity and decreasing costs of analytical techniques, pooling does not seem to be technically necessary anymore. Indeed, in some cases it is interesting to focus on individual variability to investigate which mechanisms determine plant metabolism without stress exposure. Once the plant material is pooled, the information on individual plants is irreversibly lost. Vice versa, smart experimental design allows for both—investigating variances on different levels (replicates) and detecting differences between accessions.

Several metabolomics studies examined the contribution of different variance sources to the total observed variance [2,3]. For nuclear magnetic resonance (NMR) metabolomics, Lewisetal et al. [2] found that extraction and instrumental deviations accounted for less than 10% and 1%, respectively, of the total variance in leaves of the accession L*er*-0. The substantial plant-to-plant variability of 52% in L*er*-0 could be reduced by pooling several plants to facilitate the separation of L*er*-0 from Col-0 samples. Reducing biological variability by pooling might allow for the fast detection of the effect of interest but nevertheless, it might miss subtle between-plant effects. Similar trends for extraction and instrumental variance were observed in comprehensive LC/MS-based metabolomics studies of Col-0 shoots [3]. Trutschel et al. [3] also provide a solution for how to incorporate different kinds of replicates into a powerful experimental design without the need for sample pooling.

Previous studies have investigated plant-to-plant variability during leaf development. The area of leaf six varied substantially between plants of the isogenic accession Col-0 at the same developmental stage, and this variability seems to converge in mature leaves [4]. Li et al. [5] determined there was 33%–40% plant-to-plant variability between the oil content of Col-0 seeds, and pointed out that this fact needs to be considered to draw statistically valid conclusions.

Plant-to-plant variability has neither been investigated in root metabolism nor have previous studies incorporated more than two *A. thaliana* accessions into a comprehensive root metabolic profiling analysis. Here, we analysed root metabolic profiles of 19 accessions, which were the founders of the multiparent advanced generation inter-cross (MAGIC) collection of *A. thaliana* [1,6], using a single-plant setup in a hydroponic system.

The aim of this study was to decompose the total variance of root metabolite profiles observed in untreated plants into the components attributable to (1) natural variation between accessions; (2) experimental batch; and (3) individual variability between plants. Furthermore, we investigated the relative biological variability of three important substance classes: glucosinolates (GSLs), flavonoids, and phenylpropanoids including oligolignols which seem to play a vital role in root (but not shoot) metabolism. Following the analysis of 19 accessions in their entirety, the variability of each accession was analysed to identify any particular highly or lowly variable accessions.

2. Results

2.1. Variability between Plants Is a Greater Source of Variance than Natural Variation between Accessions

Many studies on natural variation are primarily interested in differences between the accessions, and reduce plant-to-plant variability by pooling material to obtain fast results. However, to obtain a comprehensive picture of variability, the variance at each level of the experimental design should be incorporated.

The experimental setup of our study, shown in Figure 1, resulted in 222 single-plant LC/MS measurements in each electrospray ionization (ESI) mode. The alignment of chromatograms and spectra over 222 samples was performed, deviations in retention time (RT) and mass-to-charge ratio (m/z) were small across all samples (Figure S1) reflecting a sufficient quality of the measurements to analyse the effects of accession, experimental batch, and individual plant. Linear mixed models with all experimental levels as random effects were applied to decompose the total metabolic variance.

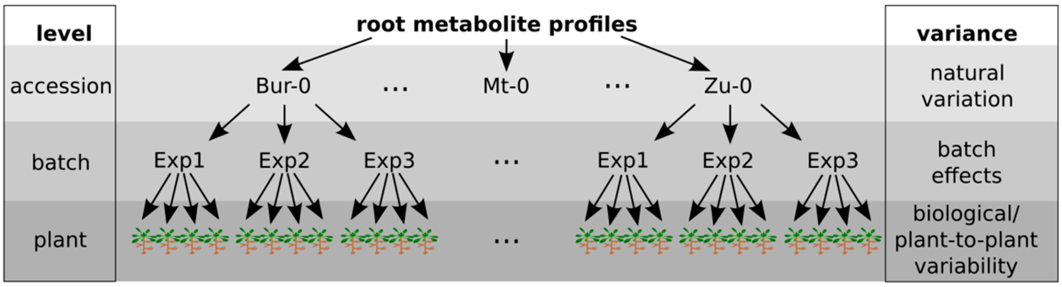


Figure 1. Nested experimental design with three levels. Each variance level had multiple replicates—to assess natural variation, 19 accessions of *Arabidopsis thaliana* (*A. thaliana*) were grown. Three independent biological experiments were performed to estimate non-biological variance derived from the experimental batch. To assess individual variability, four plants were harvested in each biological experiment for each accession. Single-plant root extracts were subjected to liquid chromatography-coupled mass spectrometry (LC/MS) and gas chromatography-coupled mass spectrometry (GC/MS) analysis.

The non-targeted metabolic profiles of the 19 accessions indicated that the between-accession variance is smaller than the plant-to-plant-variability over all features. The results for ESI(−) are shown in Figure 2a and for ESI(+) in Supplementary Figure S2.

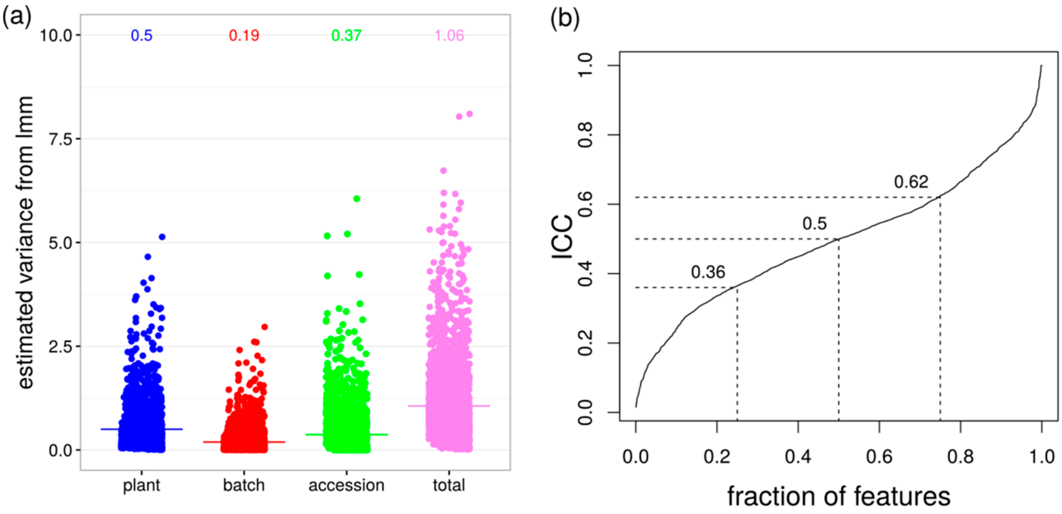


Figure 2. Variance decomposition of LC/electrospray ionization (ESI)(−) MS data set. (**a**) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot—variance of one feature, bar and number—mean variance over 2730 features; (**b**) cumulative intraclass correlation (ICC) distribution for all features ($\sigma^2_{plant}/\sigma^2_{total}$), dotted lines indicate 25%, 50% and 75% quantiles.

The mean between-plant variance $\sigma^2_{plant} = 0.50$ is 20% larger than the between-accession variance $\sigma^2_{accession} = 0.37$. The estimated mean between-experiment variation $\sigma^2_{batch} = 0.19$ is less than 40% of σ^2_{plant}. On average, plant-to-plant variability contributes to approximately half of the total variance ($\sigma^2_{plant}/\sigma^2_{total} = 0.47$). However, this biological variance has to be interpreted in the context of the total variance for comparisons across features and platforms, i.e., knowing whether the feature with the highest σ^2_{plant} also exhibits large σ^2_{total}. It may also occur that a feature with high σ^2_{plant} has low σ^2_{total}, which determines the experimental design to include more replicates on the plant level in a potential validation study.

The intraclass correlation (ICC) according to Sampson et al. [7], here $\sigma^2_{plant}/\sigma^2_{total}$, reflects which fraction of total variance is attributable to the single plant and thus, a relative biological variability.

The mean ICC ≈ 0.5 of a data set could either be representative for the majority of features (narrow interquartile range) or only for a few features if the interquartile range is broad. Figure 2b shows the cumulative ICC distribution over all features, with the fraction of features (x-axis) in increasing ICC (y-axis) order. The distribution revealed that 25%, 50%, and 75% of all these features had an ICC up to 0.36, 0.50, and 0.62. This implies that for half of the features, the plant-to-plant variability contributes to less than 50% to the total variance, and for the other half this variance level explains more than 50% of the total variance. In summary, in our non-targeted analysis of root metabolic natural variation, plant-to-plant variability seems to be larger than between-accession variance. If a broad range of metabolites are of interest, it is important to know the biological variability that is exhibited by most metabolites. If only a small subset of the non-targeted analysis is in research focus, it will be sufficient to deal with the biological variability of a certain substance class.

2.2. Plant-to-Plant Variability in Secondary Metabolism Is Substance-Class-Dependent, but Not Accession-Specific

A difficulty in non-targeted metabolomics is the assignment of the measured features to metabolites and their potential role in pathways in a living system. To facilitate the interpretation of plant-to-plant variability, three sets of annotatable compounds were quantified by integrating peak areas of the extracted ion chromatograms and analysed for their variances at each level (Table S1). In Figure 3, GSLs, flavonoids, and phenylpropanoids are indicated by circles, triangles, and squares, respectively. GSLs were the substance class with the highest plant-to-plant variability (σ^2_{plant} = 3.16, Figure 3a left, circles) compared to flavonoids and phenylpropanoids. They also showed a large deviation of the single metabolite plant variance from the mean of the substance class. Similarly, σ^2_{total} = 5.03 was highest for GSLs in the comparison to flavonoids (σ^2_{plant} = 1.63, σ^2_{total} = 2.60) and phenylpropanoids (σ^2_{plant} = 1.24, σ^2_{total} = 2.88).

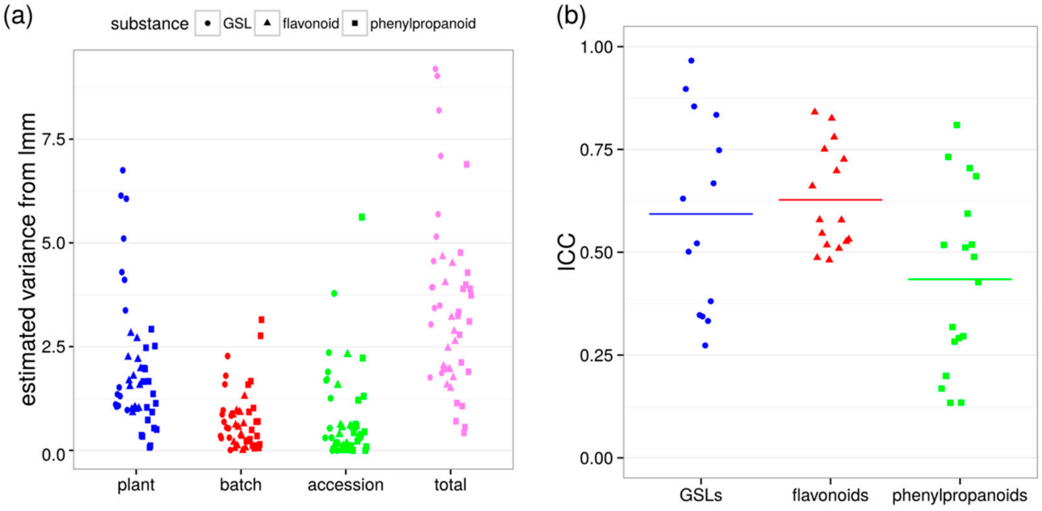


Figure 3. Biological variability of annotated secondary metabolites. (**a**) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot—variance of one metabolite; (**b**) ICCs for glucosinolates (GSLs), flavonoids, and phenylpropanoids, dot—ICC of one metabolite, bar—mean ICC for substance class.

With the current experimental setup of four plants in three batches for a total of 12 plants per accession, the minimal detectable log fold-change to distinguish between two accessions is 3.94, 2.97 and 3.24 for glucosinolates, flavonoids, and phenylpropanoids, respectively, with a power of 0.8 and a significance level of 0.05. However, plant-to-plant variability needs to be interpreted in the context of total variance to find out at which experimental level the main observation is made.

If $\sigma^2_{\text{plant}} \approx \sigma^2_{\text{total}}$, nearly all of the total variance would be caused by plant-to-plant variability and a large number of plants would be required to analyse effects beyond this experimental level, i.e., between accessions. If $\sigma^2_{\text{plant}}/\sigma^2_{\text{total}} \approx 0$, it would be sufficient to use one plant per accession. Glucosinolates and phenylpropanoids show a large range of ICCs. For flavonoid metabolites, the ICCs are rather high but similar for all analysed members of the substance class (Figure 3b). Hence, calculations with the mean ICCs like above will provide sufficient power for analyses of flavonoids, but not for all metabolites of the classes glucosinolates and phenylpropanoids.

A set of primary metabolites was also analysed for their plant-to-plant variability (Table S2) but, in comparison to secondary metabolism, the ICC distributions of carbohydrates, organic acids, amino acids, and phosphates covered a large range (Figure S3). As expected, the primary metabolism is more stable than secondary metabolism, the latter showing substance-class specific ICC distributions.

Until here, we assumed all accessions to have equal variances at the plant and batch level. In addition, we analysed if the accessions differ with regard to their plant-to-plant variability. For this purpose, linear mixed models were applied to estimate the variances of secondary metabolites for each accession separately. As shown in Figure S4, there are no clear highly and lowly variable accessions across the measured substance classes. However, Edi-0 showed relatively low ICCs for GSLs and flavonoids. Hi-0 and Sf-2 showed higher ICCs for all three compound classes.

In our analysis, taking the ICCs of secondary metabolite classes into consideration seems to be more important than the selection of accessions.

3. Discussion

Our study investigated natural variation and plant-to-plant variability of 19 key accessions in a comprehensive metabolite profiling approach. Measuring single plant extracts prevented the irreversible information loss resulting from pooling plant material and allows to distinguish between accessions and still analyse plant-to-plant variability. Environmental variation was kept to a minimum by a randomized growth regimen and selecting plants with approximately the same vigor for analyses. Both non-targeted LC/MS ionization modes indicated a higher plant-to-plant variability than natural variation between accessions and variance due to experimental batches. Plant-to-plant variability contributed to 47%–50% of the total variance, which is higher than previously reported for one particular compound class in seeds of one accession [5]. As our total variance was the sum of plant, batch and accession variance, the ICCs referring to the sum of plant and batch variance, like in the oil seed study [5], would have been larger.

Furthermore, we chose a range of secondary and primary metabolite classes for more specific analyses. Both data sets indicated that the plant-to-plant variability had the greatest contribution to the total variance of these metabolite classes. For GSLs, flavonoids and phenylpropanoids, the means of σ^2_{batch} and $\sigma^2_{\text{accession}}$ were in the same order of magnitude, whereas for primary metabolite sets $\sigma^2_{\text{accession}}$ was less pronounced with values one order of magnitude below σ^2_{batch}. The minimal detectable effects were quite large and impractical with the given experimental setup of three experiments with four plants each. Possible combinations of biological and technical replicates to reliably detect a smaller effect can be calculated with the implementation provided by Trutschel et al. [3]. All annotated substance classes displayed higher mean ICCs than the non-targeted data sets they were derived from. The higher the fraction of features with high ICCs, the higher the number of plants that is required to maintain the power in a statistical analysis. This should be taken into consideration for future experimental designs. Flavonoid metabolites have similar ICCs within their substance class and therefore, calculation with mean ICC of the substance class will be sufficient to obtain reliable results for most metabolites in this class. Contrarily, GSLs and phenylpropanoids displayed a large ICC spread and require a substance-specific estimation of variance prior to future analyses. A previous study of root exudates has demonstrated that there are substance-specific differences in some metabolite classes due to alterations in the biosynthetic pathways [8]. Since some metabolites are specifically induced during stress response, they might not have been expressed in the unperturbed physiological state that was the focus of this study. The analysis of plant-to-plant variability in each accession revealed that ICC distributions are not distinct for any of the 19 accessions with the few exceptions of Edi-0, Hi-0, and Sf-2. However, our set of 19 accessions is too small to draw a general conclusion about accession-specific plant-to-plant variability and more accessions have to be analysed in future.

There are hints that biological variability converges after development [4] and upon exposure to stress factors [9,10]. A study of *Arabidopsis* plants exposed to a biotic stress factor, namely the endophytic fungus *Piriformospora indica*, showed substantial metabolic variability in untreated control samples and only a small spread of co-cultivated samples in principal component analyses. These samples were no single plant measurements but the batch variances in both sample classes were identical and thus, the observed deviation is expected to result from plant-to-plant variability [9]. Töpfer et al. [10] found that upon abiotic stress treatment, certain metabolites were robust in their abundance from plant to plant and displayed low coefficients of variation, whereas other metabolites showed larger plant-to-plant variability.

For future natural variation studies, it might be worth considering measuring single plants and make the data available for further analyses answering research questions on a different experimental level. We have provided estimated variances for selected substances in Supplementary Tables S1 and S2. Furthermore, we provide exemplary data and the functions in an R script for variance estimation in the Supplementary Folder S1 as well as data for additional substance classes in the targeted analysis in MTBLS338 in the MetaboLights repository. This knowledge can be exploited to appropriately design an experiment prior to its conduction because it may differ between a non-targeted screen and the analysis of specific substance classes.

4. Materials and Methods

4.1. Plant Cultivation

The *A. thaliana* accessions Bur-0, Can-0, Col-0, Ct-1, Edi-0, Hi-0, Kn-0, L*er*-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 were obtained as seeds from the European Arabidopsis Stock Centre (Nottingham, UK) and surface sterilized prior to plant cultivation. All accessions were cultivated in a hydroponic system under 8 h light and 22 °C as described previously [11] and in the protocol section of MTBLS338 with four plants in each of the three independent biological experiments. All samples were rotated in the growth chamber to minimize position effects. Primary root length and root fresh weight are given in MTBLS338. Out of 228 root samples, 210 and 222 from individual plants could be used for the GC/MS and LC/MS analysis, respectively.

4.2. Liquid Chromatography/Mass Spectrometry (LC/MS)

For LC/MS analysis, 40 mg root material were extracted in 200 µL 80% methanol/water (*v*/*v*) twice according to Böttcher et al. [12] and reconstituted in 30% methanol (*v*/*v*) containing 5 µM 2,4-dichlorophenoxyacetic acid as an internal standard. Upon full loop injection into an Acquitiy UPLC system (Waters, Eschborn/Germany) mounted with a HSS T3 column (100 × 1.0 mm, 1.8 µM particle size), samples were separated at a flow rate of 150 µL/min with mixtures of A (water/0.1% formic acid) and B (acetonitrile/0.1% formic acid) with a 20 min gradient: 0–1 min isocratic 95% A, 5% B; 1–16 min linear 5%–95% B; 16–18 min isocratic 95% B; 18–18.01 min linear 95%–5% B; 18.01–20 min isocratic 5% B. Eluates were ionized using an Apollo II source (Bruker Daltonics, Billerica, MA, USA) into a MicroTOF-Q I hybrid quadrupole time-of-flight mass analyzer (Bruker Daltonics) in both ionization modes with a mass range *m*/*z* 80–1000. Mass spectrometry settings were applied as previously described [11] and in the protocol section of MTBLS338.

All LC/MS runs were acquired as centroid spectra and recalibrated with lithium formate cluster ions for each measurement. Vendor .d file formats were converted into the open standard mzData with CompassXPort (Bruker Daltonics, Billerica, MA, USA).

4.3. Gas Chromatography/Mass Spectrometry (GC/MS)

For GC/MS analysis, 40 μL of the root extract were vacuum-evaporated and subjected to a derivatization with (1) methoxyamine hydrochloride and (2) *N*,*O*-bis(trimethylsilyl)-trifluoroacetamide as previously described [13]. Derivatized samples were injected in a splitless manner into a split/splitless inlet of an Agilent 6890N GC and a ZB-5 column (30 m × 0.25 mm, 0.25 m 95% dimethyl/5% diphenyl polysiloxane film, 10 m integrated guard column, Phenomenex, Aschaffenburg, Germany) at 230 °C. An Agilent 5975 Series Mass Selective Detector (Agilent Technologies, Waldbronn, Germany) was used to detect eluting compounds from m/z 70 to 600. Vendor file format conversion and baseline correction was performed by MetAlign [14].

4.4. Data Analysis

Statistical analysis was performed using R version 3.2.0 and the Bioconductor environment [15,16]. Functions are available as an R script in the Supplementary Folder S1.

4.4.1. Raw Data Processing

All LC/MS data analysis was performed with the R packages XCMS and CAMERA [17–19]. Features were extracted with centWave (snthr = 10, ppm = 20, peakwidth = c(5,12), scanrange = c(1,3600)) and grouped (minfrac = 0.75, bw = 5, mzwid = 0.05), corrected for retention shifts and re-grouped with smaller bandwidth (bw = 2). Missing values were imputed by integration of raw data (fillPeaks) and with random numbers around the minimal intensity value across the samples.

Baseline-corrected GC/MS tags with intensities above 500 peak height were subsequently processed with TagFinder [20] and mass spectral features were grouped according to their common retention time. Clusters with at least 3 correlating tags were extracted and identified according to matching the Golm Metabolome Database [21]. In GC/MS, 15,539 tags were detected and 98 metabolites were annotated (Table S3).

All data were log-transformed to approximate a normal distribution for further statistics.

4.4.2. Targeted LC/MS Analysis

For the targeted analysis, DataAnalysis 4.2 (Bruker Daltonics, Billerica, MA, USA) was used to extract ion chromatograms, deconvolute mass spectra and determine the elemental composition. Peak areas (minimum peak area = 500) of extracted ion chromatograms were integrated with QuantAnalysis 2.0 (Bruker Daltonics, Billerica, MA, USA) to quantify compound abundances with quasi-molecular ions as listed in Table S4 [11,22]. In the LC/MS measurements, 3305 peaks ESI(+) and 2730 peaks ESI(−) were detected and all together 139 compounds could be annotated.

4.4.3. Variance Estimation with Linear Mixed Models

A linear mixed model (R package lme4, version 1.1-11, [23]) with accession, batch and plant as random effects was applied to log-transformed metabolite abundances to estimate variance contribution of each experimental level assuming equal variances for each accession. Linear mixed models with batch and plant as random effects were applied separately to each accession to examine accession-specific variances. Intraclass correlations (ICCs) were calculated as the ratio of σ^2_{plant} and σ^2_{total} according to Sampson et al. [7] and plotted as a cumulative distribution. Further analysis was constrained to known metabolites to allow for a better interpretation. The minimal detectable effect sizes were estimated with the power calculations for multilevel experiments [3].

4.5. Data Availability

All data sets including the targeted analyses are available from the MetaboLights repository under the accession number MTBLS338 [24].

5. Conclusions

This study investigated the variability in root metabolite profiles of 19 *A. thaliana* accessions. It revealed that plant-to-plant variability can be a substantial component of the overall variability in a natural variation analysis. Additionally, several selected substance classes were characterized by differing intraclass correlations. To exploit the full potential of a non-targeted metabolite profiling, single-plant measurements should be acquired and correctly integrated into the analysis. Hence, different substance classes of interest might require a customised experimental set-up.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/17/9/1565/s1.

Acknowledgments: The authors thank Christoph Böttcher and Stephan Schmidt for metabolomics advice. Expert technical assistance by Sylvia Krüger, Julia Göhricke, Siska Herklotz and Jessica Thomas is gratefully acknowledged.

Author Contributions: Dierk Scheel conceived the study. Dierk Scheel, Steffen Neumann, Lore Westphal and Nadine Strehmel supervised the study. Nadine Strehmel and Susann Mönchgesang designed and performed the experiments and analysed the data, Diana Trutschel performed statistical analysis. Susann Mönchgesang wrote the manuscript. Lore Westphal provided advice on natural variation and critical manuscript feedback. Steffen Neumann submitted the data to MetaboLights. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

EI	electron ionization
ESI	electrospray ionization
GC/MS	gas chromatography/mass spectrometry
GSL	glucosinolate
ICC	intraclass correlation
LC/MS	liquid chromatography/mass spectrometry

References

1. Gan, X.; Stegle, O.; Behr, J.; Steffen, J.G.; Drewe, P.; Hildebrand, K.L.; Lyngsoe, R.; Schultheiss, S.J.; Osborne, E.J.; Sreedharan, V.T.; et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **2011**, *477*, 419–423. [CrossRef] [PubMed]
2. Lewis, J.; Baker, J.M.; Beale, M.H.; Ward, J.L. Metabolite Profiling of GM Plants—The importance of robust experimental design and execution. In *Genomics for Biosafety in Plant Biotechnology*; Nap, J.-P., Atanassov, A., Stiekema, W.J., Eds.; IOS Press: Amsterdam, The Netherlands, 2004.
3. Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. Experiment design beyond gut feeling: Statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics* **2015**, *11*, 851–860. [CrossRef]
4. Granier, C.; Massonnet, C.; Turc, O.; Muller, B.; Chenu, K.; Tardieu, F. Individual leaf development in *Arabidopsis thaliana*: A stable thermal-time-based programme. *Ann. Bot.* **2002**, *89*, 595–604. [CrossRef] [PubMed]
5. Li, Y.; Beisson, F.; Pollard, M.; Ohlrogge, J. Oil content of *Arabidopsis* seeds: The influence of seed anatomy, light and plant-to-plant variation. *Phytochemistry* **2006**, *67*, 904–915. [CrossRef] [PubMed]
6. Kover, P.X.; Valdar, W.; Trakalo, J.; Scarcelli, N.; Ehrenreich, I.M.; Purugganan, M.D.; Durrant, C.; Mott, R. A Multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **2009**, *5*, e1000551. [CrossRef] [PubMed]
7. Sampson, J.N.; Boca, S.M.; Shu, X.O.; Stolzenberg-Solomon, R.Z.; Matthews, C.E.; Hsing, A.W.; Tan, Y.T.; Ji, B.T.; Chow, W.H.; Cai, Q.; et al. Metabolomics in epidemiology: Sources of variability in metabolite measurements and implications. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 631–640. [CrossRef] [PubMed]
8. Mönchgesang, S.; Strehmel, N.; Schmidt, S.; Westphal, L.; Taruttis, F.; Muller, E.; Herklotz, S.; Neumann, S.; Scheel, D. Natural variation of root exudates in *Arabidopsis thaliana*—Linking metabolomic and genomic data. *Sci. Rep.* **2016**, *6*, 29033. [CrossRef] [PubMed]
9. Strehmel, N.; Mönchgesang, S.; Herklotz, S.; Kruger, S.; Ziegler, J.; Scheel, D. *Piriformospora indica* Stimulates Root Metabolism of *Arabidopsis thaliana*. *Int. J. Mol. Sci.* **2016**, *17*. [CrossRef] [PubMed]

10. Töpfer, N.; Scossa, F.; Fernie, A.; Nikoloski, Z. Variability of metabolite levels is linked to differential metabolic pathways in *Arabidopsis*'s responses to abiotic stresses. *PLoS Comput. Biol.* **2014**, *10*, e1003656. [CrossRef] [PubMed]
11. Strehmel, N.; Böttcher, C.; Schmidt, S.; Scheel, D. Profiling of secondary metabolites in root exudates of *Arabidopsis thaliana*. *Phytochemistry* **2014**, *108*, 35–46. [CrossRef] [PubMed]
12. Böttcher, C.; Westphal, L.; Schmotz, C.; Prade, E.; Scheel, D.; Glawischnig, E. The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of *Arabidopsis thaliana*. *Plant Cell* **2009**, *21*, 1830–1845. [CrossRef] [PubMed]
13. Buhtz, A.; Witzel, K.; Strehmel, N.; Ziegler, J.; Abel, S.; Grosch, R. Perturbations in the Primary Metabolism of Tomato and *Arabidopsis thaliana* Plants Infected with the Soil-Borne Fungus *Verticillium dahliae*. *PLoS ONE* **2015**, *10*, e0138242. [CrossRef] [PubMed]
14. Lommen, A. MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* **2009**, *81*, 3079–3086. [CrossRef] [PubMed]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
16. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80. [CrossRef] [PubMed]
17. Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*, 504. [CrossRef] [PubMed]
18. 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. [CrossRef] [PubMed]
19. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289. [CrossRef] [PubMed]
20. Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography—Mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* **2008**, *24*, 732–737. [CrossRef] [PubMed]
21. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; et al. GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics* **2005**, *21*, 1635–1638. [CrossRef] [PubMed]
22. Lassowskat, I.; Böttcher, C.; Eschen-Lippold, L.; Scheel, D.; Lee, J. Sustained mitogen-activated protein kinase activation reprograms defense metabolism and phosphoprotein profile in *Arabidopsis thaliana*. *Front. Plant Sci.* **2014**, *5*, 554. [CrossRef] [PubMed]
23. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using LME4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
24. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class Dependent. Available online: http://www.ebi.ac.uk/metabolights/MTBLS338 (accessed on 13 September 2016).

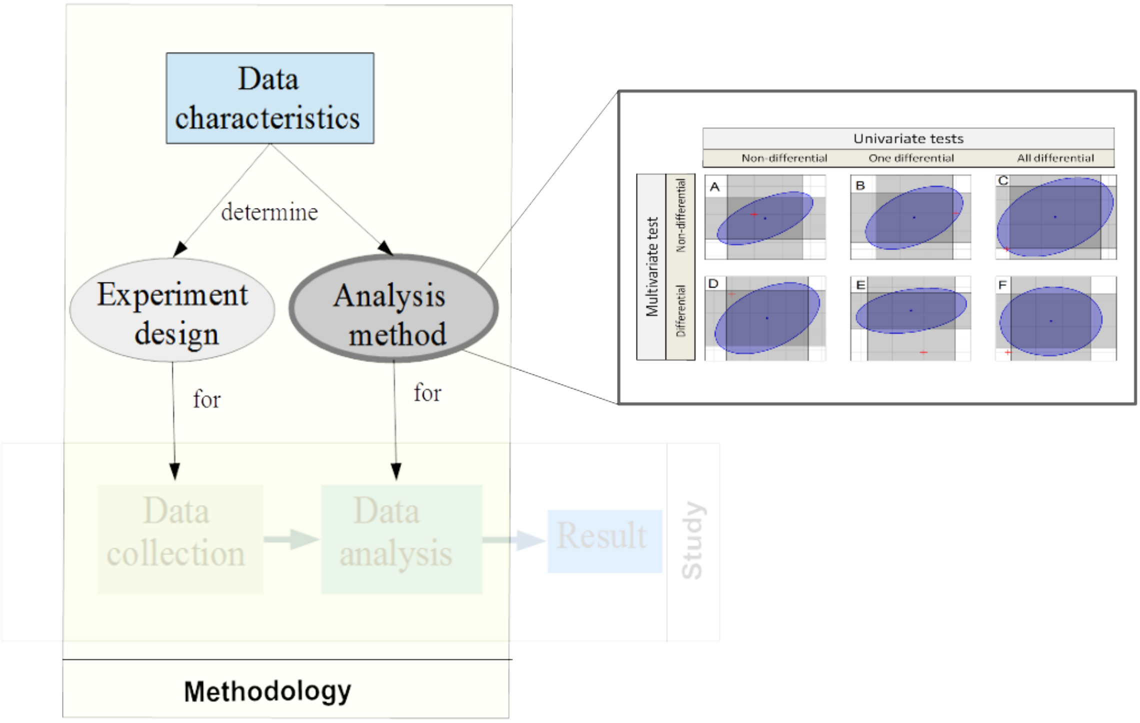
© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

6.3. Joint analysis of dependent features within compound spectra can improve detection of differential features

Front Bioeng Biotechnol., 2015 Sep 24, pp 3:129, eCollection 2015

Diana Trutschel[1,2],
Stephan Schmidt[1],
Ivo Grosse[2,3] and
Steffen Neumann[1]

[1]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Halle, Germany
[2] Martin-Luther-University Halle-Wittenberg, Institute of Computer Science, Halle, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany



frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 23 September 2015
doi: 10.3389/fbioe.2015.00129

Joint analysis of dependent features within compound spectra can improve detection of differential features

Diana Trutschel[1,2], *Stephan Schmidt*[1], *Ivo Grosse*[2,3] *and Steffen Neumann*[1]*

[1] Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany, [2] Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany, [3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

OPEN ACCESS

Edited by:
Reza M. Salek,
EMBL European Bioinformatics Institute, UK

Reviewed by:
Warwick Dunn,
University of Manchester, UK
Ron Wehrens,
Wageningen University and Research Centre, Netherlands

***Correspondence:**
Steffen Neumann,
Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, Halle 06120, Germany
sneumann@ipb-halle.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *03 April 2015*
Accepted: *13 August 2015*
Published: *23 September 2015*

Citation:
Trutschel D, Schmidt S, Grosse I and Neumann S (2015) Joint analysis of dependent features within compound spectra can improve detection of differential features. Front. Bioeng. Biotechnol. 3:129. doi: 10.3389/fbioe.2015.00129

Mass spectrometry is an important analytical technology in metabolomics. After the initial feature detection and alignment steps, the raw data processing results in a high-dimensional data matrix of mass spectral features, which is then subjected to further statistical analysis. Univariate tests like Student's *t*-test and Analysis of Variances (ANOVA) are hypothesis tests, which aim to detect differences between two or more sample classes, e.g., wildtype-mutant or between different doses of treatments. In both cases, one of the underlying assumptions is the independence between metabolic features. However, in mass spectrometry, a single metabolite usually gives rise to several mass spectral features, which are observed together and show a common behavior. This paper suggests to group the related features of metabolites with CAMERA into compound spectra, and then to use a multivariate statistical method to test whether a compound spectrum (and thus the actual metabolite) is differential between two sample classes. The multivariate method is first demonstrated with an analysis between wild-type and an over-expression line of the model plant *Arabidopsis thaliana*. For a quantitative evaluation data sets with a simulated known effect between two sample classes were analyzed. The spectra-wise analysis showed better detection results for all simulated effects.

Keywords: metabolomics, statistics, hypothesis tests, multivariate analysis, mass spectrometry

1. Introduction

Mass spectrometry is an important analytical technology in metabolomics. XCMS (Smith et al., 2006) is one of the available tools for processing mass spectrometry data. After the initial feature detection and alignment steps, the raw data processing results in a high-dimensional data matrix of mass spectral features as shown in **Table 1**, which is then subjected to further (statistical) analysis.

A typical question in metabolomics is biomarker discovery, where e.g., univariate hypothesis tests like Student's *t*-test (Student, 1908) and Analysis of Variances (ANOVA) can be used to detect differences between two or more sample classes, e.g., wildtype versus mutant or disease versus control. An example implementation is the `diffreport()` function in XCMS. Furthermore, some statistical methods can deal with more complex experimental designs with dependencies between samples (Davis, 2002; Sampson et al., 2013; Trutschel et al., 2015). But in all cases, one of the underlying assumptions is the independence between individual metabolic *features*.

TABLE 1 | A peak list of features of a two sample class MS experiment with feature group annotation mz is the mass-to-charge ratio, RT is the retention time in seconds.

mz/RT	MU 1	MU 2	...	MU 6	MU 7	WT 1	WT 2	...	WT 6	WT 7	p.uni	group.anno	p.multi
590.5/967	14.42	14.61	...	14.29	14.2	13.85	13.96	...	13.95	14.12	0.02	40	0.02
609.5/968	18.31	18.72	...	18.32	18.45	18.12	18.7	...	18.44	18.48	0.88	40	
628.5/968	17.21	17.52	...	17.17	17.21	16.95	17.49	...	17.18	17.34	0.89	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	
413.3/1106	14.92	13.23	...	14.72	14.57	14.52	14.92	...	14.52	14.27	0.65	82	0.30
538.5/1103	12.32	11.76	...	11.93	11.8	11.7	11.7	...	12.15	12.91	0.23	82	
591.5/1101	15.51	15.2	...	15.36	15.06	15.72	15.78	...	15.07	15.74	0.02	82	
592.5/1102	15.15	14.78	...	14.78	14.42	14.67	15.03	...	14.76	15.33	0.34	82	
797.5/1104	18.28	17.96	...	17.72	17.58	17.83	18.42	...	17.2	17.91	0.15	82	

Additionally, listed uni- and multivariate p-values results from univariate and multivariate tests.

However, in mass spectrometry, a single metabolite usually gives rise to several mass spectral features, e.g., isotopes, adducts, or fragments (Brown et al., 2009), which observed together and show a common behavior across samples. Another issue is that the redundant features aggravate the problem of multiple testing, and cause more type I errors (Broadhurst and Kell, 2006; Hendriks et al., 2011).

A first step to treat related features together is to group those, which originate from the same metabolite into compound spectra. Several methods for such a grouping have been developed in the last years (Ipsen et al., 2010; Alonso et al., 2011; Brown et al., 2011; Scheltema et al., 2011; Varghese et al., 2012; Kenar et al., 2014). In this paper, the grouping algorithm in the Bioconductor package CAMERA (Kuhl et al., 2012) is used, which is comprised of several steps, including compound spectra creation based on retention time, calculation of known mass differences for isotope pattern and adduct detection and a peak shape correlation analysis. This grouping then results in *compound spectra*, which contain one or more related features, which originate from the same metabolite.

A typical approach for the statistical analysis in GC/MS is to select a single *quantification ion* for each compound (Luedemann et al., 2008) for univariate tests, ignoring intensity information for the remaining mass features in a compound spectrum. On the other hand, multivariate methods like MANOVA are global approaches and analyze all features together and can take correlations into account. This has already been used in metabolomics (Steuer et al., 2007; Saccenti et al., 2014). With MANOVA, the simultaneous analysis of variables results in a better Type I error correction because of the multidimensional confidence region. In more detail, the differences in the mathematical theory between univariate and the multivariaten comparison for more than two groups (ANOVA versus MANOVA) are described in (Legendre and Anderson, 1999). The multivariate approach benefits from small signals, which contribute to the class differences, but would not be detected univariate because the effect is too small compared to the variance. However, the interpretation, which metabolites have changed, remains challenging.

Often, in metabolomics, the number of samples is much smaller than the number of features to be analyzed. Therefore, correlation and covariance structure is difficult to estimate, and requires an initial variable selection step. Often, the complex models used by global multivariate analysis are prone to the problem of overfitting with poor prediction and generalization.

In this paper, we compare the detection of differential features on the individual- and metabolites on the compound spectra level. We also introduce a multivariate analysis on the level of compound spectra instead of a global multivariate approach to determine differential metabolites, combining the benefits of uni- and multivariate analysis for biomarker detection. An advanced version of the XCMS `diffreport()` function is provided for users. This paper is structured as follows: in the next section, the metabolomics data used in this paper is briefly described, followed by the conceptual details of the statistical method. The method is applied to data from wild-type and over-expression plants. Finally, the performance of the proposed methods is compared to the univariate approach on a data set of known (simulated) effects. The implementation is provided as an R vignette in the Supplementary Material under the GPL license.

2. Materials and Methods

For the experiments, two metabolomics data sets from *Arabidopsis thaliana* (*A. th.*) were used. The first is a subset of the study available as MTBLS74, where 26 independent plant profiles and a simulated effect were used. The method is then demonstrated on a dataset of *A. th.* wildtype and a mutant line, available as MTBLS169.

2.1. Metabolite Profiling of *Arabidopsis thaliana*

2.1.1. Plant Growth and Sample Preparation

The model plant *Arabidopsis thaliana* Col-0 was used as plant material. For the genotype comparison Col-0 and the 90.32 mutant were used, a transposon-based activation tagged *A. th.* line from the TAMARA population (Schneider et al., 2005). This particular mutant has an over-expression of the AT5G55880 – AT5G55890 genetic region with unknown function. Plants were grown on soil in a growth chamber under controlled conditions as biological replicates. The frozen leaf material of each plant was ground and weighed using a cryogenics robot[1] with a weighing error $\leq$5%, and extracted with methanol. Full details are available in Supplementary Material I, Section 1 and the protocol sections of the MetaboLights studies.

[1] http://www.labman.co.uk/portfolio-type/ipb-cryogenic-grinder-and-feeder-system

2.1.2. Mass Spectrometry Analysis and Data Processing

Metabolite intensities were recorded according to (Böttcher et al., 2009). In brief, the chromatographic separation was performed on a Waters Acquity UPLC system coupled to a Bruker micrOTOF-Q mass spectrometer. Mass spectra were recorded in positive ion centroid mode with a scan rate of 3 Hz and a mass range of 100–1000 m/z. Full details are available in Supplementary Material I, Section 1 and the protocol sections of the MetaboLights studies. This experimental setup is able to routinely detect semi-polar plant metabolites from major biosynthetic classes including glucosinolates, indolic compounds, phenylpropanoids, benzenoids, flavonoids, terpenes, and fatty acid derivatives (Böttcher et al., 2011). In this paper, no metabolite identification was performed, resulting in the lowest metabolomics standards initiative (MSI) identification level (Sumner et al., 2007) MSI level four (i.e., the features are only characterized by their mass and retention time).

The measured MS data were converted to mzData with the Bruker CompassXport software. The mzData are preprocessed with the centWave feature detection algorithm (Smith et al., 2006; Tautenhahn et al., 2008) to condense the raw data to feature lists, and then aligned across samples to produce a matrix of N mass features observed in M samples. The xcms processing parameters are detailed in Supplementary Material I, Section 1, in particular, with minfrac = 1 no NA values were present in the $M \times N$ matrix to avoid any influence of a data imputation step in this evaluation. An underlying assumption of the original Student's t-test (and also ANOVA) is that the mean intensities are normally distributed. To transform the data toward more normally distributed values, all intensities were logarithmized. The related features (rows in the matrix) are grouped into compound spectra with the package CAMERA. For the remaining analyses, this CAMERA grouping is assumed to be correct. Furthermore, there is no dependency on a CAMERA based grouping, and the proposed statistical treatment can be applied to groupings from equivalent tools as well.

The raw data files, the preprocessed peak matrix, and the protocol descriptions have been submitted to the MetaboLights repository (Haug et al., 2013), and are available under the accession number MTBLS74[2]. Analogously, the second data set is available as MTBLS169[3]. All statistical calculations were performed in (R Development Core Team, 2014). The complete processing scripts are provided in the Supplementary Material I, Section 1.

2.2. Detection of Differential Features and Metabolites

The analysis for differential metabolites requires to detect intensity differences between sample classes. Here, in comparison to univariate methods to analyze features, we propose several multivariate methods to analyze compound spectra representing metabolites. First, we introduce with a graphical illustration of the different decisions from univariate and multivariate tests, then we explain the several tests. All formulas of the test are shown in detail in the Supplementary Material I, Section 3.

2.2.1. Univarate Tests

The univariate Student's t-test (Student, 1908) assumes normal distributed observations of independent features. The difference of the intensity mean between the two classes is estimated for each feature. While Student's t-test assumes equal variances of the two classes, the Welch's t-test (Welch, 1947) is a variant that allows different variances between the classes (Table S1 in Supplementary Material I, Section 3).

The confidence interval (CI) determines the accuracy of this estimation, and the CI size depends on the number of observations and the standard error (SE) of the estimated difference between means. The null hypotheses, H_0, is that no difference in means exists, the alternative H_1corresponds to a difference in means. If the CI includes the origin (zero), then the difference is considered not significant and H_0 can be accepted.

If independent univariate tests for two features in a compound spectrum are combined, the confidence interval becomes a rectangular confidence region as shown in **Figure 1**, or in general for groups with p features a p-dimensional hypercube. Even if multiple testing correction is done, the confidence region holds a hypercube.

2.2.2. Multivariate Tests

The multivariate extension of Student's t-distribution was introduced by (Hotellings, 1931). The two-sample test of unequal means with unknown and equal variances becomes in multiple dimensions the Hotelling's T^2 (c.f. Table S1 in Supplementary Material I, Section 3). For unequal covariance matrices, the extension of the Welch t-test, is the James test (Table S1 in Supplementary Material I, Section 3), introduced in James (1954).

These tests compare the difference of p-dimensional mean intensity vectors in relation to their $p \times p$ covariance matrices. Observations of features in a compound spectrum are then assumed to be multidimensional normal distributed. For this multivariate analysis, the confidence region has an ellipsoid shape.

Using the multivariate tests, this statistic requires at least $(p + 1/2)$ replicates, where p is the number of features per metabolite group, to estimate the unknown entries of each covariance matrix. For typical experiments, p easily exceeds 20 for some metabolite groups, but data sets with so many replicates are rare.

In the following, we additionally propose a variant of the multivariate methods, where only the diagonal entries of the covariance matrix are estimated, with the rest fixed to zero. This simplification ignores the correlation between features, but makes the covariance estimation more robust in the case where a compound spectrum consists of more features than samples are available to modify the idea of spectra-wise analysis on small data sets. The main axes of the ellipsoid confidence region are then parallel to the coordinate axes. The details and comparison of all tests are given in Table S1 in Supplementary Material I, Section 3.

2.2.3. Comparison of Results from Univariate and Multivariate Tests

Depending on the univariate or different multivariate test statistics different sets of metabolic compound spectra are detected as differential. The H_0 hypothesis is accepted if the assumed difference in means of zero between sample classes falls within

[2] http://www.ebi.ac.uk/metabolights/MTBLS74

[3] http://www.ebi.ac.uk/metabolights/MTBLS169

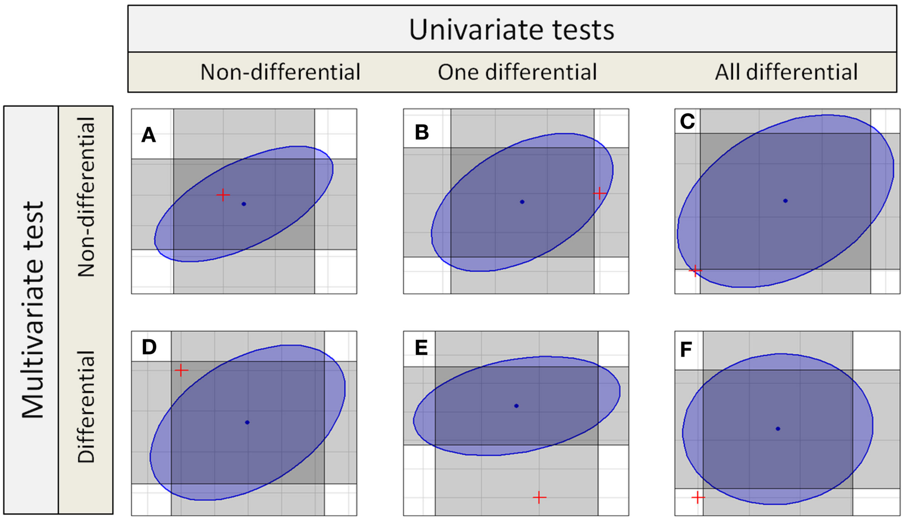


FIGURE 1 | Different decisions from univariate and multivariate test to detect differential features or compound spectra. Each gray rectangles marks the confidence interval of one test dimension, so the intersection of two rectangles marks the combined confidence region. The blue ellipse is the confidence region for a multivariate test. There are six different possibilities (six different colored spaces) for the position of the origin corresponding to the null hypotheses marked by a red "+."

the confidence interval or region. Several regions are shown in **Figure 1**.

The table also shows the different possible results for compound spectra with two features. In the simplest cases, both approaches yield the same result: in case of **Figure 1A**, no feature is differential using the univariate tests, and the compound spectrum as a whole is also not detected as differential by the multivariate test. Similarly, in **Figure 1F**, all features of the compound spectrum are differential in the univariate tests and the compound spectrum is assigned as differential by the multivariate test. But there are also cases, where the results completely differ: In **Figure 1C**, all features of the compound spectrum are differential in the univariate case, but the compound spectrum is not assigned as differential by the multivariate test, while in **Figure 1D**, none of the individual features is differential but the whole compound spectrum is detected as differential by the multivariate test. Finally, in **Figures 1B,E**, the two univariate tests for the individual features decide differently, and only one agrees with the multivariate test on the compound spectrum.

2.3. Evaluation Data and Performance Measures

The distinction between differential and non-differential can be described as a classification problem and then the typical performance measures can also be used. For the evaluation, a ground truth data set is required, where for each feature, it is known whether it is differential or not. Then, the evaluation (Algorithm 1 in the Supplementary Material I) can assess the quality of biomarker discovery with the different statistical tests by calculating the confusion matrix and the derived measures specificity and sensitivity.

The ground truth used here is a real world data set with simulated (and hence known) effect between two classes. The data set of 26 *A. th.* Col-0 wildtype plants was randomly split into two sample classes, designated as "wildtype" and "mutant," with 13 samples each.

To simulate differential features, for each compound spectrum an effect was sampled from a multivariate normal distribution with a given mean (determined by the desired effect, e.g., 0.5) and the covariance matrix that was estimated from the actual data in the 13 observations in the original "mutant class." These effects were added to the features in the "mutant class." This simulation ensures that effects are sampled for each separate compound spectrum (i.e., metabolite), rather than adding an effect to each feature individually. Thus, all compound spectra (and all its features) should be differential, and are the positive set of the ground truth. For the negative set of the ground truth, an "effect" of size zero was used.

For the simulation of the "mutant" class, only a subset of the available compound spectra can be used, since the sampling of an effect requires to estimate the covariance matrix of the compound spectra from 13 samples, which in turn is only possible for those compound spectra with a maximum of 12 features. For larger groups, it is impossible to parametrize the normal distribution used to simulate the fixed effect. Like wise, singletons (i.e., groups with only one feature) were excluded from this evaluation as the univariate and multivariate methods would give the same result.

All features are tested individually with the univariate tests, corrected for multiple-testing with Benjamini–Yekutieli procedure (Benjamini and Yekutieli, 2001) within each compound spectrum, and all compound spectra are tested with the multivariate tests.

For the comparison on the feature level, each feature in a compound spectrum that is classified as differential by the multivariate method is defined as a differential feature.

For different effects and test methods, all features are classified whether they are differential or not, and a confusion matrix can be constructed consisting of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These can be combined into sensitivity, specificity, false positive rate (FPR), and false negative rate (FNR). Repeating the prediction with different thresholds influence the performance, which can be visualized as receiver-operator curves (ROC) and summarized by the area under curve (AUC). The use of ROC curves in metabolomics is also demonstrated in Broadhurst and Kell (2006).

Finally, the evaluation can take place on the level of compound spectra (or metabolites) instead of the feature level and so compares different spectra-wise analysis approaches. This requires the definition how to interpret the multiple individual univariate decisions for a given compound spectrum. Here, all compound spectra where at least one feature was classified as differential by the univariate tests were defined as differential compound spectra. In essence, this is a two-step approach where a test on all univariate *p*-values is performed for each compound spectrum. So on the compound spectra level we can only compare the different spectra-wise analysis approaches, the two multivariate methods, which group intrinsically and the two-step approach, which uses the univariate method as the first step for spectra-wise analysis.

3. Results and Discussion

This section covers first an example for the detection of differences between a wildtype and mutant genotype experiment. Then, the analysis of the semi-synthetic ground truth dataset allows an evaluation of the statistical methods with regard to sensitivity, specificity, and area under ROC curves for multiple effects.

3.1. Analysis of an Experiment with Wildtype and Mutant Plants

First, a real dataset is analyzed. One sample class is comprised of seven *A. th.* Col-0 wildtype plants and a second class of seven samples of an *A. th.* over-expression line, a transposon based activation tagged *A. th,* line from the TAMARA population. Here, the real effect is unknown, and only a few exemplary results are described.

The data processing of the 14 samples results in a 2110 × 14 feature matrix, where CAMERA detected 335 compound spectra. The spectra with just a single feature are excluded from this comparison since the results are identical for both statistical analyses. 28% of all compound spectra have only one feature. The remaining 72% were analyzed with the both univariate and multivariate methods, except for one group with 126 features resulting from the injection peak at the beginning of the chromatography. Overall, 1891 features in 241 feature groups were analyzed.

Table 1 shows two selected compound spectra of an extended diffreport with the two compound spectra no. 40 and no. 82, the univariate *p*-value `p.uni` for each feature and the multivariate diagonal James *p*-value `p.multi` for each compound spectrum. The diagonal James test is used because of the small samples size (much smaller than the compound spectra sizes) and the assumed unequal covariance matrices between the two classes.

As shown in **Figure 2** (left), 5 features are reported exclusively by the univariate method, while the multivariate approach detected 23 features exclusively, both at a significance level of $\alpha = 0.01$.

At the compound spectra level, **Figure 2** (right) shows that 3 groups are found exclusively by the multivariate approach, which corresponds to case D in **Figure 1**. All 3 compound spectra found only by the multivariate method are compound spectra with only two or three features.

On the other hand, 4 compound spectra (one of them is a small group with only 2 features, the others have a size of 15, 17, and 35) are found that were not differential in the multivariate test, but where at least one feature was detected by the univariate approach. This corresponds to either case C where all individual features were differential, or case B where only some features were differential. Here, all 4 compound spectra were of type B.

An underlying assumption is the correctness of the CAMERA groupings, where each metabolite corresponds to one compound spectrum. In reality, it can happen that features from one metabolite are split into two (or more) compound spectra.

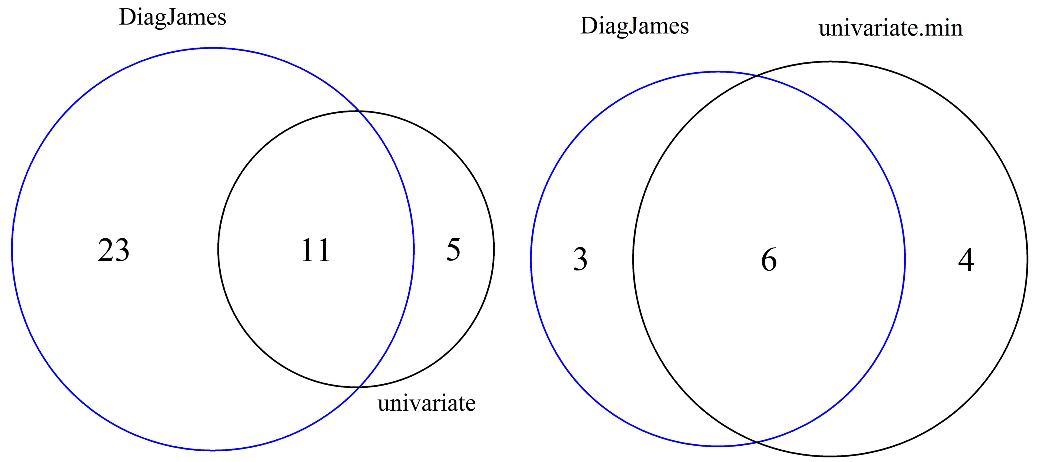


FIGURE 2 | Venn diagram of differential features and compound spectra in the wildtype-mutant experiment for the significance level of $\alpha = 0.01$. Left: number of *features* detected by univariate and multivariate method. Right: number of *compound spectra* detected by the multivariate method, compared to the number of compound spectra where at least one feature was detected univariately.

In this case, the multivariate approach looses power, and in the extreme case where a metabolite is split into many singleton spectra achieves the same results as the univariate approach. The opposite case, where two or more metabolites end up in the same compound spectrum can also have a negative influence. If, for example, a differential and a non-differential metabolite are joined, the combined "differentiality" could turn out non-significant and hide one of them.

In this experiment, the biological truth, i.e., which metabolites and features are affected by the over-expression construct is not known. For an objective evaluation, we created a semi-synthetic dataset with simulated fixed effects.

3.2. Evaluation with Multiple Simulated Fixed Effects

In this second experiment, the performance of the three statistical analysis – univariate and multivariate with both Hotellings-T^2 and the diagonal Hotellings-T^2 – was compared on a dataset of metabolite profiles from *Arabidopsis thaliana*. The xcms processing results in a matrix of 1476 features, and the CAMERA grouping reveals 282 compound spectra. As explained above, for the simulation of the "mutant" class, only a subset of 153 compound spectra with 12 or less features can be used for the ground truth.

We combined the negative set (effect 0.0) with 686 features in 153 compound spectra with the positive set consisting of the same 686 features but with the added effect. For each effect, between 0.0 and 1.4, the final ground truth dataset thus contained 306 compound spectra with a total of 1372 features.

The following exemplifies the results for the fixed effect of 0.5, corresponding to a fold change of ≈1.5 in the original, non-logarithmic data.

For a significance level of $\alpha = 0.05$, **Table 2** shows the summary of the confusion matrix for all three approaches. The multivariate approaches clearly achieve both a better sensitivity and FNR.

The Venn diagram in **Figure 3** (left) shows the 242 features are detected as differential by all three tests, 243 by both the univariate and the T^2 and 258 by both the univariate and the diagonal T^2. The Comparison of the univariate and the original T^2 shows that 16 features are found only by the univariate and 328 features only by the multivariate method. The same for the diagonal T^2 shows that only 1 feature is found only by the univariate and 253 features only by the multivariate method. Furthermore, 200 features are found by both multivariate methods. It is shown that the feature detection has more overlap between the two multivariate methods than between one of these with the univariate approach. Now, we are especially interested in cases where the multivariate methods identify compound spectra as differential, while the univariate method detects none of the features in the spectra, or cases where the univariate method detects features whose associated compound spectra are missed by the multivariate methods (**Figure 3** right). Here, only 7 compound spectra are detected by both multivariate methods, 29 by the original multivariate T^2 and 25 by the diagonal multivariate method, where any feature of this spectra is detected by univariate method. In contrast, 5 compound spectra have at least one feature, which is detected by the univariate test, but the compound spectra itself are not identified by the multivariate T^2 method and 1 compound spectrum in comparison with the diagonal multivariate T^2. 83 groups are detected by all three tests, 84 by univariate and T^2, 98 by univariate and diagonal T^2 (**Figure 3** right).

The ROC curve of the three feature detection approaches for a specific effect of 0.5 (Figure S6 in Supplementary Material II) shows the sensitivity and specificity for significance thresholds other than $\alpha = 0.05$, and confirms that the multivariate method has a higher AUC.

TABLE 2 | Comparison of performance of univariate and multivariate tests for a simulated effect of 0.5 and significance level of $\alpha = 0.05$.

Method	FP (FPR)	FN (FNR)	TP (sensitivity)	TN (specificity)
Univariate	0 (0%)	427 (62.2%)	259 (37.8%)	686 (100%)
T^2	36 (5.2%)	151 (22%)	535 (78%)	650 (94.8%)
DiagT^2	5 (0.7%)	180 (26.2%)	506 (73.8%)	681 (99.3%)

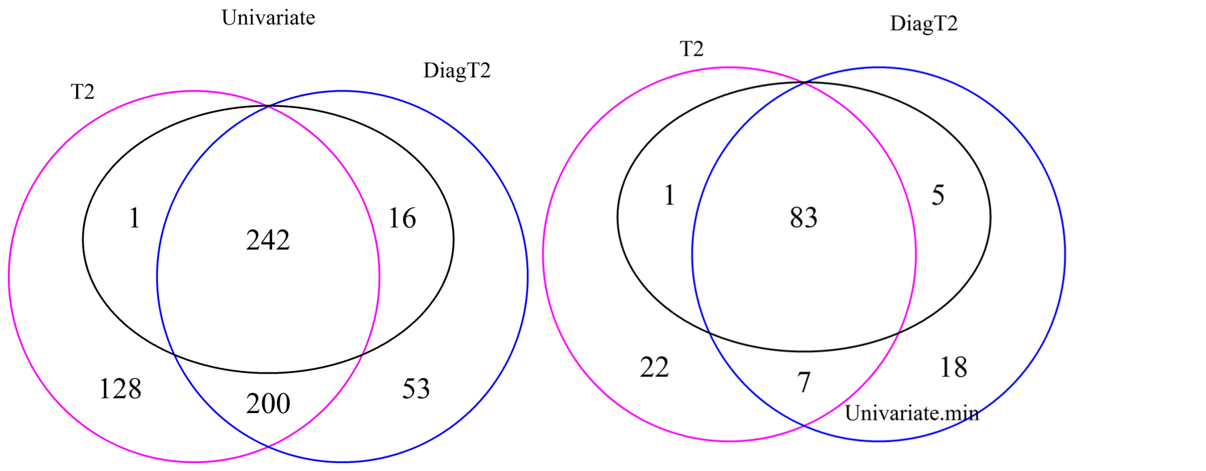


FIGURE 3 | Venn diagram of differential features and compound spectra in the simulation experiment for the simulated effect 0.5 and significance level of $\alpha = 0.05$. Left: number of *features* detected by univariate and multivariate method. Right: number of *compound spectra* detected by the multivariate method, compared to the number of compound spectra where at least one feature was detected univariately.

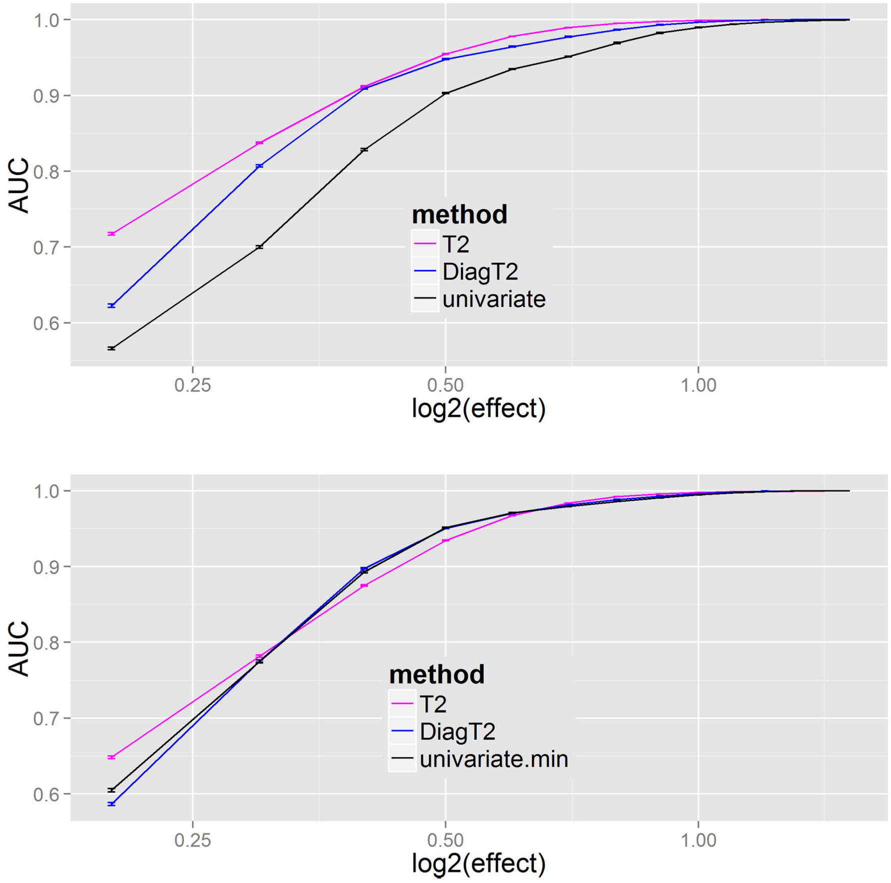


FIGURE 4 | Results of univariate and multivariate methods in feature detection are compared on the feature level (upper). At the compound spectra level (lower) the results of different grouping analysis approaches are shown. For each simulation step, several added effects of 0.2, 0.3, . . ., 1.4, 1.5 on the "mutant" class, the mean and SE of the evaluated AUCs (results from 100 repetitions) are plotted.

The next question was the behavior of the methods for different effects. The AUC was used as a summary metric of the performance. **Figure 4** shows that the multivariate T^2 as well as the diagonal T^2 method has a better AUC for the feature detection compared to the univariate approach for all effects of $0.2, 0.3, \ldots, 1.4, 1.5$. To improve the generalization, the sampling of the "mutant" data was repeated 100 times for each effect. Especially for smaller effects, the benefit of the multivariate approach is visible and also that the simplified diagonal T^2 approximates to the original T^2 for larger effects.

The results **Figure 4** (bottom) show no particular differences between the different compound spectra level (or grouping) approaches, thus the main benefit results from a joint analysis of compound spectra, while less differences are observed between the joint analysis methods. In Supplementary Material I all methods including James and diagonal James are compared in **Figure 2** on the feature level, and **Figure 3** on the compound spectra level. In a repeated sensitivity analysis (Supplementary Material III) we show that for small effects and large compound spectra Hotelling's-T^2 has an advantage over the other grouping approaches.

4. Conclusions

In mass spectrometry-based metabolomics data, metabolites (which are the objects of biological interest) will usually give rise to multiple spectral features. In recent years, methods were developed to group these related features into compound spectra. However, the statistical analysis was still based in either individual univariate tests or global multivariate analysis.

We have extended the feature-wise univariate statistic tests to a compound spectra-wise analysis. Using traditional multivariate hypothesis tests, like the Hotelling's T^2 or James test, the confidence interval becomes a multidimensional ellipsoid that resembles the joint probability for metabolites to be differential more realistically.

On real data of a comparative wildtype-mutant experiment design, the results of the univariate and multivariate tests have an overlap, while some features are detected exclusively by the univariate or multivariate test.

On the synthetic data where the actual effect was known, on the feature level, the resulting AUCs for the multivariate analysis of compound spectra were better than in the univariate case, we recommend to analyze the data compound spectra-wise for biomarker discovery in mass spectrometry metabolomics data. On the compound spectra level the advantage of T^2 over the other spectra-wise approaches is most prominent for noisy data and/or if very small effects should be detectable.

If the CAMERA grouping erronously splits a metabolite into several compound spectra the results of all spectra-wise analyses will approach the multivariate results, and false negatives can occur if a differential and a non-differential metabolite are joined by the compound spectra grouping.

While CAMERA was used in this study, the approaches are readily applicable to any data where individual features from a metabolite are grouped together. In particular, this should allow the analysis of GC/MS data, where the established data analysis typically relies on deconvoluted spectra or mass spectral tags, and where the selection of quantifier ions would have to be repeated for each sample matrix or sample type. The presented approach does not require the selection of representative ions.

The proposed joint analysis of features of a metabolite group as a spectra-wise analysis is the key idea and bridges an important gap between hypotheses tests on individual features on the one hand, and global multivariate methods, which might be more difficult to interpret on the other.

Author Contributions

SS performed the metabolomics experiment, DT performed the statistical analysis, SN and IG supervised the work. All authors contributed to the manuscript.

Funding

SS was partially funded by the BMBF (GABIProTect 0315051C). SN acknowledges funding from DFG grant NE/1396/5-1.

Acknowledgments

We thank T. Gigolashvili and U. Flügge for the *Arabidopsis thaliana* activation tagged TAMARA germplasm.

Supplementary Material

This article is accompanied by the following supplemental information which can be found online at http://journal.frontiersin.org/article/10.3389/fbioe.2015.00129:

Presentation 1 | The Supplementary Material I file contains details about the mass spectrometry setup and data processing (Section 1), additional results of the simulation experiment (Section 2), and moreover the formula of the statistical tests (Section 3).

Presentation 2 | The Supplementary Material II file shows the ROC curves for the simulation experiment.

Presentation 3 | The Supplementary Material III file shows repeated figures of the simulation experiment for different datasets.

Data Sheet 1 | The raw data to the article is available from the MetaboLights repository as accession MTBLS74[2] and MTBLS169[3]. The provided *R*-functions in file multivariateDiffreport.R and a vignette file MTBLS169analysis.Rnw is provided, which contains an example analysis on the dataset seven measurements of an *Arabidopsis thaliana* versus 7 measurements of the over-expression line. The Rdata object MTBLS169.Rdata contains the preprocessed MS peak lists and annotations.

References

Alonso, A., Juli, A., Beltran, A., Vinaixa, M., Daz, M., Ibaez, L., et al. (2011). AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* 27, 1339–1340. doi:10.1093/bioinformatics/btr138

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi:10.1186/1471-2105-9-114

Böttcher, C., von Roepenack-Lahaye, E., and Scheel, D. (2011). "Resources for metabolomics," in *Genetics and Genomics of the Brassicaceae, Crops and Models*, Vol. XII, eds I. Bancroft and R. Schmidt (Berlin: Springer-Verlag), 677.

Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., and Glawischnig, E. (2009). The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of *Arabidopsis thaliana*. *Plant Cell* 21, 1830–1845. doi:10.1105/tpc.109.066670

Broadhurst, D. I., and Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2, 171–196. doi:10.1007/s11306-006-0037-z

Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., et al. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* 134, 1322–1332. doi:10.1039/b901179j

Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., et al. (2011). Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* 27, 1108–1112. doi:10.1093/bioinformatics/btr079

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Davis, C. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. San Diego, CA: Springer.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2013). MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, D781–D786. doi:10.1093/nar/gks1004

Hendriks, M. M., Eeuwijk, F. A., Jellema, R. H., Westerhuis, J. A., Reijmers, T. H., Hoefsloot, H. C., et al. (2011). Data-processing strategies for metabolomics studies. *TrAC Trends Anal. Chem.* 30, 1685–1698 In-Vivo and On-Site Analysis {II}. doi:10.1016/j.trac.2011.04.019

Hotellings, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat.* 2, 360–378. doi:10.1214/aoms/1177732979

Ipsen, A., Want, E. J., Lindon, J. C., and Ebbels, T. M. D. (2010). A statistically rigorous test for the identification of parent-fragment pairs in LC-MS datasets. *Anal. Chem.* 82, 1766–1778. doi:10.1021/ac902361f

James, G. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika* 41, 19–43. doi:10.1093/biomet/41.1-2.19

Kenar, E., Franken, H., Forcisi, S., Wörmann, K., Häring, H.-U., Lehmann, R., et al. (2014). Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol. Cell. Proteomics* 13, 348–359. doi:10.1074/mcp.M113.031278

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, R., and Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Anal. Chem.* 84, 283–289. doi:10.1021/ac202450g

Legendre, P., and Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1–24. doi:10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2

Luedemann, A., Strassburg, K., Erban, A., and Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography – mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* 24, 732–737. doi:10.1093/bioinformatics/btn023

Saccenti, E., Hoefsloot, H., Smilde, A., Westerhuis, J., and Hendriks, M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* 10, 361–374. doi:10.1007/s11306-013-0598-6

Sampson, J. N., Boca, S. M., Shu, X. O., Stolzenberg-Solomon, R. Z., Matthews, C. E., Hsing, A. W., et al. (2013). Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol. Biomarkers Prev.* 22, 631–640. doi:10.1158/1055-9965.EPI-12-1109

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A., and Breitling, R. (2011). PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal. Chem.* 83, 2786–2793. doi:10.1021/ac2000994

Schneider, A., Kirch, T., Gigolashvili, T., Mock, H.-P., Sonnewald, U., Simon, R., et al. (2005). A transposon-based activation-tagging population in *Arabidopsis thaliana* (TAMARA) and its application in the identification of dominant developmental and metabolic mutations. *FEBS Lett.* 579, 4622–4628. doi:10.1016/j.febslet.2005.07.030

Smith, C., Want, E., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal. Chem.* 78, 779–787. doi:10.1021/ac051437y

Steuer, R., Morgenthal, K., Weckwerth, W., and Selbig, J. (2007). "A gentle guide to the analysis of metabolomic data," in *Metabolomics, Volume 358 of Methods in Molecular Biology*, ed. W. Weckwerth (New York, NY: Humana Press), 105–126.

Student. (1908). The probable error of a mean. *Biometrika* 6, 1–25. doi:10.1093/biomet/6.2-3.302

Sumner, L. W., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., et al. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3, 211–221. doi:10.1007/s11306-007-0082-2

Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9:504. doi:10.1186/1471-2105-9-504

Trutschel, D., Schmidt, S., Grosse, I., and Neumann, S. (2015). Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics* 11, 851–860. doi:10.1007/s11306-014-0742-y

Varghese, R. S., Zhou, B., Nezami Ranjbar, M. R., Zhao, Y., and Ressom, H. W. (2012). Ion annotation-assisted analysis of LC-MS based metabolomic experiment. *Proteome Sci.* 10(Suppl. 1), S8. doi:10.1186/1477-5956-10-S1-S8

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrica* 34, 28–35. doi:10.2307/2332510

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

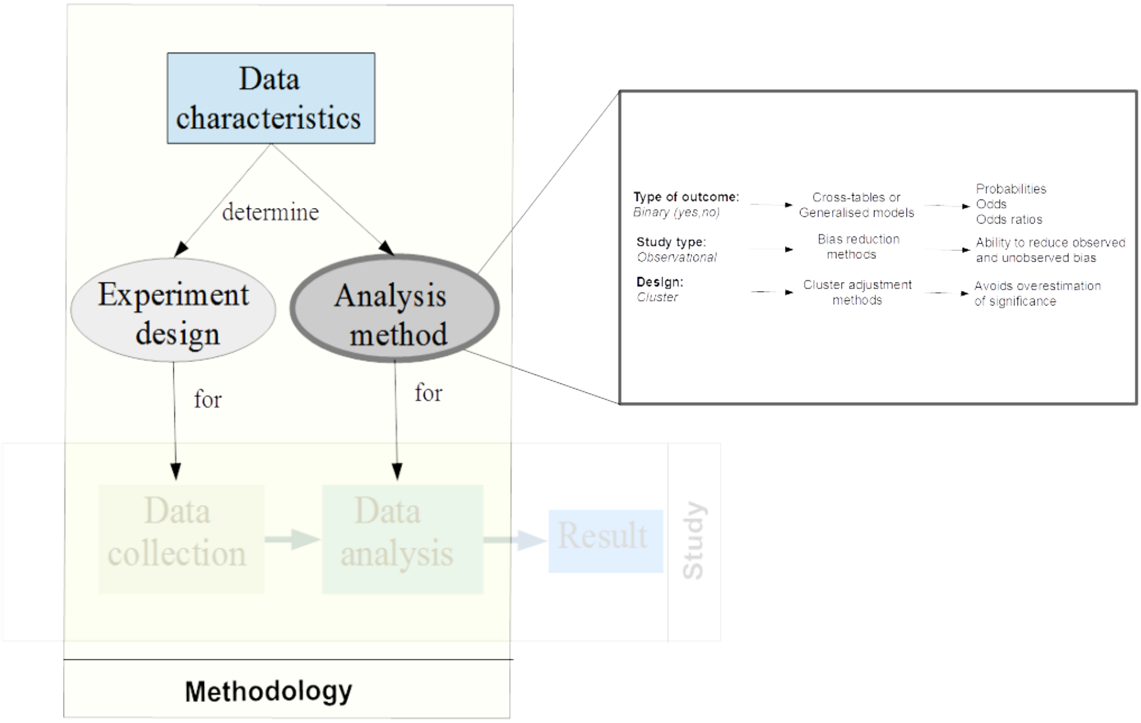
Copyright © 2015 Trutschel, Schmidt, Grosse and Neumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

6.4. Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing

International Journal of Nursing Studies, 2017, Vol. , Issue , Pages

Diana Trutschel[1,2],
Rebecca Palm[1,3], PhD
Bernhard Holle[1,3], PhD,
Michael Simon[4,5], PhD

[1]German Center for Neurodegenerative Diseases (DZNE) Site Witten, Witten, Germany
[2]Martin-Luther University Halle/Wittenberg, Institute of Informatics, Halle/Saale, Germany
[3]Department of Health, School of Nursing Science, University Witten/Herdecke, Witten, Germany
[4]Faculty of Medicine, Institute of Nursing Science, University of Basel, Basel, Switzerland [5]Nursing and Midwifery Research Unit, Inselspital Bern University Hospital, Bern, Switzerland



Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/locate/ijns

Methodological approaches in analysing observational data: A practical example on how to address clustering and selection bias

Diana Trutschel[a,b,*], Rebecca Palm[a,c], Bernhard Holle[a], Michael Simon[d,e]

[a] German Center for Neurodegenerative Diseases (DZNE), Witten, Germany
[b] Martin-Luther-University Halle-Wittenberg, Halle/Saale, Germany
[c] University Witten/Herdecke, Witten, Germany
[d] University of Basel, Basel, Switzerland
[e] University Hospital Inselspital, Bern, Switzerland

ARTICLE INFO

Keywords:
Observational study
Nonexperimental studies
Health services research
Nursing research
Propensity score
Logistic models
Multiple logistic regression

ABSTRACT

Background: Because not every scientific question on effectiveness can be answered with randomised controlled trials, research methods that minimise bias in observational studies are required. Two major concerns influence the internal validity of effect estimates: selection bias and clustering. Hence, to reduce the bias of the effect estimates, more sophisticated statistical methods are needed.
Aim: To introduce statistical approaches such as propensity score matching and mixed models into representative real-world analysis and to conduct the implementation in statistical software R to reproduce the results. Additionally, the implementation in R is presented to allow the results to be reproduced.
Method: We perform a two-level analytic strategy to address the problems of bias and clustering: (i) generalised models with different abilities to adjust for dependencies are used to analyse binary data and (ii) the genetic matching and covariate adjustment methods are used to adjust for selection bias. Hence, we analyse the data from two population samples, the sample produced by the matching method and the full sample.
Results: The different analysis methods in this article present different results but still point in the same direction. In our example, the estimate of the probability of receiving a case conference is higher in the treatment group than in the control group. Both strategies, genetic matching and covariate adjustment, have their limitations but complement each other to provide the whole picture.
Conclusion: The statistical approaches were feasible for reducing bias but were nevertheless limited by the sample used. For each study and obtained sample, the pros and cons of the different methods have to be weighted.

What is already known about the topic?

Data in nursing health services research often is observational and clustered

Clustering and selection bias can lead to biased results

What this paper adds

The paper introduces common analytical strategies to address selection bias and clustering in observational research

Providing a vignette, researchers can replicate the used analytical strategies

1. Introduction

Nursing research aims to validate, refine and generate knowledge from studies that directly and indirectly affect the delivery of nursing care (Burns and Grove, 2009). Furthermore, evaluating health services, an aim of nursing research (AACN, 2015), requires research methods that achieve the highest internal validity possible to derive unbiased effect estimates of an intervention in a certain population in real-world settings. When threats to internal validity, such as selection bias or clustering, are not addressed through the study design, statistical methods are needed to reduce the bias of the effect estimates. Two major concerns influence the internal validity of effect estimates: selection bias and clustering. These two factors are the primary focus of this article.

We are motivated by our own observational study in health services research, in which three main data characteristics need to be addressed to find a suitable analysis method. Specifically, illustrated in Fig. 1, a *dichotomous* outcome of *clustered* data in a *observational* study was

[*] Corresponding author.
E-mail address: diana.trutschel@dzne.de (D. Trutschel).

http://dx.doi.org/10.1016/j.ijnurstu.2017.06.017
Received 6 December 2016; Received in revised form 12 June 2017; Accepted 29 June 2017
0020-7489/ © 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

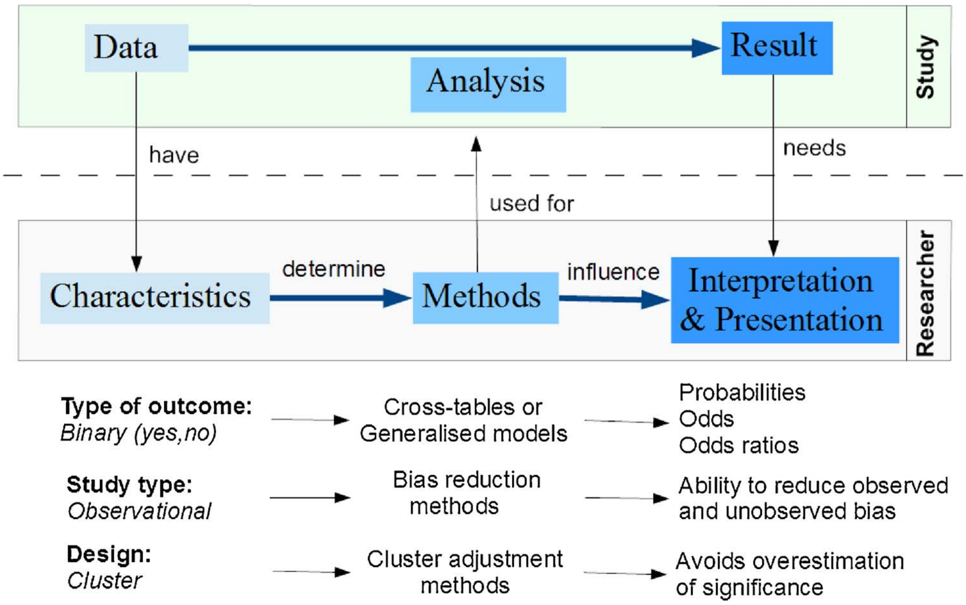

Data
Analysis
Result
Study
have
needs
used for
Characteristics
determine
Methods
influence
Interpretation & Presentation
Researcher
Type of outcome: Binary (yes,no)
Cross-tables or Generalised models
Probabilities Odds Odds ratios
Study type: Observational
Bias reduction methods
Ability to reduce observed and unobserved bias
Design: Cluster
Cluster adjustment methods
Avoids overestimation of significance

Fig. 1. In a study, the data analysis generates results. The data have their own characteristics, for example, a special outcome type, a unique study type or a specific design. These characteristics determine the choice between sophisticated methods for data analysis. Hence, the method directly influences the interpretation of the results and therefore must be carefully chosen using the skill of the researcher.

analysed.

First, the distribution of the outcome variable, which is one characteristic of our example data, influences the choice of the statistical method. Here, we analyse the use of case conferences as a binary outcome. Binary variables are summarised by probabilities, odds and odds ratios. A probability is defined as a relative frequency and can easily be understood (as a risk), whereas odds are an expression of relative probabilities – the ratio of the probability of the event occurring to the probability of no event occurring. Moreover, the odds ratio is the relation of two odds. However, because odds are not a probability, the interpretation is more difficult for practitioners (Greenland, 1987), and sometimes, odds are misinterpreted (O'Connor, 2013). Furthermore, if the model for effect estimation is not simple, then generalisable models that use link functions other than the identity functions are needed.

The second characteristic is the observational study type, which is used to collect data. In observational studies, the possibility of controlling factors that may influence the study outcome is limited to observed variables because randomisation is not part of the study design. Therefore, other options must be applied to reduce selection bias, which can contribute to over-/underestimations of the intervention effect (Starks et al., 2009). Hence, estimations of treatment effects through direct comparisons are prone to selection bias when the assignment to treatments is associated with the potential outcomes of the treatment (Ridder and Graeve, 2011).

Our example is an evaluation of special care units. Special care units serve dedicated patient populations that are in need of special care because of their health state. Special care units are implemented for conditions such as stroke, premature birth and dementia. For example, residents who reside in dementia special care units systematically differ from other residents because they are selected based on predefined criteria. Additionally, studies about dementia special care units typically have a multistage clustered data structure: residents are clustered within units, units are clustered within nursing homes, and nursing homes are clustered in provider systems. Selection bias may occur in every stage: residents in dementia special care units differ from residents in other care units, and nursing homes with dementia special care units may differ from nursing homes without dementia special care units.

Another problem that may arise in studies is the overestimation of how the significance of effects due to clustering influences the variance estimation of the effect. If more than one cluster is included in the study, a clustered or nested data structure is most likely present, and the error terms within a cluster are no longer independent. When the non-independence of the data is not accounted for in the statistical model, the odds for significant results increase. Hence, in our example, residents are clustered within nursing homes. This clustering must be considered when choosing the analysis method.

The nursing research literature contains many examples of observational studies that are necessitated to address selection bias and clustering. For instance, studies investigating the association of organisational characteristics, such as the work environment and patient or nurse outcomes, generally have to address both issues. For example, Zúñiga et al. (2015) explore the association between the work environment and care workers' perception of quality of care in 155 nursing homes in a cross-sectional study. To address selection bias, the authors employ a multilevel regression model with a range of variables as control factors (e.g., language region and unit size) and others as random effects (e.g., unit and hospital site) to address clustering.

In this article, we will introduce statistical approaches to reduce selection bias and clustering in a real-world data analysis example. We highlight the strengths and weaknesses of different methods, which are elucidated and discussed with respect to applying the methods to the chosen example study data. Additionally, we provide data and source code as a vignette (supplemental material) to show the practical implementation of the models separately and enable replicating the analysis with open-source software R (R Core Team, 2015), which might guide readers in applying the methods to their own studies and conditions.

Our aim here is not to provide a review of the methodological work within this field. Nevertheless, the following articles and books discussing propensity score (Austin et al., 2007; Belitser et al., 2011; Biondi-Zoccai et al., 2011; D'Agostino, 1998; Randolph et al., 2014; Sekhon, 2011; Stürmer et al., 2006), matching (Pimentel et al., 2015; Rosenbaum, 2002; Rubin, 2006; Stuart, 2010) and multivariate adjustment (Cepeda et al., 2003; Gelman and Hill, 2007) serve as guidance for our work.

The aim of this article is to highlight (1) why different methods should be used, (2) their application in a statistical software and (3) how to interpret the results produced by statistical methods.

2. Materials and methods

2.1. Data and research example

The provided dataset is from the observational DemenzMonitor study (Palm et al., 2014, 2015). Data from 2013 were used for the analysis. The data consist of a convenience sample of 51 nursing homes, 109 care units and 1808 residents. After residents had been excluded due to only a two-group comparison being performed and predefined exclusion criteria, we used a dataset of $n = 888$ participants from 64 care units in 36 nursing homes (available in the supplemental material). Additionally, 53 residents with missing values in any of the variables were excluded. The primary question for this analysis was whether a dementia special care unit more frequently performs case conferences than traditional care units. The outcome variable was a binary indicator for whether the condition (1) was performed or not (0). Because the study used an observational design, residents in special care units and traditional units did not necessarily share the same characteristics, thus requiring an analytical approach to address selection bias. Furthermore, the clustering of residents in nursing homes leads to non-independent observations, again requiring an analytical approach that takes this clustering into account.

2.2. Procedure

Table 1 shows the two levels of analytical strategies for addressing the problem of unequal distributions of characteristics in the condition and comparison groups and the problem of clustering: (i) different models with different abilities to adjust for dependencies to analyse binary data and (ii) different methods to adjust for selection bias. Here, both analytical problems are addressed and combined in the analysis.

We distinguish two models for obtaining inference from the binary data: a crude model and a generalised linear mixed model. In the crude model, the results are not adjusted for the hierarchical data structure (clustering) or for differences in baseline characteristics, resulting in a higher risk of false-positive results. The generalised linear mixed model, which is a multilevel model without any additional control variables, addresses the clustering issue but does not address selection bias. We describe two methods for bias reduction that can be used for analysing data with dichotomous outcomes (by crude or advance model): (1) genetic matching on samples and (2) adjustment via the common regression model. All steps of this procedure, which are shown in Table 1, can be followed and adapted for other data sets using the provided Vignette (supplemental material), which shows the implementation with the programming language R (R Core Team, 2015). In this article, we will first introduce the crude model, then adjust for clustered data with the generalised linear mixed model, and finally use this model with all methods for bias reduction (only the shaded areas in Table 1).

Table 1
Different analytical strategies for selection bias reduction and/or cluster adjustment. The crude model, generalised linear model (GLM), is not able to adjust for clustered data or reduce bias in observational studies. A generalised linear mixed model (GLMM) is essential to account for multilevel data. Selection bias can be reduced by (1) including covariates in the regression model or (2) using a matching algorithm to reach a balance on the covariates between the investigated groups.

		Bias reduction	
		No	Yes
Cluster	No	Crude (GLM)	GLM + genetic matching GLM + covariate adjustment
Adjustment	Yes	GLMM	GLMM + genetic matching GLMM + covariate adjustment

2.3. Different models and their ability to adjust for dependencies

In our example, because we analyse a binary outcome variable, common methods for normally distributed variables and statistical tests such as Student's *t*-test and ANOVA cannot be used. Testing the differences between groups is similar to testing the differences between proportions in a contingency table, which refers to the ‘crude model’. Testing the association between a dependent variable and a group of independent variables for a binary outcome requires a logistic regression model. The crude model is identical to a regression model with the group assignment (condition, control) as an independent variable without covariate adjustment. When observations are not independent, e.g., because of clustering in different nursing homes, a generalised linear mixed model is used.

2.3.1. Crude model

The crude model is a simple contingency table (upper part of Table 2) that provides an initial overview of the two-dimensional frequency distribution of cross-tabulated data – the distribution of a binary outcome variable (here, the performance of case conferences). From this table, probabilities and odds can be calculated (bottom part of Table 2). Hence, the (estimated) probability of an ‘event will take place’ can be calculated as a proportion from the frequencies in each group (see the supplemental material for equations), and differences in (estimated) probabilities between the two groups can easily be calculated. The probability of an event in a specific group is also known as risk; therefore, the risk ratio compares the probability of an event in one group to that in another (here, for example, the treatment group versus the control group). Often, the chance that something will occur is described as the odds (see the supplemental material for the equation). Although the interpretation is more difficult for practitioners (Greenland, 1987) because an odd is not a probability and sometimes is misinterpreted as a risk (O'Connor, 2013), the provided scale is indefinite and hence provides possibilities of working with other mathematical methods. The odds are the ratio of both probabilities, namely, the probability of an ‘event will take place’ versus the probability of an ‘event will not take place’ $p/(1 - p)$, and it compares how much larger one probability is relative to another in a specific group. In our case, the (estimated) probability that a case conference was conducted in the control group was 0.8, and the (estimated) probability that a case conference was not conducted was 0.2. Hence, within the control group, the (estimated) probability that a case conference was conducted is four times higher than not, which indicates an odds of 4 (0.8/0.2). The widely used odds ratio is thus the ratio of both odds, namely, the odds of the treatment group related to the odds of the control group (see the supplemental material for equations). The odds ratio compares the difference in the odds between the two groups. If the odds are equal in both groups, then the odds ratio is equal to one. In our case, the odds ratio is 2.58, which means that the odds of receiving a case conference in the condition are higher than those in the control group.

The crude model provides the same results as the logistic regression model with only one independent variable for group assignment (case versus control). The logistic regression belongs to the family of generalised linear models (GLMs), which can handle different distributions of outcome variables. Assuming that $p = P(\text{Event} = \text{yes}|X)$ is the probability that an event occurs given the predictor variables X and that p_i is this probability for one response i, the generalised linear model adapts the linear relationship using the logit function (see the supplemental material for equations). The logistic regression model with a single dichotomous predictor variable for the assignment to the group ‘Treatment’ for a binary outcome (e.g., whether a case conference was performed) is a simple example. With the maximum likelihood method, the parameters of the logistic regression model, β_0 and β_1 here, can be estimated, and then, the inverse function of the logit $logit^{-1}(x) = e^x/(1 + e^x)$ provides the ability to assess the probability values of [0,1] (see the supplemental material for equations). In our case, this means

Table 2
Upper: a contingency table of a two-group comparison for a dichotomous outcome variable, where n_{ij} is the absolute amount of outcome i in group j. Lower: parameters, their estimates calculated from the contingency table and their interpretation.

		Group		
		Treat	Control	Marginal
Outcome	No	n_{11} (22)	n_{12} (119)	$n_{1.} = n_{11} + n_{1.}$ (141)
	Yes	n_{21} (224)	n_{22} (470)	$n_{2.} = n_{21} + n_{2.}$ (694)
	Marginal	$n_{.1} = n_{11} + n_{21}$ (246)	$n_{.2} = n_{12} + n_{22}$ (589)	$N = n_{.1} + n_{.2} = n_{1.} + n_{2.}$ (835)
Interpretation	Probabilities	$p_{\text{Treat}} = \frac{n21}{n_{.1}}$ (0.91)	$p_{\text{Control}} = \frac{n22}{n_{.2}}$ (0.8)	(Risk)Diff. $= p_{\text{Treat}} - p_{\text{Control}}$ (0.11)
	Odds	$\text{Odd}_{\text{Treat}} = \frac{p_{\text{Treat}}}{1 - p_{\text{Treat}}}$ (10.18)	$\text{Odd}_{\text{Control}} = \frac{p_{\text{Control}}}{1 - p_{\text{Control}}}$ (3.95)	Odds Ratio $= \frac{\text{Odd}_{\text{Treat}}}{\text{Odd}_{\text{Control}}}$ (2.58)

Table 3
Parameters of generalised linear model with the full sample and their interpretation. Each parameter $x \in (\beta_0, \beta_1, \beta_0 + \beta_1)$ of a generalized linear model with the form of $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ can be extracted and interpreted. The left column shows the estimates of the model and are interpreted as logarithmic odds, the middle column provides a transformation into (estimated) odds, and the right provides a transformation into probabilities.

Parameter	Transformation		
	No	$e^{\text{parameter}}$	$\frac{e^{\text{parameter}}}{1 + e^{\text{parameter}}}$
β_0	$\text{logOdd}_{\text{Control}} = 1.37$	$\text{Odd}_{\text{Control}} = e^{1.37} = 3.95$	$p_{\text{Control}} = \frac{e^{1.37}}{1 + e^{1.37}} = 0.8$
$\beta_0 + \beta_1$	$\text{logOdd}_{\text{Treat}} = 2.32$	$\text{Odd}_{\text{Treat}} = e^{2.32} = 10.18$	$p_{\text{Treat}} = \frac{e^{2.32}}{1 + e^{2.32}} = 0.91$
β_1	logOdds ratio $= 0.95$	Odds ratio $= e^{0.95} = 2.58$	

that the estimated log odds and log odds ratios can support the (estimated) probabilities of receiving a case conference for each group.

In Table 3, the expressions of the estimated parameters, i.e., the estimates of the logarithmic odds and odds ratio, are listed (equations are explained in supplemental material). The exponentiated expressions of the model parameters are the odds and the odds ratio. The (estimated) probabilities of success in both groups are also given. For example, the parameter β_0 of the logistic regression model expresses the logarithmic odds of the control group, and the exponentiated value e^{β_0} is the corresponding odds, which indicates the chance that an 'event will take place' versus the opposite chance in the control group. The probability that an 'event will take place' in the control group is calculated using $e^{\beta_0}/(1 + e^{\beta_0})$. Due to the circumstances, the linear function can also be from a different family of functions; this type of model specification is the generalised linear model. Furthermore, some statistical programs provide the converted estimated values from such generalised models coincidently, but in R, these values must be calculated manually. Hence, the mathematical link and inverse link function (as shown in Table 3) must be known to provide the estimates in the required scale ((estimated) odds or probabilities in this case).

As with all regression models, the logistic regression can adjust for measured group differences (e.g., age or severity) when a binary outcome is predicted from a set of variables. Hence, proportions based on a dichotomous event are analysed using this widely used method (Ostir and Uchida, 2000).

2.3.2. Generalised linear mixed model

In our case, observations were collected from participants in care units nested in nursing homes. Therefore, one of the key assumptions of the logistic regression model – independence of observations – is violated. Because more than a half of the nursing homes (20 of 36) provided only one care unit, we use nursing home as only a cluster level (for more detail, a histogram of the number of participants in each care unit within the nursing homes is given in Fig. 2, Supplemental I). Hence, in this situation, the treatment is assigned at the individual level. In our example, the intra-class correlation coefficient (ICC) is 0.48 [0.2, 0.73], which means that 48% of the variation is explained by the variation between the nursing homes. Although a range of different estimators exists (see Wu et al., 2012 for details) [2012], here, we use the Fleiss-Cuzick estimator given by Zou and Donner (2004) to calculate the intra-class correlation coefficient on the proportional scale (see the Vignette for calculations; additionally, a model-based estimator is given).

Violation of this assumption of independence due to clustered data can lead to committing type I errors, e.g., finding an association where there is none. A solution to this problem is to apply a generalised linear mixed model. Generalised linear mixed models are an extension of the generalised linear models and are well established (see (Hardin and Hilbe, 2012; Stroup, 2012). They combine two statistical concepts: using linear mixed models to include random effects and using generalised models to model non-normal data. Hence, error terms that correspond to the different sources of variation in the data are added to the logistic regression model (Gelman and Hill, 2007), and the residual variance can be separated into components of the different involved levels (Li et al., 2011). In our example, the individual probability being statistically dependent on the nursing home where a participant lives is considered, and the variation between nursing homes and participants is quantified.

2.4. Methods for selection bias adjustment

In this section, we will introduce two basic approaches to address selection bias: (1) matching and (2) regression, and provide a very brief description for why we take this approach. Because both approaches can use all covariates or the propensity score for adjustment, we provide the definition of the propensity score first.

2.4.1. Propensity score

The propensity score was introduced by Rosenbaum and Rubin (1983) and is defined as the conditional probability of being treated given a set of covariates. The definition of the propensity score for a single subject i (Eq. (1)) is the conditional probability of assignment to the treatment group ($Z_i = 1$) given a vector of observed covariates (x_i), where Z_i is assumed to be independent. Based on the observed control variables for each subject, a propensity score for membership in the treatment group is calculated from a logistic regression. Hence, the propensity score summarises different confounding factors into one dimension and can thus be used to achieve balance (Biondi-Zoccai et al., 2011) through adjustment methods, such as matching or regression models (D'Agostino, 1998). Balancing in this context means that the baseline characteristics in the treatment and control groups are the same (matching) or that balance differences are taken into account (regression). Using propensity score methods allows estimation of unbiased treatment effects if there is no unmeasured confounder (Williamson et al., 2012). Numerous literature reports that consider the impact of the selection of the model for propensity score estimation on the ability to reduce bias through the outcome model and also balance checks after application of the propensity score are available (Arpino and Mealli, 2011; Austin et al., 2007; Belitser et al., 2011; Leyrat et al., 2014; Nguyen et al., 2017; Rosenbaum and Rubin, 1983; Stürmer et al., 2006; Williamson et al., 2012).

$$\text{Propensity score} = e(x_i) = P(Z_i = 1 | X_i = x_i) \tag{1}$$

2.4.2. *Genetic matching and hidden bias assessment*

The matching of similar individuals in the treatment group with individuals from the control group, at least theoretically, is a frequently suggested approach for balancing observed control variables in both groups (Baser, 2006). The propensity score, confounding covariates or both can be used to match members of the condition and the control group to achieve covariate balance in both groups (Sekhon, 2011). Although this approach is theoretically appealing, in practice, balance is difficult to obtain, and researchers must repeatedly specify the propensity score model to approximate covariate balance between groups (Austin, 2009). Subsequently, several balance measurements for checking before and after matching have become available (Belitser et al., 2011).

Guidance on the use of matching is given by Stuart (2010), where the different available parameters to reduce bias due to covariates by choosing well-matched samples is explained. For example, matching can be done with replacement, which means that the controls can be used as a match multiple times. If the inclusion of multiple matched control observations because one treated observation matches more than one control observation is allowed, then ties have to be handled. Furthermore, whether an exact match is required or a defined distance between individuals is possible can be specified. For the matching problem, Sekhon (2011) proposed a genetic matching algorithm that automatically maximises covariance balance.

After matching, the average treatment effect $\hat{\Theta}$ can then be estimated from the matched sample in an unbiased manner under the assumption of there being no unobserved confounder by the difference in the means of the outcomes between both groups. Eq. (2) shows that for our example, the estimated average treatment effect (provided by the R-package for matching the type of estimand that can be specified) is equal to the difference in the (estimated) probabilities (or proportions) of the 'event' between the treatment group p_{Treat} and the control group p_{Control} from the contingency table (Section 2.3.1, Table 2).

Whereas matching can address only the balance of observed variables, researchers are also interested in what the effect of unobserved variables ('unobservables') might have been. Unobservables are the key advantage of randomisation in trials because with increasing sample size, randomisation automatically balances observed and unobserved covariates. The Rosenbaum bounds are used to test the robustness of conclusions to hidden biases from unobserved confounders (Rosenbaum and Rubin, 1983). The value Γ is the odds ratio of its effect on treatment assignment – i.e., how much an unmeasured confounder would increase the odds of the measured outcome.

The use of the Rosenbaum bounds requires independent and identical distributed data. In our example, there is a lack of this independence assumption. On the one hand, within multilevel data, observations within a cluster are not independent. On the other hand, matching with replacement may result in multiple uses of controls for different match units. Hence, a more modern method is required to handle such data assessing hidden bias (Zubizarreta and Keele, 2014). However, the genetic matching approach not only provides the advantage of reducing selection bias and being able to model the propensity score without specifying an outcome but also provides a means to assess 'hidden bias' from unobserved confounders.

After applying the genetic matching approach, the matched sample can then be further analysed, e.g., using generalised linear mixed models to adjust for clustering. Because of matching, the sample size is reduced and may reduce the power for the interested effect size estimation. However, with regression analysis, multiple effects are estimated with increasing requirements per degree of freedom. Matching avoids this problem because only the effect of interest has to be estimated. Nevertheless, matching may produce data with additional 'non-independent' observations, which then should be considered through analysis.

2.4.3. *Covariate adjustment*

The most common method for reducing selection bias is likely the inclusion of independent variables (covariates) in a multiple logistic regression model for dichotomous outcomes. Hence, analysis and bias adjustment are not separated. Including covariates within the regression model subsequently provides a conditional estimate of the treatment effect (given levels of the covariates), which could differ from the marginal effects. Therefore, the estimated coefficient from the model should be interpreted with caution.

The researchers are responsible for which covariates are considered to include into the model. One possibility is to use all suspected covariates that are relevant, but covariate adjustment methods are often limited in the possible number of covariates (D'Agostino, 1998), and if models include too many variables, they may fail to converge. Convergence failure in this context means that the model cannot be estimated computationally. An alternative approach to account for different covariates is to include the propensity score in the regression equation. This idea follows the same principles as outlined earlier but without conducting a matching procedure based on the propensity score. Instead, the propensity score is included as an additional covariate in the regression model. The new variable can then be included in the regression model as one covariate rather than as an amount of covariates to control for bias and to increase the precision of the treatment effect estimate. Including one or many variables decreases the sample size for each 'cell'; thus, models including more variables have a higher risk of non-convergence. Nevertheless, this method only adjusts for bias through a regression model (not independent from the outcome), and no hidden bias assessment is possible.

$$\begin{aligned}
\hat{\Theta} &= \hat{\mu}_{\text{Treat}} - \hat{\mu}_{\text{Control}} = \frac{1}{N_{\text{Treat}}} \sum_i Y_{\text{Treat},i} - \frac{1}{N_{\text{Control}}} \sum_i Y_{\text{Control},i} \\
&= \frac{\text{no(Event=yes | Group=Treat)}}{\text{no(Group=Treat)}} - \frac{\text{no(Event=yes | Group=Control)}}{\text{no(Group=Control)}} \\
&= \text{P(Event=yes|Group=Treat)} \\
&\quad - \text{P(Event=yes|Group=Control)} \\
\Leftrightarrow \hat{\Theta} &= p_{\text{Treat}} - p_{\text{Control}}
\end{aligned} \tag{2}$$

In summary. The two methods to adjust for selection bias introduced here are matching using the genetic algorithm and adjustment within the regression model estimation. The first balances the sample independent from the outcome and provides a means of assessing 'hidden bias' from unobserved confounders. The other method adjusts for selection bias by estimating the contribution of each variable to the outcome within a regression framework. However, adding more variables can decrease statistical power in small samples because it increases the variance around the regression estimate by decreasing the number of degrees of freedom (Starks et al., 2009). Hence, for both matching and regression, the propensity score alone or in tandem can be used to achieve balanced samples. A combination of propensity score adjustment for a subset of covariates and covariate adjustment for the other is also possible.

2.5. *Estimation of treatment effect*

Although the parameter of interest is the average treatment effect during the analysis of our example study with binary outcomes, it corresponds here to the odds ratio or risk difference. Austin (2007) discussed different estimation methods in addition to the crude model and other propensity score methods being needed to assess the average treatment effect. These suggested methods have substituted using several covariates for using only the propensity score.

2.5.1. *After matching*

In addition to the marginal odds being calculated directly from the contingency table of the matched sample (for example, after propensity score matching), another possible method is model based. A logistic regression model with only one predictor variable for the assignment to the treatment group is fitted on the matched sample to estimate the impact of the treatment on the change in the odds of the outcome. This is also possible for the mixed model variant. The exponential parameter e^{β_1} from this model (Table 3) is therefore an estimate of the marginal odds ratio.

2.5.2. *After covariate adjustment*

The logistic regression model, which includes several covariates to adjust for their imbalance, provides only a conditional estimate of the treatment effect (by transformation of the coefficients as previously described), and the interpretation is in terms of adjusted changes in the corresponding covariates. Hence, the average treatment effect is available as the odds ratio/risk difference marginalised over the distribution of the included covariates. Therefore, the predicted probabilities for each individual given the confounders (sample data) are estimated under the treatment condition and under the control condition. The calculated mean probabilities $\overline{p}_{\text{treat}}$ and $\overline{p}_{\text{control}}$ can then be used to provide an estimate of the marginal odds ratio using $\frac{\overline{p}_{\text{treat}}}{1-\overline{p}_{\text{treat}}}\frac{1-\overline{p}_{\text{control}}}{\overline{p}_{\text{control}}}$ and the marginal risk difference by $\overline{p}_{\text{treat}} - \overline{p}_{\text{control}}$ (see also in the Vignette). Therefore, we use the logistic regression mixed model for our multilevel data.

3. Results

3.1. *Crude model*

According to the crude model (Table 2), 91% (n = 224) of residents in dementia special care units received a case conference, whereas only 80% (n = 470) in traditional care units received a case conference. The substantive interpretation would be that a patient is more likely to receive a case conference in dementia special care units than in traditional care units. Using the base logistic regression model with one binary predictor (dementia specific care unit or traditional care unit), i.e., the estimated model parameters β_0 (Intercept) or β_1 (dementia specific care units: treat), indicates that the model specification is not substantively different from the crude model. Hence, the retransformed values are equal to the estimates of the crude model. Table 3 shows how to obtain the transformed estimates of the logistic regression model estimates of Table 3, which were calculated as explained in Section 2.3.1.

The table shows that with the logistic regression model, the odds of obtaining the condition, i.e., the (estimated) probability of receiving a case conference versus not, is 3.95 in the control group and 10.18 in the treatment group. Hence, this results in an odds ratio of 2.58, which indicates that the odds of receiving a case conference is more than two and a half times higher in the treatment group than in the control group; in other words, being in the treatment group (relative to the control group) raises the odds of receiving a case conference. Using the inverse logit function, we can also provide the estimates in terms of probabilities (right column of Table 3) of receiving a case conference: 91% receive case conferences in the treatment group compared to 80% in the control group. These values indicate that an additional 11.3% receive a case conference in dementia special care units than in traditional care units. Table 1 in supplemental material also shows that the confidence interval of the odds ratio (corresponding to β_1) does not include 1. Hence, the difference of 11.3% between the two groups is assumed not to be random. We conclude that dementia special care units more often provide case conferences than traditional care units.

3.2. *Generalised linear mixed model and adjustment methods for bias reduction*

The results of the generalised linear mixed model are presented in Table 2 in the supplemental material. The odds ratio of 8.23 (see also in Table 4) is more than three times higher than the odds ratio in the generalised linear model ignoring the clustered data. However, the confidence intervals also increase (see the estimated confidence intervals in the Vignette), thereby increasing the p values due to the odds ratio with a covariance structure reflecting the dependencies of the observations. Although the precision of the estimates decreases, which may result from convergence problems with the estimation approach, adjustment is necessary to ensure that we do not overestimate our results.

3.2.1. *Propensity score estimation*

To address selection bias, we estimate the propensity scores for each observation. We model the group assignment using an additional generalised linear model that includes all individual related covariates (for general and health-related characteristics, see (Palm et al., 2014)) as fixed effects (no interaction terms). Fig. 3 in the supplemental material shows the unequal distributions of the estimated propensity scores between both groups and reflects the need for covariate adjustment to address selection bias.

Additionally, to account for the nested structure of the data and thus adjust for potential cluster-level unobserved confounders, we estimate the propensity scores using a generalised linear regression mixed model that includes the nursing homes as random effects (see the Vignette). However, this model failed to converge, and these estimates of the propensity scores could not be applied for further analysis, although Arpino and Mealli (2011) recommended their use as the matching variable in such multilevel settings, where the treatment is assigned at the individual level.

3.2.2. *Genetic matching and hidden bias assessment*

In this section, we show that either (1) all covariates or (2) only the propensity score as a summary of the covariates can be used for the genetic matching approach to balance the two groups. Furthermore, the matching quality due to hidden bias from unmeasured variables after matching is examined using the Rosenbaum bounds.

Initially, we used the genetic matching approach to determine the optimal covariate balance in the matched sample permitting replacements. The choice of the specific variables was based on theoretical considerations and generating a balanced sample in terms of general and health-related characteristics (see Palm et al., 2014). The algorithm samples a subset of 246 observations from each group out of the original sample, which is limited by the number of observations within the treatment group. For the outcome of interest, the estimated average treatment effect, the estimated average causal effect, and hence the (estimated) difference in probabilities between the two groups are 0.13. This value corresponds to an estimated treatment effect using the crude model on the subset of the data received from matching. According to a simple *t*-test, this difference in means is assumed to be significant. To implement a balance check after matching, the Vignette shows a variety of univariate standardised statistics being employed for each covariate proposed by Sekhon (2011), and the result shows (Vignette) that for all given covariates, balance is achieved by matching.

Using the Rosenbaum bounds at this point provides the opportunity to assess the matching quality due to hidden bias from unmeasured variables. Unfortunately, the significant *p*-value upper bound of 0.05 will be exceeded by a hidden variable with a Γ of 1.2. If we allow a *p*-value upper bound of 0.1, then it will be exceeded by a hidden variable with a Γ of only 1.3, indicating that an unobserved covariate that produces only a 1.2–1.3-fold increase in the odds of the group assignment would change the *p*-value to non-significance. Therefore, we would conclude that the matched sample is sensitive to hidden bias.

Nevertheless, we intend to use this matched sample for further analysis to compare it with the results of the unmatched sample, which is definitely biased. For the matched sample, we used the same basic generalised linear mixed model with one binary factor (dementia specific care unit or traditional care unit) that we used for the unmatched sample.

Table 4 shows the results interpreted in terms of (estimated) odds (left columns) and probabilities (middle columns) for the model adjusted for the clustered structure and adjusted for selection bias using the matched sample (third row). The odds ratio of 3.9 is half the odds ratio estimated with the same model from the unmatched sample (Table 4). The null hypothesis of no difference in the use of the conditions between the two groups would not be rejected.

Second, we use the genetic matching approach with the propensity score as the only covariate to determine the optimal propensity score balance. To check the balance after matching, the overlapping coefficient of the propensity score proposed by Belitser et al. (2011) is implemented in the provided Vignette, and the results show (supplemental material, Fig. 4) that matching based on propensity score was successful. Using the Rosenbaum bounds to check the matching quality for unobserved variables results in the significant *p*-value upper bound of 0.05 (0.1) being exceeded by a hidden variable with a Γ of only 1.8 (2). The estimates of the odds differ (Table 4), e.g., the odds of receiving a case conference in the treatment group, which for the propensity score-adjusted model are more than twice the odds from the covariate-adjusted model (46.78 vs. 40.33), and the estimated odds ratios (4.81 vs. 3.9) for both models. However, the null hypothesis of no difference between the two groups cannot be rejected.

Since Pimentel et al. (2015) and Zubizarreta and Keele (2014), accounting for multilevel structure within the matching process and balancing checks after matching are possible. These modern methods should be considered in the future when analysing multilevel data from observational studies.

3.2.3. *Covariate adjustment*

Including other relevant covariates as fixed effects instead of using the propensity score as the single indicator of group assignment within the effect estimation model to adjust for selection bias is possible. However, non-convergence occurs when there is too little data for the number of parameters or when the proposed model is not suitable for the given data. Hence, the choice of which variables should be included in the model can be based on the degree of significance in the difference between the treatment and control groups in the baseline analysis. Furthermore, variables could be included as fixed or random effects as long as the model converges. Here, we include three variables as fixed effects and two as random effects.

Nevertheless, the model needs a considerable amount of computation time to estimate the parameters and confidence intervals due to the number of included variables. The results presented in Table 4 (bottom row) show that the odds ratio of 6.99 between the two groups is not significant.

Rather than all covariates being used as fixed effects, it is possible to include only the propensity score, a continuous variable, as the single indicator of group assignment in the generalised linear mixed model in order to adjust for selection bias. The results show that the odds ratio of 6.3 between the two groups given a fixed PS value is not significant (Table 4).

The additionally estimated marginal risk differences of the generalized linear mixed model including the propensity score (in brackets, Table 4) are of comparable size to that provided by the basic generalised linear mixed model.

4. Discussion and conclusion

4.1. *In summary*

In this article, we, with the aid of a real study example, illustrate different methods to analyse data with selection bias and clustering and with a dichotomous outcome. Additionally, we provide a vignette as the supplementary material to enable readers to follow a full analysis of this study example in *R* and to adapt this method for other studies. For our study example, Table 4 presents the results for all models and methods and highlights the marked difference between the applied methods and the computed estimates. For our example, addressing the dependencies with a mixed model has a more pronounced impact on the estimation of odds ratio than adjusting for selection bias. This considerable difference can be explained by the strong clustering effect present in these data. Nevertheless, there is a greater difference in the *p*-values of testing the null hypothesis between using bias reduction methods or not than adjusting for dependent data structures.

Although the different analysis methods present different results, they at least point in the same direction, indicating that the estimated probability of receiving a case conference is higher in the treatment group than in the control group. However, in our study, when adjustment for bias and dependencies is performed, the null hypothesis of a difference in the use of condition between the two groups could not be rejected. Although there is a hint that there could be a difference, this difference could not be detected in this study due to the resulting sample and the limits of the study design.

Table 4

Estimated probabilities and differences between groups, odds and odds ratios using different models – GLM = generalised linear model or GLMM = generalised linear mixed model, and GLMM with additional methods for bias reduction, whereby 1 = genetic matching using propensity score, 2 = genetic matching using covariates, 3 = covariate adjustment using propensity score, 4 = covariate adjustment using several covariates. The table shows the marginal treatment effect, the odds ratio for both groups and the risk difference. The conditional treatment effects given by the model are also shown in brackets. The *p*-value is the probability of the Wald test statistic for the null hypothesis of no difference between the two groups.

	p_{Control}	p_{Treat}	Difference	$\text{Odd}_{\text{Control}}$	$\text{Odd}_{\text{Treat}}$	Odds ratio	p
GLM	0.80	0.91	0.11	3.95	10.18	2.58	< 0.01
GLMM	0.86	0.98	0.12	6.28	51.72	8.23	0.03
1	0.91	0.98	0.07	9.72	46.78	4.81	0.14
2	0.91	0.98	0.06	10.35	40.33	3.90	0.16
3	0.79 (0.83)	0.92 (0.97)	0.12 (0.14)	3.84 (4.98)	11.18 (31.40)	2.91 (6.30)	0.07
4	0.80 (0.73)	0.91 (0.95)	0.12 (0.22)	3.90 (2.72)	10.63 (18.98)	2.72 (6.99)	0.10

4.2. Model choice for estimation

In Fig. 1, the model choice is determined by the data characteristics, e.g., outcome type and study type and design. Our example shows that choosing the model to estimate effects within observational studies is also closely related to different key issues, such as unobserved variables, sample size and the study objectives. Here, we adjust for clustered data by using the generalised linear mixed model, where several methods are available to reduce selection bias: genetic matching and regression with propensity score or covariates. In our opinion, no single method for bias adjustment is optimal, and each approach has its own limitations and is open for discussion.

Propensity score. Since Rosenbaum and Rubin (1983), the propensity score has become increasingly popular for adjusting selection bias via matching methods or regression. However, there is a debate regarding the use of propensity scores to recover causal effects from observational studies. First, regression adjustment is not a recommended way to use the propensity score (Austin et al., 2007). Second, the propensity score is criticized for having the drawback of losing potentially useful information about predictors of outcomes (Stürmer et al., 2006). However, using the propensity score as the matching variable can circumvent the problem of having too many variables (Cepeda et al., 2003). Additionally, it has the advantage of balancing on a large number of covariates in one summarising variable, where finding matches for a large number of variables is nearly impossible (Starks et al., 2009).

Genetic matching. The genetic matching approach provides the ability to balance group allocation to imitate randomisation, such as the bias adjustment being independent from the outcome. Because only a limited number of covariates for adjustment could be used, traditional matching is often limited (D'Agostino, 1998). Hence, instead of the propensity score being used as the summary, it can be used to balance the covariates in two groups. However, regardless of whether the propensity score or covariates is employed, using the matching approaches has the trade-off of losing a large proportion of observations, which then influences the estimand (Wang, 2009). In our example, this resulted in 246 observations, some of which are sampled more than once to receive the matched sample. Furthermore, the resulting samples could be prone to hidden bias in unobserved variables, which is a general problem for all observational methods. Hence, in our example, there is a strong assumption of hidden bias after applying the methods. Ridder and Graeve (2011) stated that if hidden bias is present, then matching using the propensity score has a comparable bias, but the precision of the estimates is lower. However, matching provides balance checks and hidden bias assessment, which is not possible in a regression framework. Furthermore, after this approach, the sample can be analysed with small and simple models to estimate the interested effects. In our examples, this model has only one fixed effect and a random structure.

Regression. Bias adjustment via regression includes the adjustment variables within the estimation model; thus, the additional subsequent step of sample matching is not required. Since the simulation study of Wang (2009) found that using the propensity score provides biased effect estimates, the advantage of using the propensity score within regression to adjust for known confounders was demonstrated in small datasets by Biondi-Zoccai et al. (2011) and particularly for dichotomous outcomes (Cepeda et al., 2003). The reason for this result is that adding more variables can decrease statistical power in small samples and that using the propensity score instead produces similar estimation results with limited power (Starks et al., 2009). Using all covariates as fixed effects in the model instead of the summarised value of propensity score is also possible. Then, no information is lost, but unfortunately, the convergence of full models is often not possible or the confidence intervals of the estimates are too large because of the small number of observations in each case. This was the case in our study example, where only a subset of all covariates could be included in the regression model.

4.3. Study design and sample size in observational studies

On the one hand, in the literature, there is a demand for robustly designed observational studies to avoid as much selection bias as possible (Ellenberg, 1994), for example, a high participation rate to achieve a representative sample of the population (Hammer et al., 2009). On the other hand, if selection bias occurs and one adjustment method has to be chosen, the goal is to obtain the best method for removing bias while ensuring optimal estimation results. Therefore, a very precise estimate is not useful if it is drastically wrong, and thus, an estimate with a small bias rather than a small variance should be more convincing (Rubin, 2006). Hence, before collecting data for an observational study, two major concerns should be taken into account: (1) covariates that may obtain selection bias and hence require measurement are determined and (2) a larger sample size is needed to ensure a sufficient sample size; although there is a loss due to adjustment methods, the former concern has the most important effect. There is a demand for the ability to calculate the sample size that is needed for a sufficient estimate quality, although there is also a need to adjust for an assumed selection bias before the data are collected. Hence, further investigations should be performed to permit drawing conclusions regarding the minimum required sample size within observational studies, which has to be adjusted for bias, or, if bias appears, how much of the sample is being lost via matching.

4.4. Limitations

Several limitations of our presented model should be discussed and explained.

First, although the participants were nested in care units, which in turn were nested in nursing homes, we decided to use a 2-level instead of a 3-level mixed model for analysis, whereby the care units could also explain a part of the variation (for intra-class correlation coefficient estimates of different levels from the regression model, see the Vignette). Due to the given data structure – more than half of the nursing homes provided participants in only one care unit – we chose a simple model only using nursing homes as random effects, knowing that more than half of the variability was explained by it.

Second, we could not provide the balance of various characteristics at the different levels of the data, i.e., not at the care unit level (Leyrat et al., 2014) or at the nursing home level (Belitser et al., 2011). Due to our decision to use nursing homes as a level of clustering, an adjustment at that level has some limits. In our example, we ignore the multilevel structure in both the propensity score estimation model and the matching implementation. On the one hand, the most straightforward idea is to force matching within each cluster, provided that treated and control cases are available within each cluster. However, this approach is very difficult to realise when the cluster sizes are small and may yield a considerable loss of individuals (Arpino and Cannas, 2016). On the other hand, Arpino and Mealli (2011) proposed including the level of clustering in the propensity score estimation model if the treatment is assigned at the individual level in multilevel settings. Then, the adjustment via matching is not forced within clusters. However, in our example, this model for the propensity score estimation did not converge, and we could provide bias adjustment at only the individual level. In this context, Li et al. (2013) show that accounting for a cluster structure in at least one stage, e.g., in the propensity score estimation or within the outcome model, can greatly reduce the bias. Nevertheless, upcoming studies faced with both existing bias and multilevel data should consider applying the modern methods of optimal multilevel matching (Pimentel et al., 2015; Zubizarreta and Keele, 2014), which can close the gap.

Furthermore, we did not use any procedure for model selection (e.g., the iterative process of the estimation and balance check) for the propensity score estimation itself. This decision was pragmatic and was based on using the individual related characteristics, which were obtained in this observational study and assigned in the literature as being correlated with the circumstance of being a resident of dementia specific care units or not. Nevertheless, studies with more possible given covariates and a larger sample size and with the aim of using propensity scores primarily for bias reduction should be considered during model selection in combination with balance checks (Belitser et al., 2011).

We also did not consider the lack of independence assumption, either for the regression model after matching or for assessing hidden bias using Rosenbaum bounds, due to the matching with replacement. Solutions for this induced problem that use matching methods were reported by Stuart (2010) and Zubizarreta and Keele (2014).

Finally, for handling the missing data of the 53 participants, a statistical strategy such as multiple imputation was not conducted, nor were the results validated using a sensitivity analysis (for more details, see Carpenter and Kenward, 2013; Little and Rubin, 2002). Such a complete case analysis reduces statistical power and estimate precision; additionally, estimates can be biased in some circumstances if the missing data are not randomly distributed (Bartlett et al., 2015).

Acknowledgements

The data used for the analysis in this paper were collected in the study DemenzMonitor (2012–2014), which was funded by the DZNE Site Witten. The principal investigator of the DemenzMonitor Study is Dr. Bernhard Holle.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijnurstu.2017.06.017.

References

American Association of Colleges of Nursing, 2015. Nursing Research – Position Statement. http://www.aacn.nche.edu/publications/position/nursing-research.

Arpino, B., Cannas, M., 2016. Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. Stat. Med. 35 (12), 2074–2091.

Arpino, B., Mealli, F., 2011. The specification of the propensity score in multilevel observational studies. Comput. Stat. Data Anal. 55 (4), 1770–1780.

Austin, P.C., 2007. The performance of different propensity score methods for estimating marginal odds ratios. Stat. Med. 26 (16), 3078–3094.

Austin, P.C., 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. 28 (25), 3083–3107.

Austin, P.C., Grootendorst, P., Anderson, G.M., 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat. Med. 26 (4), 734–753.

Bartlett, J.W., Harel, O., Carpenter, J.R., 2015. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. Am. J. Epidemiol. 182 (8), 730–736.

Baser, O., 2006. Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health 9 (6), 377–385.

Belitser, S.V., Martens, E.P., Pestman, W.R., Groenwold, R.H., de Boer, A., Klungel, O.H., 2011. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol. Drug Saf. 20 (11), 1115–1129.

Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., Capodanno, D., Castagno, D., D'Ascenzo, F., Sangiorgi, G., Modena, M.G., 2011. Are propensity scores really superior to standard multivariable analysis? Contemp. Clin. Trials 32 (5), 731–740.

Burns, N., Grove, S.K., 2009. The Practice of Nursing Research. Appraisal, Synthesis and Generation of Evidence, 6th ed. Saunders Elsevier, St. Louis.

Carpenter, J.R., Kenward, M.G., 2013. Multiple Imputation and Its Application. John Wiley & Sons Ltd, Chichester, UK.

Cepeda, M.S., Boston, R., Farrar, J.T., Strom, B.L., 2003. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am. J. Epidemiol. 158 (3), 280–287.

D'Agostino, R.B., 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med. 17 (19), 2265–2281.

Ellenberg, J.H., 1994. Selection bias in observational and experimental studies. Stat. Med. 13 (5–7), 557–567.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research. New York, Cambridge University Press.

Greenland, S., 1987. Interpretation and choice of effect measures in epidemiologic analyses. Am. J. Epidemiol. 125, 761–768.

Hammer, G.P., du Prel, J.B., Blettner, M., 2009. Avoiding bias in observational studies: Part 8 in a series of articles on evaluation of scientific publications. Deutsches Ärzteblatt Int. 106 (41), 664–668.

Hardin, J., Hilbe, J., 2012. Generalized Estimating Equations, 2nd ed. CRC Press, Boca Raton, FL.

Leyrat, C., Caille, A., Donner, A., Giraudeau, B., 2014. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. Stat. Med. 33 (20), 3556–3575.

Li, B., Lingsma, H.F., Steyerberg, E.W., Lesaffre, E., 2011. Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC Med. Res. Methodol. 11 (1), 77.

Li, F., Zaslavsky, A.M., Landrum, M.B., 2013. Propensity score weighting with multilevel data. Stat. Med. 32 (19), 3373–3387.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis With Missing Data. John Wiley & Sons, Inc., Hoboken, NJ.

Nguyen, T.-L., Collins, G.S., Spence, J., Daurès, J.-P., Devereaux, P.J., Landais, P., Le Manach, Y., 2017. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC Med. Res. Methodol. 17 (1), 78.

O'Connor, A., 2013. Interpretation of odds and risk ratios. J. Vet. Intern. Med. 27 (3), 600–603.

Ostir, G.V., Uchida, T., 2000. Logistic regression: a nontechnical review. Am. J. Phys. Med. Rehabil. 79 (6), 565–572.

Palm, R., Bartholomeyczik, S., Roes, M., Holle, B., 2014. Structural characteristics of specialised living units for people with dementia: a cross-sectional study in German nursing homes. Int. J. Ment. Health Syst. 8.

Palm, R., Trutschel, D., Simon, M., Bartholomeyczik, S., Holle, B., 2015. Differences in case conferences in dementia specific vs traditional care units in German nursing homes: results from a cross-sectional study. J. Am. Med. Dir. Assoc. 17 (1), 91.e9–91.e13.

Pimentel, S.D., Yoon, F., Keele, L., 2015. Variable-ratio matching with fine balance in a study of the peer health exchange. Stat. Med. 34 (30), 4070–4082.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Randolph, J.J., Falbe, K., Manuel, A.K., Balloun, J.L., 2014. A step by step guide to propensity score matching in R. Pract. Assess. Res. Eval. 19 (18).

Ridder, A.D., Graeve, D.D., 2011. Can we account for selection bias? A comparison between bare metal and drug-eluting stents. Value Health 14 (1), 3–14.

Rosenbaum, P., 2002. Observational Studies. Springer Series in Statistics. Springer, New York.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge, UK.

Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: the Matching package for R. J. Stat. Softw. 42 (7), 1–52.

Starks, H., Diehr, P., Curtis, J.R., 2009. The challenge of selection bias and confounding in palliative care research. J. Palliat. Med. 12 (2), 181–187.

Stroup, W., 2012. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, Boca Raton, FL.

Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Stat. Sci. Rev. J. Inst. Math. Stat. 25 (1), 1–21.

Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., Schneeweiss, S., 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J. Clin. Epidemiol. 59 (5), 437.e1–437.e24.

Wang, Z., 2009. Propensity score methods to adjust for confounding in assessing treatment effects: bias and precision. Internet J. Epidemiol. 7 (2).

Williamson, E., Morley, R., Lucas, A., Carpenter, J., 2012. Propensity scores: from naive enthusiasm to intuitive understanding. Stat. Methods Med. Res. 21 (3), 273–293.

Wu, S., Crespi, C.M., Wong, W.K., 2012. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp. Clin. Trials 33 (5), 869–880.

Zou, G., Donner, A., 2004. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 60 (3), 807–811.

Zubizarreta, J.R., Keele, L., 2014. Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System. arXiv preprint arXiv:1409.8597.

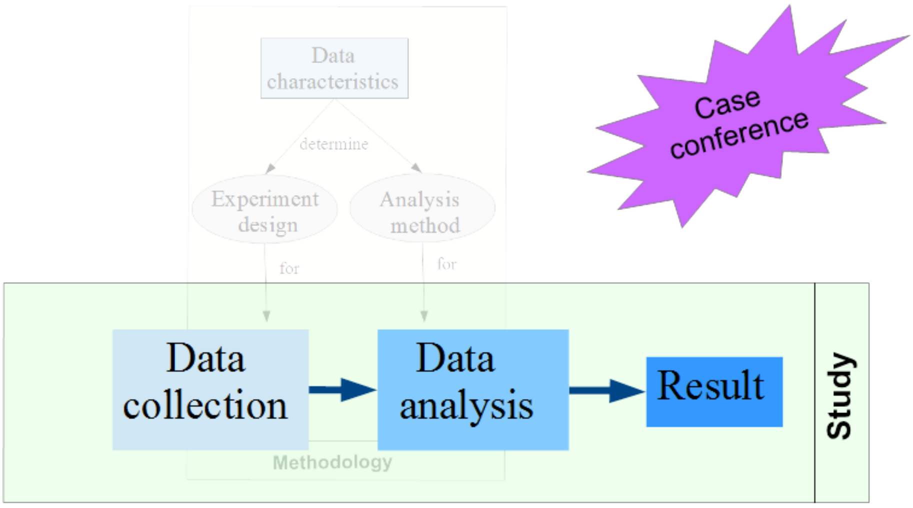
Zúñiga, F., Ausserhofer, D., Hamers, J.P., Engberg, S., Simon, M., Schwendimann, R., 2015. Are staffing, work environment, work stressors, and rationing of care related to care workers' perception of quality of care? A cross-sectional study. J. Am. Med. Dir. Assoc. 16 (10), 860–866.

6.5. Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study

JAMDA, 2016, Volume 17, Issue 1, Pages 91.e9-91.e13

Rebecca Palm[1,2], MSc,
Diana Trutschel[1,3],
Michael Simon[4,5], PhD,
Sabine Bartholomeyczik[2], PhD,
Bernhard Holle[1,2], PhD

[1]German Center for Neurodegenerative Diseases (DZNE) Site Witten, Witten, Germany
[2]Department of Health, School of Nursing Science, University Witten/Herdecke, Witten, Germany
[3]Martin-Luther University Halle/Wittenberg, Institute of Informatics, Halle/Saale, Germany
[4]Faculty of Medicine, Institute of Nursing Science, University of Basel, Basel, Switzerland [5]Nursing and Midwifery Research Unit, Inselspital Bern University Hospital, Bern, Switzerland



Abstract:

Objectives To investigate differences in the provision and performance of case conferences for people with dementia between dementia special care units (DSCUs) and traditional care units (TCUs) in nursing homes. Because DSCUs employ more staff, we expect the likelihood of the provision of case conferences to be higher in DSCUs.

Design Observational cross-sectional study. Residents from DSCUs and TCUs were compared using genetic propensity score matching over all of the observed potential covariates, including the characteristics that served as admission criteria for DSCUs. Because of the multisite structure of the data, clustering was accounted for with a generalized mixed model.

Setting DSCUs are defined as units within nursing homes that offer care exclusively to residents with dementia and that charge higher rates for the specialized care provided. TCUs are defined as care units for residents with and without dementia.

Participants A matched sample was drawn out of a convenience sample of 1808 residents from 51 nursing homes. It consisted of 264 residents from 16 DSCUs and 264 residents from 48 TCUs.

Interventions None.

Measurements Data regarding the provision of case conferences were collected by the nurses using the Dementia Care Questionnaire. Other collected data included challenging behavior (Neuropsychiatric Inventory Questionnaire), mobility (Physical Self-Maintenance Scale), cognitive impairment (Dementia Screening Scale), and sociodemographic information.

Results In the DSCU group, case conferences were provided to 91% (n = 224) of the residents; in the TCU group, 82.5% (n = 203) received a case conference. After adjusting for clustering, no significant difference between DSCUs and non-DSCUs was found. The topic of challenging behavior was discussed more often in case conferences in TCUs.

Conclusions Case conferences are a widespread intervention in German nursing homes, including both DSCUs and TCUs. The provision of a case conference is not a special feature of DSCUs.

Keywords: Dementia, dementia special care units, case conferences, nursing homes

full article on pages 80-84 is available under
`https://www.jamda.com/article/S1525-8610(15)00557-5/fulltext`

A. Appendix - Methodological details

This appendix includes three parts with more detailed information for a better understanding of the main body of this dissertation.

The first gives a short overview about the hypothesis testing theory, test quality and the performance of simulations, which are base methods for extracting information from studies including in all articles. For more detail a wide literature is published [1][2][3][4][5][6][7].

In the second, the multivariate distribution for normal distributed data in nested experiments is derived. Since this is the base distribution of the data used in almost all articles it is an important key issue of this thesis determining the applicability of tests. This known multivariate distribution is the base of the provided *samplingDataCRT* [8] - an easy tool for sampling data sets with such depended structures. Furthermore, it is shown that different approaches (traditional frequentists or likelihood ration based) provide similar test statistics for experiments to find difference between groups using nested data (used in some of the articles) and additionally their relationship to traditional test statistics for non nested data.

In the last part, for all included articles a selection of the additional informations provided by their supplemental material is given.

A.1. Hypothesis testing theory, test quality and performance of simulation studies

Hypothesis testing Statistical hypothesis testing theory is a widely used method for statistical inference. Thereby statistical hypothesis is a statement about the characteristics of random variables, a parameter or distribution of a population e.g., and represents a scientific hypothesis. Data analysis use the information gained from a sample of individuals in order to make inference about the relevant population [9]. Hence, statistical hypothesis testing is a procedure that based on parameters estimation out of a sample, a subset of the whole population, on which the statistical hypothesis should be evaluated.

The interested hypothesis can normally not confirmed directly. Hence,the opposite hypothesis, called null hypothesis, is tried to refute. Table A.1 (upper part) shows the decision situation within hypothesis testing problems. Making a decision for or against the null hypothesis most of the statistical tests use a test statistics T, a random variable calculated as a function out of the data sample or samples.

The test statistic has also a specific distribution, under null hypothesis (H_0) as well as under alternative hypothesis (H_1). Hence, a probability under distribution of the null hypothesis for the test statistic value or a even higher value can be calculated. Figure A.1 illustrates graphically the formulas of conditional probabilities, which are equal to the areas under the distribution curves for given regions. As the significance level determines the critical region

[1] Sachs, Lothar and Hedderich, Jürgen, "Angewandte Statistik", *Springer*, 2009

[2] Fahrmeir et al., "Statistik", *Springer*, 2010

[3] Hilgers, Ralf-Dieter et al., "Einführung in die Medizinische Statistik", *Springer*, 2007

[4] Rasch et al., "Quantitative Methoden 1 + 2", *Springer*, 2010

[5] Bortz, Jürgen, "Statistik für Human- und Sozialwissenschaftler", *Springer*, 2005

[6] DeGroot, Morris H., "Optimal Statistical Decisions", *Wiley*, 2004

[7] Jones et al., "Introduction to Scientific Programming and Simulation Using R", *Springer*, 2014

[8] `https://CRAN.R-project.org/package=samplingDataCRT`

[9] Ilakovac, Vesna, "Statistical hypothesis testing and some pitfalls", *Biochem Med*, 2009

		Population follows	
		H_0	H_1
Decisions out of sample	H_0	correct decision	error type II
	H_1	error type I	correct decision
Cond. Probabilities		$\mathbf{1-\boldsymbol{\alpha}} = P(T \notin CR_\alpha \mid H_0)$	$\boldsymbol{\beta} = P(T \in CR_\alpha \mid H_1)$
		$\boldsymbol{\alpha} = P(T \in CR_\alpha \mid H_0)$	$\mathbf{1-\boldsymbol{\beta}} = P(T \notin CR_\alpha \mid H_1)$

Table A.1.: Upper: Errors within testing decisions problems occur, if null hypothesis will be rejected even though it is true (α is error type I) or the null hypothesis will not rejected given the alternative hypothesis is true (β is error type II). Lower: Decisions are made using a given critical region CR_α, determined from the distribution of the test statistic under null hypothesis. Given the distributions under both hypothesis, the probabilities of errors and correct decisions can be calculated.

CR_α of the test statistic under the null hypothesis distribution, Table A.1 (lower part) shows the probability calculation of correct and error decisions corresponding to decision table (Table A.1, upper part) using the distributions illustrated in Figure A.1.

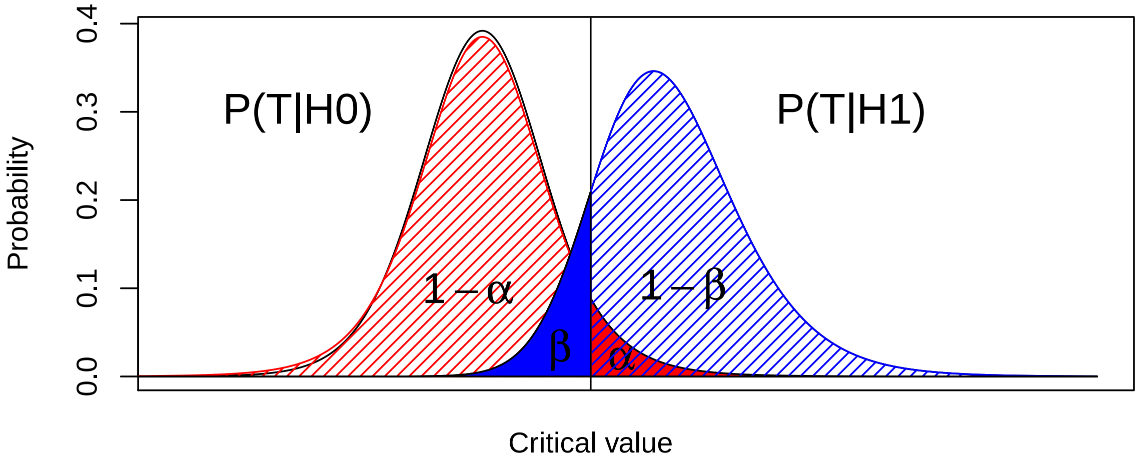


Figure A.1.: Probabilities of errors and correct decisions can be calculated as the areas under the distribution curves for given regions. The probability of the error type I α describes a conditional probability. In the opposite β is the conditional probability, that the null hypothesis will not rejected given the alternative hypothesis is true. It is known as error type II probability. The complement $1-\beta$, the probability that the the null hypothesis will be rejected given the alternative hypothesis is true, is called power.

These probabilities can also be determined empirically using a binary classification problem simulate the decision problem. Imagine n independent experiments are performed to test the same null hypothesis H_0, the confusion matrix summarizes the outcomes of this n statistical tests, shown in Table A.2. The matrix reports the number of correct and error decisions with four numbers expressed in simple counts: false positives are equivalents to type I error, false negatives are equivalents to type II error, true positives and true negatives are equivalent to correct decisions. Divide these numbers by n, the total number of tests, the confusion matrix may be expressed in relative terms [10], the joint probabilities of population and decision.

[10] Powers, David M. W., "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", *School of Informatics and Engineering, Flinders University*, 2007

While the confusion matrix (Table A.2, top) has a more descriptive role[11], the classification performance of the test (Table A.2, bottom) can then determined by dividing these numbers by the column marginals. This measured proportions are estimates of the conditional probabilities of correct and error decisions. They are known with several synonyms under different contexts (Table A.3 and hence, Table A.1 and Table A.2 can be connected.

		Truth about hypothesis			
		H_0 is true	H_0 is false		
Predicted	accept H_0	TN	FN		**Abs. num.**
	reject H_0	FP	TP		
	Marginal	n_N	n_P	n	
		TNR $= \frac{TN}{n_N}$	FNR $= \frac{FN}{n_P}$		**Est. prob.**
		FPR $= \frac{FP}{n_N}$	TPR $= \frac{TP}{n_P}$		
				ACC $= \frac{TP+TN}{n}$	

Table A.2.: Confusion matrix of binary classification problem. The columns represents the instance of truth of the hypothesis, the rows the instances of the prediction. TP = True positives, TN = True negatives, FP = False positives, TN = True negatives, TPR = True positive rate, TNR = true negative rate, FPR =False positive rate, FNR = False negative rate, ACC = Accuracy.

The most interested measures to quantifying the discrimination ability of the test are sensitivity and specificity, because the aim is to perform a test with a minimal number of false decisions. On the one hand sensitivity or empirical power is the proportion of true positives that are correctly identified by the test and on the other specificity is the proportion of true negatives that are correctly identified [12]. Hence, 1-power is the type II error or false negative rate. A a test with a high sensitivity has a low type II error rate and test with a high specificity has a low type I error rate. Determine the performance measures using a binary classification problem is a common used method to evaluating statistical tests.

Furthermore, a alternative way to visualize this performance measures has been established. The receiver operating characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers [13]. The graph combines the sensitivity (TPR) and 1-specificity (FPR) of every observed data value [14]. The first is plotted on y-axis, the second on x-axis and so offers the opportunity select the relative trade-offs between benefits and costs. The area under this ROC curve (AUC) as a global assessment of the performance of a test [14]is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [13]. The more close the AUC is to one, the better is the classifier.

The appropriate study design choice is related to statistical hypothesis testing as it is related to effect size, sample size and power. Hence, design calculations before the study is conducted, should be done.

Simulation studies As data characteristics are related to the suitable experimental design and analysis method, they should be understood before realize a real study to answer a sci-

[11] Stehman, Stephen V., "Selecting and interpreting measures of thematic classification accuracy", *Remote Sensing of Environment*, 1997

[12] Altman, D. G. and Bland, J. M., "Statistics Notes: Diagnostic tests 1: sensitivity and specificity", *BMJ*, 1994

[13] Fawcett, Tom, "An Introduction to ROC Analysis", *Pattern Recogn. Lett.*, 2006

[14] Altman, D. G. and Bland, J. M., "Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots", *BMJ*, 1994

Measure (Synonym)	Formula	Estimate
True negative rate Specificity	P(accept$\|H_0$ true) $= \frac{P(\text{accept},H_0\text{true})}{P(H_0\text{true})} = \frac{TN}{n_N}$	$1-\alpha$
False positive rate Type I error rate Fallout Significance	P(reject$\|H_0$ true) $= \frac{P(\text{reject},H_0\text{true})}{P(H_0\text{true})} = \frac{FP}{n_N}$	α
False negative rate Type II error rate Miss-rate	P(accept$\|H_0$ false) $= \frac{P(\text{accep},H_0\text{false})}{P(H_0\text{false})} = \frac{FN}{n_P}$	β
True positive rate Sensitivity Power Recall	P(reject$\|H_0$ false) $= \frac{P(\text{reject},H_0\text{false})}{P(H_0\text{false})} = \frac{TP}{n_P}$	$1-\beta$
Accuracy	$P(\text{reject}, H_0\text{false}) + P(\text{accept}, H_0\text{true}) = \frac{TP+TN}{n}$	
Error rate	$1 - \text{Accuracy} = \frac{FP+FN}{n}$	

Table A.3.: Connection of measurement and estimate of probability due to the descision problems.

entific question. Hence, pilot studies and computational simulations can be helpful to get deeper insights of data characteristics and find a powerful experiment design (used in article of section 6.1). Whereby knowing the data characteristics is essential, validation using computational simulation studies can be used to find the sufficient statistical model to analyse the data (used in article of section 6.3). Hence, computational inference is a viable and useful alternative to inference in traditional statistics. Simulation studies can be performed to evaluate and compare experimental designs or analysis methods [15]. The role of simulation as a numerical technique is to perform experiments using computer intensive procedures, which answer questions cannot be achieved with studies on real data alone [16]. Two types were classified by the type of input data: they are derived from 1) measured data of a real system or 2) via sampling from probability distribution using random numbers [17], which is known as Monte Carlo simulation. The aim is to test particular hypotheses and asses the performance of a variety of statistical methods in relation to a known truth. Several scenario should reflect the most common circumstances. Then methods can be tested and compared referring to a) accuracy of estimation method or b) the quality of hypothesis testing approaches.

The performance of methods and scenarios related to the estimation method can be evaluated with different criteria (Table A.4). The main interest of a simulation study lies in the expected estimates. Calculate the averaged estimates of all simulation repeats, then bias is, due to estimator itself, defined as derivation of the average measured from the true value of the estimate of interest, and also called systematic error [18]. Hence, the smaller the bias is

[15] Gentle, James E and Härdle, Wolfgang Karl and Mori, Yuichi, "How to computational statistics became the backbone of modern data science", *Handbook of computational statistics: concepts and methods*, Ch. I, 2012

[16] Burton, Andrea and Altman, Douglas G. and Royston, Patrick and Holder, Roger L., "The design of simulation studies in medical statistics", *Statistics in Medicine*, 2006

[17] Balci, Osman, "Guidelines for Successful Simluation Studies (Tutorial Session)", *Proceedings of the 22Nd Conference on Winter Simulation*, 1990

[18] Walther, Bruno A. and Moore, Joslin L., "The concepts of bias, precision and accuracy, and their use in

the more accuracy the method. A possible measure of overall accuracy is the mean squared error (MSE) as it includes both estimation bias and variability [16]. Therefore, the empirical standard error (SE) of the estimates of interests over all simulations of a simulation study as a measurement of variability. But in principle any measure of variability of the estimates can be used to quantify precision. Hence, bias, precision (or adverse variability) and accuracy are qualitative concepts to quantify the performance of estimators [18]. In simulation studies searching for accurate point estimators less bias and most precision is required.

Evaluate performance using the quality of hypothesis tests coverage and empirical power are measurements to control the Type I and the Type II error rates. The type I error rate for testing a null hypothesis of no effect can be controlled by the coverage. This is the proportion of times that the obtained confidence intervals over all simulation repeats contain Θ, the true value of parameter of interest [16]. Thereby, the confidence interval of each repeat i is defined as follows $\hat{\Theta}_i \pm Z_{1-\alpha/2} SE\left(\hat{\Theta}_i\right)$, where $SE\left(\hat{\Theta}_i\right)$ is the standard error of interested estimate within each repeat depending on the estimation method and $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. If two or more measures will be compared, a ROC plot is particular useful [14]. The AUC is then an indicator of the measure with the best discriminate power.

Characteristic	Formula	Measure
Mean error (ME)	$\bar{\hat{\Theta}} - \Theta$	Bias
Empirical standard error ($SE(\hat{\Theta})$)	$\sqrt{\frac{1}{S-1}\sum_{i}^{S}\left(\hat{\Theta}_i - \bar{\hat{\Theta}}\right)^2}$	Precision
Mean squared error (MSE)	$(ME)^2 + (SE)^2$	Accuracy

Table A.4.: Performance characteristics related to the estimation method can be measured within simulation studies: bias measures the derivation of the estimates average $\bar{\hat{\Theta}}$ from the true value of the paramter of interest Θ, Empirical standard error as a measurement of variability over all S number of performed simulation repeats within the study and mean squared error, which includes bias and variability.

testing the performance of species richness estimators, with a literature review of estimator performance", *Ecography*, 2005

A.2. N-dimensional distribution of observations in hierarchical experiments and the derivation of a statistical test

In experiments with nested data (used in articles of this thesis) the data follow a multivariate normal-distribution, which is shown here. This is the base of the provided *samplingDataCRT* [8] to sample data for example in cluster randomized studies.

Since the distribution of the data determines the applicability of tests, different possible test statistics for testing differences between groups are derived here.

A.2.1. Assumption

A hierarchical experiment process produces measurements at different levels. For example an 2-level hierarchical design consists of measurements X_{ij} of N observations $(j = 1, ..., N)$ at the first level, each with K observations $(k = 1, ..., K)$ at the second level, which are repeated measurements of the first level observations. The observations at the first level vary around their mean with a variance σ_1^2 and the observations below with another different variance σ_2^2. Hence, it results in total $N \times K$ (normal distributed) measurements, each vary around the class mean with a variance resulting from the sum of the two level variances $\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2$.

In this section an experiment example of two classes with such hierarchical structured data are assumed (Figure A.2). The aim is to test the mean of the two classes for equality. One example could be: within each of two gentotypes (classes) measurements of N plants or individuals (at level 1: biological replicates), each measured K times (at level 2: technical replicates) with a mass spectrometer, hence measure the amount of features/metabolites intensities (measurements).

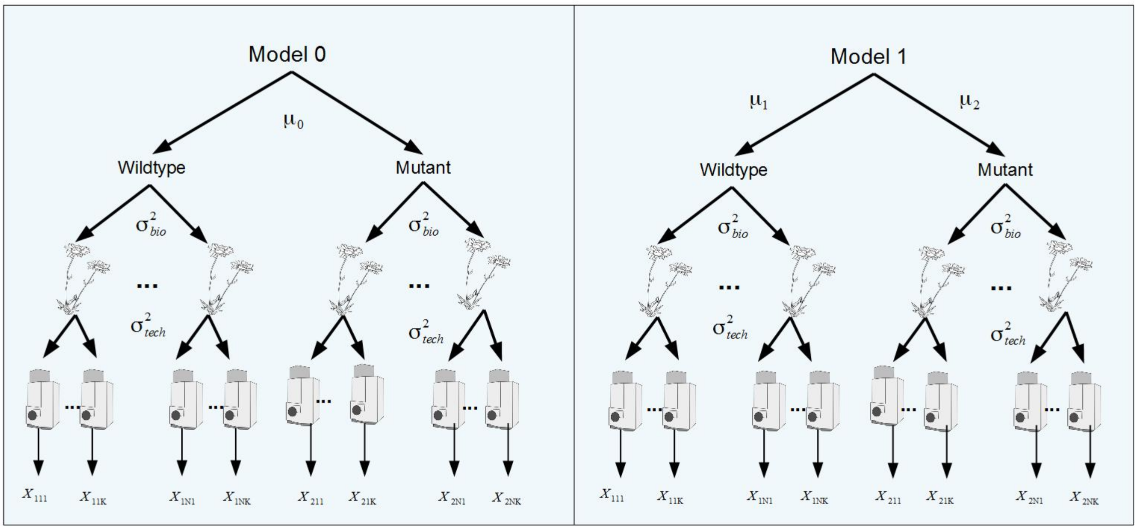


Figure A.2.: Concepts of a hierarchical model within a biological experiment

Whereby, the intensity mean of each class is unknown, two assumptions are possible: 1) the two classes have the same mean of intensities or 2) have not. The first can be understood as a reduced model (model 0), the second as a saturated (model 1). Hence, in this biological example the assumptions on the distribution of the observations X_{ijk} in one class i (wildtype or mutant) can be summarized as follows:

- observations X_{ijk} are normal distributed $X_{ijk} \sim N(\mu, \sigma_{total}^2)$ with

$$E\left(X_{ijk}\right) = \mu_i \qquad \forall j,k$$

$$Var\left(X_{ijk}\right) = \sigma^2_{total} \qquad \forall i,j,k$$

- the total variance σ^2_{total} is a sum of:
  - between individual or biological variance $\sigma^2_1 = \sigma^2_{bio}$
    (corresponds to the within class variance)
  - within individual or technical variance $\sigma^2_2 = \sigma^2_{tech}$
    (corresponds to the sum of repeated measurment + error variance)
- different measurements of an individual are dependent, measurements between different individuals are independent

$$Cov(X_{ijk}, X_{ij\tilde{k}}) = \sigma^2_{bio} \qquad \forall k \neq \tilde{k} \qquad Cov(X_{ijk}, X_{i\tilde{j}\tilde{k}}) = 0 \qquad \forall j \neq \tilde{j},$$

With this assumptions, a test statistic could be perfomed in the following, first, using the traditional frequentists approach and second, the likelihood ratio approach.

A.2.2. Traditional frequentist approach using the linear mixed model notation

The traditional frequentists approach uses the distribution of the random variable, which determines the null hypothesis H_0. If this distribution is known, the test statisitic has to be calculated from the given data. Then the null hypothesis could be rejected, if the probability of the calculated value or a smaller one do not exceed a determined threshold (see A.1 for explaination of hypothesis tests). With here given formula A.9 the test statistic used in the analysis of the first article (section 6.1) is derived. It can be found in article's appendix (section B.1) in comparison to test statistics used for non-nested data and also tests for more than two sample classes. For more details of formula derivation in such experiments see [19].

In our biological example, the normal distributed random variables $X_{ijk} \sim N\left(\mu_i, \sigma^2_b + \sigma^2_t\right)$, $(j = 1, ..., N, k = 1, .., K)$ of two classes $(i = 1, 2)$ are given. The assumptions, written in the linear mixed model notation, are the following:

- $X_{ijk} = \mu_i + \beta_{j|i} + \epsilon_{k|ij}, \qquad E\left(X_{ijk}\right) = \mu_i$
- $\beta_{j|i} \sim N\left(0, \sigma^2_b\right)$
- $\epsilon_{k|ij} \sim N\left(0, \sigma^2_t\right)$

The aim is to make a statement about the equality of the two class means μ_1 and μ_2 using the samples means $X_{1..}$ and $X_{2..}$. Following the the classical approach the distribution of these means and therefore, their expected values and variances, are needed.

1. Mean of the technical replicates (repeated measurements):

$$\begin{aligned}
X_{ij.} &= \frac{1}{K}\sum_k X_{ijk} = \frac{1}{K}\sum_k \left(\mu_i + \beta_{j|i} + \epsilon_{k|ij}\right) = \mu_i + \beta_{j|i} + \frac{1}{K}\sum_k \epsilon_{k|ij} \\
E\left(X_{ij.}\right) &= E\left(\mu_i\right) + E\left(\beta_{j|i}\right) + \frac{1}{K}\sum_k E\left(\epsilon_{k|ij}\right) = \underline{\underline{\mu_i}} \\
Var\left(X_{ij.}\right) &= Var\left(\mu_i\right) + Var\left(\beta_{j|i}\right) + \frac{1}{K^2}\sum_k Var\left(\epsilon_{k|ij}\right) = \underline{\underline{\sigma^2_b + \frac{1}{K}\sigma^2_t}}
\end{aligned}$$

[19] Ahrens, Heinz, "Varianzanalyse", *Akademie-Verlag Berlin*, 1967

2. Mean of the biological replicates:

$$\begin{aligned}
X_{i..} &= \frac{1}{J}\sum_j X_{ij.} = \mu_i + \frac{1}{J}\sum_j \beta_{j|i} + \frac{1}{J}\sum_j \frac{1}{K}\sum_k \epsilon_{k|ij} \\
E\left(X_{i..}\right) &= E\left(\mu_i\right) + \frac{1}{J}\sum_j E\left(\beta_{j|i}\right) + \frac{1}{J}\sum_j \frac{1}{K}\sum_k E\left(\epsilon_{k|ij}\right) = \underline{\underline{\mu_i}} \\
Var\left(X_{ij.}\right) &= Var\left(\mu_i\right) + \frac{1}{J^2}\sum_j Var\left(\beta_{j|i}\right) + \frac{1}{J^2}\sum_j \frac{1}{K^2}\sum_k Var\left(\epsilon_{k|ij}\right) = \underline{\underline{\frac{1}{J}\left(\sigma_b^2 + \frac{1}{K}\sigma_t^2\right)}}
\end{aligned}$$

The distribution of means of the technical replicates and the biological replicates for each class i is then given by (A.1), (A.2):

$$\boxed{\begin{aligned}
X_{ij.} &\sim N\left(\mu_i, \sigma_b^2 + \frac{1}{K}\sigma_t^2\right) \\
X_{i..} &\sim N\left(\mu_i, \frac{1}{J}\left(\sigma_b^2 + \frac{1}{K}\sigma_t^2\right)\right)
\end{aligned}} \qquad \begin{aligned}&(A.1)\\&(A.2)\end{aligned}$$

To test the equality of the class means a new random variable, the difference of the sample class means $\overline{D} = X_{1..} - X_{2..}$ ($X_{1..}$, $X_{2..}$ independent), and their corresponding distribution is required. Hence, the expected value and the variance of this difference $\overline{D}$ is derived, the distribution of $\overline{D}$ follows by (A.3):

$$\begin{aligned}
E\left(\overline{D}\right) &= E\left(X_{1..}\right) - E\left(X_{2..}\right) = \mu_1 - \mu_2 \\
Var\left(\overline{D}\right) &= Var\left(X_{1..}\right) + Var\left(X_{2..}\right) = \frac{2}{J}\left(\sigma_b^2 + \frac{1}{K}\sigma_t^2\right)
\end{aligned}$$

$$\boxed{\overline{D} = X_{1..} - X_{2..} \sim N\left(\mu_1 - \mu_2, \frac{2}{J}\left(\sigma_{bio}^2 + \frac{1}{K}\sigma_{tech}^2\right)\right)} \tag{A.3}$$

Furthermore, a standardized difference of the sample class means $\overline{D}$ is received by subtracting the expected value from the difference divided by their standard deviation ((A.4):

$$\boxed{\tilde{Z}_{\overline{D}} := \frac{\overline{D} - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{2}{J}\left(\sigma_b^2 + \frac{1}{K}\sigma_t^2\right)}} \sim N\left(0,1\right)} \tag{A.4}$$

Now, if the variances are known, the distribution of the test statistic (the standardized difference of the sample class means) given the null hypothesis is true ($H_0 : \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$) would be known (A.5):

$$\boxed{\frac{X_{1..} - X_{2..}}{\sqrt{\frac{2}{J}\left(\sigma_b^2 + \frac{1}{K}\sigma_t^2\right)}} \sim_{H_0} N\left(0,1\right)} \tag{A.5}$$

If these variances are unknown, then, it is generally valid to divide the standardized variable by a χ^2-distributed variable in order to obtain a t-distributed test statistic, using the estimates of the unknown variances drawn from the sample. Following Cochran's theorem[20], the distributions of the sum of the quadratic standardized observations of each class i V_i are known as χ^2-distributed (Equation A.6) and also $\tilde{V}_i$, if the true means are replace by the sample means (Equation A.7). It follows that the sum of both, the sum of the quadratic standardized observations of both classes, $\tilde{V} := \tilde{V}_1 + \tilde{V}_2$ is also χ^2-distributed (A.8).

$$V_i = \sum_{j=1}^{J} \tilde{Z}^2_{ij.} := \sum_{j=1}^{J} \left(\frac{X_{ij.} - \mu_1}{\sqrt{\sigma_b^2 + \frac{1}{K}\sigma_t^2}} \right)^2 \sim \chi^2_J \tag{A.6}$$

$$\tilde{V}_i = \sum_{j=1}^{J} \frac{(X_{ij.} - X_{i..})^2}{\sigma_b^2 + \frac{1}{K}\sigma_t^2} \sim \chi^2_{J-1} \tag{A.7}$$

$$\tilde{V} = \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{(X_{ij.} - X_{i..})^2}{\sigma_b^2 + \frac{1}{K}\sigma_t^2} \sim \chi^2_{I(J-1)} \tag{A.8}$$

Hence, the distribution of the sample means difference is known as t-distributed, if the standardization is done with the estimation of the variances instead of the variances itself, derived by dividing the standardized variable by the χ^2-distributed variable $\tilde{V}$:

$$\frac{\tilde{Z}_{\overline{D}}}{\sqrt{\frac{\tilde{V}}{I(J-1)}}} = \frac{\frac{\overline{D}-(\mu_1-\mu_2)}{\sqrt{\frac{2}{J}\left(\sigma_b^2+\frac{1}{K}\sigma_t^2\right)}}}{\sqrt{\frac{\sum\limits_{i=1}^{2}\sum\limits_{j=1}^{J}\frac{\left(X_{ij.}-X_{i..}\right)^2}{\sigma_b^2+\frac{1}{K}\sigma_t^2}}{I(J-1)}}} = \sqrt{\frac{J}{2}} \frac{\overline{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{I(J-1)} \sum\limits_{i=1}^{2} \sum\limits_{j=1}^{J} (X_{ij.} - X_{i..})^2}} \sim t_{I(J-1)}$$

Thus, given the null hypothesis is true, which means that the observation of both classes are obtained from distributions with the same mean, the test statistic (A.9) can be used to test the equality of that means, if the variances are unknown:

$$\boxed{\tilde{t} = \sqrt{\frac{J}{2}} \frac{X_{1..} - X_{2..}}{\sqrt{\frac{1}{I(J-1)} \sum\limits_{i=1}^{2} \sum\limits_{j=1}^{J} (X_{ij.} - X_{i..})^2}} \sim_{H_0} t_{I(J-1)}} \tag{A.9}$$

A.2.3. Likelihood-ratio approach

Using the likelihood ratio test include three steps:

1. The likelihood of the data for both models, model 1 $P\left(\underline{x}|\Theta_1 = \mu_1, \mu_2, \sigma_b^2, \sigma_t^2\right)$ and model 0 $P\left(\underline{x}|\Theta_0 = \mu_0, \sigma_b^2, \sigma_t^2\right)$, are needed.

2. Then, an estimator, usually the maximum likelihood estimator (mle), for the unknown parameter of the models $\hat{\theta_1}$, $\hat{\theta_0}$ has to be derived for inclusion into the formula of the likelihood.

3. The likelihood ratio test (A.10), a χ^2-distributed test statistic, compares then the maximum probability of the data of both models.

[20] Cochran, W. G., "The distribution of quadratic forms in a normal system, with applications to the analysis of covariances", *Mathematical Proceedings of the Cambridge Philosophical Society*, 1934

$$\boxed{LRT = 2log\frac{P(\underline{x}|\hat{\theta_1})}{P(\underline{x}|\hat{\theta_0})}} \tag{A.10}$$

The null hypothesis of equal means is the same when the data have the same probability under both models. It could be rejected, if the probability of the calculated value or a smaller one do not exceed a determined threshold (see A.1 for explanation of hypothesis test).

The following derived LRT (A.29) can then be compared with other test-statistics: 1) F-statistic of the classical nested ANOVA approach (given in Article 6.1, see Appendix B.1) , and 2) $\tilde{t}$-statistic for nested data derived by the frequentist approach.

Using the same example as before (FigureA.2), the random variables $X_{ijk} \sim N\left(\mu_i, \sigma_b^2 + \sigma_t^2\right)$ $(j = 1, ..., N, k = 1, .., K)$ of two classes $(i = 1, 2)$ are given as normal distributed. Then, a hidden variable u_{ij} corresponds to the $\mu_i + \beta_{j|i}$ of the linear mixed model from the traditional frequentist perspective and the model can be rewritten as:

$$u_{ij} \sim N(\mu_i, \sigma_b) \quad , \qquad x_{ijk}|u_{ij} \sim N(u_{ij}, \sigma_t)$$

The aim is to generate a test of the likelihood ratios of two models, at the one hand a model of two different means $\mu_1 \neq \mu_2$ for the classes $i = 1, 2$ (model 1) and at the other hand a model of the same mean $\mu_1 = \mu_2 = \mu_0$ for both classes (model 0) is assumed. Therefore, the complete data likelihood (A.11) is required.

1. Complete data likelihood: For calculating the complete data likelihood

$$P(\underline{\underline{X}}|\Theta) \quad = \quad \prod_{ij} P(\underline{x_{ij}}|\Theta), \tag{A.11}$$

first, the likelihood for one individual with K replicates $\underline{x_{ij}} = (x_{ij1}, \ldots, x_{ijK})$ is required.

$$\begin{aligned} P(\underline{x_{ij}}|\Theta) &= \int P(\underline{x_{ij}}, u_{i,j}|\Theta)du_{ij} \\ &= \int P(\underline{x_{ij}}|u_{ij}, \Theta)P(u_{i,j}|\Theta)du_{ij} \\ &= \int \left(\prod_k P(x_{ijk}|u_{ij}, \Theta)\right) P(u_{ij}|\Theta)du_{ij} \end{aligned}$$

Therefore, simplify the variables by (A.12), then it holds (A.13)

$$u_{ij} = u \rightarrow u \sim N(\mu_i, \sigma_b) \quad , \quad x_{ijk} = x_k \rightarrow x_k|u \sim N(u, \sigma_t) \tag{A.12}$$

$$P\left(u\right) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left(-\frac{1}{2}\frac{(u-\mu_i)^2}{\sigma_b^2}\right) \quad , \quad P\left(x_k|u\right) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{1}{2}\frac{(x_k-u)^2}{\sigma_t^2}\right) \tag{A.13}$$

and the likelihood of one individual in one class can be derived as follows:

$$
\begin{aligned}
P\left(\underline{x}|\Theta\right) &\overset{(A.13)}{=} \int\left(\prod_k\left(\frac{1}{\sqrt{2\pi}\sigma_t}\exp\left(-\frac{1}{2}\frac{(x_k-u)^2}{\sigma_t^2}\right)\right)\right)\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left(-\frac{1}{2}\frac{(u-\mu_i)^2}{\sigma_b^2}\right)du\\
&\overset{V_1=\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\cdot\frac{1}{\sqrt{2\pi}\sigma_b}}{=} V_1\cdot\int\exp\left(-\frac{1}{2}\left(\frac{\sum_k(x_k-u)^2}{\sigma_t^2}+\frac{(u-\mu_i)^2}{\sigma_b^2}\right)\right)du\\
&= V_1\cdot\int\exp\left(-\frac{1}{2}\left(\left(\frac{\sum_k x_k^2}{\sigma_t^2}+\frac{\mu_i^2}{\sigma_b^2}\right)-2\underbrace{\left(\frac{\sum_k x_k}{\sigma_t^2}+\frac{\mu_i}{\sigma_b^2}\right)}_{b}\cdot u+\underbrace{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}_{a}\cdot u^2\right)\right)du\\
&\overset{V_2=\exp\left(-\frac{1}{2}\left(\frac{\sum_k x_k^2}{\sigma_t^2}+\frac{\mu_i^2}{\sigma_b^2}\right)\right)}{=} V_1\cdot V_2\int\exp\left(-\frac{1}{2}\left(a\cdot u^2-2b\cdot u\right)\right)du
\end{aligned}
$$

Substitute u by (A.14) and knowing (A.15) the formula can be simplified.

$$
v=\sqrt{a}u \iff u=\frac{v}{\sqrt{a}} \tag{A.14}
$$

$$
\frac{1}{\sqrt{2\pi}\sigma}\int\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)dx=1 \iff \int\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)dx=\sqrt{2\pi}\sigma \tag{A.15}
$$

$$
\begin{aligned}
\Leftrightarrow P\left(x|\Theta\right) &\overset{(A.14)}{=} V_1\cdot V_2\cdot\int\exp\left(-\frac{1}{2}\left(v^2-2\frac{b}{\sqrt{a}}\cdot v\right)\right)dv\cdot\frac{1}{\sqrt{a}}\\
&= V_1\cdot V_2\cdot\frac{1}{\sqrt{a}}\cdot\exp\left(\frac{b^2}{2a}\right)\int\exp\left(-\frac{1}{2}\left(v^2-2\frac{b}{\sqrt{a}}\cdot v+\frac{b^2}{a}\right)\right)dv\\
&= V_1\cdot V_2\cdot\frac{1}{\sqrt{a}}\cdot\exp\left(\frac{b^2}{2a}\right)\underbrace{\int\exp\left(-\frac{1}{2}\left(v-\frac{b}{\sqrt{a}}\right)^2\right)dv}_{\sqrt{2\pi},if\,\mu=\frac{b}{\sqrt{a}},\sigma=1}\overset{(A.15)}{=} V_1\cdot V_2\cdot\frac{\sqrt{2\pi}}{\sqrt{a}}\cdot\exp\left(\frac{b^2}{2a}\right)
\end{aligned}
$$

The likelihood of one individual j (with K replicates) in one class i is now known (A.16) and shows that the observation of one individual with K replicated measurements follows a K-dimensional normal distribution (Proof see Section A.2.4).

$$
\boxed{P\left(\underline{x_{i,j}}|\Theta\right)\underset{a,b}{\overset{V_1,V_2}{=}}2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\exp\left(-\frac{1}{2}\frac{\mu_i^2}{\sigma_b^2}\right)\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_k}{\sigma_t^2}+\frac{\mu_i}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\cdot\exp\left(-\frac{1}{2}\frac{\sum_k x_k^2}{\sigma_t^2}\right)} \tag{A.16}
$$

(A.16) is used to obtain the complete data likelihood under model 0 (Eq. A.17) and under model 1 (Eq.A.18), which are both a description of a $N\times K$-dimensional normal distribution.

$$
P\left(\underline{\underline{X}}|\begin{cases}\Theta_0\\\Theta_1\end{cases}\right)\underset{(A.11)}{\overset{A.16}{=}}\prod_{i,j}\begin{cases}[2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\\ [2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\end{cases}
$$

$$\exp\left(-\frac{1}{2}\frac{\mu_0^2}{\sigma_b^2}\right)\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\mu_0}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\cdot\exp\left(-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)\Bigg] \tag{A.17}$$

$$\exp\left(-\frac{1}{2}\frac{\mu_i^2}{\sigma_b^2}\right)\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\mu_i}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\cdot\exp\left(-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)\Bigg] \tag{A.18}$$

2. Estimates of the model parameter: For the maximum likelihood estimation of the model parameters it is appropriate to use the logarithmic likelihood, so for model 0 it is derived by (Eq. A.19) and equivalent for model 1 (Eq. A.20).

$$\log P(\underline{\underline{X}}|\Theta_0)\ \overset{A.17}{=}$$

$$\sum_{i,j}\log\left[2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\cdot\exp\left(-\frac{1}{2}\frac{\mu_0^2}{\sigma_b^2}\right)\cdot\exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\mu_0}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)\cdot\exp\left(-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)\right]$$

$$\underset{V_2=\frac{1}{2}\frac{\sum_{ijk}x_{ijk}^2}{\sigma_t^2}}{\overset{V_1=2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K}{=}}\quad IJ\log V_1-V_2-\sum_{i,j}\frac{1}{2\sigma_b^2}\mu_0^2+\left(\frac{1}{2}\sum_{i,j}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\mu_0}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)$$

$$\overset{\frac{1}{K}\sum_k x_{ijk}=x_{ij.}}{=}\quad IJ\log V_1-V_2-\frac{IJ}{2\sigma_b^2}\mu_0^2+\left(\frac{1}{2}\frac{\sum_{i,j}\left(\left(\frac{K\cdot x_{ij.}}{\sigma_t^2}\right)^2+2\frac{K\cdot x_{ij.}}{\sigma_t^2}\frac{\mu_0}{\sigma_b^2}+\left(\frac{\mu_0}{\sigma_b^2}\right)^2\right)}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}\right)$$

$$\overset{\frac{1}{IJ}\sum_{i,j}x_{ij.}=x_{...}}{=}\quad IJ\log V_1-V_2+\frac{1}{2}\frac{\sum_{i,j}\left(\frac{K\cdot x_{ij.}}{\sigma_t^2}\right)^2}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}-\frac{IJ}{2\sigma_b^2}\mu_0^2+\frac{1}{2}\frac{\left(2\frac{IJK\cdot x_{...}}{\sigma_t^2\sigma_b^2}\cdot\mu_0+IJ\left(\frac{\mu_0}{\sigma_b^2}\right)^2\right)}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}$$

$$\overset{V_3=\frac{1}{2}\frac{\sum_{i,j}\left(\frac{K\cdot x_{ij.}}{\sigma_t^2}\right)^2}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}}{=}\quad IJ\log V_1-V_2+V_3-\frac{IJ}{2\sigma_b^2}\mu_0^2+\frac{IJK\cdot x_{...}}{(K\sigma_b^2+\sigma_t^2)}\cdot\mu_0+\frac{1}{2}\frac{IJ\sigma_t^2}{(K\sigma_b^2+\sigma_t^2)\,\sigma_b^2}\mu_0^2$$

$$\Leftrightarrow\log P(\underline{\underline{X}}|\Theta_0)\ =\ \underline{\underline{IJ\log V_1-V_2+V_3+\frac{IJ}{2\sigma_b^2}\left(\frac{\sigma_t^2}{K\sigma_b^2+\sigma_t^2}-1\right)\mu_0^2+\frac{IJK\cdot x_{...}}{K\sigma_b^2+\sigma_t^2}\cdot\mu_0}} \tag{A.19}$$

$$\log P(\underline{\underline{X}}|\Theta_1)\ \overset{A.18}{=}\ \underline{\underline{IJ\log V_1-V_2+V_3+\frac{J}{2\sigma_b^2}\left(\frac{\sigma_t^2}{K\sigma_b^2+\sigma_t^2}-1\right)\sum_i\mu_i^2+\frac{JK}{K\sigma_b^2+\sigma_t^2}\cdot\sum_i x_{i..}\cdot\mu_i}} \tag{A.20}$$

When the first derivation of these log-likelihoods are set to null, the maximum parameter of the functions can be obtained, called maximum likelihood estimator (mle). The mle for the parameter μ_0 within model 0 is equal to the mean of all observations of both classes (Eq. A.21), and the μ_i within model 1 to the mean of all observation in one class i (Eq. A.22).

$$\begin{aligned}
\frac{d\underline{\underline{\underline{X}}}}{d\mu_0} \log P(\underline{\underline{\underline{X}}}|\mu_0) &\overset{(A.19)}{=} IJ\left(\frac{1}{\sigma_b^2}\left(\frac{\sigma_t^2}{K\sigma_b^2+\sigma_t^2}-1\right)\mu_0 + \frac{K\cdot x_{...}}{K\sigma_b^2+\sigma_t^2}\right) = 0 \\
\Longleftrightarrow \left(\frac{\sigma_t^2-\left(K\sigma_b^2+\sigma_t^2\right)}{K\sigma_b^2+\sigma_t^2}\right)\mu_0 &= -\frac{\sigma_b^2}{K\sigma_b^2+\sigma_t^2}K\cdot x_{...} \\
\Longleftrightarrow \mu_0 &= -\frac{K\cdot\sigma_b^2}{\sigma_t^2-(K\sigma_b^2+\sigma_t^2)}\cdot x_{...}
\end{aligned}$$

$$\boxed{\Longleftrightarrow \hat{\mu_0} = x_{...} = \underline{\underline{\frac{1}{IJK}\sum_{ijk} x_{ijk}}}} \tag{A.21}$$

$$\begin{aligned}
\frac{d\underline{\underline{\underline{X}}}}{d\mu_i} \log P(\underline{\underline{\underline{X}}}|\mu_1,\mu_2) &\overset{(A.20)}{=} J\left(\frac{1}{\sigma_b^2}\left(\frac{\sigma_t^2}{K\sigma_b^2+\sigma_t^2}-1\right)\mu_i + \frac{K\cdot x_{i..}}{K\sigma_b^2+\sigma_t^2}\right) = 0 \\
\Longleftrightarrow \frac{1}{\sigma_b^2}\left(\frac{\sigma_t^2}{K\sigma_b^2+\sigma_t^2}-1\right)\mu_i &= -\frac{K\cdot x_{i..}}{K\sigma_b^2+\sigma_t^2} \\
\Longleftrightarrow \mu_i &= -\frac{K\sigma_b^2}{\sigma_t^2-(K\sigma_b^2+\sigma_t^2)}\cdot x_{i..}
\end{aligned}$$

$$\boxed{\Longleftrightarrow \hat{\mu_i} = x_{i..} = \underline{\underline{\frac{1}{JK}\sum_{jk} x_{ijk}}}} \tag{A.22}$$

Additionally, the estimates of the unknown variances within each model are required. Given the log-likelihood of model 0 (A.17) the maximum likelihood estimator for σ_b is derived by Equation A.24 as the mean quadratic deviation of all observations of both classes from the overall variance and given the model 1 (A.18) by Equation A.25 as the mean quadratic deviation of all observations of both classes from their corresponding class variance.

$$\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2} \log P(\underline{\underline{\underline{X}}}|,\hat{\mu_0},\sigma_b^2) \quad = \quad 0$$

$$\overset{A.17}{\Longleftrightarrow} \underbrace{\sum_{i,j} \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2} \log\left[2\pi\cdot\frac{1}{\sqrt{2\pi}\sigma_b}\cdot\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\right]}_{A} + \underbrace{\sum_{i,j}\frac{d\underline{X}}{d\sigma_b}\left[-\frac{1}{2}\frac{\hat{\mu_0}}{\sigma_b^2}+\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\mu_0}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right]}_{B} = 0$$

$$A = \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(2\pi\right)}_{0} + \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\frac{1}{\sqrt{2\pi}\sigma_b}\right)}_{V_1} + \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}\right)}_{V_2} + \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K\right)}_{0}$$

$$= \frac{1}{\sigma_b}\left(\frac{\sigma_t^2}{K\cdot\sigma_b^2+\sigma_t^2}-1\right) = \underline{\underline{\left(-\frac{K\cdot\sigma_b^2}{K\cdot\sigma_b^2+\sigma_t^2}\right)\frac{1}{\sigma_b}}}$$

$$\text{with}\quad V_1 = \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\frac{1}{\sqrt{2\pi}\sigma_b}\right) = \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\frac{1}{\sigma_b}\right) \overset{\text{chain rule}}{=} \sigma_b\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(\frac{1}{\sigma_b}\right) = \underline{\underline{-\frac{1}{\sigma_b}}}$$

$$V_2 = \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\log\left(\frac{1}{\sqrt{2\pi}\cdot\sqrt{\left(\frac{K\cdot\sigma_b^2+\sigma_t^2}{\sigma_t^2\cdot\sigma_b^2}\right)}}\right) = \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(-\frac{1}{2}\log\left(\frac{K\cdot\sigma_b^2+\sigma_t^2}{\sigma_t^2\cdot\sigma_b^2}\right)\right)$$

$$\overset{\text{(chain rule)}}{=} -\frac{1}{2}\frac{\sigma_t^2\cdot\sigma_b^2}{K\cdot\sigma_b^2+\sigma_t^2}\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(\frac{K\cdot\sigma_b^2+\sigma_t^2}{\sigma_t^2\cdot\sigma_b^2}\right)$$

$$\overset{\left(\frac{u}{v}\right)'=\frac{u'v-uv'}{v^2}}{=} -\frac{1}{2}\frac{\sigma_t^2\cdot\sigma_b^2}{K\cdot\sigma_b^2+\sigma_t^2}\left(\frac{2K\cdot\sigma_b\cdot\sigma_t^2\cdot\sigma_b^2-\left(K\cdot\sigma_b^2+\sigma_t^2\right)\cdot 2\sigma_t^2\cdot\sigma_b}{\sigma_t^4\cdot\sigma_b^4}\right)$$

$$= -\frac{1}{2}\frac{1}{K\cdot\sigma_b^2+\sigma_t^2}\left(-2\frac{\sigma_t^2}{\sigma_b}\right) = \underline{\underline{\frac{\sigma_t^2}{K\cdot\sigma_b^2+\sigma_t^2}\cdot\frac{1}{\sigma_b}}}$$

$$B = \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(-\frac{1}{2}\frac{\hat{\mu_0}^2}{\sigma_b^2}\right)}_{V_3=\frac{\hat{\mu_0}^2}{\sigma_b^3}} + \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right)}_{V_4} - \underbrace{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b^2}\left(\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)}_{0} = \underline{\underline{\left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}-\hat{\mu_0}\right)^2\frac{1}{\sigma_b^3}}}$$

$$\text{with}\quad V_4 = \frac{1}{2}\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b}\left(\frac{\underbrace{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)^2}_{u}}{\underbrace{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}_{v}}\right)\Bigg| \qquad u' = -4\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)\cdot\frac{\hat{\mu_0}}{\sigma_b^3}, v' = -\frac{2}{\sigma_b^3}$$

$$\overset{\frac{1}{2}\left(\frac{u}{v}\right)'=\frac{1}{2}\frac{u'v-uv'}{v^2}}{=} \frac{1}{2}\frac{-4\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)\cdot\frac{\hat{\mu_0}}{\sigma_b^3}\cdot\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)-\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)^2\cdot\left(-\frac{2}{\sigma_b^3}\right)}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)^2}$$

$$= \left(-2\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\cdot\hat{\mu_0}+\underbrace{\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)^2}}_{a^2}\right)\cdot\left(\frac{1}{\sigma_b^3}\right)$$

$$\overset{(a-b)^2-b^2=a^2-2ab}{=} \left(\left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}-\hat{\mu_0}\right)^2-\hat{\mu_0}^2\right)\cdot\left(\frac{1}{\sigma_b^3}\right) = \underline{\underline{\left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{\sigma_b^2}\right)}{\left(\frac{K\sigma_b^2+\sigma_t^2}{\sigma_t^2\sigma_b^2}\right)}-\hat{\mu_0}\right)^2\frac{1}{\sigma_b^3}-\frac{\hat{\mu_0}^2}{\sigma_b^3}}}$$

$$\longrightarrow \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b} \log P(\underline{\underline{\underline{X}}}|\hat{\mu_0}, \sigma_b^2) = IJ \cdot A + \sum_{i,j} B$$

$$= \underline{\underline{\frac{1}{\sigma_b} \left(-IJ \left(\frac{K \cdot \sigma_b^2}{K \cdot \sigma_b^2 + \sigma_t^2} \right) + \frac{1}{\sigma_b^2} \sum_{i,j} \left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2} + \frac{\hat{\mu_0}}{\sigma_b^2} \right)}{\left(\frac{K\sigma_b^2 + \sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)} - \hat{\mu_0} \right)^2 \right)}} \tag{A.23}$$

Hence, the derivation of complete data likelihood of model 0 for σ_b is known, the parameter σ_b can be derived as follows:

$$\longrightarrow \frac{d\underline{\underline{\underline{X}}}}{d\sigma_b} \log P(\underline{\underline{\underline{X}}}|\Theta_0) \overset{!}{=} 0 \underset{(A.23)}{\overset{\sigma_b \neq 0}{\iff}} IJ \left(\frac{K \cdot \sigma_b^2}{K \cdot \sigma_b^2 + \sigma_t^2} \right) = \frac{1}{\sigma_b^2} \sum_{i,j} \left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2} + \frac{\hat{\mu_0}}{\sigma_b^2} \right)}{\left(\frac{K\sigma_b^2 + \sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)} - \hat{\mu_0} \right)^2$$

$$\iff \sigma_b^4 = \frac{1}{IJK} \underbrace{\sum_{i,j} \left(K \cdot \sigma_b^2 + \sigma_t^2 \right) \left(\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2} + \frac{\hat{\mu_0}}{\sigma_b^2} \right)}{\left(\frac{K\sigma_b^2 + \sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)} - \hat{\mu_0} \right)^2}_{D} = \left(\frac{K}{IJ} \frac{1}{K\sigma_b^2 + \sigma_t^2} \sum_{i,j} (x_{ij.} - \hat{\mu_0})^2 \right) \cdot \sigma_b^4$$

$$\iff K\sigma_b^2 + \sigma_t^2 = \frac{K}{IJ} \sum_{i,j} (x_{ij.} - \hat{\mu_0})^2$$

$$\boxed{\iff {}_0\hat{\sigma_b^2} = \underline{\underline{\frac{1}{IJ} \sum_{i,j} (x_{ij.} - x_{...})^2 - \frac{\sigma_t^2}{K}}}} \tag{A.24}$$

With

$$D \overset{\sum_k x_{ijk} = K \cdot x_{ij.}}{=} \sum_{i,j} \left(\left(K \cdot \sigma_b^2 + \sigma_t^2 \right) \cdot \left(\left(\frac{\left(\frac{K \cdot x_{ij.}\sigma_b^2 + \hat{\mu_0}\sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)}{\left(\frac{K\sigma_b^2 + \sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)} \right)^2 - 2 \left(\frac{\left(\frac{K \cdot x_{ij.}\sigma_b^2 + \hat{\mu_0}\sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)}{\left(\frac{K\sigma_b^2 + \sigma_t^2}{\sigma_t^2 \sigma_b^2} \right)} \right) \cdot \hat{\mu_0} + \hat{\mu_0}^2 \right) \right)$$

$$= \sum_{i,j} \left(\frac{\left(K \cdot x_{ij.}\sigma_b^2 + \hat{\mu_0}\sigma_t^2 \right)^2}{K\sigma_b^2 + \sigma_t^2} - 2 \left(K \cdot x_{ij.}\sigma_b^2 + \hat{\mu_0}\sigma_t^2 \right) \cdot \hat{\mu_0} + \left(K \cdot \sigma_b^2 + \sigma_t^2 \right) \cdot \hat{\mu_0}^2 \right)$$

$$= \sum_{i,j} \left(\frac{1}{\sqrt{K\sigma_b^2 + \sigma_t^2}} \cdot \left(K \cdot x_{ij.}\sigma_b^2 + \hat{\mu_0}\sigma_t^2 \right) - \sqrt{K \cdot \sigma_b^2 + \sigma_t^2} \cdot \hat{\mu_0} \right)^2$$

$$= \sum_{i,j} \left(\frac{K \cdot x_{ij.}\sigma_b^2 - K \cdot \sigma_b^2 \cdot \hat{\mu_0}}{\sqrt{K\sigma_b^2 + \sigma_t^2}} \right)^2 = \underline{\underline{\sum_{i,j} K^2 \left(\frac{x_{ij.} - \hat{\mu_0}}{\sqrt{K\sigma_b^2 + \sigma_t^2}} \right)^2 \sigma_b^4}}$$

analog

$$\boxed{\frac{d\underline{\underline{\underline{X}}}}{d\sigma_b} \log P(\underline{X}|, \hat{\mu_i}, \sigma_b^2) \overset{!}{=} 0 \iff {}_1\hat{\sigma_b^2} = \underline{\underline{\frac{1}{IJ} \sum_{i,j} (x_{ij.} - x_{i...})^2 - \frac{\sigma_t^2}{K}}}} \tag{A.25}$$

3. Maximum-Likelihood-Ratio-Test: For both models the maximum log-likelihood of the complete data is required. Therefore, the mle of the parameters are included into the log-likelihood of the complete data.

$$\log P(\underline{\underline{X}}| \begin{cases} \hat{\mu_0}, {}_0\hat{\sigma_b^2}) & \overset{A.17}{=} \\ \hat{\mu_1}, \hat{\mu_2}, {}_1\hat{\sigma_b^2}) & \overset{A.18}{=} \end{cases}$$

$$\sum_{i,j}\left[\log\left(\frac{1}{{}_0\hat{\sigma}_b}\cdot\frac{1}{\sqrt{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_0\hat{\sigma_b^2}}\right)}}\right)+\underbrace{K\log\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)}_{V_1}+\left(-\frac{1}{2}\frac{\hat{\mu_0}^2}{{}_0\hat{\sigma_b^2}}\right)+\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{{}_0\hat{\sigma_b^2}}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_0\hat{\sigma_b^2}}\right)}\right)+\underbrace{\left(-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)}_{V_2}\right]$$

$$\sum_{i,j}\left[\log\left(\frac{1}{{}_1\hat{\sigma}_b}\cdot\frac{1}{\sqrt{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_1\hat{\sigma_b^2}}\right)}}\right)+\underbrace{K\log\left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)}_{V_1}+\left(-\frac{1}{2}\frac{\hat{\mu_i}^2}{{}_1\hat{\sigma_b^2}}\right)+\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_i}}{{}_1\hat{\sigma_b^2}}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_1\hat{\sigma_b^2}}\right)}\right)+\underbrace{\left(-\frac{1}{2}\frac{\sum_k x_{ijk}^2}{\sigma_t^2}\right)}_{V_2}\right]$$

Then, both functions can be compared by the maximum likelihood ratio:

$$LRT \overset{(A.10)}{=} 2\log\frac{P(\underline{\underline{X}}|\hat{\theta}_1)}{P(\underline{\underline{X}}|\hat{\theta}_0)} = 2\log P(\underline{\underline{X}}|\hat{\mu_1},\hat{\mu_2},{}_1\hat{\sigma_b^2}) - 2\log P(\underline{\underline{X}}|\hat{\mu_0},{}_0\hat{\sigma_b^2})$$

$$= \quad 2\sum_{i,j}\left(\underbrace{\log\left(\frac{{}_0\hat{\sigma}_b}{{}_1\hat{\sigma}_b}\cdot\frac{\sqrt{\left(\frac{K_0\hat{\sigma_b^2}+\sigma_t^2}{\sigma_t^2\,{}_0\hat{\sigma_b^2}}\right)}}{\sqrt{\left(\frac{K_1\hat{\sigma_b^2}+\sigma_t^2}{\sigma_t^2\,{}_1\hat{\sigma_b^2}}\right)}}\right)}_{A=\frac{1}{2}\log\frac{K_0\hat{\sigma_b^2}+\sigma_t^2}{K_1\hat{\sigma_b^2}+\sigma_t^2}}\right)+2\sum_{i,j}\left(\underbrace{\left(-\frac{1}{2}\frac{\hat{\mu_i}^2}{{}_1\hat{\sigma_b^2}}\right)+\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_i}}{{}_1\hat{\sigma_b^2}}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_1\hat{\sigma_b^2}}\right)}}_{B}\right)+2\sum_{i,j}\left(\underbrace{-\left(-\frac{1}{2}\frac{\hat{\mu_0}^2}{{}_0\hat{\sigma_b^2}}\right)-\frac{1}{2}\frac{\left(\frac{\sum_k x_{ijk}}{\sigma_t^2}+\frac{\hat{\mu_0}}{{}_0\hat{\sigma_b^2}}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{{}_0\hat{\sigma_b^2}}\right)}}_{C}\right)$$

$$\begin{aligned}
B \quad & \underset{\sum_k x_{ijk}=K\cdot x_{ij.}}{\overset{\hat{\mu_i}=x_{i..}}{=}} && \left(-\frac{1}{2}\frac{x_{i..}^2}{{}_1\hat{\sigma}_b^2}\right)+\frac{1}{2}\frac{\left(\frac{K\cdot x_{ij.}\cdot{}_1\hat{\sigma}_b^2+x_{i..}\sigma_t^2}{\sigma_t^2\,{}_1\hat{\sigma}_b^2}\right)^2}{\left(\frac{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}{\sigma_t^2\,{}_1\hat{\sigma}_b^2}\right)} \\
& = && -\frac{1}{2}\frac{x_{i..}^2}{{}_1\hat{\sigma}_b^2}\cdot\frac{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}+\frac{1}{2}\frac{1}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}\cdot\frac{\left(K\cdot x_{ij.}\cdot{}_1\hat{\sigma}_b^2+x_{i..}\sigma_t^2\right)^2}{\sigma_t^2\,{}_1\hat{\sigma}_b^2} \\
& = && \frac{1}{2}\frac{1}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}\cdot\left(\underbrace{\frac{\left(K\cdot x_{ij.}\cdot{}_1\hat{\sigma}_b^2+x_{i..}\sigma_t^2\right)^2}{\sigma_t^2\,{}_1\hat{\sigma}_b^2}-\frac{\left(K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2\right)\cdot x_{i..}^2}{{}_1\hat{\sigma}_b^2}}_{a}\right) \\
& = && \underline{\underline{\frac{1}{2}\cdot K\cdot\left(\frac{x_{ij.}^2}{\sigma_t^2}-\frac{(x_{ij.}-x_{i..})^2}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}\right)}} \\
\text{With}\quad a & = && \frac{K^2\cdot x_{ij.}^2\cdot{}_1\hat{\sigma}_b^2}{\sigma_t^2}+2\cdot K\cdot x_{ij.}\cdot x_{i..}+\frac{x_{i..}^2\sigma_t^2}{{}_1\hat{\sigma}_b^2}-K\cdot x_{i..}^2-\frac{x_{i..}^2\cdot\sigma_t^2}{{}_1\hat{\sigma}_b^2} \\
& = && K\cdot\left(\frac{K\cdot x_{ij.}^2\cdot{}_1\hat{\sigma}_b^2}{\sigma_t^2}+2\cdot x_{ij.}\cdot x_{i..}-x_{i..}^2\right) \\
& = && K\cdot\left(\frac{K\cdot x_{ij.}^2\cdot{}_1\hat{\sigma}_b^2}{\sigma_t^2}+\frac{x_{ij.}^2\cdot\sigma_t^2}{\sigma_t^2}-(x_{ij.}-x_{i..})^2\right) \\
& = && \underline{\underline{K\cdot\left(x_{ij.}^2\frac{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}{\sigma_t^2}-(x_{ij.}-x_{i..})^2\right)}} \\
\longrightarrow\text{analog}\quad C & = && \underline{\underline{-\frac{1}{2}\cdot K\cdot\left(\frac{x_{ij.}^2}{\sigma_t^2}-\frac{(x_{ij.}-x_{...})^2}{K\cdot{}_0\hat{\sigma}_b^2+\sigma_t^2}\right)}}
\end{aligned}$$

$$\boxed{\begin{aligned}
K_0\hat{\sigma}_b^2+\sigma_t^2 &\overset{A.24}{=} K\left(\frac{1}{IJ}\sum_{i,j}(x_{ij.}-x_{...})^2-\frac{\sigma_t^2}{K}\right)+\sigma_t^2=\frac{K}{IJ}\sum_{i,j}(x_{ij.}-x_{...})^2 \\
K_1\hat{\sigma}_b^2+\sigma_t^2 &\overset{A.25}{=} K\left(\frac{1}{IJ}\sum_{i,j}(x_{ij.}-x_{i...})^2-\frac{\sigma_t^2}{K}\right)+\sigma_t^2=\frac{K}{IJ}\sum_{i,j}(x_{ij.}-x_{i..})^2
\end{aligned}}\qquad\begin{aligned}(A.26)\\(A.27)\end{aligned}$$

With the equations (A.26, A.27) further simplifications are possible:

$$LRT \;=\; 2\sum_{i,j}A+2\sum_{i,j}B+2\sum_{i,j}C$$

$$
\begin{aligned}
&= \sum_{i,j} \log \frac{K_0\hat{\sigma}_b^2+\sigma_t^2}{K_1\hat{\sigma}_b^2+\sigma_t^2} + K\cdot\sum_{i,j}\left(\left(\frac{x_{ij.}^2}{\sigma_t^2} - \frac{(x_{ij.}-x_{i..})^2}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2}\right) - \left(\frac{x_{ij.}^2}{\sigma_t^2} - \frac{(x_{ij.}-x_{...})^2}{K\cdot{}_0\hat{\sigma}_b^2+\sigma_t^2}\right)\right) \\
&= \sum_{i,j} \log\log \frac{K_0\hat{\sigma}_b^2+\sigma_t^2}{K_1\hat{\sigma}_b^2+\sigma_t^2} + K\cdot\left(-\frac{\sum_{i,j}(x_{ij.}-x_{i..})^2}{K\cdot{}_1\hat{\sigma}_b^2+\sigma_t^2} + \frac{\sum_{i,j}(x_{ij.}-x_{...})^2}{K\cdot{}_0\hat{\sigma}_b^2+\sigma_t^2}\right) \\
&\overset{(A.26)}{\underset{(A.27)}{=}} \sum_{i,j} \log\left(\sum_{i,j}\frac{(x_{ij.}-x_{...})^2}{\sum_{i,j}(x_{ij.}-x_{i..})^2}\right) + K\cdot\left(-\frac{\sum_{i,j}(x_{ij.}-x_{i..})^2}{\frac{K}{IJ}\sum_{i,j}(x_{ij.}-x_{i..})^2} + \frac{\sum_{i,j}(x_{ij.}-x_{...})^2}{\frac{K}{IJ}\sum_{i,j}(x_{ij.}-x_{...})^2}\right) \\
&= \underline{\underline{IJ\cdot\log\left(\frac{\sum_{i,j}(x_{ij.}-x_{...})^2}{\sum_{i,j}(x_{ij.}-x_{i..})^2}\right)}}
\end{aligned}
$$

The following transition (A.28) holds, which means that the total sum of square can be decomposed by the sum of squares between and within the groups.

$$
\begin{aligned}
\sum_{i,j}(x_{ij.}-x_{...})^2 &= \sum_i\left(\sum_j x_{ij.}^2 - 2\cdot J\cdot x_{i..}x_{...} + J\cdot x_{...}^2\right) \\
&= \sum_i\left(\sum_j x_{ij.}^2 - J\cdot x_{i..}^2 + J\,(x_{i..}-x_{...})^2\right) \\
&= \sum_{i,j}(x_{ij.}-x_{i..})^2 + \sum_{i,j}(x_{i..}-x_{...})^2 \qquad \text{(A.28)}
\end{aligned}
$$

$$
\boxed{LRT = IJ\cdot\log\left(1+\frac{\sum_{i,j}(x_{i..}-x_{...})^2}{\sum_{i,j}(x_{ij.}-x_{i..})^2}\right) \overset{A.28}{=} IJ\cdot\log\left(\frac{\left(\sum_{i,j}(x_{ij.}-x_{i..})^2+\sum_{i,j}(x_{i..}-x_{...})^2\right)}{\sum_{i,j}(x_{ij.}-x_{i..})^2}\right)} \tag{A.29}
$$

The LRT (Eq. A.29)can be rewritten in a form that includes the F-statistic, defined in Formula A.30 (Eq. A.31 holds). Furthermore, it holds Eq. A.32, because of Eq. A.33 and analog Eq. A.34. So the LRT can be rewritten in a form that includes the previously derived $\tilde{t}$-statistic (A.32), so Eq. A.35 holds also.

$$
\boxed{F = \frac{MS_{between}}{MS_{within}} = \frac{FG_{within}}{FG_{between}}\frac{SS_{between}}{SS_{within}} = \frac{\frac{1}{I-1}\sum_{i,j}(x_{i..}-x_{...})^2}{\frac{1}{I(J-1)}\sum_{i,j}(x_{ij.}-x_{i..})^2}} \tag{A.30}
$$

$$
\boxed{LRT = IJ\cdot\log\left(1+\frac{\frac{I-1}{I-1}\sum_{i,j}(x_{i..}-x_{...})^2}{\frac{I(J-1)}{I(J-1)}\sum_{i,j}(x_{ij.}-x_{i..})^2}\right) = IJ\cdot\log\left(1+\frac{I-1}{I(J-1)}F\right)} \tag{A.31}
$$

$$
\sum_i^2 (x_{i..}-x...)^2 = \frac{1}{2}\,(x_{1..}-x2..)^2 \tag{A.32}
$$

$$\begin{aligned}
(x_{1..} - x_{...})^2 &= x_{1..}^2 - 2x_{1..}x_{...} + x_{...}{}^2 \\
&= x_{1..}^2 - 2x_{1..}\left(\frac{1}{2}x_{1..} + \frac{1}{2}x_{2..}\right) + \left(\frac{1}{2}x_{1..} + \frac{1}{2}x_{2..}\right)^2 \\
&= x_{1..}^2 - x_{1..}^2 - x_{1..}x_{2..} + \frac{1}{4}x_{1..}^2 + \frac{1}{2}x_{1..}x_{2..} + \frac{1}{4}x_{2..}^2 \\
&= \frac{1}{4}x_{1..}^2 - \frac{1}{2}x_{1..}x_{2..} + \frac{1}{4}x_{2..}^2 = \underline{\frac{1}{4}\left(x_{1..} - x_{2..}\right)^2} \qquad \text{(A.33)} \\
(x_{2..} - x_{...})^2 &= \underline{\frac{1}{4}\left(x_{1..} - x_{2..}\right)^2} \qquad \text{(A.34)}
\end{aligned}$$

$$\boxed{LRT \overset{A.32}{=} IJ\log\left(1 + \frac{\frac{J}{2}\left(x_{1..} - x_{2..}\right)^2}{\frac{I(J-1)}{I(J-1)}\sum\limits_{i,j}\left(x_{ij.} - x_{i..}\right)^2}\right) \overset{A.9}{=} \underline{\underline{IJ\log\left(1 + \frac{1}{I(J-1)}\tilde{t}^2\right)}}} \qquad \text{(A.35)}$$

In summary, the performed LRT test statistic of this example of nested data (testing the mean of two classes for equality) is an approximation of the classical F-statistic (Eq. A.31) and the performed $\tilde{t}$-test statistic for two level hierarchical experiments (Eq. A.35) Furthermore it is shown, that the $\tilde{t}$-test statistic for two level hierarchical experiments is the same as the classical test-statistic using the means of technical replicates as observations. All three test statistics can be used to test if there are differences between groups using a nested experiment design.

A.2.4. Distribution of an observation with K replicates is K-dimensional

$$\boxed{P\left(\overrightarrow{x}|\overrightarrow{\mu}, \Sigma_{p\times p}\right) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|\Sigma_{p\times p}|}} \cdot \exp\left(-\frac{1}{2}(\overrightarrow{x} - \overrightarrow{\mu})^T\Sigma_{p\times p}^{-1}(\overrightarrow{x} - \overrightarrow{\mu})\right)} \qquad \text{(A.36)}$$

A p-dimensional normal distribution is defined as (Eq. A.36). It is to show that the derived formula for the distribution of one biological observation j with K technical replicates within a two-level hierarchical experiment design (ignoring the class correspondence) is equal to the formula of a K-dimensional normal distribution. Hence, it is to proof that Eq. A.16 = Eq. A.36 (Eq. A.37) holds, whereby, $p = K$ and the formula for the likelihood of the vector of K observations $\underline{x} = (x_1, \ldots, x_K)$ (Eq. A.16) can be decomposed by a pre-exponential and an exponential part (Eq. A.38).

$$P\left(\overrightarrow{x}|\overrightarrow{\mu}, \Sigma\right) = P\left(\underline{x}|\Theta\right) \qquad \text{(A.37)}$$

$$\boxed{\begin{aligned}P\left(\underline{x}|\Theta\right) = \underbrace{2\pi \cdot \frac{1}{\sqrt{2\pi}\sigma_b} \cdot \frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}} \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K}_{\text{pre-exponential part}} \\ \underbrace{\exp\left(-\frac{1}{2}\frac{\mu^2}{\sigma_b^2}\right) \cdot \exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_k}{\sigma_t^2}+\frac{\mu}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right) \cdot \exp\left(-\frac{1}{2}\frac{\sum_k x_k^2}{\sigma_t^2}\right)}_{\text{exponential part}}\end{aligned}} \tag{A.38}$$

Corresponding to the assumptions of subsection A.2.1:

- the k-th measurement of an individual j is normal distributed $X_{jk} \sim N(\mu, \sigma^2_{total})$
- the total variance σ^2_{total} is a sum of the betwen individual (or biological) variance σ_b^2 and the within individual (or technical) variance σ_t^2
- different measurements of one individual are dependent

  $Cov(X_{jk}, X_{j\tilde{k}}) = \sigma^2_{bio} \qquad \forall k \neq \tilde{k},$

The assumption is, that a $\Sigma_{K\times K}$ dimensional Covariance-variance matrix exists. It is furthermore assumed this matrix has the following form:

$$\boxed{\Sigma = \begin{pmatrix} \sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & & & \\ & & & \sigma_b^2 \\ \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}} \tag{A.39}$$

Proof

Pre-exponential part The pre-exponential part can be simplify by:

$$2\pi \cdot \frac{1}{\sqrt{2\pi}\sigma_b} \cdot \frac{1}{\sqrt{2\pi\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}} \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_t}\right)^K = \frac{1}{\sqrt{(2\pi)^K}} \cdot \frac{1}{\underbrace{\sigma_b\sigma_t^K\sqrt{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}}_{\sqrt{|\Sigma|}?}}$$

Hence, it is to proof, that the following Equation A.40 holds for the determinant of the assumed $K \times K$ Covariance-Variance-Matrix of Eq. A.39. Whereby, the determinant of the matrix (A.39) can be derived using the following allowed transformation rules:

- the value of the determinant do not change, if rows/columns are permuted (only sign)
- the value of the determinant do not change, if a multiply of a row/column is added to a row/column

$$\sqrt{|\Sigma_{K\times K}|} \overset{?}{=} \sigma_b \sigma_t^K \sqrt{\left(\frac{K}{\sigma_t^2} + \frac{1}{\sigma_b^2}\right)} \quad \Leftrightarrow \quad |\Sigma_{K\times K}| \overset{?}{=} (\sigma_t^2)^{K-1}\left(K\sigma_b^2 + \sigma_t^2\right) \tag{A.40}$$

$$\begin{aligned}
|\Sigma| \quad &= \left| \overbrace{\begin{pmatrix} \sigma_b^2+\sigma_t^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & & \\ \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K \text{ entries}} \right| \overset{\text{1.r-2.r}}{=} \left| \overbrace{\begin{pmatrix} \sigma_t^2 & -\sigma_t^2 & 0 & ... & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & & \\ \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K \text{ entries}} \right| \\
&\overset{\text{2.c+1.c}}{=} \left| \overbrace{\begin{pmatrix} \sigma_t^2 & 0 & 0 & ... & 0 \\ \sigma_b^2 & 2\sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & & \\ \sigma_b^2 & 2\sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K\text{entries}} \right| \overset{\text{dev. 1.r}}{=} (-1)^{1+1}\sigma_t^2 \left| \overbrace{\begin{pmatrix} 2\sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & \\ 2\sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{(K-1) \text{ entries}} \right|
\end{aligned}$$

$$\begin{aligned}
\underset{(K-3)\times}{\overset{\text{repeat}}{=}} \quad & \left(\sigma_t^2\right)^{K-2} \left| \overbrace{\begin{pmatrix} (K-1)\sigma_b^2+\sigma_t^2 & \sigma_b^2 \\ (K-1)\sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K-(K-2)=2 \text{ entries}} \right| \\
\overset{\text{2-dim.}}{=} \quad & \left(\sigma_t^2\right)^{K-2} \left[\left((K-1)\sigma_b^2+\sigma_t^2\right)\left(\sigma_b^2+\sigma_t^2\right) - (K-1)\sigma_b^2\sigma_b^2\right] \\
= \quad & \left(\sigma_t^2\right)^{K-2} \left[(K-1)\sigma_b^2\sigma_b^2 + \sigma_t^2\sigma_b^2 + (K-1)\sigma_b^2\sigma_t^2 + \sigma_t^2\sigma_t^2 - (K-1)\sigma_b^2\sigma_b^2\right] \\
= \quad & \underline{\underline{\left(\sigma_t^2\right)^{K-1}\left(K\sigma_b^2+\sigma_t^2\right)}}
\end{aligned}$$

Hence, it could be shown that the pre-exponential part in both formulas, (A.16) and (A.36), is the equal, if the Variance-Covariance matrix of Eq. A.39 is assumed.

Exponential part

$$\exp\left(-\frac{1}{2}\frac{\mu^2}{\sigma_b^2}\right) \cdot \exp\left(\frac{1}{2}\frac{\left(\frac{\sum_k x_k}{\sigma_t^2} + \frac{\mu}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)}\right) \cdot \exp\left(-\frac{1}{2}\frac{\sum_k x_k^2}{\sigma_t^2}\right) = \exp\left(\underbrace{-\frac{1}{2}\frac{\mu^2}{\sigma_b^2} + \frac{1}{2}\frac{\left(\frac{\sum_k x_k}{\sigma_t^2} + \frac{\mu}{\sigma_b^2}\right)^2}{\left(\frac{K}{\sigma_t^2}+\frac{1}{\sigma_b^2}\right)} - \frac{1}{2}\frac{\sum_k x_k^2}{\sigma_t^2}}_{I}\right)$$

Simplify the inner term of the exponential as follows:

$$
\begin{aligned}
I &= -\frac{1}{2}\frac{1}{(K\sigma_b^2+\sigma_t^2)}\left(\underbrace{\frac{\left(K\sigma_b^2+\sigma_t^2\right)\mu^2}{\sigma_b^2}-\frac{\left(\sigma_b^2\sum\limits_k x_k+\sigma_t^2\mu\right)^2}{\sigma_t^2\sigma_b^2}+\frac{\left(K\sigma_b^2+\sigma_t^2\right)\sum\limits_k x_k^2}{\sigma_t^2}}_{=A}\right)\\
&\overset{A}{=} -\frac{1}{2}\frac{1}{(K\sigma_b^2+\sigma_t^2)}\left(\left(1+(K-1)\frac{\sigma_b^2}{\sigma_t^2}\right)\sum_k(x_k-\mu)^2-2\frac{\sigma_b^2}{\sigma_t^2}\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu)\right)\right)\\
&= -\frac{1}{2}\frac{1}{(\sigma_t^2)^{K-1}(K\sigma_b^2+\sigma_t^2)}(\sigma_t^2)^{K-2}\\
&\qquad\left(\left(\sigma_t^2+(K-1)\sigma_b^2\right)\sum_k(x_k-\mu)^2-2\sigma_b^2\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu\right)\right)\\
&\overset{(A.40)}{=} -\frac{1}{2}\left(\underbrace{\frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\left(\sigma_t^2+(K-1)\sigma_b^2\right)}_{\text{diagonal entries }\Sigma^{-1}}\sum_k(x_k-\mu)^2\ -2\quad\underbrace{\frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\sigma_b^2}_{\text{non diagonal entries }-\Sigma^{-1}}\quad\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu\right)\right)\\
\text{with}\quad A &= K\mu^2+\frac{\sigma_t^2}{\sigma_b^2}\mu^2-\frac{\sigma_b^2}{\sigma_t^2}\left(\sum_k x_k\right)^2-2\mu\sum_k x_k-\frac{\sigma_t^2}{\sigma_b^2}\mu^2+K\frac{\sigma_b^2}{\sigma_t^2}\sum_k x_k^2+\sum_k x_k^2\\
&= \sum_k\mu^2-2\mu\sum_k x_k+\sum_k x_k^2+\frac{\sigma_b^2}{\sigma_t^2}\left(-\left(\sum_k x_k\right)^2+K\sum_k x_k^2\right)\\
&= \sum_k(x_k-\mu)^2+\frac{\sigma_b^2}{\sigma_t^2}\left(\underbrace{-\left(\sum_k x_k\right)^2+K\sum_k x_k^2}_{=B}\right)\\
&\overset{B}{=} \sum_k(x_k-\mu)^2+\frac{\sigma_b^2}{\sigma_t^2}\left(-2\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu)\right)+(K-1)\sum_k(x_k-\mu)^2\right)\\
&= \sum_k(x_k-\mu)^2+\frac{\sigma_b^2}{\sigma_t^2}(K-1)\sum_k(x_k-\mu)^2-2\frac{\sigma_b^2}{\sigma_t^2}\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu\right)\\
&= \underline{\left(1+(K-1)\frac{\sigma_b^2}{\sigma_t^2}\right)\sum_k(x_k-\mu)^2-2\frac{\sigma_b^2}{\sigma_t^2}\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu\right)}\\
\text{and}\quad B &= -\left(\sum_k x_k\right)^2+K\sum_k x_k^2=-\left(\sum_k x_k^2+2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}x_kx_{k'}\right)+K\sum_{k=1}^{K}x_k^2\\
&= -2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}x_kx_{k'}+(K-1)\sum_{k=1}^{K}x_k^2\\
&= -2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}x_kx_{k'}+(K-1)\sum_{k=1}^{K}\left(x_k^2-2x_k\mu+\mu^2+2x_k\mu-\mu^2\right)\\
&= -2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}x_kx_{k'}+(K-1)\sum_k\left((x_k-\mu)^2+2x_k\mu-\mu^2\right)\\
&= -2\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}x_kx_{k'}-(K-1)\sum_k x_k\mu+\frac{K(K-1)}{2}\mu^2\right)+(K-1)\sum_k(x_k-\mu)^2\\
&= \underline{-2\left(\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}(x_k-\mu)(x_{k'}-\mu)\right)+(K-1)\sum_k(x_k-\mu)^2}
\end{aligned}
$$

It is to proof, if for the diagonal and non diagonal entries of the inverse of the assumend Covariance-Variance matrix $\Sigma_{K\times K}^{-1}$ (Eq. A.39) the Eq. (A.41) and (A.42) are hold.

$$\left({}_{K\times K}\Sigma^{-1}\right)_{ij} = \frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\left(\sigma_t^2 + (K-1)\sigma_b^2\right), \qquad \forall i = j \tag{A.41}$$

$$-\left({}_{K\times K}\Sigma^{-1}\right) = \frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\sigma_b^2, \qquad \forall i \neq j \tag{A.42}$$

Using minors of a matrix, for a regular $n \times n$-dimensional matrix A it exists exactly one inverse matrix A^{-1}, which is constructed as follows:

$$A_{n\times n}^{-1} = \frac{1}{|A|}\begin{pmatrix} A_{11} & A_{21} & ... & A_{n1} \\ A_{12} & A_{22} & ... & A_{n2} \\ & & & \\ A_{1n} & A_{2n} & ... & A_{nn} \end{pmatrix}$$

whereby, A_{ij} is defined as $A_{ij} = (-1)^{i+j} D_{ij}$ and D_{ij} as the $(n-1) \times (n-1)$-dimensional minor of $A_{n\times n}$ by excluding row i and column j.

Because the Covariance-Variance matrix $\Sigma_{K\times K}$ (Eq. A.39) is symmetric and have the same entries in the diagonals or in the non diagonals (and hence, receive the same minors, each), for the inverse of the assumed matrix $\Sigma_{K\times K}$ Eq. A.43 can be followed.

$$\Sigma_{K\times K}^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} (-1)^{1+1}\Sigma_{11} & (-1)^{1+2}\Sigma_{12} & ... & (-1)^{1+2}\Sigma_{12} \\ (-1)^{1+2}\Sigma_{12} & (-1)^{1+1}\Sigma_{11} & ... & (-1)^{1+2}\Sigma_{12} \\ & & & \\ (-1)^{1+2}\Sigma_{12} & (-1)^{1+2}\sigma_{12} & ... & (-1)^{1+1}\Sigma_{11} \end{pmatrix} \tag{A.43}$$

Thus, it satisfies to derive Σ_{11} and Σ_{12}, whereby the determinant for a matrix $\Sigma_{K\times K}$ (Eq. A.39) is known by (A.40) and for a matrix $\tilde{\Sigma}_{K\times K}$ which has the form of Eq. A.44 it can be shown that the determinant is given by Eq. A.45

$$\tilde{\Sigma}_{K\times K} = \begin{pmatrix} \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_t^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & & & \\ & & & & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 & \sigma_b^2 + \sigma_t^2 \end{pmatrix} \tag{A.44}$$

$$\left|\tilde{\Sigma}_{K\times K}\right| = \sigma_b^2\left(\sigma_t^2\right)^{K-1} \tag{A.45}$$

Whereby, the determinante is derived by th following:

$$\left|\tilde{\Sigma}_{K\times K}\right| \overset{\text{1.Z. - 2.Z.}}{=} \left|\overbrace{\begin{pmatrix} 0 & -\sigma_t^2 & 0 & 0 & ... & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & & & \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K \text{ entries}}\right|$$

$$\overset{\text{dev. 1.r}}{=} \left(-\sigma_t^2\right)(-1)^{1+2} \left|\overbrace{\begin{pmatrix} \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_t^2 & \sigma_b^2 & ... & \sigma_b^2 \\ & & & & & \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{(K-1) \text{ entries}}\right|$$

$$\overset{\text{rep.}}{=} \left(-\sigma_t^2\right)^{K-2}(-1)^{K-2} \left|\overbrace{\begin{pmatrix} \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_t^2 \end{pmatrix}}^{K-(K-2)=2 \text{ entries}}\right| \overset{\text{Det. 2-dim.}}{=} \left(\sigma_t^2\right)^{K-2}\left(\sigma_b^2\left(\sigma_b^2+\sigma_t^2\right)-\sigma_b^2\sigma_b^2\right) = \underline{\underline{\left(\sigma_t^2\right)^{K-1}\sigma_b^2}}$$

Hence, for the inverse of the Covariance-Variance matrix Σ the diagonal entries are derived by (Eq. A.46) and analog for the non-diagonal entries by (Eq. A.47).

$$\begin{aligned} \frac{1}{|\Sigma_{K\times K}|}(-1)^{(1+1)}\left[\Sigma_{K\times K}\right]_{11} = \frac{1}{|\Sigma_{K\times K}|}\left|\Sigma_{(K-1)\times(K-1)}\right| &\overset{A.40}{=} \frac{1}{|\Sigma_{K\times K}|}(\sigma_t^2)^{K-2}\left(\sigma_t^2+(K-1)\sigma_b^2\right) \\ &= \underline{\underline{\frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\left(\sigma_t^2+(K-1)\sigma_b^2\right)}} \qquad \text{(A.46)} \\ \frac{1}{|\Sigma_{K\times K}|}(-1)^{(1+2)}\left[\Sigma_{K\times K}\right]_{12} = -\frac{1}{|\Sigma_{K\times K}|}\left|\tilde{\Sigma}_{(K-1)\times(K-1)}\right| &\overset{A.45}{=} -\frac{1}{|\Sigma_{K\times K}|}\sigma_b^2 \\ &= \underline{\underline{-\frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\sigma_b^2}} \qquad \text{(A.47)} \end{aligned}$$

Thus, it could be shown, that Σ^{-1} has the form (A.48). The diagonal and non diagonal entries of the inverse of the assumend Covariance-Variance matrix $\Sigma_{K\times K}$ (Eq. A.39) follows by the Eq. (A.41) and (A.42). Hence, the distribution for a biological individual with K technical replicates follows a K-dimensional normal distribution, because Eq. A.37 holds. The complete data of N biological individuals with K technical replicates each for both classes (mutant, wildtype) follows then a $(2 \times N \times K)$-normal distribution.

$$\left[\Sigma_{K\times K}\right]^{-1} \overset{(A.43)}{=} \underline{\underline{\frac{(\sigma_t^2)^{K-2}}{|\Sigma_{K\times K}|}\begin{pmatrix} \left(\sigma_t^2+K\sigma_b^2\right) & -\sigma_b^2 & ... & -\sigma_b^2 \\ -\sigma_b^2 & \left(\sigma_t^2+K\sigma_b^2\right) & ... & -\sigma_b^2 \\ & & & \\ -\sigma_b^2 & -\sigma_b^2 & ... & \left(\sigma_t^2+K\sigma_b^2\right) \end{pmatrix}}} \checkmark \qquad \text{(A.48)}$$

B. Appendix - Supplemental material to the publications

Here, only some detailed information given by the supplemental of each article are selected. For example mathematical details of used methods, which could help to fully understand the theories of the articles, but also some additional analysis results to help completing the picture.

B.1. Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data

B.1.1. Variance estimation

Only the overall variance, i.e. the sum of technical and biological variances, can be estimated directly from the dataset. To estimate variances at all hierarchical levels, a linear nested regression model was used for each detected feature. Each observed feature intensity in each sample was approximated by modelling the effects of the instrumental, preparation and biological variances.

$$Y_{nei} = \mu + \beta_n + \gamma_{ne} + \delta_{nei}$$

Y_{nei} observed measurement of plant n, extraction e, injection i

μ overall mean of population

β_n independent random effect of plants on level n

γ_{ne} independent random effect of extraction on level e in plants on level n

δ_{nei} independent random effect of injection on level i in extraction on level e in plants on level n

Mean Square Deviation	unbiased estimator
$MSQ_{tot} = \frac{1}{NEI-1} \sum_n^N \sum_e^E \sum_i^I (y_{nei} - y_{...})^2$	$\hat{\sigma}^2_{tot} = MSQ_{tot}$
$MSQ_{biol} = \frac{1}{(N-1)} EI \sum_n^N (y_{n..} - y_{...})^2$	$\hat{\sigma}^2_{biol} = \frac{(MSQ_{biol} - MSQ_{prep})}{EI}$
$MSQ_{prep} = \frac{1}{N(E-1)} I \sum_n^N \sum_e^E (y_{ne.} - y_{n..})^2$	$\hat{\sigma}^2_{prep} = \frac{(MSQ_{prep} - MSQ_{instr})}{I}$
$MSQ_{instr} = \frac{1}{NE(I-1)} \sum_n^N \sum_e^E \sum_i^I (y_{nei} - y_{ne.})^2$	$\hat{\sigma}^2_{instr} = MSQ_{instr}$

Figure B.1.: Linear hierarchical model with 3 levels and deriving unbiased estimation of all variance levels in hierarchical experiment. The random effects β_n, γ_{ne}, δ_{nei} are independent among each other. The mean squared deviation of all observations Y_{nei} in every level leads to an unbiased estimator for all variance levels. y_{nei} = observation of injection i of preparation e of plant n. $y_{ne.}$ = mean of all observation of preparation e of plant n, . $y_{n..}$ = mean of all observation of plant n. $y_{...}$ = overall mean. N = number of plants, E = number of preparations of each plant, I = number of injection of each plant and preparation.

Using the linear nested regression model (Equation 3.1 in the main article) the mean squared deviation of observations can be used as an estimator of several variances, but this is biased. Correcting the mean squared deviation in every level as shown in the table in Fig. B.1 leads to an unbiased estimator for all variance levels: instrumental variance σ^2_{instr}, preparation variance σ^2_{prep}, biological variance σ^2_{biol}, and total variance σ^2_{tot}. We used the data of the pilot study and preprocessing as described in the main article to calculate these values.

B.1.2. Test statistics and their distribution of hierarchical and non-hierarchical models

With a setup of normal-distributed data, which can be sampled hierarchical and non hierarchical, with unknown means μ_i and unknown, but equal variances σ, we derive the test statistics to compare the unknown means of the sampled classes, shown in Table B.1. It results from the assumption of equal means of the different classes in the model of the null hypothesis in comparison with unequal means in the model of the alternative hypothesis. The test with non hierarchical data is the Student's t-test, if only two classes are used and the ANOVA, if more than two classes are available. Using hierarchical models performs the same test statistics, but the observation of the second level are averaged. In contrast to the non hierarchical test the distributions of the alternative hypothesis include the second level variance (technical variance) besides the one level variance (biological variance).

Using derivation of the Maximum Likelihood Ratio of the two models of null and alternative hypothesis to identify the best fitting model leads to test statistics, which is a logarithmic term of the shown $t, \tilde{t}, F, \tilde{F}$.

two sample classes ($C = 2$)	more than two sample classes ($C > 2$)
$t = \frac{\sqrt{\frac{N}{C}} x_{1.} - x_{2.}}{\sqrt{\frac{1}{C(N-1)} \sum_{c,n} (x_{cn} - x_{c.})^2}} \begin{cases} \sim_{H_0} t_{C(N-1)} \\ \sim_{H_1} t_{C(N-1), \sqrt{\frac{N}{C}} \frac{\mu_1 - \mu_2}{\sigma}} \end{cases}$	$F = \frac{C(N-1)}{N-1} \frac{\sum_{c,n} (x_{c.} - x_{..})^2}{\sum_{c,n} (x_{cn} - x_{c.})^2} \begin{cases} \sim_{H_0} F_{N-1, C(N-1)} \\ \sim_{H_1} F_{N-1, C(N-1)}, \frac{\sum_c (\mu_c - \mu)^2}{\sigma} \end{cases}$
$\tilde{t} = \frac{\sqrt{\frac{N}{C}} x_{1..} - x_{2..}}{\sqrt{\frac{1}{C(N-1)} \sum_{c,n} (x_{cn.} - x_{c..})^2}} \begin{cases} \sim_{H_0} t_{C(N-1)} \\ \sim_{H_1} t_{C(N-1)}, \frac{\sqrt{\frac{N}{C}} \mu_1 - \mu_2}{\sqrt{\sigma_{bio}^2 + \frac{\sigma_{tech}^2}{E}}} \end{cases}$	$\tilde{F} = \frac{C(N-1)}{N-1} \frac{\sum_{c,n} (x_{c..} - x_{...})^2}{\sum_{c,n} (x_{cn.} - x_{i..})^2} \begin{cases} \sim_{H_0} F_{N-1, C(N-1)} \\ \sim_{H_1} F_{N-1, C(N-1)}, \frac{\sum_c (\mu_c - \mu)^2}{\sigma_{bio}^2 + \frac{\sigma_{tech}^2}{E}} \end{cases}$

Table B.1.: Test statistic for hypotheses tests and their distributions. If observations of two or more classes are normal distributed, where equal variances (homoscedasticity) are assumed, standard statistical methods can be used to testing the equality of means of the several classes with C number of classes, N number of biological replicates and E number of technical replicates. Upper row describe the test statistics of the non-hierarchical model with $s^2 = \sqrt{\frac{1}{C(N-1)} \sum_{c,n} (x_{cn} - x_{c.})^2}$ the adjusted sample variance of the non-hierarchical model. The lower row describe the test statistics of the hierarchical model with $s_{bio}^2 := \frac{1}{C(N-1)} \sum_{c,n} (x_{cn.} - x_{c..})^2 - \frac{\sigma_{tech}^2}{E}$ the adjusted biological sample variance and $s_{tech}^2 := \frac{1}{CN(E-1)} \sum_{c,n,e} (x_{cne} - x_{cn.})^2$ adjusted technical sample variance of the hierarchical model, using E number of technical replicates of each biological. Normal distributions N_{μ,σ^2} are determined with parameters of μ (position of the distribution) and σ^2 (shape of the distribution), t-distributions $t_{df,ncp}$ and χ^2-distributions $\chi_{df,ncp}$ with the number of degrees of freedom DoF and noncentrality parameter ncp.

B.2. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent

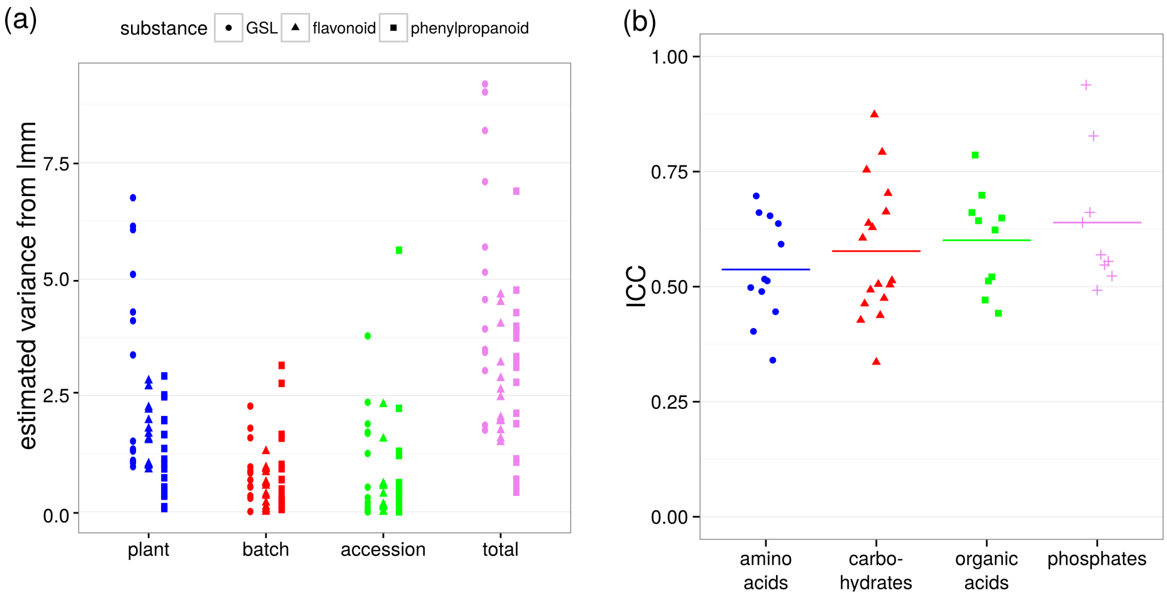


Figure B.2.: Biological variability of annotated primary metabolites. (a) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot—variance of one metabolite; (b) ICCs for carbohydrates, organic acids, amino acids and phosphates, dot—ICC of one metabolite, bar—mean ICC for substance class. GSL = glucosinolate.

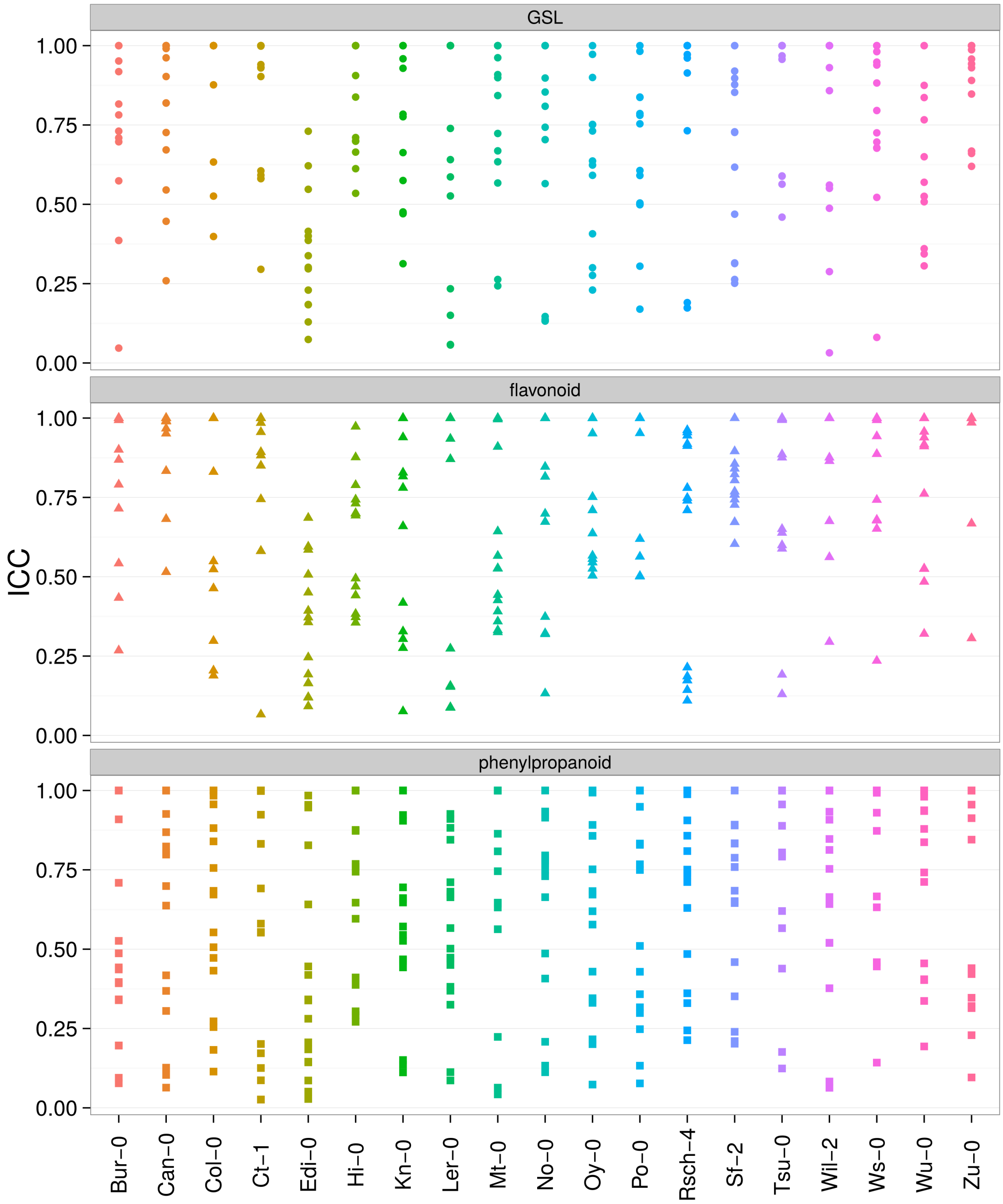


Figure B.3.: Accession-specific ICCs of secondary metabolites. ICCs were determined as $\sigma^2_{plant}/\sigma^2_{total}$ from 19 linear mixed models. GSL = glucosinolate.

B.3. Joint analysis of dependent features within compound spectra can improve detection of differential features

Feature grouping to compound spectra. The rows in the matrix are annotated and the grouping of related features into compound spectra was performed, both with the package CAMERA Kuhl et al., 2012. At first the function xsAnnotate() with all samples, then for grouping the function groupFWHM() was used (see all in Algorithm 1). Figure B.4 shows, that the size of the identified compound spectra varies from 1 to 123, although it has been noted that the compound spectra of size 123 is an injection peak and would not be analysed.

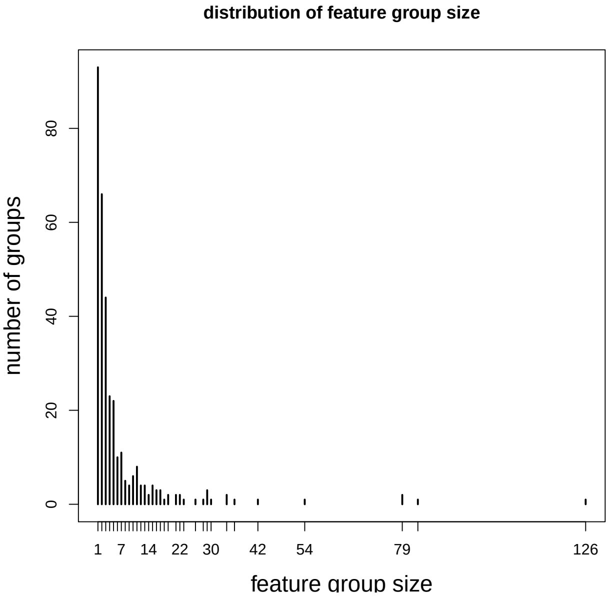


Figure B.4.: Distribution of size of compound spectra identified with CAMERA annotation for the wildtype-mutant experiment.

Algorithm 1 Simulation of a gold-standard dataset on a real dataset to determine the quality of univariate and multivariate tests to detect differentiell features or compound spectra for several numbers effects eff_i.

```
INPUT: DS = dataset of MS measurements of plants of one genotype
INPUT: CSA = compound spectra annotation corresponding to the DS
$F$ #number of row of DS (number of features)
$N$ #number of columns of DS (number of samples)
$C$ #number of compound spectra
# split DS into two classes
for DS do
    class WT ← DS[,1 : $\frac{N}{2}$] # 1 : $\frac{N}{2}$ samples of DS
    class MU.ref ← DS[,$\frac{N}{2}$ + 1 : $N$] # $\frac{N}{2}$ + 1 : $N$ samples of DS
    for all compound spectra j=1,..,C do
        estimate covariance matrix $\Sigma_j$ out of MU.ref
for all eff $\in c(0.2, 0.3, ..., 1.5)$ do
    while $iter < 1000$ do
        for all compound spectrum $j \in 1, ..., C$ do
            # combine Negative dataset as all samples came from the same genotype
            MU(j) ← Mu.ref(j) + X(j),
            X(j)~ $N(0, \sigma_j)$ for each compound spectra j, j=1,..,C
            class MU←c(MU(1),...,MU(C))
            Negatives ← c(WT, MU)
            calculate p.Hotellings, p.DiagHottellings
            for ∀ feature $k \in 1, ..., K_c$ within compound spectrum $j$ do
                calculate p.univariate
                overtaken p.Hotellings, p.DiagHottellings
                if $p < 0.05$ then
                    FP
                else
                    TN
            # combine Positive dataset as an effect is added in one class
            MU(j) ← Mu.ref(j) + X(j),
            X(j)~ $N(eff, \sigma_j)$for each compound spectra j, j=1,..,C
            class MU ←c(MU(1),...,MU(C))
            Positives ← c(WT, MU)
            calculate p.Hotellings, p.DiagHottellings
            for ∀ feature $k \in 1, ..., K_c$ within compound spectrum $j$ do
                calculate p.univariate
                overtaken p.Hotellings, p.DiagHottellings
                if $p < 0.05$ then
                    TP
                else
                    FN
Calculate the number of TP, TN, FP, FN, AUC of all 1000 repeats
```

A.	univariate Student's t-test	multivariate Hotelling's T^2
test statistic	$t = \frac{\bar{x_1}-\bar{x_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $s_p = \frac{(n_1-1)s_1+(n_2-1)s_2}{n_1+n_2-2}$	$T^2 = \frac{n_1 n_2}{n_1+n_2}\left(\bar{X_1}-\bar{X_2}\right)^T S_p^{-1}\left(\bar{X_1}-\bar{X_2}\right),$ $S_p = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$
H_0	$t \le t_{\frac{\alpha}{2},n_1+n_2-2},$	$\frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p}T^2 \le F_{\alpha,p,n_1+n_2-p-1}$

B.	univariate Welch's t-test	multivariate James test
test statistic	$t = \frac{\bar{x_1}-\bar{x_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$T_u^2 = \left(\bar{X_1}-\bar{X_2}\right)^T S^{-1}\left(\bar{X_1}-\bar{X_2}\right),$ $S = \frac{S_1}{n_1}+\frac{S_2}{n_2}$
H_0	$t \le t_{\frac{\alpha}{2},\nu},$ $\nu \approx \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1}+\frac{s_2^4}{n_2^2\nu_2}}$	$T_u^2 \le \chi^2_{\alpha,A+B\chi^2_{1-\alpha,p}}$ $A = 1+\frac{1}{2p}\sum_{i=1}^{2}\frac{tr(S^{-1}S_i)^2}{n_i-1}$, tr=trace of matrix $B = \frac{1}{p(p+2)}\left[\frac{1}{2}\sum_{i=1}^{2}\frac{tr(S^{-1}S_i)^2}{n_i-1}+\frac{1}{2}\sum_{i=1}^{2}\frac{(trS^{-1}S_i)^2}{n_i-1}\right]$

Table B.2.: Formula of the univariate Student's t test Student, 1908, Welch's t-test Welch, 1947, the multivariate Hotelling's T^2 test Mardia, Kent, and Bibby, 2003 in comparison with the James test James, 1954 for two sample classes, where n_i number of observations of sample class i. The Welch's t-test is used to compare two univariate samples on difference in means $\bar{x_1}, \bar{x_2}$ with the assumption of unequal variances s_1^2, s_2^2, where ν_1 are the degrees of freedom associated with the variance estimate of sample i. The Hotelling's T^2 test is used compare two p-dimensional samples on difference in mean vectors $\bar{X_1}, \bar{X_2}$ with the assumption of unknown, but equal covariance matrices S_1, S_2, so the the pooled covariance matrix S_p is used creating the test statistic. The multivariate equivalent to the univariate Welch's t-test is the James test. Here, unknown and unequal covariance matrices S_1, S_2 are assumed. The proposed uncorrelated type of James test uses covariance matrices with only the diagonal entries, the variances.

B.4. Analysing observational data: methodological challenges to address clustering and selection bias, a practical example in health services research in nursing

Propbilities

$$
\begin{aligned}
\text{Risk}_{\text{Control}} = p_{\text{Control}} &= P(\text{Event=yes}|\text{Group=Control}) \\
&= \frac{\text{no(Event=yes, Group=Control)}}{\text{no(Group=Control)}} \\
&= \frac{n_{22}}{n_{.2}} \\
\text{Risk}_{\text{Treat}} = p_{\text{Treat}} &= P(\text{Event=yes}|\text{Event=Treat}) \\
&= \frac{\text{no(Event=yes, Group=Treat)}}{\text{no(Group=Treat)}} \\
&= \frac{n_{21}}{n_{.1}} \\
\text{Risk Ratio}\frac{p_{\text{Treat}}}{p_{\text{Control}}} &= \frac{n_{22}}{n_{21}}\frac{n_{.1}}{n_{.2}}
\end{aligned}
$$

Propbilities in term of odds ratios

$$
\begin{aligned}
\text{Odd(Group=x)} &= \frac{p_x}{1-p_x} \\
&= \frac{P(\text{Event=yes}|\text{Group=x})}{P(\text{Event=no}|\text{Group=x})} \\
&= \frac{P(\text{Event=yes}|\text{Group=x})}{1-P(\text{Event=yes}|\text{Group=x})} \\
&= \frac{\text{Risk}_{\text{Group}}}{1-\text{Risk}_{\text{Group}}} \\
\text{Odds Ratio (OR)} &= \frac{\text{Odd(Group=Treat)}}{\text{Odd(Group=Control)}} \\
&= \frac{\frac{P(\text{Event=yes}|\text{Group=Treat})}{P(\text{Event=no}|\text{Group=Treat})}}{\frac{P(\text{Event=yes}|\text{Group=Control})}{P(\text{Event=no}|\text{Group=Control})}} \\
&= \frac{p_{\text{Treat}}}{1-p_{\text{Treat}}}\frac{1-p_{\text{Control}}}{p_{\text{Control}}}
\end{aligned}
$$

Link function for outcomes using a logistic regression model with a dichtomeous predictor variable

$$
\begin{aligned}
\mathit{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\text{Treat}_i \\
\Leftrightarrow p_i &= \mathit{logit}^{-1}\left(\beta_0 + \beta_1\text{Treat}_i\right) = \frac{e^{(\beta_0+\beta_1\text{Treat}_i)}}{1+e^{(\beta_0+\beta_1\text{Treat}_i)}}
\end{aligned}
$$

Parameter of a logistic regression model with a dichtomeous predictor variable

$$
\begin{aligned}
\beta_0 &= \log\left(\frac{P(\text{Event=yes}|\text{Group=Control})}{1-P(\text{Event=yes}|\text{Group=Control})}\right)\\
&= \log \text{Odd}(\text{Group=Control})\\
\beta_1 &= \log\left(\frac{\frac{P(\text{Event=yes}|\text{Group=Treat})}{1-P(\text{Event=yes}|\text{Group=Treat})}}{\frac{P(\text{Event=yes}|\text{Group=Control})}{1-P(\text{Event=yes}|\text{Group=Control})}}\right)\\
&= \log\left(\frac{\text{Odd}(\text{Group=Treat})}{\text{Odd}(\text{Group=Control})}\right)\\
&= \log \text{Odds Ratio (OR)}\\
\beta_0+\beta_1 &= \log\left(\text{Odd}(\text{Group=Control})\right)+\log\left(\frac{\text{Odd}(\text{Group=Treat})}{\text{Odd}(\text{Group=Control})}\right)\\
&= \log\left(\text{Odd}(\text{Group=Control})\left(\frac{\text{Odd}(\text{Group=Treat})}{\text{Odd}(\text{Group=Control})}\right)\right)\\
&= \log\left(\text{Odd}(\text{Group=Treat})\right)
\end{aligned}
$$

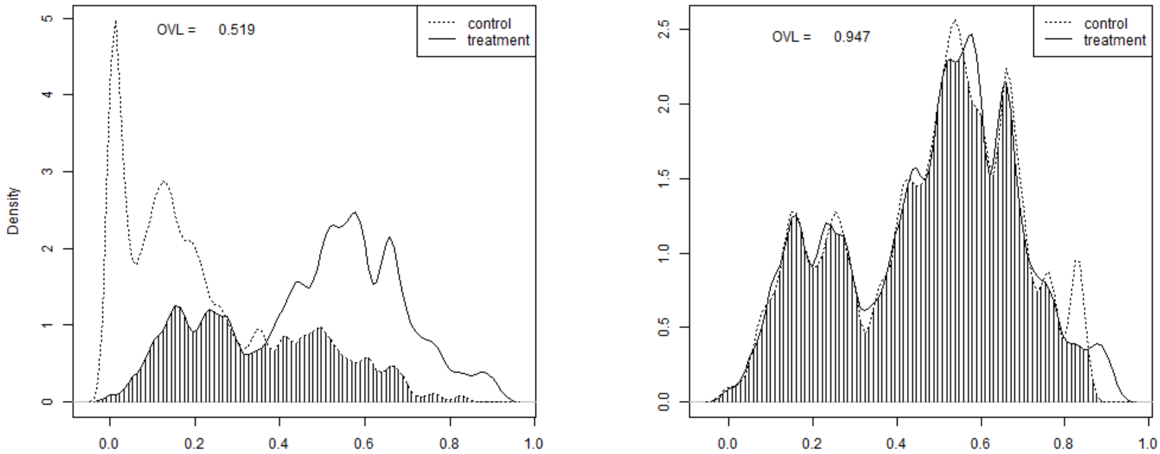


Figure B.5.: Balance check by the overlapping coefficient of propensity score before (left) and after (right) matching on propensity score. The plots show the density of the propensity score within each group and the dashed lines assign the overlap of both.

Diana Trutschel

Experience

since 07/2014 **Research assistant**, *Deutsches Zentrum für Neurodegenerative Erkrankungen e.V, Standort Witten.*

Statistical analysis of health care research data, Implementation of new statistical methods, Statistical guidance

10/2008-11/2013 **Research assistant**, *Leibniz Institute of Plant Biochemistry, Bioinformatics and Mass Spectrometry*, Halle/Saale.

Statistical analysis of mass spectrometry data, Development of experimental design, Statistical guidance

04/2008-08/2008 **Research assistant**, *Group of Prof. Dr. Ivo Grosse (ivo.grosse@informatik-uni.halle.de), Martin-Luther-University*, Halle/Saale.

Sequence analysis, Research and teaching

05/2006-10/2007 **Student Assistent**, *Institut für Pflanzengenetik und Kulturpflanzenforschung*, Gatersleben.

Sequence analysis

10/2000-08/2002 **Nurse**, *emergency hospitalisation and urology*, Halle/Saale.

Education

04/2008 **Diploma in bioinformatics**, *Martin-Luther-University*, Halle/Saale, (*1.6*).

Diploma thesis "Modifizierung des MDD-Algorithmus zur Erkennung von Donorstellen", *Martin-Luther-University*, Halle/Saale, Prof.Dr. Ivo Große. (*1.7*)

10/2001-04/2008 **University studies in bioinformatics**, *Martin-Luther-University*, Halle/Saale.

09/2000 **Graduated nurse**, *Städtisches Krankenhaus Martha-Maria*, Halle/Saale, (*1.3*).

10/1997-09/2000 **Training as a nurse**, *Städtisches Krankenhaus Martha-Maria*, Halle/Saale, (*1.3*).

07/1997 **Final secondary-school examinations**, *Agricolagymnasium*, Hohenmölsen, (*1.4*).

Diana Trutschel

Diana Trutschel

Scientific publications

2018 **“Quality of life in people with severe dementia and its association with the environment in nursing homes: an observational study“**, *Rebecca Palm, Diana Trutschel, Christian G. G. Schwab, Martin N. Dichter, Burkhard Haastert, Bernhard Holle*, The Gerontologist, Vol. .
DOI:

2017 **“Methodological approaches in analysing observational data: a practical example on how to address clustering and selection bias“**, *Diana Trutschel, Rebecca Palm, Bernhard Holle, Michel Simon*, International Journal of Nursing Studies, Vol. 76.
DOI:10.1016/j.ijnurstu.2017.06.017

2017 **“Dementia Care Mapping: Effects on nursing home caregiver attitudes, job satisfaction and burnout. A quasi-experimental trial“**, *Martin Nikolaus Dichter, Diana Trutschel, Christian Günter Georg Schwab, Burkhard Haastert, Tina Quasdorf, Margareta Halek*, International Psychogeriatrics, Vol. 29.
DOI:10.1017/S104161021700148X

2017 **“Feasibility and effectiveness of a telephone-based social support intervention for informal caregivers of people with dementia: Study protocol of the TALKING TIME project“**, *Martin Berwig, Martin Nikolaus Dichter, Bernd Albers, Katharina Wermke, Diana Trutschel, Swantje Seismann-Petersen, Margareta Halek*, BMC Health Services Research, Vol. 17.
DOI:10.1186/s12913-017-2231-2

2017 **“Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance“**, *Rami Al Shweiki, Susann Mönchgesang, Petra Majovsky, Domenika Thieme, Diana Trutschel, Wolfgang Hoehenwarter*, Journal of Proteome Research, Vol. 16.
DOI:10.1021/acs.jproteome.6b00645

2016 **“Differences in Case Conferences in Dementia Specific vs Traditional Care Units in German Nursing Homes: Results from a Cross-Sectional Study“**, *Rebecca Palm, Diana Trutschel, Michael Simon, Sabine Bartholomeyczik, Bernhard Holle*, Journal of the American Medical Directors Association, Vol. 17.
DOI:10.1016/j.jamda.2015.08.018

2016 **“Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent“**, *Susann Mönchgesang, Nadine Strehmel, Diana Trutschel, Lore Westphal, Steffen Neumann, Dierk Scheel*, International Journal of Molecular Sciences, Vol. 17.
DOI:10.3390/ijms17091565

2015 **“Joint analysis of dependent features within compound spectra can improve detection of differential features “**, *Diana Trutschel, Stephan Schmidt, Steffen Neumann, Ivo Grosse*, Frontiers in Bioengineering and Biotechnology, Vol. 3.
DOI:10.3389/fbioe.2015.00129

2015 **“Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data “**, *Diana Trutschel, Stephan Schmidt, Steffen Neumann, Ivo Grosse*, Metabolomics, Vol. 11.
DOI:10.1007/s11306-014-0742-y

2015 **“Dementia care mapping: Effects on residents’ quality of life and challenging behavior in German nursing homes. A quasi-experimental trial“**, *Martin Nikolaus Dichter, Tina Quasdorf, Christian Günter Georg Schwab, Diana Trutschel, Burkhard Haastert, Christine Riesner, Sabine Bartholomeyczik, Margareta Halek*, International Psychogeriatrics, Vol. 27.
DOI: 10.1017/S1041610215000927

Open source software

- **samplingDataCRT**: sampling data for SWD and other study designs in R, available at CRAN: `https://cran.r-project.org/web/packages/samplingDataCRT/index.html`

Diana Trutschel