# Medicometrics

PhD thesis by
Morten Arendt Rasmussen
2012

UNIVERSITY OF COPENHAGEN

# PhD thesis

Morten Arendt Rasmussen

# Medicometrics

Quality and Technology ▪ Department of Food Science

Faculty of Science ▪ University of Copenhagen

**Title**
MEDICOMETRICS

**Submission**
February 1st 2012

**Defence**
May 23rd 2012

**Supervisors**
Professor Rasmus Bro
Department of Food Science, Faculty of Science
University of Copenhagen

Lasse Tengbjerg Hansen
Department Manager for Biomarkers in Global Development
Novo Nordisk A/S

Morten Colding-Jørgensen
Scientific Director
Novo Nordisk A/S

**Opponents**
Professor Torben Martinussen (Chairman)
Department of Basic Science and Environment
University of Copenhagen

Claus Andersson
Partner in Life Science Venture Fund of 500 MUSD

Professor Age Smilde
Swammerdam Institute for Life Sciences
University of Amsterdam

Cover illustration by Sonja Vils Rasmussen


ISBN: 978-87-7611-499-2
Printed by SL Grafik, Frederiksberg C, Denmark

### Medicometrics

Medicometrics is the science of integrating different sources of measurements related to a pathological system. It is an interfacial discipline utilizing elements from applied mathematics, multivariate statistics, chemometrics, pharmacometrics, medicine, biology, biochemistry, etc. The philosophy of Medicometrics is highly exploratory and holistic, aiming at multivariate patterns oppose to single factor associations.

# Preface

This thesis is the outcome of a collaborative work between Quality and Technology (at Department of Food Science, Faculty of Life sciences, University of Copenhagen), an industrial partner Novo Nordisk A/S, research units at hospitals (Danish Pediatric Asthma Center at Gentofte University Hospital, Department of Clinical Biochemistry at University of Southern Denmark in Esbjerg and Pediatric Department at Glostrup University Hospital) and the National Institute for Health Data and Disease Control in Denmark. The aim of the thesis is to develop new methods for handling of data, originating from human intervention trials and patient cohorts. It was (and is) the intention, that these models should be developed in collaboration with- and communicated to the medical doctors, biologist etc. Therefore, most of the published work appear in medical journals.

The present 70 pages is written so that a person with some statistical, mathematical and chemometrical background will be able to fully grasp the content. In case of limited technical knowledge, it should however be possible to follow the argumentation.

I am grateful for the support I have received from all the collaborators in understanding the physiology and biology to an extend that is relevant for development of useful models. Specifically I would like to thank

Lasse Tengbjerg Hansen and Morten Colding Jørgensen, Novo Nordisk A/S for co-supervision, Sinan B. Sarac and Christian H. Rasmussen from the Biomodeling team at Novo Nordisk A/S, Hans Bisgaard, Nilo Følsgaard and Charlotte G. Carson from Danish Pediatric Asthma Center at Gentofte University hospital, Jane Skov from Department of Clinical Biochemistry at University of Southern Denmark in Esbjerg. Marie Louise C. M. Andersen from Pediatric Department at Glostrup University hospital and Sjurdur Olsen and Sesilje B. Petersen from the National Institute for Health Data and Disease Control in Denmark.

Furthermore, I am grateful for the always inspiring supervision I have received from my main supervisor Rasmus Bro. Also I would like to thank my colleagues at Quality and Technology for a very pleasant, casual and professional working environment.

Thanks to Robert Tibshirani for supervision during my 4 month stay at Department of Statistics - Stanford University.

Thanks to Jane Skov and Pernille Vils for proof reading, and Sonja Vils Rasmussen for the front page.

The final appreciations goes to my family and friends for support. Special thanks to Pernille and Sonja for filling my life with stuff more important than algebra.

Morten Arendt Rasmussen
Copenhagen, February 2012

# Abstract

In biological, medical and pharmacological science cause and effect is seldom of binary nature. In order to fully understand these systems a multivariate approach is needed.

The aim of the present work was to develop new multivariate methods for handling of data from clinical trials and patient cohorts with special emphasis on graphical presentation of results. Presented here are stages of the analysis process in a generic form, with focus on obstacles and benefits related to the methods applied.

It is shown, that variation related to different sources can be handled effectively by orthogonalization techniques. For instance irrelevant patient specific variation can be removed prior to modeling. The idea of compressing data into latent component models is shown to unravel intuitive biological patterns. These patterns are furthermore shown to be less uncertain than the raw data. Usage of pattern recognition techniques, such as PCA and PLS, in connection with visualization, is demonstrated as a general framework useful for better data understanding. This leads to biologically intuitive visualizations, where the complexity is much easier to grasp compared to presentation in tables.

It is concluded that novel data analytical techniques, developed in collaboration between physicians, biologists, pharmacologists and data analysts potentially can bring even more new knowledge and that the current data analytical state of the art, is rather conservative and leaves substantial room for improvements.

# Resumé

Årsag og virkning er sjældent binært indenfor biologiske, medicinske og farmakologiske videnskaber. For fuldt ud at forstå disse komplekse biologiske systemer er det nødvendigt med en multivariat tilgang.

Formålet med denne afhandling har været at udvikle nye multivariate metoder til håndtering af data fra kliniske forsøg og patient korhorter, med særlig vægt på grafisk præsentation af resultater. Afhandlingen præsenterer, i en generisk form, stadier af den dataanalytiske proces, med fokus på fordele og ulemper relateret til de anvendte metoder.

Det vises at variation relateret til forskellige faktorer effektivt kan håndteres via ortogonaliserings teknikker. For eksempel kan irrelevant patient specifik variation således fjernes før modelering. Komprimering af multivariate data til latente komponenter vises at resultere i intuitive biologiske mønstre. Disse mønstre viser sig endvidere at være mindre usikre end de rå data. Mønstergenkendelse teknikker, så som PCA og PLS, er i kombination med visualisering, er vist nyttige som en generisk ramme til bedre data forståelse. Disse teknikker fører til biologisk intuitive mønstre, hvor kompleksiteten er lettere at forstå og kommunikere i forhold til for eksempel tabulering.

Det konkluderes, at nye dataanalytiske teknikker, udviklet i samarbejde mellem læger, biologer, farmaceuter og dataanalytikere potentielt kan tilvejebringe ny viden, og at den nuværende *state of the art* inden for biologisk data analyse er temmelig konservativ og efterlader betydelige plads til forbedringer.

# List of publications

Paper I

**Morten A. Rasmussen**, Lasse T. Hansen, Morten C. Jørgensen, Rasmus Bro

Multivariate evaluation of pharmacological responses in early clinical trials - a study of rIL-21 in the treatment of patients with metastatic melanoma

*British Journal of Clinical Pharmacology,* ***69*** *(2010), 379-390*

Paper II

**Morten A. Rasmussen**, Jane Skov, Else Bladbjerg, Johannes Sidelmann, Marianne Vamosi, Jørgen Jespersen

Multivariate analysis of the relation between diet and warfarin dose

*European Journal of Clinical Pharmacology,* ***68*** *(2012), 321-328*

## Paper III

Jane Skov, Else Bladbjerg, **Morten A. Rasmussen**, Johannes Sidelmann, Anja Leppin, Jørgen Jespersen
Genetic, clinical and behavioural determinants of vitamin K-antagonist dose - Explored through multivariable modelling and visualization
*Basic & Clinical Pharmacology & Toxicology,* ***110*** *(2012), 193-198*

## Paper IV

Sesilje B. Petersen, **Morten A. Rasmussen**, Marin Strøm, Thorhallur I. Halldorsson, Sjurdur F. Olsen
Socio-demographic characteristics and food habits of organic consumers: A study from the Danish National Birth Cohort
*Submitted for Public Health Nutrition*

## Paper V

Nilo Følsgaard, Bo Chawes, **Morten A. Rasmussen**, Anne L Bischoff, Charlotte G Carson, Jakob Stokholm, Louise Pedersen, Trevor T Hansel, Klaus Bønnelykke, Susanne Brix, Hans Bisgaard
Neonatal Cytokine Profile in the Airway Mucosal Lining Fluid is skewed by Maternal Atopy
*American Journal of Respiratory and Critical Care Medicine,* ***185*** *(2012), 275-280*

## Paper VI

Charlotte G. Carson, **Morten A. Rasmussen**, Jonas P. Thyssen, Torkil Menné, Hans Bisgaard
Endotyping Atopic Dermatitis Children by Filaggrin Gene Mutation Status in a Prospective Cohort Study.
*Submitted for British Journal of Dermatology*

## Paper VII

Karin Kjeldahl, Morten A. Rasmussen, Annelouise Hasselbalch, Kirsten O. Kyvik, Lene Christiansen, Emma Per-Trepat, Serge Rezzi, Sunil Kochhar, Torkild I. A. Sørensen, Rasmus Bro
No genetic footprints of the fat mass and obesity associated (FTO) gene in human plasma $^{1}$H CPMG NMR metabolic profiles
*in preparation*

## Other publications by the author

### Paper VIII

Kristoffer Laursen, Ulla Justesen, **Morten A. Rasmussen**
Enhanced Monitoring and Detection of Unknown Impurities by LC-MS
*Journal of Chromatography A,* ***1218*** *(2011), 4340-4348*

### Paper IX

Kristoffer Laursen, **Morten A. Rasmussen**, Rasmus Bro
Comprehensive control charting applied to chromatography
*Chemometrics and Intelligent Laboratory Systems,* ***107*** *(2011) 215-225*

### Paper X

MarieLouise C. M. Andersen, **Morten A. Rasmussen**, Sven Pörksen, Jannet Svensson, Jennifer V. Jørgensen, Jane Thomsen, Niels T. Hertel, Flemming Pociot, Jacob S. Petersen, Lars Hansen, Henrik B. Mortensen, Lotte B. Nielsen
Identification of baseline patterns and genetic fingerprints related to beta-cell destruction and ZnT8 autoantibody profiles in Danish children with new onset type I diabetes
*in preparation*

Paper XI

Christian L. Hansen, Frans W. van den Berg, **Morten A. Rasmussen**, Søren B. Engelsen, Steve Holroyd
Detecting variation in ultrafiltrated milk permeates - Infrared spectroscopy signatures and external factor orthogonalization
*Chemometrics and Intelligent Laboratory Systems,* ***104*** *(2010), 243-248*

**Morten A. Rasmussen**, Thomas Janhøj, Richard Ipsen
Effect of fat, protein and shear on graininess, viscosity and syneresis in low-fat stirred yoghurt
*Milchwissenshaft,* ***62*** *(2007), 54-58*

**Morten A. Rasmussen**, Rasmus Bro
Sparse Models - A tutorial
*Submitted for Chemometrics and Intelligent Laboratory Systems*

Sinan B. Sarac, Christian H. Rasmussen, **Morten A. Rasmussen**, Christine E. Hallgreen, Tue Søeborg, Morten C. Jørgensen, Per K. Christensen, Steffen Thirstrup, Erik Mosekilde
A Comprehensive Approach to Benefit-Risk Assessment in Drug Development
*Basic & Clinical Pharmacology & Toxicology, (DOI: 10.1111/j.1742-7843.2012.00871.x)*

Morten Allesø, Sitaram Velaga, Amjad Alhalaweh, Claus Cornett, **Morten A. Rasmussen**, Frans van den Berg, Heidi Lopez de Diego, Jukka Rantanen
Near-Infrared Spectroscopy for Cocrystal Screening. A Comparative Study with Raman Spectroscopy
*Analytical Chemistry,* ***80*** *(2008), 7755-7764*

Ravn, L. S., Andersen, N. K., **Rasmussen, M. A.**, Christensen, M., Edwards, S. A., Guy, J. H., Henckel, P. and Harrison, A. P.
De Electricitatis Catholici Musculari - concerning the electrical properties of muscles; with emphasis on meat quality
*Meat Science,* ***80****, (2008) 423-430*

# List of Nomenclature and Abbreviations

| | |
|---|---|
| $\cos(\cdot)$ | cosine operator |
| $\langle \cdot, \cdot \rangle$ | scalar product |
| $\lceil \cdot \rceil$ | ceil operator |
| $\mathbf{X}$ | matrix |
| $\mathbf{x}$ | vector |
| $\mathcal{N}(\cdot, \cdot)$ | Gaussian distribution. |
| $\mathcal{N}_p(\cdot, \cdot)$ | Gaussian distribution in $p$ dimensions |
| $\mathcal{U}(\cdot, \cdot)$ | Uniform distribution |
| $\overline{\mathbf{x}}$ | a mean vector |
| $\| \cdot \|_p$, $\| \cdot \|_p$ | the $p$ norm of a vector (in this work $p = 1, 2$) |
| $\Phi(\cdot)$ | is the Gaussian operator with mean 0 and variance 1 |
| $\pi$ | prior probability |
| $\rho$ | correlation coefficient |

| | |
|---|---|
| $E(\cdot)$ | Expectation operator |
| $F(\cdot)$ | cumulative distribution function |
| $f(\cdot)$ | probability distribution function |
| $FDR$ | False Discovery Rate |
| $FWER$ | Family Wise Error Rate |
| $L_1$ norm | sum of absolute values of a vector $\lvert \cdot \rvert_1$ |
| $L_2$ norm | (square root of) sum of squared values of a vector $\lvert \cdot \rvert_2$ |
| $SN$ | Signal to Noise |
| $x$ | scalar |
| AdaLasso | Adaptive Lasso |
| Allsub | All subsets regression |
| ANOVA | ANalysis Of Variance |
| ASCA | Anova Simultaneous Component Analysis |
| BE | Backward Elimination |
| DO | Direct Orthogonalisation |
| EPO | External Parameter Orthogonalisation |
| EV | Explained Variance |
| FDA | Food and Drug Administration (US) |
| FSR | Forward Stepwise Regression |
| iPLS | interval Partial Least Squares |
| LARS | Least Angle RegreSsion |
| Lasso | Least Absolute Shrinkage and Selection Operator |
| LDA | Linear Discriminant Analysis |
| MLR | Multiple Linear Regression |

| | |
|---|---|
| MSC | Multiplicative Scatter Correction |
| MSPC | Multivariate Statistical Process Control |
| OLS | Ordinary Least Squares |
| OPLS | Orthogonal PLS |
| OSC | Orthogonal Signal Correction |
| PARAFAC | PARAllel FACtor analysis |
| PARAFASCA | PARAllel Factor Anova Simultaneous Component Analysis |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PCR | Principal Component Regression |
| QSAR | Quantitative Structure Activity Relationship |
| RA | Rheumatoid Arthritis |
| RR | Ridge Regression |
| sCD25 | soluble CD25 |
| SLE | Systemic Lupus Erythematosus |
| SR | Selectivity Ratio |
| SSe | Sum of Squared errors |
| SVD | Singular Value Decomposition |
| VIP | Variable Important for Prediction |
| VKA | Vitamin K Antagonist |

# List of Figures

# Contents

Chapter 1

# Introduction

*You cannot look in one direction. In order to see reality, (you) have to see in three or four or seven dimensions.*

*- Dalai Lama 2008*

During the last decades a huge amount of high throughput methods have been developed, and it is now possible to screen the entire genome [97], to analyze the metabolic profile [64] from e.g. human plasma, to estimate gene transcription activity from microarrays [80] etc. At reasonable cost almost all thinkable biological markers can be quantified. From a scientific point of view these techniques seed a hope for a future with more detailed understanding of health, pathology and life as such.

These novel data structures demand new analytical methods [76], and a number of disciplines coupling mathematics and statistics with the relevant scientific field are emerging.

Traditionally, biological experiments were setup to test a few well defined hypotheses from a few measurable well defined metrics. "An idea about how nature works and a confirmatory test of this". This is sometimes referred to as deductive reasoning. This process is objective, rigorous and adds solidity to current knowledge but does not extend it. An example of such could be a clinical phase III trial. Most research aims at extending the current knowledge which corresponds to the opposite process "Observe how nature phenomenologically manifests itself and try to figure out how it works". This process is sometimes referred to as inductive reasoning. An example of such could be a first-human phase I trial exploring efficacy and safety.

One obstacle is, that studies suited for inductive reasoning are being analyzed by deductive methods. One of the most common techniques to support evidence is the use of a $p$-value, which reflects the probability of the observed data under some hypothesis, often a *null* hypothesis of no difference. Within the statistical community there has been some debate on how appropriate the $p$-value as a single statistics is as support for objective evidence [42, 50, 73]. The use of $p$-values to support evidence in a strictly deductive fashion seems appropriate, at least if supplied with some measure of effect size. Contrary, the use of a $p$-value for inductive reasoning is a fallacy. That is: "We observe a difference for some covariate, hence this difference is caused by the covariate" this is known as the problem of induction [61]. R.A. Fisher, the father of modern statistics, discusses this issue in a paper from 1935 [36]:

> *Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. This was at first assumed; but once the distinction*

> *between the proposition and its converse is clearly stated, it is seen to be an assumption, and a hazardous one. The inferences of the classical theory of probability are all deductive in character. They are statements about the behavior of individuals, or samples, or sequences of samples, drawn from populations which are fully known. ... The fact that the concept of probability is adequate for the specification of the nature and extent of uncertainty in these deductive arguments is no guarantee of its adequacy for reasoning of a genuinely inductive kind.*

It is a fact that a large amount of data is collected and analyzed by scientists with core competences *different* from data analysis. This advocates for optimal reasoning about what is relevant to measure, handling of instruments and how to setup and conduct experiments. On the other hand, software is being produced for analysis of data, without any need for in depth understanding of the mathematical computation and the statistical assumptions behind. The misuse of the $p$-value as a measure for objective evidence is an example of a scientific separation between the data generation and data analytical process, and in its extreme, an obsession of a dichotomous representation of nature.

The present work is based on a merger of statistical and chemometric methodology. The aim is to understand the implication of the different methods and to somehow bridge data analysis and the scientific field for a better extraction of information, more biologically correct modeling and a more complex holistic representation of results as opposed to a black and white perception of nature.

The rest of the introduction is organized with a brief historical overview of statistics and chemometrics followed by four motivating

examples of questions and the data supplied for answering these questions.

## 1.1 Elements of statistical thinking

Statistical reasoning is founded in theory of probability. The concept of logic involves elements of probability and can hence be tracked back to the work of Aristotle (384BC - 322BC), where he in *Prior Analytics* defines probable as

> *"A probability is a generally approved proposition: That men know to happen or not to happen, to be or not to be, for the most part thus and thus" (70a, 5. Cf1357a, 35)"* [63].

Aristotle used probability in development of logical theory for causal reasoning, but did not formalize the concepts in terms of mathematics. In 1654 Blaise Pascal (1623 - 1662) initiated this scientific field as a mathematical discipline. A field which have had great contribution from a number of thinkers. Jakob Bernoulli (1654 - 1705), who with *Ars Conjectandi I-IV* (1713) extend the work from primarily considering games into a much wider area of civil, moral and economic matters [81]. Pierre-Simon Laplace (1749-1827) who in *Théorie analytique des probabilités* (1812) thoughtfully formalized a wide range of theorems related to approximation, finite sample size, central limit theorem, inverse probability, etc. Thomas Bayes (1702 - 1761) a British mathematician formulated Bayes theorem which is the core of conditional probability. In the same period of time mathematical methods for solving least squares was developed by Johann Carl Friedrich Gauss (1777 - 1855) and Adrien-Marie Legendre (1752 - 1833) and the ideas of correlation and regression to the mean was brought to life by Francis Galton (1822 - 1911) [8].

In the beginning of the 20$^{\text{th}}$ century frequentist statistics as a statistical subfield was founded. With the possibility of oversimplification, data is a realization of an unknown distribution and the frequentist theory formalizes methods that based on data estimates this distribution. Karl Pearson (1857 - 1936), a student of Francis Galton and head of the first university department of statistics (University College London), formalized the chi-squared test for evaluation of frequency data and developed the Pearson's correlation coefficient, extending Galton's work on correlation, measuring linear dependence between entities [87]. In 1908 William Sealy Gosset (1876 - 1937) under the alias "Student" developed the t-test for evaluation of unknown finite normal samples, a method that has had huge impact on applied statistics [89]. In 1918 Ronald A. Fisher (1890 - 1962), an evolutionary biologist and the father of modern statistics founded the analysis of variance (ANOVA) [35]. ANOVA is probably his most significant contribution to experimental science, but also statistical concepts like maximum likelihood, sufficiency, ancillarity and Fishers Information and methods like Fishers linear discriminant analysis (LDA) and permutation tests are significant contributions by Fisher. Jerzy Neyman (1894 - 1981) and Egon Pearson (1895-1980) (son of Karl Pearson) in 1930's developed the concepts of type II error [70], power of a test [69] and confidence intervals [68]. Frequentist methodology is the primary classical methods used today for analysis of experimental data within medical, pharmacological and biological science.

## 1.2 Elements of chemometrical thinking

Chemometrics is an application emerging discipline emerging from analytical chemistry. As a scientific field chemometrics was initiated by the analytical chemists Svante Wold and Bruce Kowalski in the early 1970's. With the development of analytical instruments with multi-

variate output, such as spectrometers, additional tools for handling of data were at demand. The aim was to develop a framework based on mathematics and statistics for handling of data with chemical origin. That is, to mathematically formulate chemically relevant models using terms that makes chemical sense and hereby unravel the black box methodology of mathematical models, with a reformulation using a more intuitive vocabulary. One example is the chemometrical work-horse Principal Component Analysis (PCA). PCA is mathematically related to eigen decomposition of (two) covariance matrix's. This methodology was identified by e.g. Leonard Euler (1707 - 1783) [34] and Joseph Louis Lagrange (1736 - 1813) [84] and extended into singular value decomposition (SVD) by Camille Jordan (1838 - 1921), Eugenio Beltrami (1835 - 1899), James Joseph Sylvester (1814 - 1897), Erhard Schmidt (1876 - 1959) and Hermann Weyl (1885 - 1955) [86, 5]. In 1901 Karl Pearson published eigenvector decomposition of real matrices with information truncation in few components and hence produced the first example of the PCA as we know it today [74]. PCA was introduced to analysis of chemical data by e.g. Bruce Kowalski in 1973 [56]. Mathematically PCA refers to eigenvalues, eigenvectors, rotated basis etc. Chemometrically PCA is characterized by latent components, scores, loadings, Hotellings, residuals etc. These are e.g. used for exploration of sample distribution in high dimensions and inter variable correlation structure as a mean for scientific reasoning. Numerically the two approaches are similar but different in the way the model is characterized and used. Parallel Factor Analysis (PARAFAC) is another example of an existing mathematical model, which were rediscovered and translated into chemometrics. PARAFAC is a method for component decomposition of higher order arrays identified by Richard A. Harshman (1943 - 2008) [47] and Carroll & Chang [14] in the area of mathematical psychology (psychometrics) and translated to analytical chemistry by C. J. Appellof and E. R. Davidson in

1981 [4] and further refined by e.g. Rasmus Bro [11]. In addition to reformulation of existing mathematical models, methods have been developed within the chemometrical society where Partial Least Squares (PLS) for regression purposes is the most remarkable contribution [100].
Chemometrics as a data analytical discipline rely on mathematical modeling and visualization. While mathematical modeling is the foundation for all data analytical disciplines, the use of distinct visualization tools such as score- and loading plots for model interpretation and residuals versus Hotellings $T^2$ plot for outlier detection is a methodology developed within chemometrics extending the ideas of data visualization from e.g. John Tukey (1915 - 2000) [96] and Edward Tufte (1943 - ) [94].

## 1.3 Four Cases

In order to specify the context of this thesis in terms of specific scientific questions and the data provided for unraveling these questions, four examples are presented. The examples correspond to real questions and data I have been analysing, but are here presented in a generalized form. When relevant, I will throughout the thesis refer to these cases either directly or as the paper where they appear.

### Early exploratory clinical trials

Drug development is extremely costly, why the number of patients enrolled in trials is attempted to be reduced to a minimum. Early clinical trials (Phase I) have the primary purpose of 1) addressing safety / side effects issues related to treatment and 2) revealing the pharmaco- kinetic and dynamic profile of the drug. Most side effects can have multiple causes. This in combination with low number of patients and high patient to patient variation complicates assessment

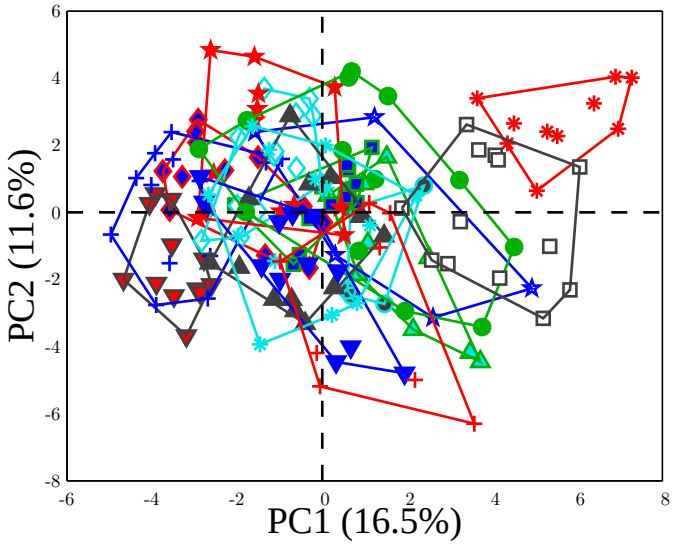



Figure 1.1: Score plot of 16 patients evaluated at 15 time points on 43 biomarkers. The samples are grouped according to patient.

of whether the effect is truly systematic treatment related or if it is a sporadic event. Figure 1.1 shows a score plot from a Principal Component Analysis (PCA) on 16 patients evaluated with respect to 43 safety and efficacy biomarkers examined at app. 15 time points. This score plot highlights the huge inter patient variation. This variation is not of primary interest and hence represent an obstacle for extraction of relevant information. Furthermore, the 43 biomarkers are obviously related physiologically, which can be utilized in model building.

Unraveling questions related to drug action rely on methods for handling of variation sources, derived responses and exploration. These questions are treated in Paper I [77].

### Disease progression

It is of utmost importance to understand the biological behavior of chronical diseases, such as diabetes, heamostasis, gout etc. Understanding the interplay between disease and behaviorial parameters such as treatment, physical activity, diet etc. and fixed parameters such as genetics and gender can help medical guidance for a better life. Observational study cohorts, where a patient population are examined cross sectional or longitudinal, are often the basis for better disease and/or treatment understanding. Such data comprise of factors related to lifestyle, general health, genetics, physical activity, diet etc. and often consist of a few hundred patients. The questions of interest relates to interplay between genetical-, environmental factors and treatment. Paper II [78], Paper III [82] and Paper X are examples of work relevant in this context.

### Inflammation markers

Through the development of micro arrays, expression levels of a large amount of different genes can be assessed and a fair amount of case control studies tries to explore transcriptomic disease patterns in order to enhance the pathologic understanding and e.g. develop targeted drugs. These studies often only examines a small group of patients ($\sim 100 - 200$) compared to the number of genes ($> 10^5$). Univariate screening of such a high number of variables will by chance return false associations at even rather high significance threshold levels, why issues related to feature selection and control of false discovery is of crucial importance.

**Organic food consumption during pregnancy related to prevalence of childhood asthma**

Addressing factors relating offspring health and the mothers lifestyle, diet, physical activity etc. during pregnancy is a rather complicated task. One way to reveal these connections is through large observational epidemiological studies. Denmark have the largest (in the world) birth cohort with $\sim 50.000$ women [71]. Organic foods are assumed health beneficial due to different production conditions avoiding use of artificial fertilizers and pesticides compared to conventional production [17]. However, there is a strong association between the healthiness of the diet and the proportion of organic foods in such an observational data material. Revealing causality between organic food consumption during pregnancy and childhood asthma addresses questions related to confounding and complexity understanding. Paper IV [75] document this confounder challenge based on the Danish National Birth Cohort.

These examples highlight the complexity of the nature that we for example through data try to reveal. In the following I will present different approaches to three stages of the data analytical process. The first stage called *Basis truncation* refer to the handling of variation sources through projection, stratification or variable selection and is here seen as an initial step to focus the model. The second stage called *Uncertainty estimation* refer to estimation of the relevant uncertainty related to the parameters of interest, and especially the issue of multiple testing is presented. The third stage called *Interpretation* refers the process of inferring biological connections and relevance from the model. Here issues of bias and visualization are discussed. The three parts are connected as a data analytical chain, but are here presented as independent sections. Throughout the thesis small examples based on simulated data are presented in gray boxes in order to substantiate

the argumentation. Results based on real data are presented in the published work and are grossly omitted from this thesis.

*Chapter 2*

# Basis truncation

One task in data analysis is to focus on the *relevant* part of the variation. Focusing is here seen as the process of including specific variation while excluding the rest. This process can be approached from both the sample and the variable direction, and can hence be seen as a row (sample) space basis truncation or a column (variable) space basis truncation. Handling of confounders or covariates is a row space basis truncation, where stratification is truncation of the diagonal basis, and handling through projections is a truncation of a rotated basis. Imagine for example gender as a confounding variable. Narrowing the analysis to only include e.g. women, corresponds to sample stratification, while subtracting the response difference related to gender corresponds to collapsing the gender dimension and is hence truncation by projection. Likewise variable selection is truncation of the diagonal basis in the variable space. Projections to latent structures (such as PCA and PLS) are examples of truncation of a rotated basis in both sample and variable space. In the following different truncation techniques are presented, and some advantages are discussed.

## 2.1 Row space truncation

Identification and estimation of different sources of variation in relation to a certain response is a discipline anchored in the analysis of variance (ANOVA) [35]. These ideas have been extensively elaborated within the statistical community and models like analysis of covariance, mixed effect models, split plot designs etc. are today *of the shelf* methods implemented in statistical software and used widely for analysis of research data from various fields. Within the field of chemometrics similar approaches are used in methods like ANOVA Simultaneous Component Analysis (ASCA) [83], PARAFASCA [55] for $\mathbf{X}$ block only models and External Parameter Orthogonalization (EPO) [79], Orthogonal Signal Correction (OSC) [101], Direct Orthogonalization (DO) [3] for calibration optimization.

Source of variation has basically two faces. Take a scalar response $y$ and let it be normally distributed. Let $y$ be dependent of two (dichotomous) factors $x_1$ and $x_2$ such that $y \sim \mathcal{N}(\mu(x_1, x_2), \sigma^2(x_1, x_2))$. Imagine a setup where $y$ reflects the weight of an animal, $x_1$ reflects the age of the animal with the levels $10days$ and $1year$ and $x_2$ correspond to some treatment. Deviance between means ($\mu(x_1, x_2)$) is referred to as systematic offset. Differences in covariance structure ($\sigma^2(x_1, x_2)$) is a random effect. For this example it is likely that the variance or uncertainty in weight for the animals of age $1year$ is higher than variance for the animals of age $10days$. This is reflected as differences in the diagonal elements of $\sigma^2$ and is referred to as heteroschedasticity.

Methods based on projections are solely related to handling and estimation of variation due to offsets related to systematic effects. That could for example be handling of differences between gender or differences between subjects in cases of repeated measures. In such cases the variation related to gender or subject can be removed by collapsing of the dimension related to gender or subject respectively.

Differences in variance structure can sometimes be handled by pretreatment of data e.g. logarithm transformation. This field is however not as extensively elaborated as handling and estimation of variation related to systematic offset. In the following a walk through of variation handling via projections is given.

### Orthogonalization

The core of orthogonalization splitting procedures is based on projection of a vector onto two orthogonal subspaces. The projection step is a variable wise operation. Therefore, it is sufficient to consider the univariate case. In order to make life simple, imagine a response ($\mathbf{y}$) dependent of two factors ($\mathbf{x_1}$ and $\mathbf{x_2}$). Using the setup from above, the dependent variable ($\mathbf{y}$) can be expressed as an additive model with homoschedastic error.

$$\mathbf{y} = \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \mathbf{e} \tag{2.1}$$

Where $\mathbf{e} = [e_1, \ldots, e_n] \sim \mathcal{N}(0, \sigma^2)$ and $\beta_1$, $\beta_2$ and $\sigma^2$ are the model parameters. $\mathbf{y}$, $\mathbf{x_1}$ and $\mathbf{x_2}$ are centered.

Let $\mathbf{x_1}$ be the factor of interest (e.g. treatment) and $\mathbf{x_2}$ be a nuisance factor. Correction of $\mathbf{x_2}$ on $\mathbf{y}$ can be done through a projection of $\mathbf{y}$ onto the *null* space of $\mathbf{x_2}$. That is:

$$\mathbf{y}_o = \left( I - \mathbf{D_{x_2}} \left( \mathbf{D_{x_2}}^T \mathbf{D_{x_2}} \right)^{-1} \mathbf{D_{x_2}}^T \right) \mathbf{y} \tag{2.2}$$

where $\mathbf{D_{x_2}}$ correspond to a design matrix for categorical $\mathbf{x_2}$. If $\mathbf{x_2}$ is continuous and assumed linearly related to $\mathbf{y}$, $\mathbf{D_{x_2}} = \mathbf{x_2}$. The part of $\mathbf{y}$ related to $\mathbf{x_2}$ is hence:

$$\mathbf{y}_{x_2} = \mathbf{y} - \mathbf{y}_o = \left( \mathbf{D_{x_2}} \left( \mathbf{D_{x_2}}^T \mathbf{D_{x_2}} \right)^{-1} \mathbf{D_{x_2}}^T \right) \mathbf{y} \tag{2.3}$$

If $\mathbf{x_1}$ and $\mathbf{x_2}$ are orthogonal e.g. from an experimental design, $\mathbf{y}_{x_2}$ will be the least squares estimates for $\beta_2$ in equation 2.1 and regression of $\mathbf{y}_o$ on $\mathbf{x_1}$ will likewise produce the least squares estimates for $\beta_1$.

Under such optimal circumstances data can be interpreted with respect to one factor at a time. ASCA, PARAFASCA and EPO rely on this design orthogonality condition.

The orthogonality condition is however not always fulfilled. Examples of such could be: Observational studies and designed experiments with exclusion of samples.

If $\mathbf{x_1}$ and $\mathbf{x_2}$ are correlated ($\langle \mathbf{x_1}, \mathbf{x_2} \rangle \neq 0$ for centered vectors) the projection in equation 2.2 will unintentionally remove information from $\mathbf{y}$ related to $\mathbf{x_1}$ corresponding to the degree of correlation between $\mathbf{x_1}$ and $\mathbf{x_2}$. This complicates interpretation from such a filter and potentially damage the predictable performance in a calibration setting. An alternative is to correct $\mathbf{y}$ only using the *unique* part of $\mathbf{x_2}$. That is to sustain *all* the information related to $\mathbf{x_1}$. This can be achieved through two projection steps first estimating the part of $\mathbf{x_2}$ in the *null* space of $\mathbf{x_1}$.

$$\mathbf{x_{2o1}} = \left( I - \mathbf{D}_{\mathbf{x}_1} \left( \mathbf{D}_{\mathbf{x}_1}{}^T \mathbf{D}_{\mathbf{x}_1} \right)^{-1} \mathbf{D}_{\mathbf{x}_1}{}^T \right) \mathbf{x_2} \tag{2.4}$$

Then use $\mathbf{x_{2o1}}$ for projection of the response variable.

$$\mathbf{y}_o = \left( I - \mathbf{x_{2o1}} \left( \mathbf{x_{2o1}}{}^T \mathbf{x_{2o1}} \right)^{-1} \mathbf{x_{2o1}}{}^T \right) \mathbf{y} \tag{2.5}$$

$\mathbf{x_{2o1}}$ is orthogonal to $\mathbf{x_1}$ and shorter than $\mathbf{x_2}$ ($|\mathbf{x_2}|_2 > |\mathbf{x_{2o1}}|_2$), hence the variation removed is smaller and information related to $\mathbf{x_2}$ still occurs in $\mathbf{y}$. I refer to this as *restricted orthogonalization*. For a graphical illustration see Figure 2.1

In Figure 2.2 are shown how much of the *relevant* variation, that is; the part of the response ($\mathbf{y}$) that is related to the factor of interest

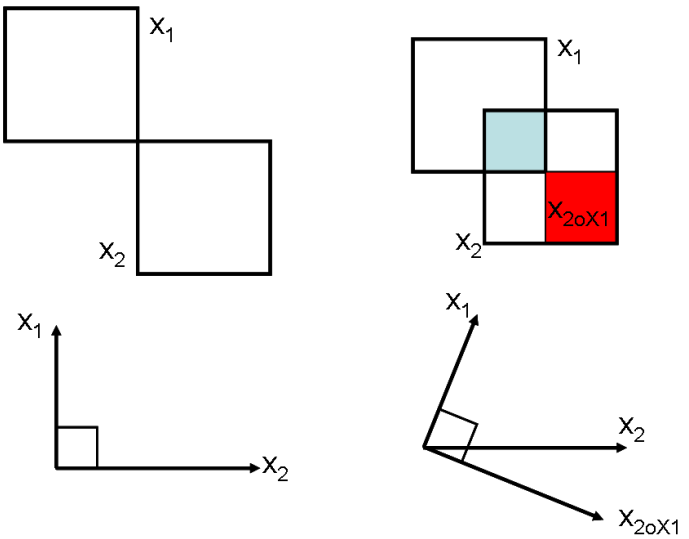



Figure 2.1: Left panel: Two orthogonal factors $\mathbf{x_1}$ and $\mathbf{x_2}$ with an empty space intersection. Right panel: Two correlated factors $\mathbf{x_1}$ and $\mathbf{x_2}$, with an intersection. The *unique* part of $\mathbf{x_2}$ is shown as the vector/space $\mathbf{x_{2ox1}}$

($\mathbf{x_1}$), that is removed by orthogonalization with the nuisance factor ($\mathbf{x_2}$) for full (-) and restricted (-) orthogonalization. It is evident that the higher the correlation between $\mathbf{x_1}$ and $\mathbf{x_2}$ the choice of orthogonalization impacts the results. In Hansen *et al* (2010) (Paper XI) and Kjeldahl *et al* (2011) (Paper VII) are shown applications of *restricted orthogonalization* [46]. In Paper I (Case I) full orthogonalization is used for handling of variation related to differences between patients in a longitudinal phase I trial examining safety biomarkers.

In terms of interpretability the choice of either total- or partly deflation of variation related to a nuisance factor (here $\mathbf{x_2}$) is difficult, and it is important to emphasize that single factor conclusion is related

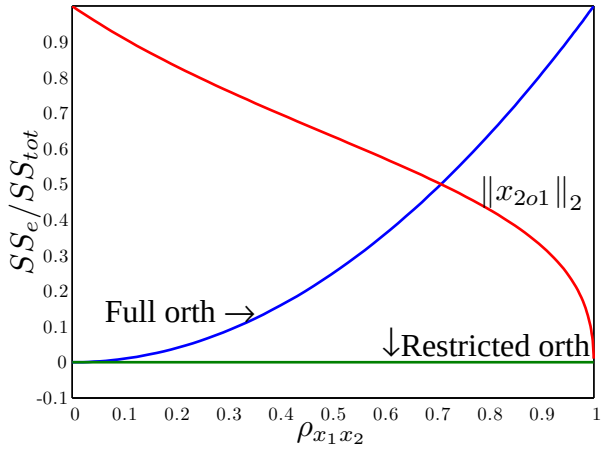



Figure 2.2: Variation removed ($SS_e/SS_{tot}$) by orthogonalization as function of correlation between $\mathbf{x_1}$ (factor of interest) and $\mathbf{x_2}$ (nuisance factor) ($\rho_{\mathbf{x_1},\mathbf{x_2}}$)). $|x_{2o1}|_2$ correspond to the *length* of the vector used for the orthogonalization (see equation 2.5)

to the uniqueness and hence the subspace overlap/correlation between other factors, measured or not.

### Relation to Least Squares

For orthogonal factors, the corresponding least squares estimates of $\beta_1$ and $\beta_2$ in equation 2.1 can be calculated from a joint model, via (two) univariate regressions or by successively orthogonalization followed by regression. All of these approaches will produce exact identical results. For correlated factors ($\rho_{\mathbf{x_1},\mathbf{x_2}} \neq 0$) this is not the case. In comparison with the two orthogonalization strategies described above, where either the length of the factor to be removed or the length of the remaining factor is preserved, estimation by least squares is a compromise between the two. The intersect between the two spaces spanned

by $\mathbf{x_1}$ and $\mathbf{x_2}$ (the blue square in Figure 2.1) is split between the two parameters $\beta_1$ and $\beta_2$.

### Stratification for handling of confounders

Confounders are defined as intermediate factors that modulate the relation between response and dependent variable. An example of a confounder could be the general healthiness of the diet in a study comparing the degree of organic food consumption with some outcome. Confounders handled via model expansion (i.e. including the confounder as independent variable in a linear model) rely on correct representation of confounder-response association (linear, log, polynomial, ordinal, categorical etc.). Furthermore, the response pattern for the variable of interest should be homogeneous across the entire confounder space. Imagine that we wish to estimate the impact of broccoli intake in relation to the risk of strokes from an observational cohort. Here, a natural confounder could be the general healthiness of the diet such that healthy diet is strongly correlated with high intake of broccoli. In this context a homogenous response pattern means, that regardless of how healthy you are, the impact on risk of strokes of broccoli intake is the same. This is referred to as exchangeability. Exchangeability is the idea that for *one factor* the individuals can exchange groups without changing the covariance in relation to the parameter of interest. In the present context it means that all the healthy becomes unhealthy (exchange) and vice versa without changing the correlation between intake of broccoli and healthiness. An obvious demand is that intake of broccoli and healthiness are independent and uncorrelated. Optimally the covariance structure is broken by an experimental design, from which answers concerning broccoli and strokes can be drawn unambiguous in relation to healthiness [44, 43, 67]. If these assumptions are violated the estimation of relevant effect size is biased depending on the degree of confounding [98]. Sample stratifi-

cation is a way to surpass these issues.

Stratification is the process of partitioning the distribution to get a more homogeneous (sub-) sample. For example, patients for a clinical trial have to meet certain criterions based on age, comorbidity, BMI etc. This can be seen as a priory stratification. Another example is gender stratification where e.g. cases are compared with controls for males and females respectively. In the small example above it means that risk of strokes are compared between high intake of broccoli and low intake of broccoli only for subjects with the *same* level of healthiness. Mathematically effect estimation by stratification can be formulated as:

$$y_i = \beta_C x_i + e_i \tag{2.6}$$

Where $i = I_C$ (the indicator for stratum $C$), $y_i$ is the response variable, $x_i$ is the variable of interest and $\beta_C$ the corresponding effect estimate. $e_i \sim \mathcal{N}(0, \sigma_C^2)$

**Example: Stratification vs. model expansion**

This example compares stratification with model expansion for handling of confounder related variation.

Let $\mathbf{x}$ be the variable of interest and $\mathbf{z}$ be a confounder variable. The "correlation" between the two variables are between zero and one ($0 < \rho_{x,z} < 1$).The response variable ($\mathbf{y}$) is constructed such that it is related to $\mathbf{z}$ through a second order polynomial but *not* related to $\mathbf{x}$. Both $\mathbf{x}$ and $\mathbf{z}$ are truly continuous, but due to practical circumstances measured as two level

categorical variables. Formally data are constructed as follows with 500 observations:

$$\mathbf{x} \sim \mathcal{U}(0, 2)$$
$$\mathbf{e} \sim \mathcal{U}(0, 2)$$
$$\mathbf{z} = \rho\mathbf{x} + (1 - \rho) \cdot \mathbf{e}$$
$$\mathbf{e_y} \sim \mathcal{N}(0, 1.5^2)$$
$$\mathbf{y} = \mathbf{x} \cdot 0 + \mathbf{z}^2 \cdot 3 + \mathbf{e_y}$$

The observations we actually see:

$$\mathbf{x}^* = \lceil \mathbf{x} \rceil$$
$$\mathbf{z}^* = \lceil \mathbf{z} \rceil$$

Where $\lceil \cdot \rceil$ is the ceiling operator.

In Figure 2.3 are shown the effect estimate calculated using normal confounder adjustment (equation 2.7) and stratification based on the level of the confounder (equation 2.8).

Linear model

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 z_i^* + e_i \tag{2.7}$$

for $i = 1, \ldots, 500$

Stratified models

$$y_i = \beta_0 + \beta_1 x_i^* + e_i \tag{2.8}$$

for $i = I(z^* = 1)$ respectively $i = I(z^* = 2)$

$\beta_0, \beta_1$ and $\beta_2$ are estimated by least squares $(\min(\|\mathbf{e}\|_2^2))$.

From Figure 2.3 it is registered that normal confounder adjustment assuming exchangeability, i.e. independence between confounder and variable of interest, when this is not the case, bias the solution. On the other hand when the assumption is valid the estimates are less uncertain.

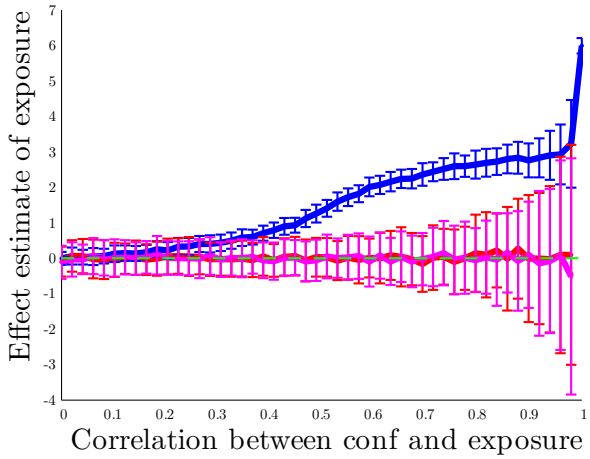



Figure 2.3: Effect size estimation (mean ± std) based on confounder adjustment and confounder stratification (one for each strata) (number of simulation = 100)

Here an example with a single confounder is presented. These ideas can be extended to the multivariate confounder case using methods for

clustering of samples in order to construct strata. An example where these issues are relevant can be found in Paper IV (Case IV) [75], where effect of organic food consumption is highly confounded with the general healthiness of the diet. This complicates inference estimation between degree of organic food consumption during pregnancy and some relevant outcome e.g. development of childhood asthma for the offspring.

## 2.2 Variable selection

Variable selection has two main purposes. To improve predictive performance and to enhance interpretability. All prediction models are based on minimization of an error term and hence have a well defined cost function. Improvement of performance is straightforward, as model selection based on a valid error term (e.g. by cross- or test set validation) will have enhanced performance. Interpretability on the other hand do not have a well defined cost function and can not be formalized as a pure mathematical optimization problem. Therefore, caution must be taken when interpreting performance optimality as causality.

Outlined, variable selection techniques like Forward Step wise Regression, Lasso, iPLS, Genetic Algorithms etc. results in two groups: Active and passive variables. The group of active variables consists of variables related (directly or indirectly) to the outcome and variables that by chance have spurious correlations with the outcome, either alone or in combination with other active variables. The passive group consists of variables irrelevant for prediction of the outcome and relevant but redundant variables. Vaguely formalized the cost function, with respect to interpretability, should result in only removal of the irrelevant and spurious correlated variables.

With the change of oversimplification selection techniques can be grossly divided into four categories:

**Forward techniques** which successively add variables that increase the performance. These are for example: Forward Stepwise Regression, iPLS (with forward selection), the different Lasso types (see section 2.2), etc. Common for these techniques is, that redundant information in the passive set will remain passive.

**Backward techniques**. Based on a model using all variables, successively elimination of single variables or variable groups are done as long as performance increases. Examples of algorithms are: Backward Elimination and iPLS (with backward elimination). These techniques remove irrelevant variables but retain the spurious.

**Model assessment**. Based on a model using all variables, variables are excluded via assessment of model characteristics such as hard thresholding of (small) regression coefficients, Variables Important for Prediction (VIP), Selectivity Ratio (SR) etc. [2].

**Model space search paths**. Assessment of all combination of variables (All Subsets) is practically impossible for more than app. 20 variables ($2^{20} > 10^6$). Furthermore, exhaustive search strategies like *All Subsets* suffer from increased selection bias (see section 4.3). Alternative methods searching the model space is e.g. Genetic Algorithms, where parent models are mated producing children model. The parent model performance is used as a prior probability of selection, such that good models are up weighted and bad models are down weighted in the breeding process through model generations. For further details on genetic algorithms see Leardi and Gonzalez (1998) [60]. Forward Selection and Backward Elimination are highly restricted search paths and are hence a subcategory of model space search paths techniques.

For most of these generic methods, the ability to deselect irrelevant and spurious variables while sustaining redundant information is sacrificed in the optimization process. From an algorithmic point of view backward elimination techniques are the only techniques intrinsically sustaining redundant information while excluding irrelevant variables. None of the methods are capable of handling spuriously correlated variables.

### A note on the Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso) is a biased regression method that embeds variable selection. Since it was discovered in 1996 by R. Tibshirani it has received a lot of attention, and might be one of the most influential statistical methods of the last twenty years [92].

Lasso is a penalized regression operator [91] constraining the Ordinary Least Squares (OLS) solution. The OLS solution can be expressed as regression of $\mathbf{y}$ $(n \times 1)$ on $\mathbf{X}$ $(n \times p)$ minimizing the squared residuals:

$$\hat{\beta}^{OLS} = \text{argmin}_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \langle x_i \beta \rangle)^2 \tag{2.9}$$

This produces the unbiased solution. However future prediction performance can be improved from a biased solution adding a penalty to the minimization problem in equation 2.9 [54]. Examples of some popular biased solutions are Ridge Regression (RR), Principal Component Regression (PCR) and Partial Least Squares regression (PLS). Ridge Regression puts a bound on the $L_2$ norm of the regression coefficients ($\|\beta\|_2^2 \leq \lambda$) [49], PCR truncates the rank of $\mathbf{X}$ to only include the largest principal directions [39] and PLS reduces the dimension of $\mathbf{X}$ using the basis estimated as eigenvectors to the matrix $\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$, successively replacing $\mathbf{y}$ with the current residuals [19]. The Lasso penalizes the $L_1$ norm of the regression coefficients ($\|\beta\|_1^1 \leq \lambda$). For feasible values of $\lambda$ this penalty produces sparse solutions (i.e. a number of the regression coefficients are exactly zero) and hence embeds variable selection [91]. Several variants of the Lasso are known/used.

**Lasso**. The OLS solution is scale invariant, as variables with high variance obtain correspondingly lower regression coefficients. Due to the bound on the $L_1$ norm of the regression coefficients, the Lasso is scale variante, and under normal circumstances the dependent variables (columns of $\mathbf{X}$) are hence scaled to equal variance.

**Adaptive Lasso**. In the normal formulation the predictors are scaled to equal variance in order not to favor selection of variables with high variance. The Adaptive Lasso puts prior emphasis on the variables with high OLS regression coefficients through a priory scaling of the predictors with $\left|\hat{\beta}^{OLS}\right|$ followed by Lasso on the scaled predictors [103].

**Fused Lasso**. The Fused Lasso utilizes any ordering of the variables, that is, if $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ is related in e.g. time or as wavelength, m/z ratios etc. Under such circumstances it might make sense to bound $\beta_i - \beta_{i+1}$ $(i = 1, \ldots, p-1)$. Bounding the $L_1$ norm of the gradient of $\beta$ $(\sum_1^{p-1} |\beta_i - \beta_{i+1}|_1^1 \leq \lambda)$ is referred to as the Fused Lasso [93].

**Group Lasso**. The Lasso is unable to select two highly correlated predictors which is known to be an advantage for e.g. spectral data. The intuition is that, inclusion of two highly correlated variables have a high price in terms *two* regression coefficients compared to just one. The grouped Lasso combines the selection properties of the Lasso and the shrinkage properties of RR and keeps or kills predefined groups of variables in the estimation [62].

For estimation, the Least Angle Regression (LARS) algorithm efficiently produce the entire solution from $\hat{\beta}^{Lasso} = \overrightarrow{\mathbf{0}}$ to $\hat{\beta}^{Lasso} = \hat{\beta}^{OLS}$ for Lasso and the Adaptive Lasso [29]. Cyclical coordinate descend algorithms including a soft threshold operator can efficiently compute

$L_1$ penalized solutions for a grid of penalties and is hence a more general approach for solving different Lasso type of problems [40].

**Is the Lasso a forward- or a backward procedure?**

In the estimation of solution path by LARS or coordinate descend, the algorithms start at $\hat{\beta}^{Lasso} = \overrightarrow{\mathbf{0}}$ and then increase the penalty ($\lambda$), which results in inclusion of more and more variables. From this perspective it can hence be seen as a forward procedure. Nevertheless this is only due to computational circumstances. For $n > p$ problems, starting at the OLS solution ($\lambda = \left\|\beta^{OLS}\right\|_1^1$) followed by decrease of the penalty, would produce the exact same solution path, as the Lasso criterion is a convex optimization problem, which has an unique solution. Due to the similarities with Forward Stagewise Regression (see Efron *et al* (2004) for details on this technique), and the fact that the *largest* solution only contains $r$ active parameters ($r = \text{rank}(\mathbf{X})$), the Lasso is more similar to forward- than backward procedures [29].

**Example: Variable selection techniques**

This is example is included in order to compare different variable selection techniques for linear models with respect to performance and which variables that are included for different models.

The setup is $\mathbf{X}$ $(n \times p)$ with $n = 50$ and $p = 30$ and $\mathbf{y}$ $(n \times 1)$.

$$\mathbf{X} \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}_y$$

Both $\mu$ and $\beta$ are drawn from a Gaussian with mean 0 and variance 1 and $diag(\mathbf{\Sigma}) = 1$, $offdiag(\mathbf{\Sigma}) = 0.3$. The error on the dependent variable ($\mathbf{e}_y$) is defined as: $\mathbf{e}_y \sim \mathcal{N}(0, 10^2)$

Model search paths (a sequence of models ranging from sparse models with few active variables to dense models with all variables being active) are estimated using five different techniques - Three MLR methods: Forward Stepwise Regression, Backward Elimination and All Subsets and two shrinkage methods: Lasso and Adaptive Lasso.

The different models are evaluated with respect to performance on a test set with 5000 samples (see Figure 2.4c). For comparison of search paths, the regression vectors are arranged in a three-way array with the dimensions:

1. The five algorithms
2. The regression coefficients $(\beta_1, \ldots, \beta_{30})$
3. How many active variables $(1, \ldots, 30)$

As some of the algorithms produce several models with the same number of active variables, the following restriction is incorporated. For All Subsets, the best performing solution with $v$ active variables ($v = 1, .., 30$) is chosen. For Lasso and Adaptive Lasso the shortest solution (in terms of the $L_1$ norm) with $v$ active variables ($v = 1, .., 30$) is chosen.

This array is decomposed by Tucker3 [11] using centering in first and second mode and scaling in second mode. The first mode components are plotted for exploration of search path similarities (see Figure 2.4a).

Forward Stepwise Regression was programmed using the Statistics Toolbox™ (v7.4) for Matlab®(v7.11.0.584 R2010b). Backward Elimination was programmed using *in house* algorithm for Matlab®. All Subsets Selection was estimated using the R package `leaps` in R (v2.11.1). Lasso and Adaptive Lasso were calculated using `glmnet` package for Matlab®. Tucker3 decomposition was calculated using the PLS toolbox (v6.0.1) for Matlab®.

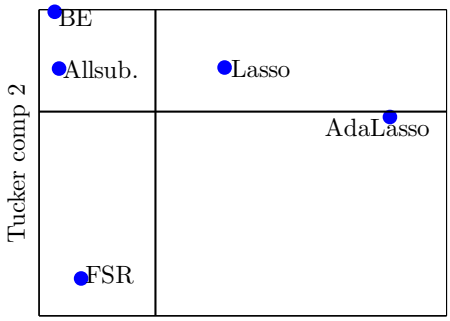
The results indicate that the five methods can be partitioned into two groups both depending on predictive performance (see Figure 2.4c) and regression coefficient estimates (see Figure 2.4a). The MLR group with least squares solution for the active variables (Backward Elimination, All Subsets and Forward Stepwise Regression) and the shrinkage methods, with shrunken (compared to least squares) estimates for the active variables (Adaptive Lasso and Lasso). Obviously, the start and end of the solution path are similar across all five methods. Investigation of the first component in the third mode of the Tucker model, related to *how many* variables that are included, suggests that it is especially models with 15 to 25 active variables that are different between the two groups (MLR vs. shrunken), where the MLR methods

obtain larger regression coefficients (see Figure 4.2b). The difference between the MLR models (Forward Stepwise Regression vs. All Subsets and Backward Elimination) seen in the second component (first mode) is likewise mostly related to first component, third mode. There do not seem to be a pattern reflecting *which* regression coefficients (small or large) that cause this difference (second mode of Tucker model, results not shown).

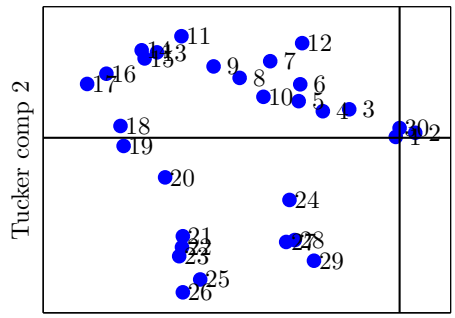
From this it seems to make a difference which variable selection strategy that is used. Although different approaches produce models with equally good predictive performance, the single regression coefficient estimates and variables included, might differ significantly.

### Interpretability

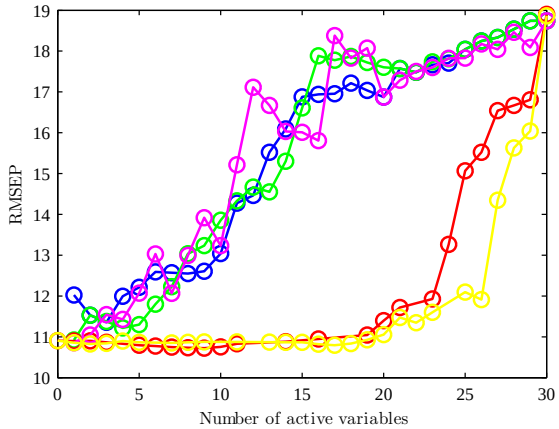
Beyond predictive performance, enhancement of interpretability by variable selection is partly related to the ability of the human brain to grasp complex patterns. In this context variable selection serves as a method to condensate the complexity into a few factors which the human brain can grasp. Imagine for example that we wish to compare the degree of physical activity with the change in a high number of different pathways of the metabolism. An oracle tells us, that actually *everything* is affected by change in the level of physical activity and furthermore that the different pathways also affect each other. Some of the pathways might only show small changes, whereas others show high dependency but none of them are independent of change in level




(a) First mode




(b) Third mode




(c) Performance

Figure 2.4: Results from comparison of different variable selection techniques. (a) and (b) First- and third mode scatter plot of Tucker3 decomposition of regression coefficients from; Forward Stepwise Regression (FSR), Backward Elimination (BE), All Subsets (Allsub), Adaptive Lasso (AdaLasso) and Lasso. (c) Test set performance versus number of active variables for different variable selection techniques; Forward stepwise regression, Backward elimination, All subsets, Lasso and Adaptive lasso.

of physical activity. If the entire system is analyzed in accordance with the oracle knowledge, the inference might be extremely hard to interpret. In order to reduce the complexity and focus only on the pathways that are highly affected, a model with variable selection can be used. This reductionist process is not necessarily to get a more correct model of the system but merely a way to surpass the deficiency of cognitive capacity[1].

[1] Cognitive capacity is the total amount of information the brain is capable of retaining at any particular moment.

## 2.3 Derived responses

In latent component models like PCA and PLS the idea of compressing the $n$ by $p$ matrix into a few components ($k << \min(n, p)$) can be seen as a way to surpass mathematical problems in fitting a regression vector and is in this perspective a similar approach as other multivariate regression techniques such as Ridge Regression (RR) [49]. From a predictive point of view, RR and regression through principal components (Principal Component Regression (PCR)) are fairly similar. Where RR shrinks all the eigenvalues with a constant term, which have highest impact for low variance (eigenvalue) components, PCR deselects these components. This results in fairly similar models in terms of predictability [39]. Boulesteix and Strimmer (2006) discuss the advantages of PLS modeling in a genomic setting, with respect to statistical efficiency, diversity of applications (survival, classification etc.), computational cost etc. [9]. Beyond these advantages latent component models have *the components* as additional features. In most medical research the aim is firstly to obtain higher degree of knowledge and secondly to build models with good predictive performance.

Interpretation of single components or combinations of components, through graphical visualization may lead to simple and meaningful patterns. Eriksson *et al* (2006) use PLS to analyze Quantitative Structure Activity Relationship (QSAR) from which the derived latent variables could be interpreted as fundamental chemical characteristics as lipophilicity, polarity etc. [33]. In the work of Rasmussen *et al* (2012) (Paper II) PLS is equally used for estimation of maintenance dose of vitamin K antagonists (VKA) for treatment of patients at risk for arterial or venous thrombosis. The predictors consisted of selected relevant genetic polymorphisms and measures of lifestyle, health status and diet. Compression of data into components revealed

patterns describing the relation between health related behavior and diet. Interpretation of these patterns leads to a biological intuitive understanding of the interplay between the factors relevant for estimation of the maintenance dose of VKA [78]. Rasmussen *et al* (2010) (Paper I) and Carson *et al* (2012) (Paper VI) present work where PCA has been used to derive interpretable and medically meaningful components, which in addition are *less* uncertain than the individual variables [77, 15].

## 2.4 Power enhancement

Statistical power refers to the probability of finding a difference (rejecting the *null* hypothesis) given that there is a difference. Statistical power is hence directly related to effect size, residual variance and number of test samples. Increasing the number of hypothesis tests in a single trial, increases the probability of rejecting a true *null* hypothesis at a given significance level by chance. As a consequence the significance level for rejection is lowered with resulting loss of power. Prosaically speaking, testing costs. Hence, as the number of variables/tests conducted increase, the power decrease.

Consider $\mathbf{X}$ being a matrix of $p$ variables and $n$ samples. Of interest is, which variables that can be considered significantly different for some covariate. The setup is hence the same as for multiple testing (see section 3.1). From these methods (False Discovery Rate ($FDR$) control and Family Wise Error Rate ($FWER$) control), it is evident that $p$ (the number of variables) has impact on the significance level used for splitting the variables into *nulls* and *non nulls*. These methods are subjected to the assumption of independence between variables. In several applications this is seldom the case. E.g. for spectroscopic data this is never the case, as several variables reflects signal from the same analyte, why the relevant rank ($r$) is (almost) always less than the number of variables ($r < p$). Even for biological data, different chemical compounds are connected through pathways such as the immune response system or certain metabolic pathways. It is evident, that different phenomena reflected by the system are *a combination* of different variables. The relevant rank for such a system might be just as high as the number of variables, but is not uncovered by independent univariate tests across the $p$ variables.

Control of the type I error, assuming independence between variables, results in too conservative error rates, when data are dependent.

To circumvent this problem, we have considered two different strategies. One, which enhances the standardized effect size through *error reduction* by derived components and one, which aims at the *relevant rank*, i.e. the number of tests conducted.

### Standardized effect size enhancement

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be two *connected* effect responses measured across $n$ individuals. For computational ease let them both be Gaussian with the same variance ($\sigma^2$) and with a correlation of $\rho$. Of interest is the population effect estimate ($\mu_x$ and $\mu_y$) and especially if these can be rejected to be equal to zero. The test power is here related to $\sigma$, or its estimate ($\hat{\sigma}$), estimated effect size ($\hat{\mu_x}$ and $\hat{\mu_y}$) and the number of samples $n$. The scalars of interest are:

$$\frac{\hat{\mu}_x}{\hat{\sigma}\sqrt{n}} \qquad \text{and} \qquad \frac{\hat{\mu}_y}{\hat{\sigma}\sqrt{n}}$$

Lets assume that $x$ and $y$ are reflecting to some extend the same condition, e.g. different biomarkers related to the same pathway. From this perspective it makes sense to construct a single derived response ($z$) based on both $x$ and $y$ from a linear combination/weighted average of the two. Let

$$z_i = \alpha x_i + (1 - \alpha) y_i$$

where $0 < \alpha < 1$.

From calculus $z_1, \ldots, z_n$ is normal distributed with mean:

$$\mu_z = \alpha \mu_x + (1 - \alpha) \mu_y$$

and variance

$$\begin{aligned}\sigma_z^2 &= \alpha^2\sigma^2 + (1-\alpha)^2\sigma^2 + 2\alpha(1-\alpha)\sigma^2\rho \\ &= \sigma^2(1 + 2(1-\rho)(\alpha^2 - \alpha)) \\ &\leq \sigma^2.\end{aligned}$$

As $\mu_z$ is between $\mu_x$ and $\mu_y$ the expectation of the test measure $E\left(\frac{\hat{\mu}_z}{\hat{\sigma}_z\sqrt{n}}\right)$ will be more powerful than at least one of the test measure for $x$ or $y$ [45].

In addition to a more powerful test, the number of tests conducted is reduced (from two to one) leading to more power due to issues related to multiple testing [21].

The derived measure ($z$) is here set by a defined combination of $x$ and $y$. If the derived measure for example is based on results from a principal component analysis via scores, the same arguments are valid.

**Example: Power of components**

A small simulation study on a Gaussian mixture with 2 to 1000 variables is setup to investigate the association between correlation ($\rho_{x_1x_2}$) and test power ($\hat{\mu}/\hat{\sigma}_{\hat{\mu}}$). The setup is a paired test and the data are constructed as follows:

$$\mathbf{X_0} \sim \mathcal{N}_p(\mu_\mathbf{0}, \mathbf{\Sigma}) + \mathcal{N}_p(\mathbf{0}, \mathbf{I})$$
$$\mathbf{X_1} \sim \mathcal{N}_p(\mu_\mathbf{1}, \mathbf{\Sigma}) + \mathcal{N}_p(\mathbf{0}, \mathbf{I})$$

where $\mu_\mathbf{0} = \overrightarrow{\mathbf{0}}$, $\mu_\mathbf{1} = \overrightarrow{\mathbf{0.5}}$, $diag(\mathbf{\Sigma}) = 1$ and $offdiag(\mathbf{\Sigma}) = \rho$.

**Univariate** test is a paired t-test.

**Multivariate decomposition via SVD**

$$\begin{bmatrix} \mathbf{X_0} \\ \mathbf{X_1} \end{bmatrix} = \mathbf{1} \cdot \overline{\mathbf{x}}^T + \begin{bmatrix} \mathbf{U_0} \\ \mathbf{U_1} \end{bmatrix} \mathbf{DV^T}$$

Where $\overline{\mathbf{x}}$ is the mean of $\begin{bmatrix} \mathbf{X_0} \\ \mathbf{X_1} \end{bmatrix}$, $\mathbf{D}$ is diagonal, $\mathbf{V}$ is the right eigenvectors and $\begin{bmatrix} \mathbf{U_0} \\ \mathbf{U_1} \end{bmatrix}$ is the left eigenvectors.

The first component, i.e. the first coloum of $\mathbf{U_0}$ and $\mathbf{U_1}$ is used as single variable entries in a paired t-test. The results are shown in Figure 2.5. These results confirm that principal components effectively reduce uncertainty and enhance test power compared to single variables especially when the data covaries.

Paper I (Case I) examines safety biomarker profiles from a clinical phase I study via principal component analysis. Here, five biomarkers related to liver toxicity clearly exhibit similar patterns [77]. Paper V examine levels of 18 different cytokines and chemomkines in neonates. Grossly these 18 different biomarkers exhibit the same pattern, and are hence sufficiently described by a few principal components [37].

### Power through effective rank

In multiple testing (see chapter 3.1) the aim is to control the type I error, i.e. the number of false rejections. Under assumptions concerning (in-)dependency etc. analytical methods can be derived for estimation of the relevant threshold [7]. In the work by Laursen *et al* (2011) (Paper IX) on multivariate statistical process control (MSPC), we use

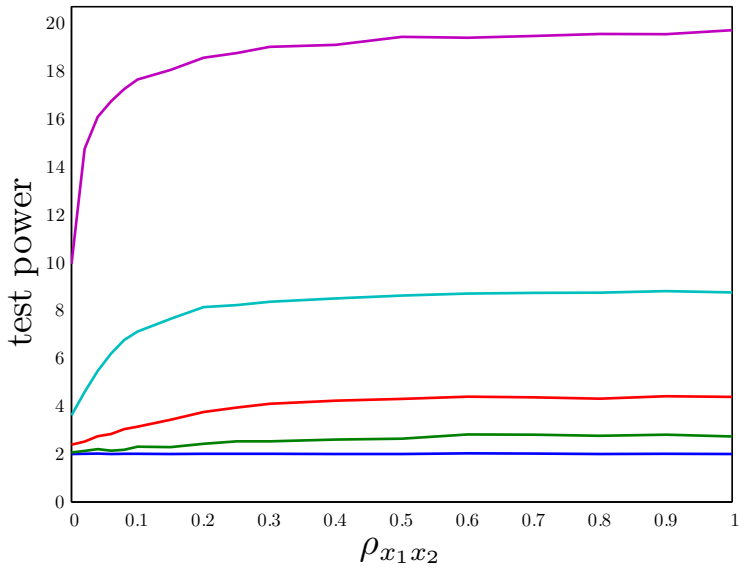



Figure 2.5: Test of difference, based on raw data or on first component from SVD with 2, 5, 20 and 100 variables. Covariance matrix is 1 in diagonal and $\rho_{x_1 x_2}$ outside diagonal. Number of simulations = 1000

a data driven bootstrap procedure to estimate an entire distribution of statistical control charts at different threshold limits [24]. Examination of these distribution on independent *null* samples reveals a distribution of false discovery frequencies in relation to threshold limits. Recursively the threshold bound can be estimated. In this application the data driven procedure produces less conservative thresholds, compared to the analytical derived (assuming independence), and highlights that the rank of the system is less than the number of variables measured.

The two approaches (*Standardized effect size enhancement* and

*Power through effective rank*) use the covariance structure to effectively increase the power and hence decrease the error rate.

*Chapter 3*

# Uncertainty estimation

Models preferably should be accommodated with some sort of relevant uncertainty estimate. That could e.g. be a hypothesis test probability of some relevant hypothesis or a confidence interval for the accuracy of the estimate. Uncertainty estimation in most of the work conducted for this thesis relies on widely accepted uncertainty estimation techniques like e.g. cross validation for component selection, analytical methods for test probability and confidence intervals or permutation testing for test probabilities where there is no simple analytical expression for such. In the work by Laursen *et al* (2011a, 2011b) (Paper VIII and IX) and Kjeldahl *et al* (2011) (Paper VII) we use a multiple testing framework for uncertainty estimation [59, 58]. As this is quite novel in the scientific field of chemometrics, this section solely focuses on the methodology of type I error control in case of multiple testing.

## 3.1 Multiple testing

In the beginning of the $20^{\text{th}}$ century, agriculture experimentation led Fisher to develop analysis of variance for testing of single response association with design factors [35]. High throughput techniques like microarrays today produce numerous (often $> 10^4$) responses simultaneously. This data structure is far from optimally suited for classical frequentist theory, as developed by Neyman, Pearson and Fisher [28]. This fact has in parallel emerged a statistical discipline for testing of multiple hypothesis simultaneously.

Let $\mathbf{X}$ $(n \times p)$ be a matrix of $p$ variables and $n$ samples. Of interest is *which* variables can be considered significantly different for some covariate ($\mathbf{y}$). A normal approach is to produce a statistic for each of the $p$ variables. Depending on the nature of the covariate (categorical, e.g. treatment and control or numerical) and the distribution of the $p$ variables, *of the shelf* methods like unpaired t-tests, regression with the possibility of including covariates or similar test for the non parametric cases can be used to produce one statistic ($z$) for each of the $p$ variables $(z_1, \ldots, z_p)$. The variables, obtaining the largest statistic in terms of *distance* to the *null* hypothesis, are considered the most interesting. The end goal is to split the $p$ variables into a significant part and a non significant part solely based on $z_i$ $(i = 1, \ldots, p)$. This split can be done controlling different error rates. The two most common types are control of the False Discovery Rate ($FDR$), where the rate of wrong rejections is controlled, or Family Wise Error Rate ($FWER$) control, where the possibility of at least *one* wrong rejection is controlled.

### Control of type I error

Let $(p_1, \ldots, p_p)$ reflect *null* hypothesis test probabilities for the $p$ variables and let $z_i = \Phi^{-1}(p_i)$ be the corresponding $z$ value ($\Phi(\cdot)$ is the Gaussian operator with mean 0 and variance 1). The largest $z_i's$ (positive and negative) are reflecting the *most* interesting variables, whereas $z_i's$ close to 0 do not show deviation from the *null* hypothesis.

#### Family Wise Error Rate

For ease, let $z_1, \ldots, z_k, z_{k+1}, \ldots, z_p$ and $p_1, \ldots, p_k, p_{k+1}, \ldots, p_p$ be ordered such that $z_1$ and $p_1$ reflect the *most* significant- and $z_p$ and $p_p$ reflect the *least* significant variable. $z_k$, $p_k$ and $z_{k+1}$, $p_{k+1}$ reflect the statistics for the smallest significant and the largest non significant variable respectively. Olive Jean Dunn developed the well known Bonferoni method for controlling the $FWER$ [22]:

$$FWER = P(X \leq 0)$$

Here, $X$ is the number of falsely rejected hypothesis. This leads to rejecting all the hypothesis for which $p_i \leq \alpha/p$, where $FWER \leq \alpha$.

The Bonferoni correction is conservative, where the trade off between type I and type II error is in favor of type I, as the number of falsely accepted hypothesis is high. This is especially relevant when the number of tests is high, and is primary due to the prior assumption of all hypothesis being *null* and independent [6, 88].

**False Discovery Rate**

An alternative to $FWER$ control is control of the false discovery rate ($FDR$). $FDR$ is defined as:

$$FDR = E(X)/n_{rej}$$

where $X$ is the number of falsely rejected hypothesis and $n_{rej}$ is the number of rejected hypothesis.

A number of methods calculate the $FDR$, especially Benjamini and Hochbergs method is widely used [6]. This might be due to the simple analytical form. Let $p_1 < p_2 < \ldots < p_p$ be ordered *null* hypothesis test statistics: Let $q$ be a fixed $FDR$ and $N$ the number of tests. Then find the largest $i$ for which:

$$p_i \leq i/Nq$$

Among others Efron (2004) [26] and Storey (2002) [88] have derived methods for calculating the $FDR$, assuming that the $p$ test statistics $(z_1, \ldots, z_p)$ come from a mixture density:

$$F(z) = \pi_0 F_0(z) + \pi_1 F_1(z)$$

where $\pi_0$ and $F_0(\cdot)$ refer to the proportion of true *null* hypothesis and its cumulative distribution respectively. $\pi_1$ and $F_1(\cdot)$ are the corresponding for the true non *null* hypothesis. For further details see [26, 88]. Laursen *et al* (2011a, 2011b) (Paper VIII and IX) and Kjeldahl *et al* (2011) (Paper VII) incorporate this methodology for inference estimation in cases of multiple testing [59, 58].

## 3.2 Multiple testing vs. multivariate models

Let $\mathbf{X}$ $(n \times p)$ be a matrix of $p$ variables and $n$ samples and $\mathbf{y}$ $(n \times 1)$ be a response. Of interest is:

- Is there information in $\mathbf{X}$ which is relevant for $\mathbf{y}$?
- Which of the $p$ variables can be considered significant?

Two approaches revealing this association are considered:

1. $p$ univariate test.
2. One multivariate model which include variable selection.

Do these two approaches produce similar results in terms of interesting variables? In order to answer this question an example is conducted on real microarray data for investigation of inflammatory diseases (Case III).

**Example: Multiple testing vs. Multivariate model**

This example is included to compare the set of selected variables based on either multivariate modeling or multiple testing methodology.

Two datasets with microarray samples from white blood cells from patients suffering from the inflammatory diseases *systemic lupus erythematosus* (SLE) (#48 case #21 control) and *rheumatoid arthritis* (RA) (#34 case #21 control) are used. The data have 37202 and

17491 variables respectively.

**Univariate model**

$p$ two sided t-test were conducted. The probability under the *null* hypothesis of no difference was used for sorting the $p$ variables.

**Multivariate model**

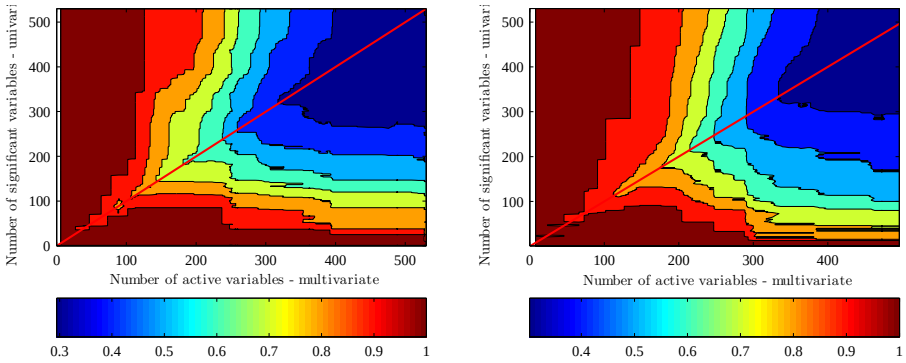
A logistic regression penalized by elastic net. Elastic net is a combination of two penalties on the regression coefficients: A ridge penalty bounding the $L_2$ norm of the $\beta$ which handles problems due to colinearity and a Lasso penalty bounding the $L_1$ norm of $\beta$ for variable selection (see section 2.2) [104]. In this example the balance between the two penalties was set to 1:9 ($L_1$:$L_2$), as colinearity, due to the #variable/#sample ratio, dominates these data.

Figure 3.1 shows the relative overlap between the two methods (in terms of variables selected) as a function of how many variables that are included. Let $S_{multi}$ and $S_{uni}$ denote the set of included variables for the multivariate and univariate model respectively. Let $|S|$ be the number of active variables for $S$. The relative overlap $RO$ is defined as:

$$RO = \frac{|(S_{multi} \cap S_{uni})|}{min(|S_{multi}|, |S_{uni}|)}$$

Comparing the selection path for the two methods, it is obvious and beyond any doubt that these are not independent. Nevertheless, for models including more

than 30-50 variables, the search path deviates and hence produces different results. The model spaces highlighted here correspond to non overfitted models (tested by 10 fold cross validation for the multivariate case) and $FDR < 10^{-8}$ (SLE) and $FDR < 10^{-6}$ (RA) (tested by Benjamini and Hochbergs method [6]). Elastic net was computed using `glmnet` package for Matlab®.




(a) SLE (b) RA

Figure 3.1: Comparison of two strategies for selecting important variables; A univariate (t-test) and a multivariate (elastic net). The proportion of common variables ($RO$) by the two strategies is reflected by the color. (a) corresponds to microarrays comparing cases of *systemic lupus erythematosus* with healthy controls. (b) corresponds to microarrays comparing cases of *rheumatoid arthritis* with healthy controls.

*Chapter 4*

# Interpretation

After modeling and uncertainty estimation, the results are interpreted in terms of scientific findings. This task relies on understanding the model restrictions and assumption limitations. Hence, this process can be a cradle for introduction of biases.

Interpretation of results is a task that can be supported by proper representation of results, e.g. through visualization.

This chapter is constructed with a discussion on pit-falls in interpretation of single estimates from a multivariate model. Following this, a section discussing data visualization is presented. The chapter is closed with discussion of possible biases induced throughout the scientific process.

## 4.1 Interpretation of regression coefficients

Multivariate models with several independent variables ($\mathbf{X}$) for estimation of a single response ($\mathbf{y}$) are often used for estimation of inference between single independent variables and response simply by evaluation of the corresponding regression coefficient. This effect is formulated as the marginal effect (on the response), and relies on the assumption, that the independent variables are independent.

Brown and Green (2009) investigated stability of regression coefficients from multivariate PLS models with dependent ($\mathbf{X}$) variables. Their results indicate that regression vector estimates based on consecutive draws from the *same* population diverge both in magnitude and sign, while having fairly similar predictive performance [13]. It is hence the minimization problem that becomes constant, but not necessarily the regression vector estimates. That means that all the models work equally well, but interpretation of the different regression vectors leads, wrongly, to different conclusions. The study of Brown and Green (2009) simulates a fairly simple system, like vibrational spectroscopy where the signal ($\mathbf{X}$) is indeed a sum of the pure signals. Even under such idealized conditions regression vector interpretation is shown problematic [13].

In cases where a linear model is known to be a simplistic model of the truth, where e.g. interactions and non measurable quantities are omitted from the model, these issues might be even more limiting with respect to single variable interpretation.

In the work by Rasmussen *et al* (2012) (Paper II) [78] the aim is to explore the relation between health related behavior with particular emphasis on diet and genotypes for determination of vitamin K antagonist (VKA) treatment. Especially the relation between intake of vitamin K from the diet and the dose of VKA is of interest. Here

the regression coefficient (the marginal effect) for dietary vitamin K intake from a joint model is negative, which is intuitively non-sense from a cause and effect point of view. Nevertheless it can be shown that dietary vitamin K intake is related to physical activity and body weight. This explains the observed results, where low intake of vitamin K is connected to higher body weight. Also it highlights that it is impossible to make marginal conclusions from multi factorial studies based on observational data, and that truncation of results into *one* dimension (a regression vector, a single OPLS component, etc.) can be hazardous in terms of interpretability when crucial data assumptions concerning independence are violated [31, 32].

## 4.2 Complexity understanding - Exploration and visualization

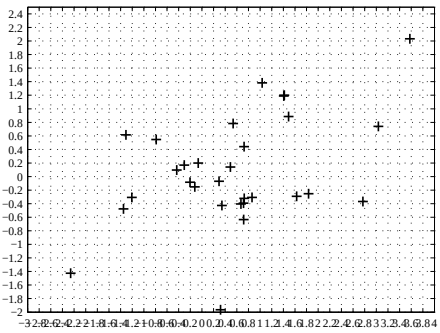
The discipline of proper examination of data, in order to understand behavior between response-, design- and covariate variables, distribution of samples and samples with outlying behavior, is a process which goes ahead of modeling and parameter estimation. This task is often suppressed in reporting of results and might be considered as low-fi science, but is nevertheless crucial, as model choices depend on this. Dealing with few variables, this task can be conducted, using simple graphical techniques as scatter- and line plots with color/marker labels according to categorical variables [48]. In modern science, the number of variables measured can be extremely high, which complicates usage of simple scatter plots for data investigation. E.g. microarrays, reflecting the transcriptional level of genes, often come in vectors of $10000 - 50000$ numbers per sample. Scatter plots of all combinations exceeds $10^7$ and are therefore not a manageable option. Furthermore, subsequent to modeling the number of parameter estimates from such studies can be $> 100$.

This obviously demands techniques that can enhance the interpretability. Visualization of empirical data has been used for at least 200 years, mostly in forms of graphs and spatial representation of information [99, 10]. Information visualization techniques within computer science is a growing field, where the subfield concerned with empirical data has publications tracking back to the early 1960's. Especially the work by J. Tukey, *The Future of Data analysis* (1962) and *Exploratory data analysis* (1977), has had major impact [95, 96]. In the exploratory data analysis, visualization is one, but likely the most powerful, technique, and whole journals are devoted to the subject of visualization (e.g. *Information visualization* - from 2002). Data graphics as a scientific discipline is thoughtfully reviewed and further developed by E.

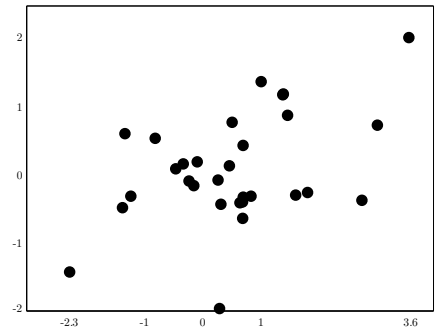
R. Tufte, a statistician devoted to proper presentation of data. One contribution by Tufte is a list of principles for graphical excellence (see box below) [94]. Among a variety of concepts, Tufte talks about *Data Ink Ratio* which derive from splitting graphics into *data ink* - the graphical part with information and *non data ink* - a disturbing part which should be minimized. In Figure 4.1 an example of two presentations of the same data is given. One with high data ink ratio and one with low data ink ratio [94].

In Figure 4.1 A) non data ink is used for a large number of axis labels, grid borders etc. In B) the data markers have appropriate size, tick marks are used for highlighting of the data range and there is no disturbing grid.

Data mining techniques, such as PCA and PLS in combination with graphics, offer a window into data as a starting point in the data analytical process, guiding sample characterization and model selection. In chemometrics, visualization of estimated components by simple scatter plots, e.g. score plot for sample distribution and loading plot for covariance structure between variables, is the most fundamental visualization technique. From a chemical, biological or pharmaceutical point of view these graphics reveal intuitive valid patterns and highlights novel connections. From a strictly mathematical point of view, 2d score- and loading plots are information truncation only highlighting the variation accounted for by the components, which is seldom 100%, while the undescribed part often is a combination of systematic variation (true rank $> 2$) and a random noise part. Hence, conclusion based on such graphics can correctly be argued as incomplete, with a demand of total systematic variation exploration, e.g. by exploration of higher components, but incorrectly be accused for data stretching and overfitting as component models should be properly validated like every other data derived statistics such as mean differences, odds ratios etc.



(a) Bad




(b) Good

Figure 4.1: Two representation of the same data. a) Low *data ink ratio*. b) High *data ink ratio*

Figure 4.2 shows three exactly similar loading plots based on the work in Rasmussen *et al* (2010) (Paper I). Only 30% of the total variation is described. So, are these results of any significance? In Figure 4.2 A, the different variables are denoted by numbers, which, from a purely mathematical point of view, is sufficient. It can be concluded that e.g. variable 1 and 41 have a correlation coefficient at $r^2 \sim 0.3$

($r^2 = \cos(\alpha) \cdot EV$, where $\alpha$ is the angle between the two vectors, and $EV$ is the variance explained by the two components). This is not much, and arguably close to random. The data originate from a clinical trial with variables being efficacy and safety biomarkers, so in order to biologically utilize the results, some labeling can be considered (see Figure 4.2 B). Prior mechanistic understanding of the system suggests that *sCD25*, a surrogate marker for the drug activity and *C reactive protein*, an acute inflammatory precursor, should exhibit similar patterns. This is in accordance with the results, as these are closely positioned in the loading plot. Sub grouping of the biomarkers based on medical knowledge further reveals an ordered clustering pattern, where immune activation biomarkers cluster opposite to biomarkers for white blood cells and electrolytes etc. (see Figure 4.2 C). This simple incorporation of prior knowledge on the graphical presentation strengthens the interpretability and helps the biological validation of the results.

In connection to Tufte's principle of graphical excellence (see box below) and as an extension, this loading plot (and score plot); handles large data sets, presents many numbers in small space, encourages the eye to compare different pieces of data etc. Tufte discusses the issue of *graphical redundancy* as e.g. presenting a bar plot with legends corresponding to the height of the bars. In PCA *mathematical redundancy* between variables is handled by low rank representation of data. Arguably this lowers the number of plots needed, and hence improves the clarity.

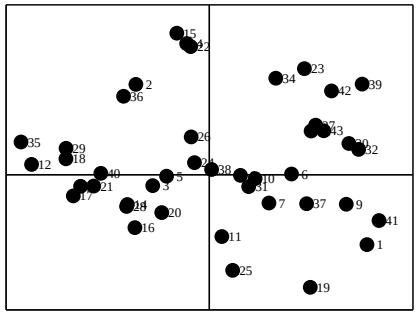
**Tufte's principle of graphical excellence**

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

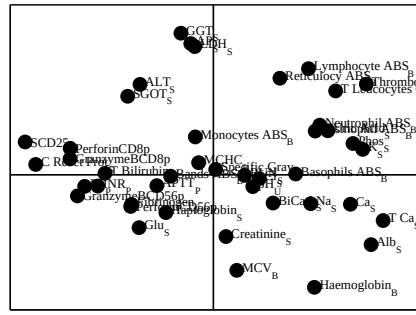
[94]

Case II concerning disease understanding and progression are treated in Paper II, Paper III, Paper VI and Paper X [78, 82, 15]. In these papers loading plots from PLS and PCA models are used for visualization of the covariance structure between variables of different na-

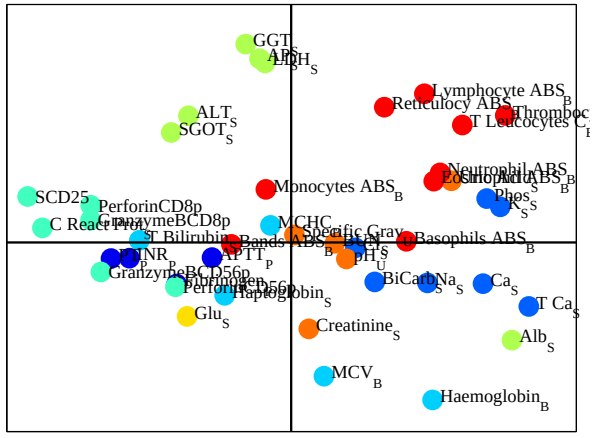
ture (biomarkers, genetic polymorphism, physical activity, food intake etc.). Indeed this leads to intuitive biological patterns that easily communicate a rather complex story.



(a) Loading plot A



(b) Loading plot B



(c) Loading plot C

Figure 4.2: Three similar loading plot with different labeling (Loading plot A: Numbers, Loading plot B and C: Variable names), and coloring (Loading plot A and B: No coloring, Loading plot C: Colored according to variable subgroups).

## 4.3 Bias

Bias is the systematic deviation from the true distribution parameter. Several sources of bias exist originating from different stages of the scientific procedure; formulation of questions, experimental design, collection of data, analysis of data, selection of selling points, submission and publication.

An objective and rigorous approach throughout the scientific procedure is the upfront stated methodology. Nevertheless, when financial or personal career issues are involved, some stretching of data is observed. It is a documented fact, that numerous high impact discoveries are proven misleading in terms of effect size [51]. *Winner's curse*, where the *low lying fruits* are selected, is a well known source of bias in case of large number of simultaneous hypothesis tests (see section 3.1), and is therefore bias related to the formulation and analysis of multiple questions. Especially in genome wide studies this is an obstacle in estimation of the effect size [65], but this problem is also significant within other fields as e.g. metabolomics [12]. The problem is, that the significant variables are selected based on a large *observed* effect size, which is partly due to a true association and partly by chance. In the below example *Winners curse*, the relation between the number of tests conducted (number of variables) and the difference between the observed and the true effect size for *the most* significant hypothesis, is shown for a simulation study.

**Example: Winners curse**

This example is included to show the bias that occurs when multiple hypothesis is tested simultaneously.
The setup is $p$ *null* hypothesis tested with unpaired t-tests from independent data $\mathbf{X_0}(n \times p)$ and $\mathbf{X_1}(n \times p)$.

$$\mathbf{X_0} \sim \mathcal{N}_p(\mu_\mathbf{0}, \mathbf{I})$$
$$\mathbf{X_1} \sim \mathcal{N}_p(\mu_\mathbf{1}, \mathbf{I})$$

where $\mu_\mathbf{0} = \overrightarrow{\mathbf{0}}$ and $\mu_\mathbf{1} = [\underbrace{0, \frac{1}{p-1}, \ldots, \frac{p-1}{p-1}}_{p}]$

with the t-test statistics:

$$t_i = \frac{\bar{x}_{1i} - \bar{x}_{0i}}{s_{pooled_i}\sqrt{\frac{2}{n}}}$$

where

$$s^2_{pooled_i} = \frac{2n}{n-2}\left(\sum_1^n (x_{0ij} - \bar{x}_{0i})^2 + \sum_1^n (x_{1ij} - \bar{x}_{1i})^2\right)$$

for $i = 1, \ldots, p$

The largest absolute value of $t_i$ is considered the most interesting. In Figure 4.3, the true effect size $(\mu_\mathbf{1} - \mu_\mathbf{0})$ is compared with the observed effect size $(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{0})$ for this element.

In this simulation $n = 30$ and $p = (2, 10, 40, 100, 400, 10^3, 4 \cdot 10^3, 10^5)$. 1000 simulations were conducted for each combination.

The presence of selection bias due to multiple testing is obvious and transparent as highlighted by this small example (see Figure 4.3), and different mathematical methods for unbiased effect size estimation in such situations have been developed [102, 90]. Bias due to the inflation of human obsession of success is more of a delicate matter. Scientific work is, like everything else, based on optimization of

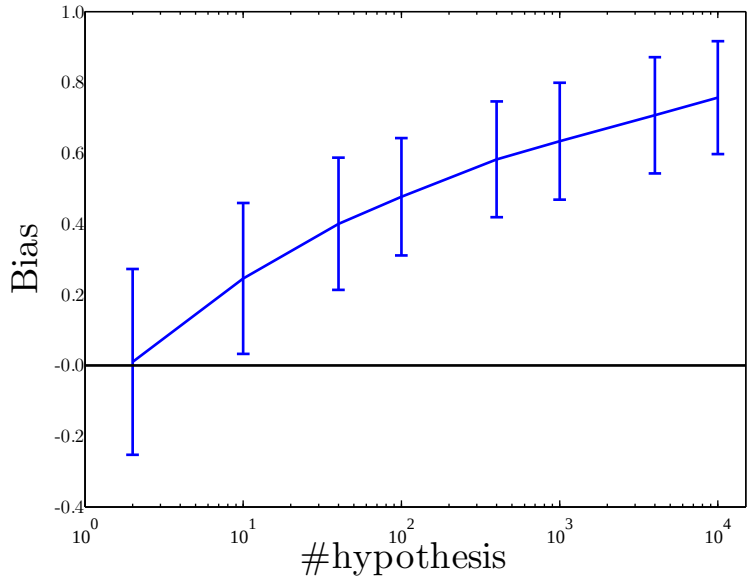



Figure 4.3: Bias (mean±std) for most significant variable as function of number of tests (number of simulations = 1000)

some cost function. This cost function could be, for example: Number of publications, number of publications in high impact journals, personal h-index, high academic positions, working with nice people, good student evaluations and so forth. In breeding of e.g. animals, it is known that optimizing certain criteria will consequently decrease performance for other criteria. Realizing this fact, in the frame of scientific work, seeds for bias.

### Publication bias

Publication bias is heavily reviewed by J.P.A. Ioannidis [53, 52, 57]. Ioannidis (2005) [51] shows that 31% (14/45) of highly cited results (> 1000) presented in the major clinical journals (*New England Journal of*

*Medicine*, *JAMA* and *Lancet*) are concluded to reflect too optimistic effect size estimates. Dickersin *et al* (1987) and Dwan *et al* (2008) [20, 23] have investigated publication bias within clinical trials. Their findings show higher degree of negative studies among *unpublished* results, compared to published. The cause is primary identified as an author responsibility, as studies investigating review bias propose no difference between journals willingness to publish negative results compared to positive results [72].

### Question bias

Question bias relates to the initial stage of the scientific process, where relevant questions are raised followed by design of an experiment that is powered to answer these questions. From a commercial or personal point of view, questions can be split into three types;

1. Need to know.
2. Nice to know.
3. Don't need to know.

The first category includes safety evaluation in a clinical trial, where registration of phenotypic characteristics as nausea, vomiting, head ache etc. and dichotomous registration of safety biomarkers according to reference interval reflecting normal-/non range in a purely descriptive fashion are used for addressing questions related to safety issues and site effects. An example of such is given in [18].

The second category could in a clinical trial setting comprise efficacy and related beneficial effects of treatment, and is addressed with relevant statistics based on pre clinical discoveries.

In the last category, "Don't need to know", it is my personal experience, from work within clinical drug development, that e.g. systematic

response patterns of safety biomarkers are positioned. It is less commercially severe to report a drug having sporadic safety signals, than a systematic pattern, even though biochemical knowledge can support such findings. This is not deliberately withholding of information, but more likely conscious lack of eagerness to pursue such questions in connection with excuses as *statistically underpowered data*. From a pure data analytical point of view there is no difference between surrogate markers for safety and efficacy. Efficacy on the other hand is in the category of "nice to know", and great effort is put into pursuing biological mechanisms. For an example see e.g. Elishmereni *et al* (2011) [30]. In any case, both areas can rightly be accused of question and publication bias.

*Chapter 5*

# Discussion

The presented methods in this thesis are only a subset of what is available for proper analysis of data. What is presented, is what relates to (some of) the work I have been doing for this thesis. The discussion will focus on issues related to choice of methods and the probability of success given the data and the questions.

## 5.1 Selecting the model search path

The data analytical approach in divergent scientific areas is different, and depending on who a given data set with given questions is presented for, different techniques are applied. This data analytical discrepancy is related to educational background, geographic regions and tools made available within software. For example statistical data analysis within clinical drug development is often conducted in SAS®. For advanced analysis, combining several standard tools, this platform requires highly specialized users. Furthermore, the graphical presentation of results in SAS® fails in terms of Tufte's principles of graphical excellence (see section 4.2). There is no golden standard of which

tools or methods to use, unless defined by an oracle such as regulatory authorities, and different approaches might reveal different interesting trends or obstacles to be handled in data, due to the different nature of the tools. Bayesian statistics, an area of statistics roughly speaking opposite to frequentist statistics, uses prior knowledge concerning parameters of interest, in data based estimation. The philosophy of Bayesian statistics is, that any additional valid knowledge concerning the system will improve the estimation. The different methods presented here are just a minor subset of methods capable of handling the same kind of data structures, and the choice of method will eventually inflate the results. Based on the questions asked and the prior knowledge concerning the system, selection of an appropriate algorithm will enhance the probability of success, just as prior distributional knowledge will improve results [85, 54, 27, 25]. In order to exemplify the difference between solutions from different algorithms, linear regression techniques are explored.

**Example: Model search path**

In order to compare *how* the choice of model/algortihm inflate the subsequent results, six different model search path algorithms (PLS, Ridge Regression, Lasso, Adaptive Lasso, Elastic net and Forward Stepwise Regression) are compared on the same data. The aim is to explore method similarities in terms of regression vector estimates.
The setup is:

$$\mathbf{X} = \mathbf{X}^* + \mathbf{E}$$
$$\mathbf{X}^* \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$$

$$\mathbf{E} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}\epsilon_X^2)$$
$$\mathbf{y} = \mathbf{X}^*\beta + \mathbf{e}_y$$
$$\mathbf{e}_y \sim \mathcal{N}(0, \epsilon_y^2)$$

where the covariance matrix of $\mathbf{X}$, $\mathbf{\Sigma}$ has 1's in the main diagonal and 0.3 everywhere else. The regression coefficients $\beta$ are drawn from a gaussian with zero mean and variance one. The noise parameters are: $\epsilon_X = 1$ and $\epsilon_y = 5$ which correspond to a signal to noise ratio of $SN_y = 0.7$ and $SN_X = 1.1$. The number of samples and parameters are $n = 30$ and $p = 40$. Data is centered.

The aim is to estimate $\mathbf{b}$ from:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

The methods are described in section 2.2 (Ridge Regression, Lasso, Adaptive Lasso and Forward Stepwise Regression). The Elastic net penalty is split 1:9 between $L_1$ and $L_2$. For details on Elastic net see Zou and Hastie (2005) [104]. For details on PLS see de Jong (1993) [19].

Each method produces a model search path (a sequence of regression vectors ($\hat{\mathbf{b}}$)) ranging from a highly penalized models (one component PLS, high penalty on the norm of the regression vectors for Ridge Regression, Lasso, Adaptive Lasso and Elastic net, and one variable selected for Forward Stepwise Regression) to *close to* unconstrained models (full rank PLS, non active bounds on the regression vector norm and OLS solution for Forward Stepwise Regression).

The model search paths are arranged in a matrix as stacked regression vectors, and analyzed by PCA.

In Figure 5.1, the score plots for the first three components are shown. The first component in general describes *how much* the **X** data is used, i.e. a measure of degrees of freedom. This variation between different solutions is therefor trivial. Forward Stepwise Regression and to some extend the Adaptive Lasso obtain results quite different from the other models. In PC3 difference between the Lasso and the group consisting of PLS, Ridge Regression and Elastic net, is registered. PLS and Ridge Regression obtain comparable solutions in accordance with [39].

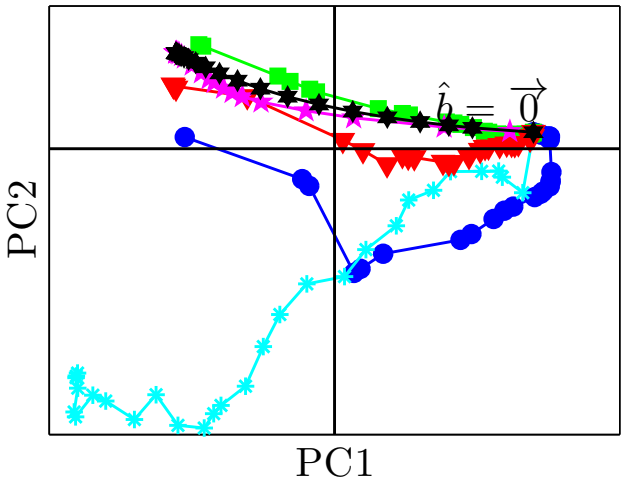
The example highlights the differences between methods for estimation of a regression vector. A comparison with methods beyond linear regression, such as regression trees, neural networks, support vector machines etc. might reveal even higher discrepancy between results. Hence, some thoughts concerning the choice of algorithm in relation to variable selection, shrinkage of coefficients, underlying latent structure, normality, non-linearity etc. based on system knowledge, and to some extend a justification of these choices seem appropriate in order to fully utilize the data. Stepping even further back in the scientific process, system knowledge can help selecting the *operational level*.

## 5.2 Selecting the operational level

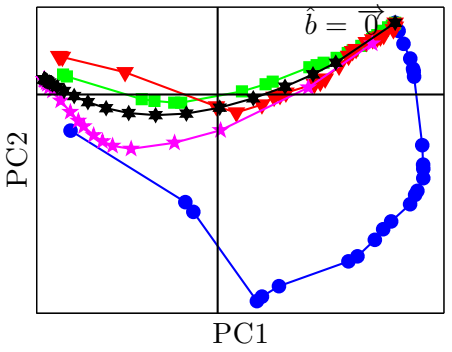
In 2001, the human genome was fully sequenced [97], which prospectively raised the possibility of uncovering relation between genetics and health. This led to high expectations in terms of development of better drugs for the benefit of the entire humanity. Evaluation of the achievements so far leads to the conclusion that there is still a large degree of unexplained variation, and a shift in the drug development paradigm is delayed [38]. The work of Kjeldahl *et al* (2011) (Paper VII) represents an example of how difficult it is to compare genetics with environmental factors from an uncontrolled observational cohort. The aim of the work was to investigate the relation between plasma profiles and a selected single nucleotides polymorphism (SNP), known to be related to obesity. The response variable is a SNP (rs9939609) that is related to genetics and is therefor constant throughout life. On the other hand the markers in plasma indicate a snap shot of the dynamic metabolism, which reflects a number of variation sources such as gender and age, but also environmental variables related to eating habits, physical activity etc. Hence, it is evident that there is a long *distance*, in terms of a chain of cause and effects, between the response variable and the predictors. Elaborating on the genetic part, it is the level of gene expression, in terms of translation to RNA and further to specific proteins/enzymes, that is likely to cause a change in plasma profiles. Hence, the higher the degree of people with non active genes, the more the effect size difference (between different genotypes) is shrunken. Likewise quantifying plasma profiles by $^{1}$H-NMR have limitations in terms of sensitivity of low concentration compounds and signal suppression for certain molecules. Furthermore, the profiles are inflated by a number of quantifiable (e.g. gender, age, etc.) and inquantifiable (e.g. diet prior to sampling) factors. These factors in combination reduce the power of such data. One way to surpass the issue of power is to include a higher number of samples. This, however,

only guarantees validity under the assumption of independence, which is highly unlikely as environmental- and genetic factors are connected.

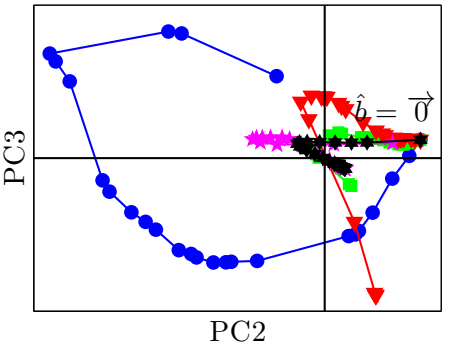
In order to increase the probability of success there is a need for incorporation of detailed system knowledge in planning, collection and evaluation of the data.




(a) Score plot (all six algorithms)




(b) Score plot (five algorithms)




(c) Score plot (five algorithms)

Figure 5.1: Score plot of model search paths for six models: PLS, Ridge Regression, Lasso, Adaptive Lasso, Elastic net and Forward Stepwise Regression. (a) PC1/PC2 for all six search paths, (b) PC1/PC2 for five search paths (Forward Stepwise Regression removed), (c) as (b) for PC2/PC3.

*Chapter 6*

# Perspectives in Drug development

In 2004, the US Food and Drug Administration (FDA) announced a paper "Innovation or Stagnation: Challenge and Opportunity on the Critical path to New Medical Products" [38] quoting, that despite the revolution in biomedical science with increase in high throughput molecular techniques, such as microarrays, gene sequencing, etc., yielding ever higher number of basic findings, these are not translated into new drugs. In FDA's opinion the problems is addressed as:

> *What is the problem? In FDA's view, the applied sciences needed for medical product development have not kept pace with the tremendous advances in the basic sciences. The new science is not being used to guide the technology development process in the same way that it is accelerating the technology discovery process. For medical technology, performance is measured in terms of product safety and effectiveness. Not enough applied scientific work has*

> *been done to create new tools to get fundamentally better answers about how the safety and effectiveness of new products can be demonstrated, in faster time frames, with more certainty, and at lower costs. In many cases, developers have no choice but to use the tools and concepts of the last century to assess this century's candidates. As a result, the vast majority of investigational products that enter clinical trials fail. Often, product development programs must be abandoned after extensive investment of time and resources. This high failure rate drives up costs, and developers are forced to use the profits from a decreasing number of successful products to subsidize a growing number of expensive failures. Finally, the path to market even for successful candidates is long, costly, and inefficient, due in large part to the current reliance on cumbersome assessment methods.* [38]

The new tools demanded here are e.g. biomarkers for early detection of safety and efficacy in combination with mechanistic and data driven models for effectively integration of current knowledge [1, 66]. In recent years more advanced mathematical methods have been applied to explore and confirm drug action [41, 16]. This development is highly appreciated with a demand for further development [16]. Rasmussen *et al* (2010) (Paper I) is an example of an explorative data analytical approach that reveals systematic safety issues related to liver toxicity early in phase I [77].

From a scientific point of view these methodological tendencies are a natural consequence of novel data types produced and the multiplex of questions addressed. Commercially the bundle of techniques can be rather scary, as it is now possible to uncover delicate questions related to side effects. The (un-) willingness to pursue certain type of questions consequently results in a bias (see section 4.3).

## 6.1 Who have the responsibility? - The *Ph* part of *PhD*

Classically drug development aims at confirming direct medical utility and estimating degree of registered side effects. This results in a data structure optimally suited for analysis by classical frequentist methods. With development of biomedical platforms, surrogate biomarkers for both efficacy and safety are nowadays used as early signs of safety and efficacy. In combination with screening for e.g. relevant genetic polymorphisms, these data potentially characterize the heterogeneity of the patient population, both initially (at baseline) and in relation to treatment. Obeying the complex nature of biology by applying complex data analytical tools, potentially reveal the systematic patterns of this system. This is nevertheless not trivial, as choices concerning modeling and presentation of results are complex, compared to results from e.g. an unpaired t-test. In order to fully achieve the potential, more acceptance and knowledge of the complex bioanalytical platforms and the advanced mathematical methods for information extraction, have to be present throughout the entire drug development chain. The pharmaceutical companies have a responsibility of pursuing relevant issues throughout the entire drug history, in order to enhance understanding of drug action for better labeling. In connection the regulatory authorities have a responsibility of accepting, that higher degree of knowledge results in early potential warnings, which might be a major obstacle, but maybe is an optimally trade off between benefits and risks. Academia has a responsibility of communicating the biological complexity and multi factorial relations between causes and effects in order to clarify, that life is a gray balance and a highly constrained optimization problem, and not dichotomous black and white.

# Bibliography


[1] G. An, J. Bartels, and Y. Vodovotz. In silico augmentation of the drug development pipeline: examples from the study of acute inflammation. *Drug Development Research*, 72:1–14, 2010.

[2] CM Andersen and R. Bro. Variable selection in regression - a tutorial. *Journal of Chemometrics*, 24:728–737, 2010.

[3] C.A. Andersson. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, 47(1):51–63, 1999.

[4] C. J. Appellof and E. R. Davidson. Strategies for Analyzing Data from Video Fluorometric Monitoring of Liquid Chromatographic Effluents. *Analytical Chemistry*, 53:2053–2056, 1981.

[5] E. Beltrami. Sulle funzioni bilineari [on bilinear functions]. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, 11:98–106, 1873.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing.

*Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[8] J.M. Bland and D.G. Altman. Regression towards the mean. *BMJ: British Medical Journal*, 308(6942):1499, 1994.

[9] A. Boulesteix and K. Strimmer. Short title: Partial Least Squares for Genomics Analyses. 2006.

[10] W.C. Brinton. *Graphic methods for presenting facts*. The Engineering magazine company, 1914.

[11] R. Bro. PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[12] D.I. Broadhurst and D.B. Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171–196, 2006.

[13] C.D. Brown and R.L. Green. Critical factors limiting the interpretation of regression vectors in multivariate calibration. *TrAC Trends in Analytical Chemistry*, 28(4):506–514, 2009.

[14] J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[15] C.G. Carson, M.A. Rasmussen, J.P. Thyssen, T. Menné, and H. Bisgaard. Phenotyping atopic dermatitis in children by filaggrin status. *Submitted for British Journal of Dermatology*, 2011.

[16] A.F. Cohen, Y.K. Loke, A. Ferro, L.D. Lewis, A. Somogy, and J.M. Ritter. Editors' report, November 2010. *British Journal of Clinical Pharmacology*, 71(1):1–2, 2011.

[17] A.D. Dangour, K. Lock, A. Hayter, A. Aikenhead, E. Allen, and R. Uauy. Nutrition-related health effects of organic foods: a systematic review. *The American journal of clinical nutrition*, 92(1):203, 2010.

[18] I.D. Davis, B.K. Skrumsager, J. Cebon, T. Nicholaou, J.W. Barlow, N.P.H. Moller, K. Skak, D. Lundsgaard, K.S. Frederiksen, P. Thygesen, et al. An open-label, two-arm, phase I trial of recombinant human interleukin-21 in patients with metastatic melanoma. *Clinical cancer research*, 13(12):3630, 2007.

[19] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[20] K. Dickersin, SS Chan, TC Chalmersx, HS Sacks, and H. Smith Jr. Publication bias and clinical trials. *Controlled Clinical Trials*, 8(4):343–353, 1987.

[21] J. Dien, D.J. Beal, and P. Berg. Optimizing principal components analysis of event-related potentials: matrix type, factor loading weighting, extraction, and rotations. *Clinical neurophysiology*, 116(8):1808–1825, 2005.

[22] O.J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, pages 52–64, 1961.

[23] K. Dwan, D.G. Altman, J.A. Arnaiz, J. Bloom, A.W. Chan, E. Cronin, E. Decullier, P.J. Easterbrook, E. Von Elm, C. Gamble, et al. Systematic review of the empirical evidence of

study publication bias and outcome reporting bias. *PLoS One*, 3(8):3081, 2008.

[24] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26, 1979.

[25] B. Efron. Why isn't everyone a bayesian? *American Statistician*, pages 1–5, 1986.

[26] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[27] B. Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5, 2005.

[28] B. Efron. *Large-Scale Simultaneous Inference*. 2010.

[29] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[30] M. Elishmereni, Y. Kheifetz, H. Søndergaard, R.V. Overgaard, and Z. Agur. An integrated disease/pharmacokinetic/pharmacodynamic model suggests improved interleukin-21 regimens validated prospectively for mouse solid cancers. *PLoS Computational Biology*, 7(9):e1002206, 2011.

[31] R. Ergon. PLS post-processing by similarity transformation (PLS+ ST): a simple alternative to OPLS. *Journal of chemometrics*, 19(1):1–4, 2005.

[32] R. Ergon. Finding Y-relevant part of X by use of PCR and PLSR model reduction methods. *Journal of Chemometrics*, 21(12):537–546, 2007.

[33] L. Eriksson, P.L. Andersson, E. Johansson, and M. Tysklind. Megavariate analysis of environmental QSAR data. Part I–A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular diversity*, 10(2):169–186, 2006.

[34] L. Euler. *A complete theory of the construction and properties of vessels: with practical conclusions for the management of ships, made easy to navigators. Translated from Théorie complette de la construction et de la manœuvre des vaissaux, of the celebrated Leonard Euler, by Henry Watson, Esq.* printed for J. Sewell, 1790.

[35] R.A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[36] R.A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935.

[37] N.V. Folsgaard, B.L. Chawes, M.A. Rasmussen, A.L. Bischoff, C.G. Carson, J. Stokholm, L. Pedersen, T.T. Hansel, K. Bonnelykke, S. Brix, and Bisgaard H. Neonatal Cytokine Profile in the Airway Mucosal Lining Fluid is skewed by Maternal Atopy. *American journal of respiratory and critical care medicine*, 185:275–280, 2012.

[38] US Food. Drug Administration: Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. *US Dept of Health and Human Services*, 2004.

[39] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[40] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[41] J.V.S. Gobburu. Pharmacometrics 2020. *The Journal of Clinical Pharmacology*, 50(9 suppl):151S, 2010.

[42] S.N. Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130(12):995, 1999.

[43] S. Greenland. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, 79(3):340, 1989.

[44] S. Greenland and J.M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413, 1986.

[45] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[46] C.L. Hansen, F. den Berg, M.A. Rasmussen, S.B. Engelsen, and S. Holroyd. Detecting variation in ultrafiltrated milk permeates-Infrared spectroscopy signatures and external factor orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, 2010.

[47] R.A. Harshman. *Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis*. Citeseer, 1970.

[48] J.A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213, 1975.

[49] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[50] R. Hubbard and R.M. Lindsay. Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1):69, 2008.

[51] J. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA: the journal of the American Medical Association*, 294(2):218, 2005.

[52] J. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):696, 2005.

[53] J. Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640, 2008.

[54] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: held at the Statistical Laboratory, University of California, June 20-July 30, 1960*, page 361. Univ of California Press, 1961.

[55] J.J. Jansen, R. Bro, H.C.J. Hoefsloot, F.W.J. van den Berg, J.A. Westerhuis, and A.K. Smilde. PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. *Journal of Chemometrics*, 22(2):114–121, 2008.

[56] B.R. Kowalski and C.F. Bender. Pattern recognition. II. Linear and nonlinear methods for displaying chemical data. *Journal of the American Chemical Society*, 95(3):686–693, 1973.

[57] P.A. Kyzas, D. Denaxa-Kyza, and J. Ioannidis. Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer*, 43(17):2559–2579, 2007.

[58] K. Laursen, U. Justesen, and M.A. Rasmussen. Enhanced monitoring of biopharmaceutical product purity using liquid chromatography-mass spectrometry. *Journal of Chromatography A*, 2011.

[59] K. Laursen, M.A. Rasmussen, and R. Bro. Comprehensive control charting applied to chromatography. *Chemometrics and Intelligent Laboratory Systems*, 2011.

[60] R. Leardi and A. Lupianez Gonzalez. Genetic algorithms applied to feature selection in pls regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, 41(2):195–207, 1998.

[61] E.L. Lehmann. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249, 1993.

[62] Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5):2272–2297, 2006.

[63] E.H. Madden. Aristotle's treatment of probability and signs. *Philosophy of Science*, 24(2):167–172, 1957.

[64] G. Maria et al. Metabolomic Profiling for Identification of Novel Potential Biomarkers in Cardiovascular Diseases. *Journal of Biomedicine and Biotechnology*, 2011, 2011.

[65] M.I. McCarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein, J. Little, J.P.A. Ioannidis, and J.N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[66] Q. Mi, N.Y.K. Li, C. Ziraldo, A. Ghuma, M. Mikheev, R. Squires, D.O. Okonkwo, K. Verdolini-Abbott, G. Constantine, G. An, et al. Translational systems biology of inflammation: potential applications to personalized medicine. *Personalized Medicine*, 7(5):549–559, 2010.

[67] S.L. Morgan and C. Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge Univ Pr, 2007.

[68] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333, 1937.

[69] J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289, 1933.

[70] J. Neyman and E.S. Pearson. The testing of statistical hypotheses in relation to probabilities a priori. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 29, pages 492–510. Cambridge Univ Press, 1933.

[71] J. Olsen, M. Melbye, S.F. Olsen, T.I.A. Sørensen, P. Aaby, A.M. Nybo Andersen, D. Taxbøl, K.D. Hansen, M. Juhl, T.B. Schow, et al. The Danish National Birth Cohort-its background, structure and aim. *Scandinavian journal of public health*, 29(4):300, 2001.

[72] C.M. Olson, D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J.W. Hogan, Q. Zhu, J. Reiling, and B. Pace. Publication bias in editorial decision making. *Jama*, 287(21):2825, 2002.

[73] D.B. Panagiotakos. The value of p-value in biomedical research. *The Open Cardiovascular Medicine Journal*, 2:97, 2008.

[74] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[75] S.B. Petersen, M.A. Rasmussen, H. Torjusen, M. Strøm, T.I. Halldorsson, and S.F. Olsen. Socio-demographic characteristics and food habits of organic consumers: A study from the Danish National Birth Cohort. *Submitted for Public Health Nutrition*, 2011.

[76] J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

[77] M.A. Rasmussen, M. Colding-Jørgensen, L.T. Hansen, and R. Bro. Multivariate evaluation of pharmacological responses in early clinical trials–a study of rIL-21 in the treatment of patients with metastatic melanoma. *British journal of clinical pharmacology*, 69(4):379–390, 2010.

[78] M.A. Rasmussen, J. Skov, E.M. Bladbjerg, J.J. Sidelmann, M. Vamosi, and J. Jespersen. Multivariate analysis of the relation between diet and warfarin dose. *European Journal of Clinical Pharmacology*, 68:321–328, 2012.

[79] J.M. Roger, F. Chauchard, and V. Bellon-Maurel. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and intelligent laboratory systems*, 66(2):191–204, 2003.

[80] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467, 1995.

[81] G. Shafer. The significance of Jacob Bernoulli's Ars Conjectandi for the philosophy of probability today* 1. *Journal of econometrics*, 75(1):15–32, 1996.

[82] J. Skov, E.M. Bladbjerg, M.A. Rasmussen, J. Sidelmann, A. Leppin, and J. Jespersen. Genetic, Clinical and Behavioural Determinants of Vitamin K-Antagonist Dose-Explored through Multivariable Modelling and Visualization. *Basic & Clinical Pharmacology & Toxicology*, 110:193–198, 2012.

[83] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. Van Der Greef, and M.E. Timmerman. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043, 2005.

[84] L.A. Steen. Highlights in the history of spectral theory. *The American Mathematical Monthly*, 80(4):359–381, 1973.

[85] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206. Berkeley, University of California Press, 1956.

[86] G.W. Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[87] S.M. Stigler. Francis Galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.

[88] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[89] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[90] L. Sun, A. Dimitromanolakis, L.L. Faye, A.D. Paterson, D. Waggott, and S.B. Bull. BR-squared: a practical solution to the winner's curse in genome-wide scans. *Human Genetics*, pages 1–8, 2011.

[91] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[92] R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[93] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[94] E.R. Tufte and G. Howard. *The visual display of quantitative information*, volume 16. Graphics press, 1983.

[95] J.W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[96] J.W. Tukey. Exploratory data analysis. *Reading, MA*, 1977.

[97] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A.

Holt, et al. The sequence of the human genome. *science*, 291(5507):1304, 2001.

[98] H.I. Weisberg. *Bias and Causation: Models and Judgment for Valid Comparisons*. Wiley, 2010.

[99] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[100] H. Wold. Soft modeling by latent variables: The nonlinear iterative partial least squares approach. *Perspectives in probability and statistics, papers in honour of MS Bartlett*, pages 520–540, 1975.

[101] S. Wold, H. Antti, F. Lindgren, and J.
"Ohman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2):175–185, 1998.

[102] S. Zöllner and J.K. Pritchard. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615, 2007.

[103] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[104] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

# Paper I

Multivariate evaluation of pharmacological responses in early clinical trials - a study of rIL-21 in the treatment of patients with metastatic melanoma

Morten A. Rasmussen, Lasse T. Hansen, Morten C. Jørgensen, Rasmus Bro

*British Journal of Clinical Pharmacology,* ***69*** *(2010), 379-390*

BJCP British Journal of Clinical Pharmacology

DOI:10.1111/j.1365-2125.2009.03600.x

# Multivariate evaluation of pharmacological responses in early clinical trials – a study of rIL-21 in the treatment of patients with metastatic melanoma

**Morten Arendt Rasmussen, Morten Colding-Jørgensen,[1] Lasse Tengbjerg Hansen[2] & Rasmus Bro[3]**

*Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Frederiksberg C, [1]Development Projects Management, Novo Nordisk A/S and [2]Novo Nordisk A/S, Bagsvaerd, and [3]Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Frederiksberg C, Denmark*

**Correspondence**
Mr Rasmus Bro, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
Tel: +45 3533 3296
Fax: +45 3533 3245
E-mail: rb@life.ku.dk

**Keywords**
chemometrics, IL-21, malignant melanoma, multivariate, orthogonalization, PCA, principal components

**Received**
16 December 2008

**Accepted**
31 October 2009

## WHAT IS ALREADY KNOWN ABOUT THIS SUBJECT

- Analysis of data from clinical trials is often performed using univariate statistics.
- In early phases of clinical drug development, interpretation of rare clinical events can be difficult by univariate methods.
- Principal component analysis has proven successful within related scientific areas such as, for example, genomics and metabonomics, where compression of data and extraction of maximum information are of utmost importance.

## WHAT THIS STUDY ADDS

- This study reveals that multivariate chemometric methods coupled with visualization gives a comprehensive overview of early clinical trial data to guide dose and regimen selection and provides additional findings overlooked by traditional univariate methods.
- This method revealed novel pharmacological patterns in the treatment of metastatic melanoma with recombinant interleukin-21.

## AIMS

Evaluation of the utility of multivariate data analysis in early clinical drug development.

## METHODS

A multivariate chemometric approach was developed and applied for evaluating clinical laboratory parameters and biomarkers obtained from two clinical trials investigating recombinant human interleukin-21 (rIL-21) in the treatment of patients with malignant melanoma. The Phase I trial was an open-label, first-human dose escalation safety and tolerability trial with two separate dosing regimens; six cycles of thrice weekly (3/w) *vs.* three cycles of daily dosing for 5 days followed by 9 days of rest (5+9) in a total of 29 patients. The Phase II trial investigated efficacy and safety of the '5+9' regimen in 24 patients.

## RESULTS

From the Phase I trial, separate pharmacological patterns were observed for each regimen, clearly reflecting distinct properties of the two regimens. Relations between individual laboratory parameters were visualized and shown to be responsive to rIL-21 dosing. In particular, novel systematic pharmacological effects on liver function parameters as well as a bell-shaped dose–response relationship of the overall pharmacological effects were depicted. In validation of the method, multivariate pharmacological patterns discovered in the Phase I trial could be reproduced by the dataset from the Phase II trial, but not from univariate exploration of the Phase I trial.

## CONCLUSIONS

The new data analytical approach visualized novel correlations between laboratory parameters that points to specific pharmacological properties. This multivariate chemometric data analysis offers a novel robust, comprehensive and intuitive tool to reveal early pharmacological responses and guide selection of dose regimens.

## Introduction

Clinical drug development is a stepwise, time-consuming and complex process during which an increasing amount of data is collected across numerous trials with different end-points and aims. In controlled clinical trials, subjects are carefully evaluated with respect to predefined clinical and laboratory end-points and closely monitored with respect to unexpected effects. In clinical drug development it is important to extract as much information as early as possible to establish the foundation for the selection of regimen, dose, and patient population for subsequent large-scale clinical trials. Traditional analysis of clinical data is often univariate in nature. Multivariate analysis is capable of finding patterns that are only revealed by relations between variables. Chemometric data analysis tools coupled with visualization provide an opportunity for a rapid and comprehensive overview of a given biological experiment such as a clinical trial. For example, principal component analysis (PCA) allows visualization of a multivariate dataset through so-called principal components. The principal components are variables that are weighted averages of the original variables and found in such a way that they optimally (in a least squares sense) represent the major part of the variation in the data in as few components as possible. Each component can be considered as a descriptive fingerprint of the intrinsic underlying latent variations of the data in the sense that it contains information from all variables simultaneously (for review, refer to Wold *et al.* [1] and Christie [2]). Unlike more deductive approaches that typically need verified or hypothesized relevant measurements [3], PCA makes it possible to perform an exploratory analysis including many variables, even those that are not *a priori* known to be relevant. The exploratory analysis can then provide means for assessing to which degree such variables are indeed relevant. Chemometric methods have historically been developed for chemical analysis, but have in recent years proven valuable in other areas, such as genomics and metabonomics [4–7].

In the present study we applied PCA in early clinical pharmacology trials investigating recombinant human interleukin-21 (rIL-21) in the treatment of malignant melanoma. IL-21 is a cytokine with pronounced antineoplastic properties, primarily exerted by stimulation of natural killer (NK) cells and cytotoxic T-cell subsets to kill tumour cells (for review, see Skak *et al.* [8]). Currently, rIL-21 is in the development for the treatment of various neoplastic conditions, including malignant melanoma. Early clinical investigations have revealed that rIL-21 was generally well tolerated with signs of antineoplastic effects in a subset of patients [9, 10].

The most common adverse events encountered in the first-human dose trial were fatigue, fever, nausea, and headache and the maximal tolerable dose was declared to be 30 μg kg$^{-1}$ [9]. Moreover, serum levels of soluble CD25 (sCD25) were shown to reflect rIL-21-mediated systemic immune activation and distinct pharmacodynamic responses of individual dosing regimens [9, 11]. In addition, several other molecular biomarkers of NK and T-cell activation have been investigated, including mRNA expression of the effector molecules granzyme B and perforin in CD56+ NK cells and CD8+ T cells [9, 12].

Here we propose a multivariate approach to analyse clinical laboratory parameters that can provide valuable information on pharmacological responses complementary to univariate methods when applied in early clinical pharmacology trials.

## Methods

### Data material

The present study is based on an early clinical development programme including a Phase I and Phase II trial. The Phase I data are from an open-label, two-armed dose escalation study, investigating the safety and tolerability, biomarkers, pharmacokinetics, and efficacy of increasing doses of rIL-21 administrated as an intravenous (i.v.) bolus injection in two different dose regimens: dosing at three times weekly (3/w) (Monday, Wednesday and Friday) in a period of 6 weeks (a total of 18 doses across six cycles) with four different dose levels (1, 3, 10 and 30 μg kg$^{-1}$) and daily dosing for 5 days followed by 9 days without treatment (5+9) in a period of 6 weeks (a total of 15 doses across three cycles) with six different dose levels (1, 3, 10, 30, 50 and 100 μg kg$^{-1}$). In the Phase I trial, a total of 29 patients with histologically confirmed surgically incurable metastatic stage IV malignant melanoma were enrolled [9]. The Phase II trial was an open-label, single-armed, fixed-dose study investigating the efficacy, safety and biomarkers of 30 μg kg$^{-1}$ rIL-21 administered as i.v. bolus injection in the '5+9' dose regimen for a period 6 weeks. A total of 24 patients were enrolled and 12 patients were continued on extension treatment for assessment of progression-free survival [10]. According to the Phase I and Phase II protocols, a total of 43 variables encompassing clinical laboratory parameters and biomarkers were assessed in both trials. The half-life of rIL-21 is approximately 1–4 h [9]. The plasma levels of rIL-21 were hence not detectable at the time points when the majority of samples for laboratory parameters and biomarkers were collected, and pharmacokinetic data were therefore not included in the present analysis.

All patients were treated at the Austin Hospital, the Peter MacCallum Cancer Centre, the Royal Melbourne Hospital, Cabrini Health (all in Melbourne, Australia), Westmead Hosital (Sydney, Australia) or Sir Charles Gairdner Hospital (Perth, Australia). All patients provided written informed consent before any study-specific procedures. The trial protocols were approved by the Human Research Ethics Committees of the participating hospitals and were imple-

**Table 1**

Overview of removed outliers. If identifiable, the diverging measurements are listed

| Patient ID | Removed data points | Outlier values | Number of recorded values (out of 43) | Clinical observations |
|---|---|---|---|---|
| **105 – '3/w'** | All data points at day 2 | | 2 | No clinical observations observed |
| **112 – '5+9'** | Band Abs at day 8 | Band Abs – 0.8 | 12 | No clinical observations observed |

mented under the Australian Therapeutic Goods Administration Clinical Trials Notification scheme. The clinical trials were sponsored by Novo Nordisk A/S.

## *Data analysis*

The data were analysed using a modified version of PCA. PCA is a model where the multivariate dataset is compressed into a few orthogonal/uncorrelated principal components (PC) holding the systematic variation of the dataset. Each component is simply a new variable computed as a weighted average of all the original variables and can be considered a descriptive fingerprint in the sense that it contains information from all variables simultaneously. The weights are determined so that the first component explains as much as possible of all the variables. Subsequent components are determined similarly, explaining as much as possible of the yet unexplained part of the variation in data. Results from PCA are presented as components, each containing a score and a loading vector. Each loading vector has as many elements as variables and the elements are the weights for calculating the new variables – the scores. Hence, a numerically high weight implies that the specific variable is important for the component. The sample specific scores explain how the observation behaves with respect to the component. A high score value means that the specific observation has high values on the variables with high loading elements [1].

As expected from clinical trial data, the analysed datasets contain substantial variation across patients (data not shown). The influence of this variation is not of primary concern initially as the focus is on the overall treatment effect within each dose regimen. In order to focus on treatment effects in the PCA analysis, filtering is performed by removing the average level from each subject and each variable. This way, all individual patient data will have the same average level for every variable. Mathematically, removing patient-specific effects is done by orthogonalization and can be considered as a way to focus the analysis on the part of the data specific to treatment-related effects. As a consequence of this, no patient covariates was included in the analysis, as these effects would be removed in the orthogonalization step. Note that no use is made of treatment information in the orthogonalization, which is important in order to avoid spurious correlations (for details see Appendix).

The PCA solution is a least squares solution over all the variables and can therefore be overly influenced by individual variables that are given in numerically large numbers. The variables are hence scaled (and centred) to have equal variance prior to PCA. According to the protocol, not all laboratory or biomarker parameters were assessed on all trial visits. Values for such visits, i.e. for samples that were not obtained, are in PCA defined as 'missing values'. These data are hence not deviating from the trial protocol but merely just collected as planned, i.e. at different time points compared with most other variables. Expectation maximization provides means for fitting the model without introducing biased estimates due to such 'missing values' [13]. Using expectation maximization, the amount of missing values in the data is irrelevant, as these do not affect the resulting model. Only the amount of determined information present in the data is critical.

## *Outliers*

Outliers are single measurements or samples identified as statistically irregular in the model, i.e. samples that influence the model in a way that is potentially detrimental to use of the model. The outliers are determined by model residuals (Q-residuals) and distance to model centre (Hotellings $T^2$), and further examined for general pattern deviation and/or single variable measurement deviation [1]. One out of 212 data points was removed as an outlier from the Phase I '3/w' dataset. A single extreme measurement was removed from the Phase I '5+9' dataset. No outliers were removed from the Phase II dataset. The characteristics for the outliers are listed in Table 1. It is of utmost importance to emphasize that outliers are not necessarily wrong and hence should be medically evaluated for clinical relevance and eventually be examined properly by other methods. However, this is not described further as it is beyond the scope of the present paper.

## *Model post-processing*

The result of the first step is a PCA model of filtered data that contains a score and a loading vector for each component. The score vector has as many elements as there are time–patient points. The model can be further elaborated by re-arrangement of the scores into a matrix with as many rows as doses and as many columns as time points as written in step 4 of the Model description (Appendix, Model description). For each score vector a matrix is made

where the average score for a given dose and time is given in the corresponding element. This matrix provides information on the time–dose information of that particular component and is analysed with a subsequent PCA model providing an even further condensed approximation of the variation in the data useful for understanding, e.g. the time-dependent variation. This will be exemplified in the Results section. An algorithm for the complete data analysis is described in the Appendix.

All calculations are conducted using in-house algorithms written in MATLAB® ver. 7.6.0.324. The function can be downloaded from www.models.life.ku.dk.

## Results

### *Pharmacological responses of the '5+9' dose regimen*

By applying PCA to the '5+9' datasets, two significant ($P < 0.001$) components appeared. For the '5+9' regimen, components 1 and 2 describe 20.0% and 9.7%, respectively, of the total variation across all variables (Figure 1a,b). The structure of the pharmacological response described by the scores in the first component clearly reflected an underlying signature of the '5+9' dose regimen (Figure 1a). During the course of 5 days of treatment, a pronounced decrease in score values was observed in each treatment cycle. This response was fully reverted during 9 days of treatment pause. The pattern described by the second component did not display a similar signature of the '5+9' dose regimen. However, the score values increased to some extent during the course of three treatment cycles, indicating an accumulation of the pharmacological effect reflected by the variables expressed in this component (Figure 1b). Tentatively, and based on the shape of the curves, the two components describe the acute and the cumulative pharmacological effects, respectively. In order to illustrate dose-dependency of these distinct pharmacological responses, score values as function of day and dose were analysed with a subsequent one-component PCA

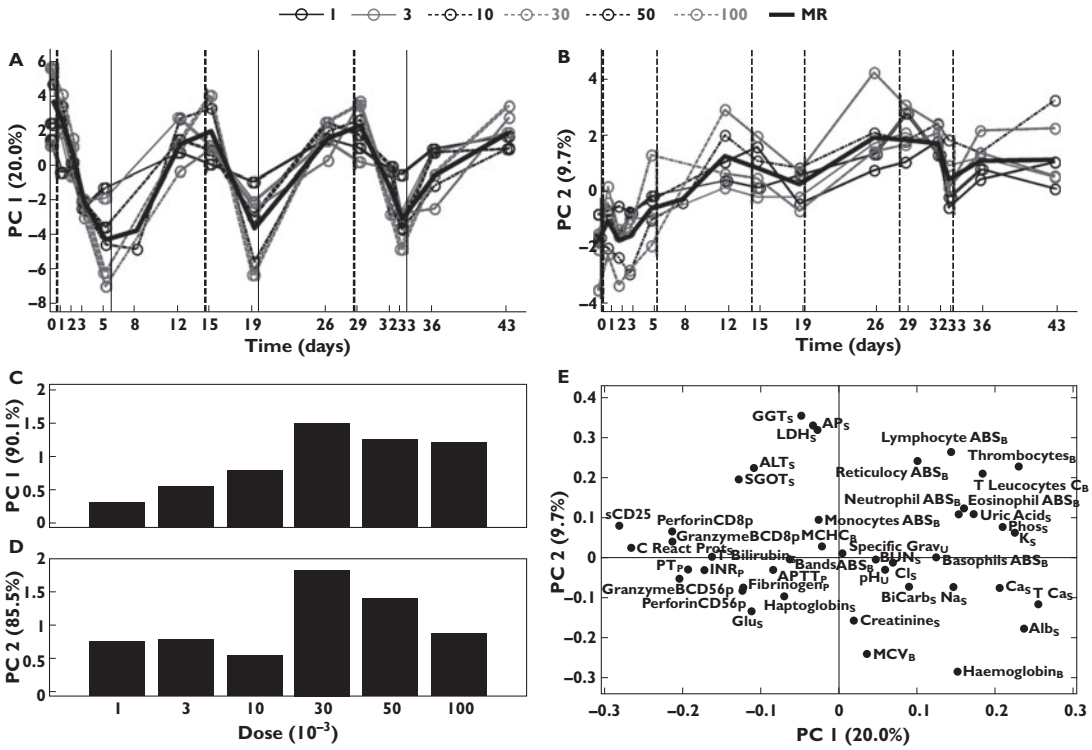



**Figure 1**

Results from two component models of data obtained from regimen '5+9'. (a) Score values for component 1 plotted *vs.* day for each of the six doses (μg kg$^{-1}$) and the mean response (MR) of those. (b) Score values for component 2 plotted *vs.* day. (c) Scores from one-component principal component analysis (PCA) model of scores from component 1. (d) Scores from one-component PCA model of scores from component 2. (e) Loading plot of component 1 *vs.* component 2. Subscripts: B, blood; P, plasma; S, serum; U, urine

as described in the Appendix (Figure 1c,d). For both components of this subsequent PCA model, a bell-shaped dependency of dose peaking at 30 μg kg$^{-1}$ was observed, indicating a maximal pharmacological response at this dose level. The observation of a bell-shaped pharmacology is novel and was not found in the original univariate analyses of these data [9].

Potential relationships between the pharmacological responses of the individual laboratory parameters and biomarkers were assessed by a scatter plot of the first and second loading vector of the initial PCA model across all dose levels (Figure 1e). Based on the positions, the individual parameters could be divided into several categories. Parameters far from the origin are those most influential in the model and parameters close to each other are correlated with respect to the variation reflected by the components. One group of laboratory parameters (upper right quadrant) was composed of haematology parameters, e.g. peripheral blood lymphocyte counts (Lymphocyte ABS$_B$). These parameters showed high scores for both components and hence behaved as a combination of the two, i.e. decreased during the 5 days of treatment, increased during the 9 days of rest, and slightly accumulated during the 6 weeks of treatment. Opposite to this group (lower left quadrant) was a group composed of biomarkers of NK cell activation, e.g. perforin mRNA expression of purified peripheral blood CD56$^+$ NK cells (PerforinCD56p). These parameters were negatively correlated to the group in the upper right quadrant and hence increased during treatment, decreased during rest, and slightly decreased during the 6 weeks of treatment, reflecting rIL-21-mediated effects on NK cell function. In general, variables with high positive or negative loading values for the first component, e.g. peripheral blood lymphocyte counts (Lymphocyte ABS$_B$) and sCD25, respectively, reflected the signature of the '5+9' dose regimen and were thus qualified as biomarkers for the overall pharmacological effects. Pharmacological effects on these NK and T-cell activation biomarkers is an expected finding and was also described in the original reports of these data [9, 12]. However, the inverse relation between peripheral blood lymphocyte and T leucocyte counts *vs.* NK cell and T-cell activation markers that is clearly reflected in Figure 1e has not previously been described and supports rIL-21-mediated immune activation as the primary cause of changes in these blood cell counts.

Variables with high (positive or negative) loading values for the second component indicated an accumulation/decreasing pattern registered during the 6 weeks of treatment. Parameters with numerically low loading values for both components, e.g. mean corpuscular haemoglobin concentration (MCHC$_B$) did not match the overall pattern of the dataset and were hence inadequately described by this model. However, this does not imply absence of clinical relevance of such lab parameters, but merely indicated that the overall patterns observed in the two components were not reflected in these parameters.

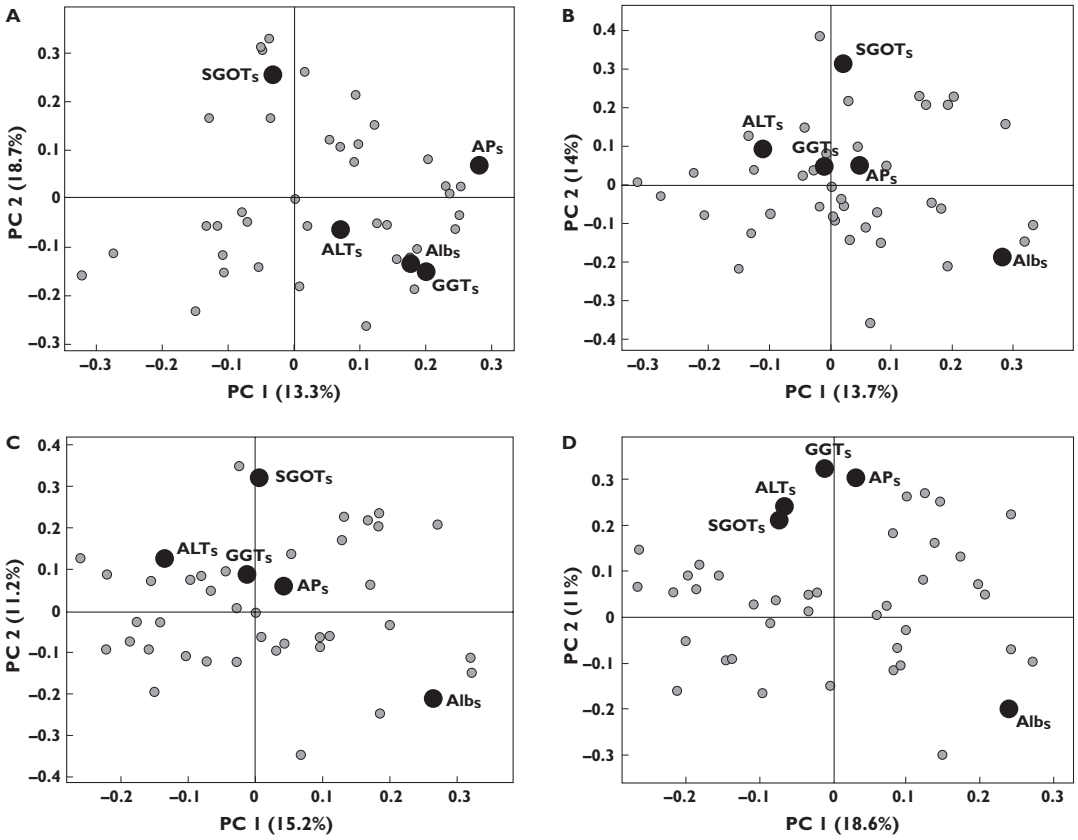
Parameters with high loading values for the second component and low loading values for the first, such as the liver function parameters, e.g. gamma glutamyl transferase (GGT$_S$) behaved opposite to parameters with low second component and high first component loading values such as serum albumin (Alb$_S$). The finding that these and other liver function parameters systematically decreased and increased, respectively, during treatment cycles is novel and was not described in the original univariate analyses of this dataset [9]. Moreover, a subsequent PCA clearly reveals that this underlying pharmacological effect on liver function parameters is visible already at the third dose-level (10 μg kg$^{-1}$) tested during the dose-escalation part of the Phase I trial (Figure 2).

## *Pharmacological responses of the '3/w' dose regimen*

Components 1 and 2 of the '3/w' regimen described 12.7% and 13.2% of the total variation, respectively (Figure 3a,b). In contrast to the cyclic signature of the '5+9' dose regimen, a more continuous pharmacological effect was observed for the '3/w' regimen. This difference in pharmacological effects between the regimens has previously only been described by univariate analysis of sCD25 and hence not as a general phenomenon across all assessed laboratory parameters [9]. In the '3/w' regimen the first component showed large variability in the first few days of treatment, followed by a decrease to a constant yet fluctuating level from day 5 and onwards. Component 2 showed ascending score values for the first 2–3 weeks followed by a stable plateau for the rest of the treatment period, indicating that steady state was reached for the pattern of variables described in component 2. As for the '5+9' regimen, maximal pharmacological responses were observed for the first component at 30 μg kg$^{-1}$ (Figure 3c). However, due to dose-limiting toxicities, dose levels >30 μg kg$^{-1}$ were not tested with the '3/w' regimen [9]. For the second component the pharmacological effect was observed to peak at the 3 μg kg$^{-1}$ dose level (Figure 3d).

## *Comparisons of pharmacological responses between dose regimens*

For the '3/w' dose regimen both differences from and similarities to the '5+9' regimen were observed for the individual laboratory parameters and biomarkers when presented as a scatter plot of loading values corresponding to components 1 and 2 (Figure 3e). As for the '5+9' regimen, the liver function parameters, e.g. GGT$_S$, clustered opposite to Alb$_S$ in the direction of component 2, indicating an impact on the liver function for both regimens. However, for the '3/w' regiment the distance between serum albumin and other liver parameters such as GGT and alkaline phosphatase (AP$_S$) was less pronounced, indicating similar but a more gradual effect on the liver function with the '3/w' regimen compared with the '5+9' regimen. For activation of NK cells, differences were




## Figure 2

Loading plot of component 1 *vs.* component 2 from principal component analysis model from regimen '5+9'. (a) Dose level 1 μg kg$^{-1}$. (b) Dose level 1 and 3 μg kg$^{-1}$. (c) Dose level 1, 3 and 10 μg kg$^{-1}$. (d) Dose level 1, 3, 10 and 30 μg kg$^{-1}$. Only labels for liver functionality parameters are included. Subscript: S, serum

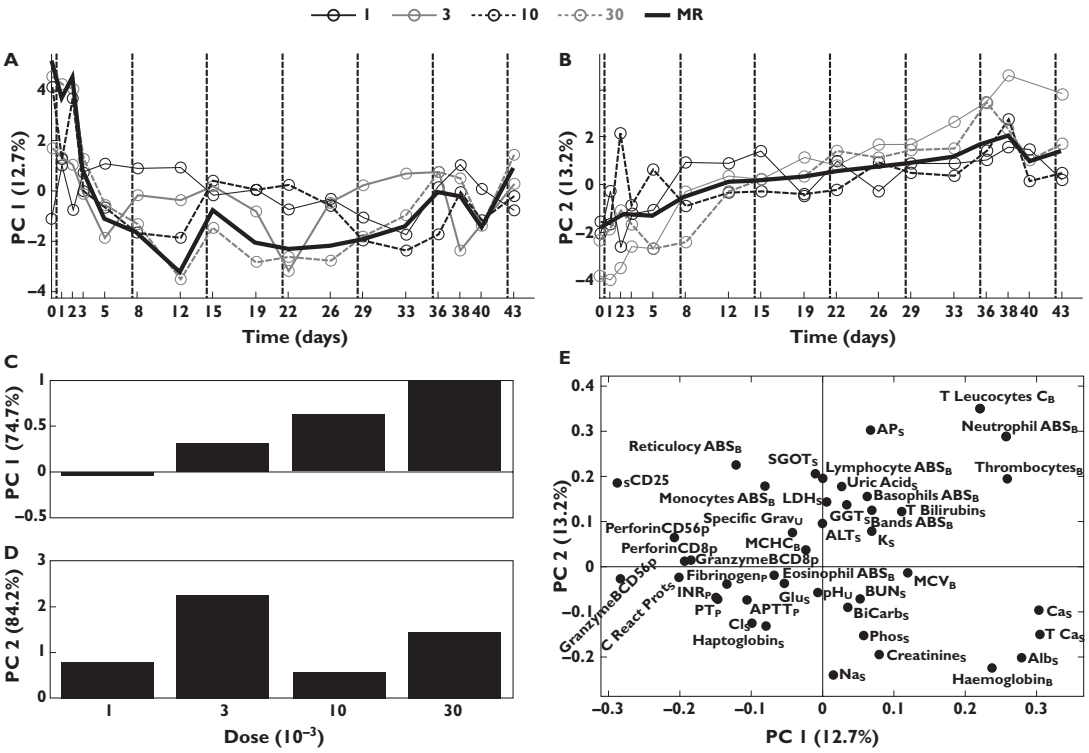
observed between the two dose regimens. In the '3/w' regimen, perforin mRNA expression of purified peripheral blood CD56$^{+}$ NK-cells (PerforinCD56p) and lymphocyte $ABS_B$ clustered in a single quadrant (Figure 3e, upper left). In the '5+9' regimen, these parameters were clearly separated in opposite directions, indicating a more pronounced effect on NK cells compared with the '3/w' regimen (Figure 1e). None of these pharmacological differences between the regimens was found by the univariate methods in the original reports of these data [9, 12].

### *Clinical efficacy and adverse events*

During treatment, tumour size was registered as a secondary end-point for efficacy. In this trial, sporadic antitumour responses were observed in <10% of patients [9]. Analysis of change in tumour size *vs.* score values (for both components and regimens) did not reveal systematic variation in laboratory parameters and biomarkers that correlated with tumour shrinkage (data not shown). A similar analysis for the two most commonly reported adverse events, i.e. fatigue and pyrexia, revealed a positive correlation between score values for component 1 and number of fatigue events ($P < 0.001$) and pyrexia events ($P = 0.02$) for regimen '5+9', indicating a rIL-21 treatment-related response (Figure 4 and Appendix, Analysis of adverse events).

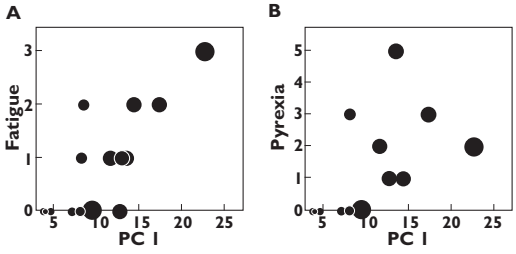
### *Validation of the model in an independent clinical trial*

A total of 43 variables were included in the '5+9' regimen datasets from both the dose-escalation Phase I trial and the fixed-dose Phase II trial. A training model composed of the '5+9' Phase I dataset and a validation model of the Phase II dataset was built. The training model from Phase I




## Figure 3

Results from two-component model of data obtained from regime '3/w'. (a) Score values for component 1 plotted *vs.* day for each of the four doses (μg kg$^{-1}$) and the mean response (MR) of those. (b) Score values for component 2 plotted *vs.* day. (c) Scores from one-component principal component analysis (PCA) model of scores from component 1. (d) Scores from one-component PCA model of scores from component 2. (e) Loading plot of component 1 *vs.* component 2. Subscripts: B, blood; P, plasma; S, serum; U, urine




## Figure 4

Number of adverse events *vs.* score values from a subsequent one-component model of initial principal component analysis results for regimen '5+9'. The size of the point reflects dose level for the respective patient. (a) Number of fatigue events. (b) Number of pyrexia events

was analysed for the ability to predict score values from the Phase II data (Figure 5a,b). Although there were minor differences in magnitude of the pharmacological responses, with a trend towards higher responses in the Phase I trial, the pattern reflecting the signature of the '5+9' regimen was clearly sustained for component 1 (Figure 5a) and to some extent also for the less descriptive second component (Figure 5b). Significant correlations were found between Phase I and II scores for each component (PC1, $R^2 = 0.95$, $P < 0.0001$; PC2, $R^2 = 0.75$, $P = 0.02$).[1] Similarities between the Phase I and Phase II models were further verified by visual comparison of loading values from two independent PCA models (Figure 6). Mathematically, the loadings were rotated and superimposed in order to see if the two models capture the same variation. This revealed that most parameters clustered close together

[1]Calculated on scores from days represented in both trials.

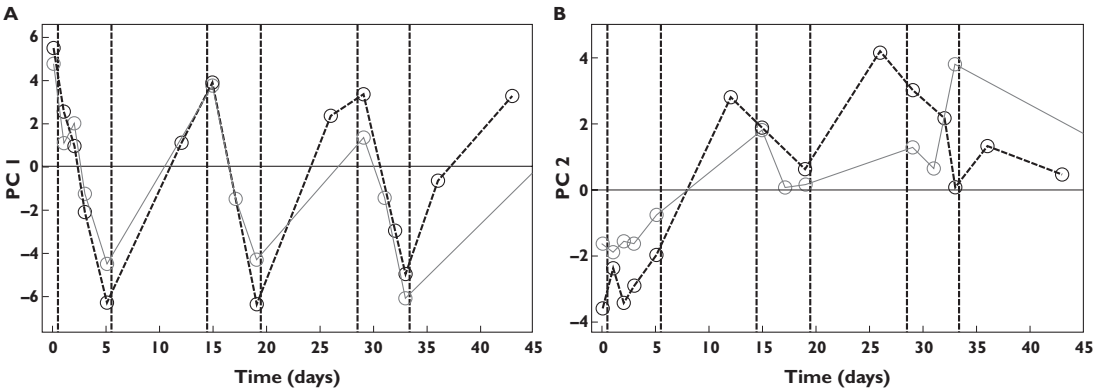



**Figure 5**

Validation and comparison of results from Phase I and Phase II dose regime '5+9'. Calibrated (Phase I) and validated (Phase II) score values for dose 30 μg kg$^{-1}$. (a) PC1. (b) PC2. (A) Phase I Expvar 20.0% (—⊖—); Phase II Expvar 9.1% (- - ⊖ - -); (B) Phase I Expvar 9.7% (—⊖—); Phase II Expvar 3.9% (- - ⊖ - -)

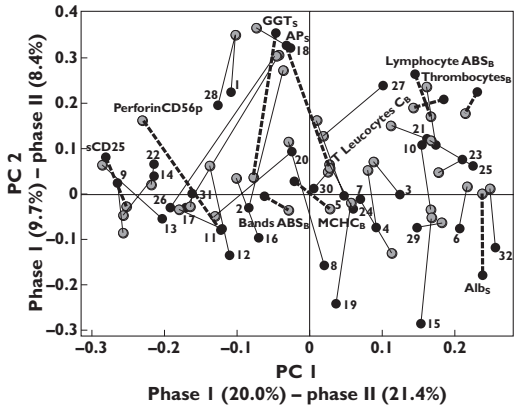



**Figure 6**

Validation and comparison of results from Phase I and Phase II dose regime '5+9'. Comparison of loading plots from two individual models from Phase I (●) and Phase II (●), respectively. Matching variables are connected with lines. Only labels for discussed variables are shown with highlighted lines ('– –'). Subscripts: B, blood; P, plasma; S, serum; U, urine. The remaining variables are numbered, labels are listed in Appendix. Correlation between loadings are $R^2_{(PC\ 1)} = 0.94$ and $R^2_{(PC\ 2)} = 0.40$

within the individual quadrants. Loading values for both components were found to be significantly correlated between the two trials (PC1, $R^2 = 0.94$, $P < 0.0001$; PC2, $R^2 = 0.40$, $P = 0.008$). However, some parameters shifted between the trials such as PerforinCD56p, GGT and albumin, indicating an effect of the dose-escalation *vs.* a fixed-dose trial.

## Discussion

Analysis of individual laboratory parameters from a single or groups of trials can be an exhaustive process in which rare clinical events and/or unexpected pharmacological responses may not be clearly emphasized until relatively late in the development process. Moreover, such responses may be more clearly reflected in a pattern of several variables rather than individual ones. We hypothesized that multivariate chemometric data analysis coupled with visualization would provide an opportunity for a rapid and comprehensive overview of clinical trial data in revealing novel pharmacological findings. The nature of multivariate data analysis is to grasp the common variation of the data and hence the variation common across several variables. We developed a multivariate chemometric data analysis tool based on PCA for visual inspection of pharmacological responses in clinical pharmacology trials. The tool is an unsupervised data analysis tool, which in an assumption-free manner enables visualization of the main variation present in data. It combines methods for handling data collected at different time points as well as methods for focusing on variations relating to dose–time effects rather than interindividual effects. We utilized the tool to show distinct signatures of pharmacological responses of clinical laboratory and biomarker parameters in a Phase I and a Phase II trial investigating rIL-21 in the treatment of patients with malignant melanoma. The responses reflected acute (first component) and cumulative (second component) effects of two different dosing regimens. For both regimens, a maximal pharmacological response was for the first component observed at 30 μg kg$^{-1}$ and supports the notion that additional pharmacological effect cannot be achieved at doses >30 μg kg$^{-1}$. However, for the

second component of the '3/w' regimen the response was observed to peak at the 3 μg $kg^{-1}$ dose level (Figure 3d). Moreover, the response observed for this component gradually increased throughout the trial, indicating accumulation of some of the variables with the '3/w' regimen (Figure 3b). Interestingly, dose-limiting toxicities were observed only at the 30 μg $kg^{-1}$ dose level with the '3/w' regimen and not the '5+9' regimen, suggesting different safety profiles of the two regimens [9]. Nevertheless, the 30 μg $kg^{-1}$ dose level was during dose escalation in the Phase I trial declared by clinical criteria as the maximal tolerable dose for both regimens, supporting the utility of chemometric approaches for selection of dose levels [9]. Moreover, the bell-shaped appearance of the pharmacological effects observed with the '5+9' regimen (Figure 1c,d) has not previously been described and indicates that additional pharmacological activity may not be achievable at dose levels >30 μg $kg^{-1}$. However, whether doses >30 μg $kg^{-1}$ will be feasible is still under consideration [11].

In previous clinical investigations of rIL-21, serum levels of sCD25 have been shown to reflect rIL-21-mediated systemic immune activation and distinct pharmacodynamic responses of individual dosing regimens [9, 11]. These univariate sCD25 responses clearly resemble the observed multivariate patterns of acute (first component) responses across all the assessed laboratory and biomarker parameters and point to sCD25 as a robust surrogate for acute pharmacological responses to rIL-21 (Figures 1e and 3e). Other biomarkers also closely linked to the rIL-21 mechanism of action such as perforin mRNA expression of NK cells and numbers of peripheral blood lymphocytes have also previously been described by univariate methods [9, 12]. Here, we have shown more distinct responses between the two regimens for these biomarkers, indicating more pronounced effects with the '5+9' regimen on NK cell activation. This difference between the two regimens is a novel finding that was not identified during the original analyses of the datasets based on univariate methods and supports the utility of the multivariate approach in regimen selection for subsequent late-stage clinical trials [9, 12]. However, these differences may at least in part also be related to different sample time points and inclusion of higher dose levels in the '5+9' regimen.

In addition to mechanistic patterns of biomarker responses, the multivariate approach also revealed novel correlations between clinical laboratory parameters related to the safety of this novel compound. In particular, rIL-21 induced changes in liver function laboratory parameters that were clearly visualized by inverse clustering of $Alb_s$ and liver enzymes [$GGT_s$, serum glutamic oxaloacetic transaminase ($SGOT_s$), serum alanine aminotransferase ($ALT_s$) and $AP_s$]. The observation that these liver function parameters systematically changed and inversely clustered is novel and was not detected during the original univariate analyses of these data. However, sporadic events of elevated liver enzymes were originally reported by univariate analysis in a subset of the patients [9]. By applying the multivariate approach it clearly becomes visible that these sporadic adverse events reflect a systematic and more general underlying pharmacological adverse effect on liver function parameters. Simultaneous analysis of available data as they are reported reveals that this trend, i.e. similar clustering of $GGT_s$, $SGOT_s$, $ALT_s$ and $AP_s$ *vs.* $Alb_s$ as illustrated in Figure 1e, becomes evident already at the third dose level (10 μg $kg^{-1}$) in the regimen '5+9' of the dose-escalation Phase I trial (Figure 2c). This clearly supports the notion that the multivariate approach provides additional information to univariate methods and this at an earlier stage in the clinical development process. Moreover, differences between the dose regimens may indicate a more gradual effect on liver function with the '3/w' regimen compared with the '5+9' regimen. This important information was also not captured in the original univariate analyses of the data, further supporting the utility of multivariate approaches for regimen selection [9].

The systematic variation explained by the models is approximately 30% for both regimens in two components. There are 43 variables and hence if these were completely independent of each other (orthogonal) each component in a PCA would explain 2.33% of the variation. That 30% is explained in two components therefore directly implies that the variables have a large common underlying structure. It is also expected that the percentage of variance explained is well below 100%. Otherwise, the measured variables would have been completely redundant and all information could have been implicitly obtained from measuring just two variables rather than the present 43. Hence, there is additional information in these variables probably reflecting idiosyncratic phenomena that are not related to treatment.

Variables that do not correlate with the variation in the rest of the data material, i.e. data with low loading values for both component 1 and 2, are insufficiently described by the models. For example, the $MCHC_B$ was inadequately described by the models. Evaluation of these types of variable should be supplemented by additional univariate statistical analysis in order to ascertain whether such variables have a different important clinical relevance.

According to the clinical trial protocols the individual laboratory and biomarker assessments were collected at different sample time points. For example, laboratory biochemistry and urinalysis were assessed on days 1, 2, 3 and 5 in the first week of dosing, whereas sCD25 was assessed on days 1, 2 and 5. In chemometric terms such sample collection schemes create 'missing values'. In the model, such values are estimated by expectation maximization (see Methods). Consequently, variables with a high degree of 'missing values' are estimated with higher uncertainty. For this reason, pharmacokinetic data were not included in the present analysis. As for other cytokines, the half-life of rIL-21 is very short, i.e. approximately

1–4 h [9]. Hence, plasma levels of rIL-21 were undetectable at the time points when all other laboratory parameters were assessed. Future investigations of the proposed multivariate approach in clinical pharmacology trials investigating compounds with longer half-lives will reveal any potential value of integrating pharmacokinetic data into this model.

The robustness of the multivariate approach was tested by comparing a Phase I data model with an independent Phase II data model. This validation revealed that the pharmacological pattern depicted in the Phase I trial was strikingly reproducible and predictive of the response in a subsequent and independent trial. This finding further supports the utility of the multivariate approach in extracting additional information from early Phase I trials to facilitate decision making and planning for late-stage clinical development.

The presented multivariate models were not able to reveal correlations between laboratory variables and clinical end-points for efficacy using the present datasets. Also by univariate methods, parameters predictive of efficacy have not previously been reported and may at least in part be related to the fact that the overall proportion of patients that experienced measurable clinical antitumour responses in these early trials was <10% [9, 10]. In addition, these limited data were rather variable among the individual patients, and lack of association to the systematic PCA components is therefore not surprising. In other therapeutic areas such as diabetes, where clinical efficacy end-points are more frequently encountered, the multivariate model may have a higher potential in revealing correlations to efficacy. In support, the model did reveal significant correlations between the first component for the regimen '5+9' and the more frequently encountered clinical safety end-points for fatigue and pyrexia, emphasizing that these events are related to the acute (component 1) treatment response. By univariate methods, increased levels of serum IL-10 have actually previously been reported in patients experiencing dose-limiting toxicities [14]. Since measurements of IL-10 and a large number of other serum biomarkers were not included in the Phase II trial, these data were not included in the present analysis. However, the lack of strong correlations to infrequently observed clinical end-points illustrates limitations in the utility of the presented multivariate model. This points to the notion that multivariate models should be applied only in conjunction with univariate methods in order to capture all relevant information. Future studies will reveal if multivariate models can be applied in other areas, e.g. in diabetes where the trial design, efficacy rates and safety measurements are very different from what is used in oncology. The presented findings warrant further investigations of the utility of multivariate chemometric approaches in early- and late-stage clinical drug development across therapeutic areas.

## Conclusion

Multivariate chemometric data analysis offers a comprehensive and intuitive tool to reveal early pharmacological responses in clinical pharmacology trials. The multivariate nature of the method allows simultaneous analysis of many parameters and allows more detailed findings than traditional univariate approaches. The present study has revealed novel correlations between laboratory parameters related to liver function and biomarkers exploring the pharmacological properties of rIL-21. Furthermore, the presented multivariate approach can be used as guidance in dose and regimen selections for subsequent studies.

## Competing interests

M.C-J. and L.T.H. are employees of Novo Nordisk A/S.

*Ulrik Mouritzen, Steen Hvass Ingerwersen and Per Knud Christensen are acknowledged for their scientific support and critical comments.*

## Appendix

### Model description

In the following, the complete sequence of steps in our multivariate approach is explained in detail. The basis for the algorithm is a data matrix **X** of size *I* (rows) times *J* (columns). Each column holds the measurements of one variable and each row contains the data for one subject measured at one instance. Generically, the modelling is conducted as the following:

**1** Missing individual observations (elements of **X**) are imputed using expectation maximization [13]. This means that the model is determined in a least squares sense given the observed data without any need to exclude either rows or columns, which would be wasteful and potentially critical given that only few subjects are usually available. Imputation works by iteratively fitting the model to complete data initialized with suitable numbers where missing data occur and then at each iteration replacing the elements that are missing with estimates of the data obtained from the model.

**2** The data matrix **X** (*I X J*) is orthogonalized in order to remove systematic irrelevant variation, here due to subject-specific variation. Orthogonalization can be written formally as $\mathbf{X}_{\mathbf{ort}} = (\mathbf{I} - \mathbf{D}\mathbf{D}^{+})\mathbf{X}$ where $\mathbf{X}_{\mathbf{ort}}$ is the filtered data, **I** (*I X I*) is the identity matrix, and **D** (*I X* $n_{subj}$) is the design matrix with respect to subject. $\mathbf{D}^{+}$ refers to the pseudoinverse of **D**, $n_{subj}$ refers to the total number of subjects/patients. The matrix **D** is a dummy matrix that contains ones in column *n* in the rows of subject *n*. This orthogonalization step removes any differences in level between subjects, i.e. similar to what can be termed the subject-effect in analysis of variance [15].

Orthogonalization is described in detail in the latter part of this Appendix. Removing this variation is essential in order to filter off subject variation that is not related to the effect of the treatment. Note that other types of irrelevant variation can also be removed using several orthogonalization steps if needed. Also, note that the residuals of the orthogonalization contain the subject-specific variation that, if needed, can be further scrutinized.

**3** Having removed the subject-specific variation, a PCA model [16] is determined on the orthogonalized data. As opposed to traditional PCA, it is crucial to use the correct number of components, because the missing data imputation in step 1 depends on the number of components. In practice, several numbers of components are tested and evaluated using the explained variance *vs.* number of components plot as is usual in PCA [17], and number of iterations for determination of missing values. The statistical relevance of the model is tested by permutation testing using 1999 random permutations [18].

**4** Post-processing is performed on the result of the PCA model in order to enhance the visualization. Specifically, the scores of the PCA model are averaged across day-dose. Hence, instead of presenting a score value for each subject, the scores are shown as averages for a specific day and dose.

## The algorithm

According to the above, the formal algorithm can be written as follows.

**1** Missing data elements are imputed with mean values of the respective variables.

**2** Data are orthogonalized into; $\mathbf{X}_{ort}$ (matrix with information orthogonal to **D**) and $\mathbf{X}_{rest}$ (matrix with information linear correlated to **D**)

$$\mathbf{X}_{ort} = (\mathbf{I} - \mathbf{D}\mathbf{D}^{+})\mathbf{X}$$

$$\mathbf{X}_{rest} = \mathbf{D}\mathbf{D}^{+}\mathbf{X}$$

where $\mathbf{D}^{+} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$

**3** Data are autoscaled to equal variance [19].

**4** PCA on orthogonalized and autoscaled data, decomposing $\mathbf{X}_{ort}$ into a score matrix (**T**), a loading matrix (**P**) and a residual matrix (**E**).

$$\mathbf{X}_{ort} = \mathbf{T}\mathbf{P}' + \mathbf{E}$$

**5** PCA data approximation calculated

$$\mathbf{X}_{model\text{-}ort} = \mathbf{T}\mathbf{P}'$$

**6** Approximation de-orthogonalized

$$\mathbf{X}_{model} = \mathbf{X}_{model\text{-}ort} + \mathbf{X}_{rest}$$

**7** Approximations backscaled

**8** Missing data are imputed with backscaled approximations.

Step 2–8 is repeated until convergence of approximation results.

## Orthogonalization

**X** *(I X J)* is a data matrix, **D** *(I X k)* is matrix (or vector) with external information such as patient id. The purpose of orthogonalization is to split **X** into a part linear related to **D** ($\mathbf{X}_{D}$) and a part orthogonal/perpendicular to **D** ($\mathbf{X}_{OD}$).

Examine the linear regression problem

$$\mathbf{X} = \mathbf{D}\mathbf{B} + \mathbf{E}$$

where **B** is some regression matrix and **E** is the part of **X** not explained by **D**. From this **B** can be extracted as:

$$(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} = (\mathbf{D}'\mathbf{D})^{-1}(\mathbf{D}'\mathbf{D})\mathbf{B} = \mathbf{B}$$

and $\mathbf{X}_{D}$ can be estimated to:

$$\mathbf{X}_{D} = \mathbf{D}\mathbf{B} = \mathbf{D}\left((\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X}\right)$$

and $\mathbf{X}_{OD}$ to:

$$\begin{aligned}\mathbf{X}_{OD} &= \mathbf{X} - \mathbf{X}_{D} = \mathbf{X} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} \\ &= \left(\mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\right)\mathbf{X} = (\mathbf{I} - \mathbf{D}\mathbf{D}^{+})\mathbf{X}\end{aligned}$$

with $\mathbf{D}^{+} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ as the pseudo inverse of **D**.

## Analysis of adverse events

The numbers obtained throughout the trial period of the most frequent adverse events (fatigue and pyrexia) were compared with score values from a one-component PCA model on the initial results. In Figure 4 data are shown for regimen '5+9'.

## Abbreviations for Figure 6

**1** ALT (alanine aminotransferase) serum
**2** APTT (activated partial thromboplastin time) plasma
**3** Basophils ABS blood
**4** BiCarbonate serum
**5** BUN (blood urea nitrogen) serum
**6** Calcium serum
**7** Chloride serum
**8** Creatinine serum
**9** C-reactive protein serum
**10** Eosinophil ABS blood
**11** Fibrinogen plasma
**12** Glucose serum
**13** GranzymeBCD56p
**14** GranzymeBCD8p
**15** Haemoglobin blood
**16** Haptoglobin serum
**17** INR (International Normalized Ratio) plasma
**18** LDH (lactate dehydrogenase) serum

**19** MCV (mean corpuscular volume) blood
**20** Monocytes ABS blood
**21** Neutrophil ABS blood
**22** PerforinCD8p
**23** Phosphorous serum
**24** pH urine
**25** Potassium serum
**26** PT plasma
**27** Reticulocyte ABS blood
**28** SGOT (serum glutamic oxaloacetic transaminase) serum
**29** Sodium serum
**30** Specific Grav urine
**31** T Bilirubin serum
**32** T Calcium serum
**33** Uric acid serum

## REFERENCES


**1** Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab 1987; 2: 37–52.

**2** Christie OHJ. Introduction to multivariate methodology. An alternative way? Chemometr Intell Lab 1995; 29: 177–88.

**3** Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. Br J Anaesth 2006; 97: 64–8.

**4** Albanese J, Martens K, Karkanitsa LV, Dainiak N. Multivariate analysis of low-dose radiation-associated changes in cytokine gene expression profiles using microarray technology. Exp Hematol 2007; 35: 47–54.

**5** Liszka-Hackzell JJ, Schött U. Presentation of laboratory and sonoclot variables using principal component analysis: identification of hypo- and hypercoagulation in the HELLP syndrome. J Clin Monit Comput 2004; 18: 247–52.

**6** Keun HC. Metabonomic modeling of drug toxicity. Pharmacol Ther 2006; 109: 92–106.

**7** Keun HC, Athersuch TJ. Application of Metabonomics in drug development. Pharmacogenomics 2007; 8: 731–41.

**8** Skak K, Frederiksen KS, Lundsgaard D. Interleukin-21 activates human natural killer cells and modulates their surface receptor expression. Immunology 2008; 123: 575–83.

**9** Davis ID, Skrumsager BK, Cebon J, Nicholaou T, Barlow JW, Moller NPH, Skak K, Lundsgaard D, Frederiksen KS, Thygesen P, McArthur GA. An open-label, two-arm, phase I trial of recombinant human interleukin-21 in patients with metastatic melanoma. Clin Cancer Res 2007; 13: 3630–6.

**10** Davis ID, Brady B, Kefford RF, Millward M, Cebon J, Skrumsager BK, Mouritzen U, Hansen LT, Skak K, Lundsgaard D, Frederiksen KS, Kristjansen PEG, McAthur GA. Clinical and biological efficacy of recombinant human interleukin-21 (rIL-21) in patients with stage 4 malignant melanoma without prior treatment: a phase 2a trial. Clin Cancer Res 2009; 15: 2123–9.

**11** Thompson JA, Curti BD, Redman BG, Bhatia S, Weber JS, Agarwala SS, Sievers EL, Hughes SD, DeVries TA, Hausman DF. Phase I study of recombinant interleukin-21 in patients with metastatic melanoma and renal cell carcinoma. J Clin Oncol 2008; 26: 2034–9.

**12** Frederiksen KS, Lundsgaard D, Freeman JA, Hughed SD, Holm TL, Skrumsager BK, Petri A, Hansen LT, McArthur GA, Davis ID, Skak K. IL-21 induces *in vivo* immune activation of NK cells and $CD8^+$ T cells in patients with metastatic melanoma and renal cell carcinoma. Cancer Immunol Immunother 2008; 57: 1439–49.

**13** Krijnen WP, Kiers HAL. An efficient algorithm for weighted PCA. Comput Stat 1995; 10: 299–306.

**14** Dodds MG, Frederiksen KS, Skak K, Hansen LT, Lundsgaard D, Thompson JA, Hughes SD. Immune activation in advanced cancer patients treated with recombinant IL-21: multianalyte profiling of serum proteins. Cancer Immunol Immunother 2009; 58: 843–54.

**15** Box EP, Hunter WG, Hunter JS. Statistics for Experimenters. New York: John Wiley and Sons, 1978.

**16** Jackson JE. Principal components and factor analysis: part I – principal components. J Qual Technol 1980; 12: 201–13.

**17** Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem 2008; 390: 1241–51.

**18** Ledauphin S, Hanafi M, Qannari EM. Simplification and signification of principal components. Chemometrics Int Laboratory Systems 2004; 74: 277–81.

**19** Bro R, Smilde AK. Centering and scaling in component analysis. J Chemometrics 2003; 17: 16–33.

# Paper II

Multivariate analysis of the relation between diet and warfarin dose

Morten A. Rasmussen, Jane Skov, Else Bladbjerg, Johannes Sidelmann, Marianne Vamosi, Jørgen Jespersen

*European Journal of Clinical Pharmacology,* ***68*** *(2012), 321-328*

Eur J Clin Pharmacol (2012) 68:321–328
DOI 10.1007/s00228-011-1123-3

PHARMACOEPIDEMIOLOGY AND PRESCRIPTION

# Multivariate analysis of the relation between diet and warfarin dose

**Morten Arendt Rasmussen · Jane Skov · Else-Marie Bladbjerg · Johannes J. Sidelmann · Marianne Vamosi · Jørgen Jespersen**

Received: 12 July 2011 / Accepted: 29 August 2011 / Published online: 21 September 2011


**Abstract**
*Purpose* The vitamin K antagonist (VKA) warfarin is effective for the prevention of thromboembolisms. Maintenance doses differ greatly among patients and are known to be primarily determined by genetic polymorphisms. The relative impact of dietary vitamin K intake is still a matter of debate. We hypothesize that a multivariate model is more suitable for exploring the relation between dietary intake of vitamin K and warfarin dose than conventional uni- or bivariate analyses.
*Methods* In a cross-sectional study, we interviewed 244 patients in the maintenance phase of warfarin therapy and detected polymorphisms in the *VKORC1* and *CYP2C9* genes. Dietary vitamin K intake was estimated from food frequency questionnaires.
*Results* A univariate correlation analysis and the regression coefficient from the multivariate model showed a small but significant negative relation between vitamin K intake and warfarin dose. A loading plot of the partial least squares regression model illustrated this counter-intuitive observation, which might be explained by the latent structure between variables. The variation in warfarin dose could be divided into two significant latent variables, the so-called components. In component one, pharmacogenetics explained 52% of dose variation. Component two described health-related behavior (diet, physical activity and body weight) and explained 8% of dose variation. Here, vitamin K intake positively correlated with warfarin dose.
*Discussion* This study highlights the importance of choosing a statistical method that reflects the complexity of data for interpretation of results from observational studies. The multivariate model appears to be well suited to describe the complex relationship between vitamin K intake and VKA dose.

**Keywords** Partial least squares regression · Vitamin K · Warfarin · Diet

Morten Arendt Rasmussen and Jane Skov contributed equally.

M. A. Rasmussen
Department of Food Science/Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Copenhagen, Denmark

J. Skov · E.-M. Bladbjerg · J. J. Sidelmann · J. Jespersen
Unit for Thrombosis Research, Institute of Public Health, University of Southern Denmark, Esbjerg, Denmark

M. Vamosi
Unit for Health Promotion, Institute of Public Health, University of Southern Denmark, Esbjerg, Denmark

J. Skov (✉)
Department of Clinical Biochemistry, Hospital of South West Denmark, Unit for Thrombosis Research, University of Southern Denmark,
Finsensgade 35,
DK-6700 Esbjerg, Denmark
e-mail: skov_jane@yahoo.com

## Introduction

Vitamin K antagonists (VKA) have been widely used to prevent thromboembolisms for several decades [1]. They have a narrow therapeutic window, and the response to treatment can be measured in terms of the international normalized ratio (INR). Patients treated with VKA require frequent and careful monitoring, often in the setting of specialized anticoagulant clinics [2]. Despite recent advances, e.g., in the standardization of thromboplastins [3], clinicians responsible for VKA treatment are still faced by

numerous challenges. One of these challenges is the large intra-individual variation in VKA dose demands to achieve INR values within the therapeutic range, often necessitating adjustments of the VKA dose. One potential cause of INR fluctuations is dietary vitamin K intake [4, 5]. Further complicating this matter, therapeutic dose requirement between patients may differ by more than a factor 10 [6].

The mechanism of action and metabolism of VKA explains some of the large dose variation: vitamin K serves as a cofactor in a hepatic transformation of glutamic acid residues into γ-carboxyglutamic acid. This post-translational modification is necessary for the normal function of the vitamin K-dependent coagulation factors (factors II, VII, IX and X and proteins C, S and Z). During the reaction vitamin K is oxidized and must be reduced by the enzyme vitamin K epoxide reductase (VKOR) before it can function as a cofactor again. VKA treatment inhibits VKOR and leads to reduced recycling of vitamin K and thereby a reduction in the amount of functional coagulation factors [7]. The type of VKA mostly used in Denmark and worldwide is warfarin [1], of which the most active enantiomer S-warfarin is metabolized by the hepatic enzyme cytochrome P450 2C9 [8]. This enzyme is responsible for metabolism of a wide range of xenobiotics [9] and can be inhibited or induced by several classes of pharmaceuticals [10]. Accordingly, about half of the inter-individual variation in the therapeutic dose of warfarin can be explained by single nucleotide polymorphisms (SNPs) in the highly polymorphic *VKORC1* and *CYP2C9* genes [11]. Clinical variables such as age, weight, ethnicity and certain co-medications are also responsible for dose variation, but to a smaller degree [6].

Interestingly, patients with a low dietary intake of vitamin K may be prone to INR fluctuations [12], and recent studies have indicated that vitamin K supplementation can improve the stability of VKA treatment [5, 13]. In light of these findings, it seems highly relevant to explore the impact of vitamin K intake on VKA treatment.

Intuitively, it is expected that patients who have a large dietary intake of vitamin K require a larger dose of VKA than patients with a smaller dietary intake. Nevertheless, this association was either absent [14] or shown to be very slight [15] in two cross-sectional studies of more than 300 patients in the maintenance phase of anticoagulant therapy, and a smaller study showed no correlation between maintenance dose of warfarin and vitamin K intake averaged over 28 days [16].

Conventional linear regression analyses such as those employed in these studies are useful for determining the quantitatively most important variables, but are not necessarily the best choice for investigating relations that may prove more complex, such as the relation between diet and warfarin dose. The following report employs a multivariate technique with visualization features to investigate the relation between dietary intake of vitamin K and a maintenance dose of warfarin. This method enables accounting for sources of variability that directly or indirectly affect one or both of these variables, in this report including genetic polymorphisms, selected clinical variables and physical activity level. Thereby, we aimed to gain a deeper understanding of the relation between diet and warfarin maintenance dose.

## Materials and methods

### Design and study participants

The participants in this cross-sectional study were patients in the maintenance phase (duration of treatment>3 weeks at the time of inclusion) of VKA treatment at the Department of Clinical Biochemistry at the Hospital of Southwest Denmark. At this nurse-managed and physician-supervised anticoagulant clinic, dosage of VKA is computer-assisted using the DAWN anticoagulant therapy software (4S Information Systems Ltd, Cumbria, UK) [17]. All patients attending the anticoagulant clinic were eligible for inclusion in the project if they were above 18 years of age and provided written consent. Patients were excluded if treatment with heparin, warfarin or phenprocoumon was contraindicated or they were undergoing cancer chemotherapy. A total of 250 consecutive patients were enrolled between May 2009 and January 2010. Because of the explorative design of the study, no power analyses were carried out, but the resulting population size is comparable to those of other studies on diet and VKA dose [14, 15]. In addition to the 250 patients included, a total of 116 patients declined to participate. The non-participants were older and more likely to be diagnosed with atrial fibrillation than the participants [18].

Six patients were treated with phenprocoumon, while the remaining took warfarin. For the present analyses, the patients taking phenprocoumon were excluded, resulting in a total of 244 patients. The characteristics of this ethnically homogeneous population (except for one Asian and one Italian, all were Scandinavian) are displayed in Table 1.

In dealing with ethical aspects the Declaration of Helsinki was followed throughout the study and approval was obtained from the regional Ethics Committee (project-ID S-20080020).

### Data sampling

On the day of enrollment, a structured interview was conducted. Following the interview, anthropometric measurements were taken by the interviewer. Information about VKA dose, INR, and INR target was extracted from DAWN.

**Table 1** Baseline characteristics of study participants ($n$=244)

| Characteristic | Data |
|---|---|
| Age[a] | 68 years (60–75) |
| Sex | 134 males (65%) |
| BMI[a] | 28.5 kg/m$^2$ (25.9–32.1) |
| Indication for warfarin treatment | Atrial fibrillation 140 (57%) |
| | Venous thrombosis 67 (28%) |
| | Mechanical heart valve 29 (12%) |
| | Other 8 (3%) |
| Duration of warfarin treatment>12 months | 134 (55%) |
| INR target | 2.5 (range 2.0–3.0): 234 (96%) |
| | 3.0 (range 2.5–3.5): 10 (4%) |
| INR within therapeutic range | 161 (66%) |
| Warfarin dose[a] | 31.9 mg/week (22.5–47.5) |
| Vitamin K score[a] (arbitrary) | 65 (37–109) |
| VKORC1 genotype (rs9934438) | CC 80 (33%) |
| | CT 124 (51%) |
| | TT 40 (16%) |
| CYP2C9 genotype (2* and 3*) | 1.1 159 (65%) |
| | 1.2 48 (20%) |
| | 1.3 31 (13%) |
| | 2.3 or 3.3 6 (3%) |
| Polypharmacy patients (≥5 medications) | 128 (53%) |
| Amiodarone users | 21 (9%) |

INR, international normalized ratio

[a] Median (quartiles)

## VKORC1 and CYP2C9 polymorphism detection

On the day of the interview venous blood was sampled from each patient in tubes containing EDTA and subsequently used for DNA analysis. Genomic DNA was extracted from peripheral blood cells following centrifugation, removal of plasma and lysis of erythrocytes.

*CYP2C9* genotyping was performed by PCR-RFLP methods for detection of four SNPs rs1856908, rs9332113, rs1934968, and rs9332238. For rs1856908, the following primers were used: 5′CTG GGA TTG CAT GTT GGT TT and 5`GGA ATT TTC TCA GGC AGA TCA, for rs9332113 5′GTT AGA CGG AGA CGA TCA CGT and 5′AGG AGA GTT CCT TTG AGG CCA GGT, for rs1934968 5′GAT GAT GTT AAT CTG TCA ACT TTG C and 5′ACA AGG ATC CCC ACT GTC AC, for rs9332238 5′CCC ATC CAC CCA TCT ATC TC and 5′CCG TTT TCC TGA AAA TAG CAA. Primer sequences were found using the primer3 program [19]. Amplification took place in a final volume of 25 μl containing 1 U of Taq polymerase (Roche Applied Sciences, Mannheim, Germany), 25 ng of each primer, 1.25 nmol of each dNTP (Stratagene 200415, Agilent Technologies, Santa Clara, CA, USA), and 2.5 μl 10* reaction buffer (Roche Applied Sciences, Mannheim, Germany). Cycles ($n$=30) of 92°C for 30 s, 62°C for 50 s, and 72°C for 30 s were carried out for all reactions.

The rs9332238 amplicon was digested by the Tfi-I restriction endonuclease (at a temperature of 65°C for 2 h), the rs1934968 amplicon by Bsm-I (65°C, 2 h), the rs9332113 amplicon by Spe-I (37°C, 3 h), and the rs1856908 amplicon by Bsr-I (65°C, 2 h).

A multiplex PCR-RFLP was used for the detection of the *2 (rs1799853) and *3 (rs1057910) allele variants [8].

Real-time PCR followed by melting curve analysis was performed on the Lightcycler Instrument 2.0 (Roche Diagnostics) for detection of the *VKORC1* SNP 1173C/T (rs9934438) [20]. RFLP methods were used for detection of the −1639G/A (rs9923231), the 689C/T (rs17708472), and the 9041G/A (rs7294) SNP. We used the primer sequences and restriction enzymes reported by Sipeky et al. [21].

We re-analyzed 5% of the samples (at random) and the results were confirmed. Our internal quality control program for DNA analysis is described in detail elsewhere [22].

## Statistical analyses

Univariate correlation analysis was performed using Spearman's correlation.

We used 37 variables considered to be predictors for estimation of VKA dose for the multivariate analyses. In our attempt to investigate the correlation between dietary vitamin K intake and the maintenance dose of VKA, we chose to include genetic polymorphisms, which are known to have a significant effect on VKA dose, and clinical variables, some of which are known to influence VKA dose, in addition to the variables describing diet and physical activity. We entered all ten SNPs tested. The number of variant alleles was considered continuous (in terms of the effect on the maintenance dose of VKA) and the SNP genotypes were hence coded 1 (wild type), 2 (one variant allele), and 3 (two variant alleles). The continuity assumption was verified by individual bar plots of VKA dose for each SNP (results not shown). The clinical variables were age, sex, height, weight, BMI, hip circumference, waist circumference, INR target (a dichotomous variable describing whether the target was 2.5 or 3.0), number of medications, and use of amiodarone.

Dietary habits were evaluated by a food frequency questionnaire with 80 items, covering common components of a traditional Danish diet such as rye bread or potatoes and items with high vitamin K content, i.e., broccoli, spinach, and liver. Vitamin K intake during the previous 4 weeks was estimated from the food frequency question-

naire focusing on 22 specific foods with a moderate to high vitamin K content. Each food was assigned a unit value by dividing the vitamin K content in micrograms per 100 grams by 100 [15]. For example, broccoli has a vitamin K content of 260 μg/100 g and was assigned the unit value 2.6. The unit value was multiplied by an estimate of the number of times the patients had ingested the food during the last month ("a few times"=2, "a couple of times a week"=8, "once daily"=25 and "more than once a day"=50). These values were summed to give an estimated vitamin K score [15]. The foods and unit values were broccoli (2.6), watercress (2.5), beans (1.7), spring onions (2.07), kale (2.5), white cabbage (0.59), wheat bran (0.83), endive (2.31), chick peas (2.64), liver (1.04), mayonnaise (0.75), parsley (7.9), chives (3.1), Brussels sprouts (2.5), red cabbage (1.49), green salad (1.3), celery root (1), soy oil (5.4), pointed cabbage (1.7), spinach (3.4), seaweed (13.85) and peas (0.7).

In addition to vitamin K score, the food items were divided into seven groups: grains and starches, fruit and vegetables, meat, dairy products, sources of fat, sources of sugar, and drinks. For each food group two principal components were extracted describing the main systematic variation of data. These 14 components were used as a condensed measure of the 80 primary food variables.

Physical activities were described by the continuous medical emergency team (MET) score calculated from the short form of the International Physical Activity Questionnaire [23] and the patient's overall classification of spare time activities, denoted "physical activity level" (1=sedentary, 2=moderately active, 3=participated in exercise activities. The last category, 4=participated in competitive sports, was not chosen by any of the patients).

In summary, the 37 variables can be divided into several groups: 10 polymorphisms, 10 clinical variables (age, sex, weight, height, BMI, waist circumference, hip circumference, INR target, number of medications, and amiodarone), food components (14), physical activity (2), and vitamin K score.

Partial least squares regression (PLS) was used to decompose the information from these 37 variables from 244 patients into a few underlying significant latent variables (which are called components in the following text) relevant for prediction of warfarin dose. Overfitting was avoided by inclusion of only statistically significant components, the number of which was estimated by cross-validation (random segmentation into 10 segments, repeated five times), evaluating the cross-validated mean squared error for different numbers of components.

In addition to the prediction ability, PLS models also have additional visualization features. The PLS model decomposes data into scores and loadings [24]. The so-called "scores" are related to the samples, i.e., the patients, and describe how different or similar they are in terms of the variation relevant for prediction. The "loadings" are related to variables and reflect the pattern of variables that are responsible for the prediction. In this report, scatter plots of component one (*x*-axis) and component two (*y*-axis) have been used to illustrate loadings. Closely related variables are placed closely together, while negative associations are reflected by variables being placed opposite each other in the horizontal or vertical direction. For example, two variables that are placed closely together in the direction of the *x*-axis are positively correlated in component one, while variables placed on opposite ends of the *y*-axis negatively are correlated in component two. Variables clustering in the plot are therefore positively correlated in both components, while other variables may have a positive correlation through one component and a negative in the other. In the present study, the focus of the multivariate analysis is to use loading plots for interpretation of the relations between variables.

## Results

Both vitamin K score and warfarin dose displayed great inter-individual variation and both deviated from a normal distribution (see Table 1). The median vitamin K score ranged between 0 and 329 and warfarin dose between 5 mg/week and 117.5 mg/week. The relation between vitamin K score and warfarin dose is shown in Fig. 1. From this crude analysis we found that there was a

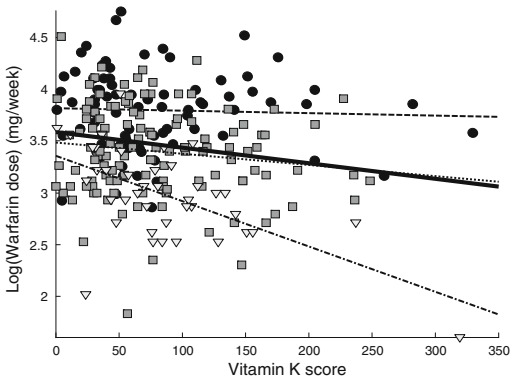



**Fig. 1** Scatter plot of vitamin K score versus maintenance dose of vitamin K antagonist ($n$=244). The *bold line* shows the best fit. Spearman's correlation coefficient is $r$=−0.16, $p$=0.012. The three *thinner lines* show the relation for patients with no (*upper line*, the patients are indicated by *circles*), one (*middle line*, the patients are indicated by *squares*) or two (*lower line*, the patients are indicated by *triangles*) variant alleles of the VKORC1 polymorphism rs9923231/rs9934438. The data are unadjusted and should be evaluated cautiously

modest but significant negative association between the two variables ($r=-0.16$, $p=0.012$). This result is unadjusted for any confounders known to have an influence on warfarin dose. We stratified the patients by the VKORC1 polymorphism rs9923231/rs9934438 and found that all three groups had a negative relation between vitamin K score and warfarin dose, and that this was most pronounced for patients with two variant alleles and thereby the lowest dose requirements.

Next, we included relevant confounders such as weight, height, food groups, physical activity level, genetic polymorphisms and variables relating to VKA treatment in a multivariate PLS model with warfarin dose as the outcome and the remaining variables as predictors. From this confounder-adjusted analysis, the regression coefficient for dietary vitamin K intake was still negative, albeit rather small. The largest coefficients were observed for the two *VKORC1* polymorphisms in complete linkage disequilibrium rs9923231 and rs9934438, the *CYP2C9* polymorphism rs1057910 (*3) and age; all were negative (results not shown).

To further investigate the apparent counter-intuitive relation between vitamin K score and warfarin dose, we scrutinized the loading plot of the multivariate PLS model. In the present model, two components (representing latent variables, see also Materials and methods) were found to have statistical significance. The loadings for the 37 predictors and the outcome variable are illustrated in a scatter plot of component one versus component two (Fig. 2a). The present data contain some redundancy of certain predictor variables, i.e. BMI is derived from height and weight, and BMI and waist circumference are also expected to increase in some synchrony. Therefore, it makes good sense that weight, BMI, hip circumference and waist circumference are clustered closely together in Fig. 2. When including only the patients with an INR measurement with the designated target, the loading plot was similar to that of all patients (Fig. 2b).

Both plots (Fig. 2a and b) show that the main variation in warfarin dose is described by the two polymorphisms in complete linkage disequilibrium *VKORC1* SNPs rs9923231 and rs9934438, clinical variables such as weight, BMI, waist circumference, hip circumference and to a smaller extent the food group "drinks component 2". These measures span the first and most significant component in terms of prediction of warfarin dose and comprise the extreme *x*-values in the plot. In this direction, a negative relation between vitamin K score and warfarin dose was observed. Component one explained 52% of the variation in warfarin dose.

The second component is primarily described by the clinical variables weight, BMI, waist circumference, and hip circumference. This component is spanned by the above measures on one side and vitamin K score, "fruit and vegetables component 1," and physical activity level on the other side and make up the *y*-axis of the figure. In this direction, a negative correlation between body weight and vitamin K intake was demonstrated, while warfarin dose was positively associated with vitamin K score. Component two explained an additional 8% of warfarin dose variation.

Age and the *CYP2C9* polymorphism rs1057910 (*3) contribute to both component one and two and negatively correlated with warfarin dose in both.

## Discussion

The present study describes a surprising negative relation between vitamin K intake and maintenance dose of warfarin. This effect was most prominent in the group of patients with the lowest dose requirements (Fig. 1). When adjusting for a number of possible confounders in a multivariate analysis, we confirmed that maintenance dose of warfarin primarily depends on pharmacogenetics, importantly polymorphisms in *VKORC1* and *CYP2C9*, and clinical variables such as body weight and age [15, 25]. The presence of variant alleles of the rs9923231/rs9934438 or rs1057910 and higher age showed a negative correlation with VKA dose, while body weight correlated with a higher warfarin dose in component one and a lower dose in component two (Fig. 2a or b). The correlations with SNPs and age dominate the results and are hence described in latent component one, visualized as the *x*-axis of the loading plots (Fig. 2a or b). Component one can be considered a mainly "pharmacogenetic axis" when attempting to explain warfarin dose and did, not surprisingly, explain the greatest part of dose variation (52%).

As opposed to the paradoxical negative relation between vitamin K score and warfarin maintenance dose initially observed, we found a positive correlation in component two, illustrated by the *y*-axis on scatter plots of the loadings from the PLS model (Fig. 2a and b). In other words, we observed the expected positive relation between vitamin K score and maintenance dose in component two. Component two cannot purely be viewed as the "health-related behavioral axis" since pharmacogenetic and clinical variables also contribute. Nevertheless, it is in this direction that the influence of vitamin K intake on warfarin dose is most clearly illustrated, in addition to the correlations between this variable, body weight, and physical activity. Component two explains a much smaller portion of dose variation than component one, again pointing out the huge importance of pharmacogenetics for VKA dose prediction. It may seem a little too convenient to search for the component showing the desired result. In the present example, however, the pharmacogenetic variables completely overshadow the influence of diet in quantitatively most

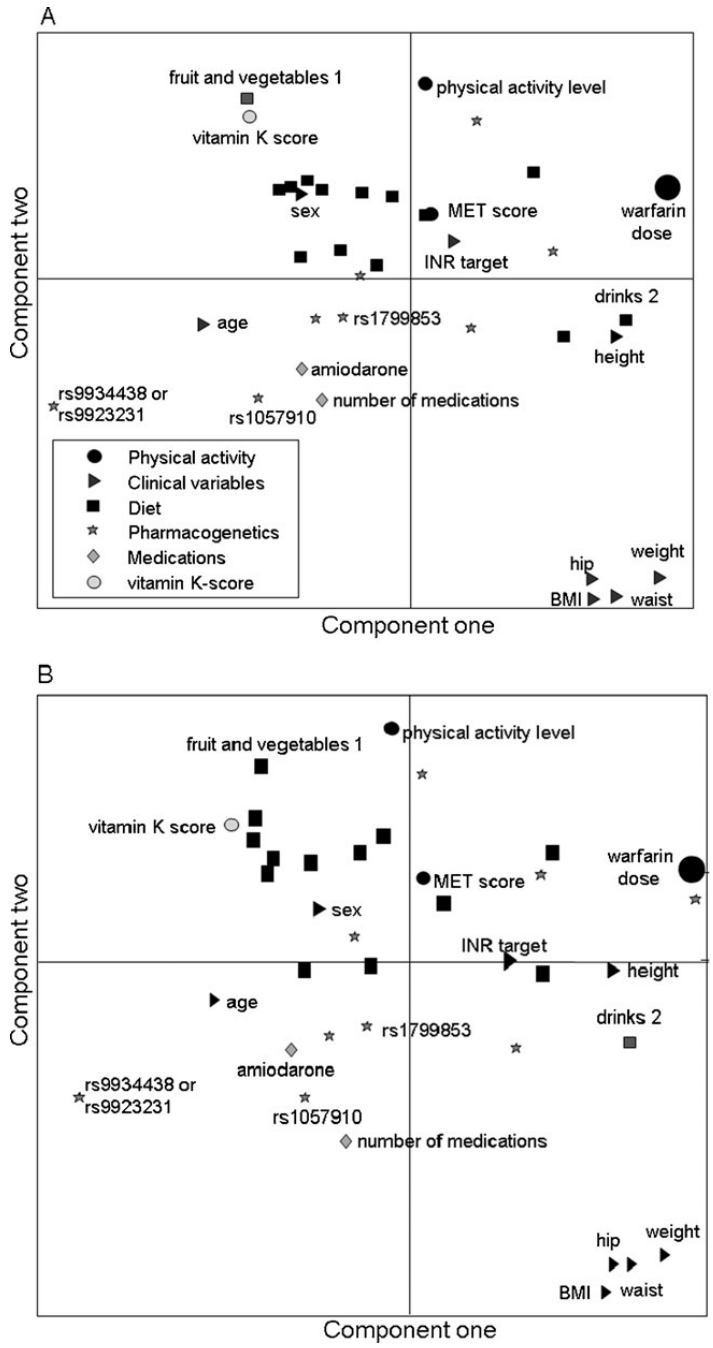



**Fig. 2** Loading plot of component one versus component two from a partial least squares regression model predicting maintenance dose of warfarin from 37 predictor variables, consisting of selected clinical variables ($n$=10), food components ($n$=14, describing consumption of grains and starches, fruit and vegetables, meat, dairy products, sources of fat, sources of sugar, and drinks, two components for each), overall physical activity level classification (dividing the patients into sedentary, moderately active or participants in exercise activities) and medical emergency team (MET) score (reflecting time spent on physical activities), polymorphisms in *VKORC1* and *CYP2C9* ($n$=10), and vitamin K score. Labels are included for all the clinical variables, but only the major contributing genetic polymorphisms and food components. **a** is a loading plot for all patients ($n$=244), while **b** shows the 161 patients who had an international normalized ratio (INR) within the therapeutic range at the date of inclusion in the project. A loading plot is a visual representation of the relation structure between variables, where long distances in either the vertical or horizontal direction signify a negative relation, while the clustering of variables indicates a positive relation between them

important component one. Therefore, one has to turn to component two before the relations between vitamin K score, physical activity, and other health-related variables can be evaluated.

Vitamin K score obviously directly correlates with diet. This variable was derived from the amount of intake of certain vitamin K-rich foods, mainly leafy green vegetables [26]. Accordingly, the loading plots highlight that "fruit and

vegetables 1" is closely positioned with vitamin K score. The composition of diet is connected to the group of related clinical variables including weight, BMI, hip circumference, and waist circumference. Thus, a diet rich in fruit and vegetables correlates with a lower body weight. Furthermore, the patients' self-reported level of physical activity was positioned close to intake of fruit and vegetables, and opposite body weight. Since these data are cross-sectional, the question of cause and effect among weight, physical activity level, and composition of diet cannot be answered, only speculated upon. The observed relations are, however, fully compatible with what is expected from common sense.

Number of medications and use of amiodarone negatively correlated with warfarin dose in both components and show some correlation with age. These two variables were chosen to describe use of medications because previous studies indicated that they affected VKA dose [18]. Including other types of conventional and alternative medications did not influence the loading plots in any significant way (unpublished), which is why they have been omitted here, for the sake of clarity.

Similar loading plots were observed for patients who had an INR inside or outside the therapeutic range on the day of enrolment (compare Fig. 2a and b). This result might reflect that all participants were patients in long-term therapy, in whom the majority of fluctuations were small, or a limitation of the cross-sectional design. From the present data, we could not evaluate how large a percentage of time each patient spent within the therapeutic range.

The study is limited by the estimation of vitamin K intake based on food frequency questionnaires, which contain no information on serving size. We were therefore unable to directly calculate the amount of vitamin K ingested and have to rely on the arbitrary vitamin K score. We acknowledge that this semi-quantitative approach does not provide exact results, but we do not believe it is prone to systematic errors. The large majority of vitamin K intake in the present population came from vegetables such as broccoli and kale, while liver was consumed much less frequently. Since the study population was highly homogeneous regarding age, ethnicity, and socio-demographics (residents of a rural or small-town region on the Danish west coast), it may be speculated that the serving sizes of these vegetables showed little variation, hopefully reducing the impact of the limitations of the vitamin K score. The vitamin K score should be viewed as arbitrary, and the observed values cannot be extrapolated to other populations. Furthermore, we were unable to estimate the contribution from vitamin K produced by intestinal bacteria [26].

In the specific case of dietary vitamin K intake and warfarin dose, we conclude that the counter-intuitive negative (present data) or non-existing [14, 16] correlation could very well result from an overly simplistic approach to evaluating the problem at hand. In component two of the PLS model, we observed the expected relation, where a larger dietary intake of vitamin K correlates with a larger maintenance dose of warfarin. This relation is, however, complexly intercorrelated with body weight and physical activity, which in themselves influence warfarin dose [6, 27]. In light of recent studies [13], we foresee a renewed interest in vitamin K intake and vitamin K supplementation for anticoagulated patients. The present results suggest that it is not sufficient to evaluate dietary intake of vitamin K without taking other variables, such as physical activity level and body weight, into consideration. Future prospective studies should be directed at determining whether certain patterns of health-related behavior, as opposed to single indicators, are connected to the stability of anticoagulant treatment.

**Acknowledgements** The authors would like to thank Gunhild Andreasen, Anders Vestergaard Fournaise, Tanja Graff, Annette Larsen, Bodil Leed, Pernille Tandrup Nielsen, Asta Nørregaard and Katrine Overgaard for technical assistance. We are also very grateful for helpful suggestions and comments from Professor Rasmus Bro and consultant physician, Associate Professor Jørgen Gram.

**Financial support** The project was supported by Grant number 09–063088 from The Danish Council for Strategic Research.

## References

1. Pirmohamed M (2006) Warfarin: almost 60 years old and still causing problems. Br J Clin Pharmacol 62:509–511
2. Njaastad AM, Abildgaard U, Lassen JF (2006) Gains and losses of warfarin therapy as performed in an anticoagulation clinic. J Intern Med 259:296–304
3. Poller L (2004) International Normalized Ratios (INR): the first 20 years. J Thromb Haemost 2:849–860
4. Rombouts EK, Rosendaal FR, van der Meer FJ (2010) Influence of dietary vitamin K intake on subtherapeutic oral anticoagulant therapy. Br J Haematol 149:598–605
5. Sconce E, Avery P, Wynne H, Kamali F (2007) Vitamin K, supplementation can improve stability of anticoagulation for patients with unexplained variability in response to warfarin. Blood 109:2419–2423
6. Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT, Limdi NA, Page D, Roden DM, Wagner MJ, Caldwell MD, Johnson JA (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 360:753–764
7. Ford SK, Moll S (2008) Vitamin K supplementation to decrease variability of International Normalized Ratio in patients on vitamin K antagonists: a literature review. Curr Opin Hematol 15:504–508
8. Moridani M, Fu L, Selby R, Yun F, Sukovic T, Wong B, Cole DE (2006) Frequency of CYP2C9 polymorphisms affecting warfarin metabolism in a large anticoagulant clinic cohort. Clin Biochem 39:606–612
9. Van Booven D, Marsh S, McLeod H, Carrillo MW, SangKuhl K, Klein TE, Altman RB (2010) Cytochrome P450 2C9-CYP2C9. Pharmacogenet Genomics 20:277–281

10. Juurlink DN (2007) Drug interactions with warfarin: what clinicians need to know. CMAJ 177:369–371
11. Lubitz SA, Scott SA, Rothlauf EB, Agarwal A, Peter I, Doheny D, van der See S, Jaremko M, Yoo C, Desnick RJ, Halperin JL (2010) Comparative performance of gene-based warfarin dosing algorithms in a multiethnic population. J Thromb Haemost 8:1018–1026
12. Sconce E, Khan T, Mason J, Noble F, Wynne H, Kamali F (2005) Patients with unstable control have a poorer dietary intake of vitamin K compared to patients with stable control of anticoagulation. Thromb Haemost 93:872–875
13. Gebuis EP, Rosendaal FR, van Meegen E, van der Meer FJ (2011) Vitamin K1 supplementation to improve the stability of anticoagulation therapy with vitamin K antagonists: a dose-finding study. Haematologica 96:583–589
14. Gage BF, Eby C, Milligan PE, Banet GA, Duncan JR, McLeod HL (2004) Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. Thromb Haemost 91:87–94
15. Aquilante CL, Langaee TY, Lopez LM, Yarandi HM, Tromberg JS, Mohuczy D, Gaston KL, Waddell CD, Chirico MJ, Johnson JA (2006) Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. Clin Pharmacol Ther 79:291–302
16. Khan T, Wynne H, Wood P, Torrance A, Hankey C, Avery P, Kesteven P, Kamali F (2004) Dietary vitamin K influences intra-individual variability in anticoagulant response to warfarin. Br J Haematol 124:348–354
17. Poller L, Keown M, Ibrahim S, Lowe G, Moia M, Turpie AG, Roberts C, van der Besselaar AM, van der Meer FJ, Tripodi A, Palaretti G, Shiach C, Bryan S, Samama M, Burgess-Wilson M, Heagerty A, Maccallum P, Wright D, Jespersen J (2009) A multicentre randomised assessment of the DAWN AC computer-assisted oral anticoagulant dosage program. Thromb Haemost 101:487–494
18. Skov J, Bladbjerg EM, Sidelmann J, Vamosi M, Jespersen J. Plenty of pills: polypharmacy prevails in patients of a Danish anticoagulant clinic (2011) *Eur J Clin Pharmacol* doi:10.1007/s00228-011-1045-0 Online First™
19. Rozen S, Skaletsky HJ (2000) Primer 3 on the WWW for general uses and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, NJ, pp 365–386
20. Hatch E, Sconce EA, Daly AK, Kamali F (2006) A rapid genotyping method for the vitamin K epoxide reductase complex subunit 1 (VKORC1) gene. J Thromb Haemost 4:1158–1159
21. Sipeky C, Csongei V, Jaromi L, Safrany E, Polgar N, Lakner L, Szabo M, Takacs I, Melegh M (2009) Vitamin K epoxide reductase complex 1 (VKORC1) haplotypes in healthy Hungarian and Roma population samples. Pharmacogenomics 10:1025–1032
22. Bladbjerg EM, Gram J, Jespersen J, de Maat MP (2002) Internal quality control of PCR-based genotyping methods in research studies and patient diagnostics. Thromb Haemost 87:812–816
23. Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P (2003) International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 35:1381–1395
24. Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
25. Gage BF, Eby C, Johnson JA, Deych E, Rieder MJ, Ridker PM, Milligan PE, Grice G, Lenzini P, Rettie AE, Aquilante CL, Grosso L, Marsh S, Langaee T, Farnett LE, Voora D, Venstra DL, Glynn RJ, Barett A, McLeod HL (2008) Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clin Pharmacol Ther 84:326–331
26. Shearer MJ, Newman P (2008) Metabolism and cell biology of vitamin K. Thromb Haemost 100:530–547
27. Shibata Y, Hashimoto H, Kurata C, Ohno R, Kazui T, Takinami M (1998) Influence of physical activity on warfarin therapy. Thromb Haemost 80:203–204

# Paper III

Genetic, clinical and behavioural determinants of vitamin K-antagonist dose - Explored through multivariable modelling and visualization

Jane Skov, Else Bladbjerg, Morten A. Rasmussen, Johannes Sidelmann, Anja Leppin, Jørgen Jespersen

*Basic & Clinical Pharmacology & Toxicology,* ***110*** *(2012), 193-198*

BCPT Basic & Clinical Pharmacology & Toxicology

*Basic & Clinical Pharmacology & Toxicology*, **110**, 193–198 Doi: 10.1111/j.1742-7843.2011.00789.x

# Genetic, Clinical and Behavioural Determinants of Vitamin K-Antagonist Dose – Explored Through Multivariable Modelling and Visualization

**Jane Skov[1], Else-Marie Bladbjerg[1], Morten A. Rasmussen[3], Johannes J. Sidelmann[1], Anja Leppin[2] and Jørgen Jespersen[1]**

[1]Unit for Thrombosis Research, Institute of Public Health, University of Southern Denmark, Esbjerg, Denmark, [2]Unit for Health Promotion, Institute of Public Health, University of Southern Denmark, Esbjerg, Denmark, and [3]Department of Food Science/Quality and Technology, Faculty of Life Science, University of Copenhagen, Copenhagen, Denmark

(Received 1 June 2011; Accepted 11 August 2011)

*Abstract:* Vitamin K antagonists (VKA) are highly effective anticoagulants but their use is hampered by multiple interactions with food and medicine and a narrow therapeutic range. The large variation in dose requirements has led to the development of several dosing algorithms based on pharmacogenetic and clinical variables. In contrast, evidence about the influence of behavioural (i.e. diet and exercise) and socio-psychological factors is sparse. To investigate the impact of pharmacogenetic, clinical, behavioural and socio-psychological factors on maintenance dose of VKA. In a cross-sectional study, we interviewed 250 consecutive patients from an anticoagulant clinic and subsequently measured pharmacogenetic and anthropometric variables. Statistical analyses were carried out using linear regression and multivariable models with visualization features. In both types of analyses, the strongest determinants of VKA dose were polymorphisms in the *VKORC1* and *CYP2C9* genes and age. Half of the variation in VKA dose could be explained by a linear regression model including four variables, while a multivariable model with 20 pharmacogenetic and clinical variables explained 60%. A multivariable model including 94 predictor variables was not notably better regarding predictive performance, but visualization of this model offered information about the correlation structure between predictor variables. The strongest determinants of VKA dose are well-known pharmacogenetic variables and age. The variables describing health-related behaviour and socio-psychological factors are strongly inter-correlated and not useful in dosing algorithms.

Vitamin K antagonists (VKA) are effective and widely used oral anticoagulants [1]. In Denmark, warfarin and less commonly phenprocoumon are prescribed for the prevention of thromboembolic complications in patients with mechanical heart valves or atrial fibrillation (AF) combined with co-morbid conditions [2,3] or for preventing the recurrence of venous thromboembolic disorders. VKA treatment has to be monitored frequently and carefully [4], because fluctuations of the international normalized ratio (INR) [5] outside the designated range are strongly correlated with haemorrhagic or thromboembolic events [6–8]. On the population level, INR control [9,10] can be improved and the number of bleedings or thromboses reduced [11] by computer-assisted dosage of VKA in the setting of specialized anticoagulant clinics [12].

One of the serious challenges associated with the management of VKA treatment is a large inter-individual variation in maintenance dose. Much of the variability can be ascribed to genetic polymorphisms, importantly those in the *VKORC1* gene [13], coding for the enzyme inhibited by VKA, and the *CYP2C9* gene [14] coding for the cytochrome P450 enzyme mainly responsible for warfarin clearance. Furthermore, clinical factors such as age, weight and use of certain co-medications influence maintenance dose in a robust and predictable manner. Multiple dosing algorithms have been developed based on genetic and clinical information [15,16]. These algorithms are able to explain between a third and half of the variation in dose requirements [16], much more than algorithms based solely on clinical data [15,17].

In contrast to the abundant evidence for genetic and clinical influences [15], little is known about the impact of individual behavioural factors on VKA dose. For example, solid evidence regarding the impact of diet [17–19] and exercise [20,21] is sparse. Nevertheless, concerns about diet can play a prominent role in the everyday life of an anticoagulated patient, and a perception of life-style limitations caused by VKA might decrease the motivation to comply with treatment. Therefore, it is of high importance that clinicians are able to make recommendations regarding nutrition, physical activity, smoking and alcohol.

In this study, the influence of a wide range of genetic, anthropometric, behavioural and socio-psychological variables on VKA maintenance dose in patients from a Danish anticoagulant clinic was investigated. The non-genetic variables were expected to show strong inter-correlations in a complex manner. By employing not only a conventional

Address for correspondence: Jane Skov, Unit for Thrombosis Research, Department of Clinical Biochemistry, Hospital of South West Denmark, University of Southern Denmark, Finsensgade 35, DK-6700 Esbjerg, Denmark (fax +45 79182430, e-mail jskov@health.sdu.dk).

linear regression model, but also a multivariable approach with additional visualization features, we attempted to estimate the contribution from such variables. The study was aimed at generating knowledge that can promote individualization of VKA treatment.

## Materials and Methods

*Design and study participants.* The participants in this cross-sectional study were patients in the maintenance phase (duration of treatment >3 weeks at the time of inclusion) of treatment with a vitamin K antagonist (VKA) at the Department of Clinical Biochemistry at the Hospital of South West Denmark. At this nurse-managed and physician-supervised anticoagulation clinic, dosage of VKA is computer-assisted using the DAWN anticoagulation therapy software (4S Information Systems Ltd, Cumbria, UK) [22]. The patients attending the anticoagulant clinic were eligible for inclusion if they were above 18 years of age and provided written consent. Patients were excluded if treatment with heparin, warfarin or phenprocoumon was contraindicated or they were undergoing cancer chemotherapy. A total of 250 consecutive patients were enrolled between May 2009 and January 2010, while 117 eligible patients did not participate in the project. Characteristics of study participants are presented in table 1. The majority of the patients were treated with warfarin, but six patients took phenprocoumon. These were not found to be dissimilar from the remaining patients and were therefore included in subsequent analyses. The majority of the patients had an INR target of 2.5 (therapeutic range, 2.0–3.0), but 11 patients had a target of 3.0 (range, 2.5–3.5). On the day of inclusion, 37 patients (14.8%) had an INR below therapeutic range, while 46 (18.4%) had an INR above therapeutic range. The lowest INR was 1.1 and the highest 7.8.

The non-participants were older than the participants (median age 71 years *versus* 68 years, Mann–Whitney *U* test: $p < 0.01$) and more likely to have been in VKA treatment for more than six months (97% *versus* 76%, $\chi^2$-test: $p < 0.01$), but there were no gender differences between the two groups. Participation rates were lower for patients diagnosed with AF or mechanical heart valves than venous thrombotic disorders ($\chi^2$-test: $p = 0.027$).

In dealing with ethical aspects, the Declaration of Helsinki was followed throughout the study and approval was obtained from the regional Ethics Committee (project-ID S-20080020).

*Interview and assessment of clinical variables.* On the day of enrolment, an interview (see Data S1 for further details) using a standardised questionnaire was conducted. Following the interview, the patients' height, weight, waist- and hip-circumference, systolic and diastolic arterial blood pressure and heart rate were measured by the interviewer.

*Table 1.*
Baseline characteristics of study participants (n = 250).

| | | |
|---|---|---|
| Age–years[1] | 68 (59–75) | |
| Male sex | 65.2% | |
| Indication for VKA treatment | Atrial fibrillation | 56.8% |
| | Mechanical heart valve | 12.0% |
| | Deep Vein Thrombosis | 17.2% |
| | Pulmonary Embolism | 10.8% |
| | Other | 3.2% |
| Duration of VKA treatment at the time of enrolment | ≤3 months | 13.7% |
| | ≥12 months | 55.8% |
| BMI (kg/m$^2$)[2] | 29.0 (28.3–29.7) | |
| VKA dose[2] | 31.4 mg/week (29.4–33.5) | Range 5–117 mg/week |

VKA, vitamin K antagonists.
[1]Median (quartiles).
[2]Geometric mean (95% confidence interval).

*Blood sampling and analyses.* Five tubes of venous blood were sampled from each patient. Three samples were collected in tubes containing 0.5 ml 0.109 M sodium citrate (4.5 ml each) and two samples in tubes containing EDTA (one 5 ml and one 10 ml). One tube of citrated blood was used for the determination of INR, and the 5-ml tube of EDTA blood was used for the measurement of haemoglobin.

The tubes containing citrated blood were centrifuged within an hour at 2000 × *g* for 20 min. at 20°C, and the plasma was collected and stored at −80°C in portions of 350 µl. Citrated plasma was used for ELISA detection (Dako, Glostrup, Denmark) of von Willebrand factor antigen (vWF), ultra-sensitive detection of C-reactive protein (CRP) by nephelometry (Siemens Healthcare Diagnostics, Marburg, Germany) and measurement of D-dimer by a latex-based immunoassay using the STA-R equipment (Stago Diagnostica, Asniéres-sur-Seine, France).

The 10-ml tube containing EDTA blood was used for DNA analyses (see below).

*VKORC1 and CYP2C9 polymorphism detection.* Genomic DNA was extracted from peripheral blood cells following centrifugation, removal of plasma and lysis of erythrocytes.

*CYP2C9* genotyping was performed by PCR/RFLP methods for the detection of four SNPs rs1856908, rs9332113, rs1934968 and rs9332238 (see Data S1). A multiplex PCR/RFLP was used for the detection of the *2 (rs1799853) and *3 (rs1057910) allele variants [14].

Real-time PCR followed by melting curve analysis was performed on the Lightcycler Instrument 2.0 (Roche Diagnostics, Mannheim, Germany) for the detection of the *VKORC1* SNP 1173C/T (rs9934438) [23]. RFLP methods were used for the detection of the −1639 G/A (rs9923231), the 689 C/T (rs17708472) and the 9041 G/A (rs7294) SNP. We used the primer sequences and restriction enzymes reported by Sipeky *et al.* [24].

The SNPs tested were selected from international databases to cover as much as reasonably possible of the variation in the *CYP2C9* and the *VKORC1* genes. Two SNPs (rs9934438 and rs9923231) expected to be fully correlated were included as a measure of the reliability of our genotyping. We also re-analysed 5% of the samples (at random), and the results were confirmed. Our internal quality control programme for DNA analysis is described in detail elsewhere [25].

*Statistical analyses.* Statistical analyses were performed in SPSS version 18 (SPSS Inc., Chicago, IL, USA), and a significance level of 0.05 was used for all tests.

We inspected Q–Q plots to determine whether variables followed a normal distribution. Results are given as geometric mean (95% confidence interval) for lognormal variables or median (quartiles) for the remaining variables.

Chi-square-tests were used to assess deviations from Hardy–Weinberg equilibrium at the SNP sites tested. Differences among SNP groups were tested using the non-parametric Kruskal–Wallis test.

To sort the variables according to importance for the estimation of therapeutic VKA dose, we used forward stepwise regression, which successively includes variables contributing the most to the predictive performance. The stop criterion was estimated by tenfold cross-validation. Because of the explorative design of the study, no power analyses were carried out.

*Multivariable pattern analysis.* We used 94 variables for the initial multivariable analyses. VKA treatment was described by weekly dose (used as the outcome variable), INR, indication [five groups: AF or mechanical heart valves, deep vein thrombosis (DVT) or pulmonary embolism (PE) and 'other'], use of warfarin or phenprocoumon, INR target and duration of treatment at the time of inclusion. Additionally, a dichotomous variable determined whether the indication

for VKA treatment was 'arterial' (AF and mechanical heart valve) or 'venous' (DVT and PE). The clinical variables entered were as follows: age, sex, BMI, height, weight, waist circumference, hip circumference, waist-to-hip ratio, heart rate, systolic and diastolic blood pressure. The number of medications taken by each participant was entered, so was a dichotomous variable defining whether the patient took a specific medication or not. Compliance was described by a dichotomous variable (had the patient missed one or more doses or not), so was polypharmacy (use of five or more preparations [26]). Use or non-use of each category of the most common dietary supplements was defined by six dichotomous variables. Dichotomous variables indicated the presence of each $CHADS_2$ risk factor except age.

Patients were defined as current, former or non-smokers, and for the current smokers, the daily tobacco consumption was entered. Three variables described alcohol intake: Alcohol consumption frequency during the last 12 months was scored from 1 = 'never' to 8 = 'almost daily' and the estimate of units of alcohol during the last 4 weeks was entered. A dichotomous variable indicated whether the participant had consumed more than three units of alcohol at one occasion. Two variables were included for physical activities: the number of MET minutes calculated from the short form International Physical Activity Questionnaire, and the overall classification of spare time activities scored from 1 = sedentary to 4 = competitive sports. The food items were divided into seven groups: grains and starches, fruit and vegetables, meat, dairy products, sources of fat, sources of sugar and drinks. For each food group, a principal components analysis (PCA) was conducted. This is an explorative data compression technique that can be used to decompose the variation in a large data matrix into fewer underlying *latent variables*, also called principal components (PCs). The data matrix is divided into so-called *scores*, relating to the samples, *loadings*, relating to variables, and *residuals*, the unsystematic variation not explained by the model. The loadings can be plotted in a scatter diagram, thereby illustrating the grouping patterns of variables. From each food group, two PCs were extracted. The above principles are illustrated in fig. 1, which is a loading plot of the first two PCs of the group 'drinks'. Variables placed close to each other (in either the horizontal or the vertical direction in PC one or PC two, respectively) are positively correlated, while a long distance implies a negative correlation. Thus, in fig. 1, the first PC (shown on the *x*-axis) is spanned by diet drinks and coffee in one direction and tea in the opposite one. The second PC (*y*-axis) is spanned by water as one extreme and diet drinks as the other.

The resulting 14 components (seven food groups, two PCs from each) were used as a condensed measure of the 80 primary food variables. In addition, vitamin K intake during the previous month was assessed by the vitamin K score (see Data S1).

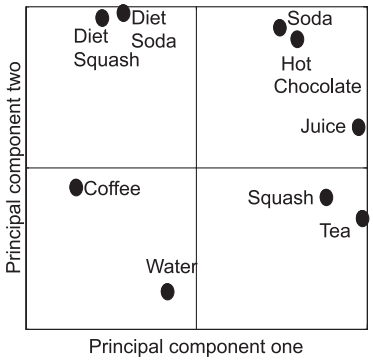



Fig. 1. Loading plot of principal component one *versus* principal component two for the principal components analysis of the food group 'drinks'.

The three socio-demographic variables described whether the patient lived alone or not, his or her level of education (below or above 7 years of primary school) and whether they worked or had retired. Four subjective indicators of well-being were included: The patients' answer to question 1 of the SF-12-v1 questionnaire ('In general, would you say your health is excellent, very good, good, fair or poor?'), the SF-12 Physical Health Composite Scale Score, SF-12 Mental Health Composite Scale Score and the total score on the Perceived Stress Scale.

Finally, we entered haemoglobin, plasma concentrations of vWF, CRP and D-dimer, and all SNPs tested. For the latter, the number of variant alleles was considered additive and the SNP genotypes were coded 1 (wild type), 2 (one variant allele) and 3 (two variant alleles).

We used partial least squares regression (PLS) to estimate the therapeutic VKA dose from the above groups of independent variables. PLS is, similarly to PCA, a technique that decomposes the data into a few latent variables. Similarly to PCA, the data matrix is divided into *scores*, *loadings* and *residuals*. Unlike PCA, in this type of model, the contribution of a single variable to the response is not estimated as it is in conventional regression models. Instead, the number of significant components is found, to which each variable can contribute more or less. By comprising the data matrix into a few latent structures (components), PLS models show great robustness against such problems as multicollinearity between variables. In this study, the models were evaluated as root mean squared error of cross-validation to find the optimal number of components. In addition to predictive performance, PLS models offer visualization features. In Scatter plots of loadings (so-called loading plots), the correlation structure between variables is illuminated.

Additionally, a 'minimal' PLS model with only genetic polymorphisms and selected clinical variables (20 predictor variables) was built as a reference model for more direct comparison between the performance of a conventional linear and a multivariable model.

## Results

*Genetic polymorphisms.*

A vast majority (N = 245) of the study participants reported their ethnicity as 'Danish'. The remaining five were as follows: 'German' (N = 1), 'Faroe Islander' (N = 2), 'Sri Lankan' (N = 1) and 'Italian' (N = 1).

No deviations from Hardy–Weinberg equilibrium were detected at the SNP sites tested, and the genotypes for *VKORC1* polymorphisms rs9934438 and rs9923231 were 100% correlated. We found that VKA dose varied significantly among genotype groups for the rs9934438/rs923231, rs17708472, rs7294 (all *VKORC1*), rs1856908, rs1799853 and rs1057910 (*CYP2C9*) SNPs (table S1).

*Linear regression analysis.*

The forward stepwise linear regression model estimating VKA dose included four variables, three of them SNPs. The first variable to enter the regression was the *VKORC1* polymorphism rs9934438 or rs9923231. The second variable to enter the model was age, and finally, the two *CYP2C9* SNPs were rs1057910 (*3) and rs1799853 (*2). The total adjusted $R^2$ for this model was 50% (table 2).

*Multivariable pattern analyses.*

The PLS model had two statistically significant latent variables or components. *Component one* explains 51% of the variation in VKA dose, while the first two components in

*Table 2.*
Stepwise linear regression model of VKA dose, forward selection.

| Variable | Estimated effect on VKA dose (confidence interval) | Adjusted cumulative $R^2$ | *p*-value |
|---|---|---|---|
| rs9934438 (*VKORC1*) | −0.25 (−0.20 to −0.29) | 0.28 | <0.01 |
| Age | −0.18 (−0.13 to −0.22) | 0.40 | <0.01 |
| rs1057910 (*CYP2C9*) | −0.16 (−0.12 to −0.21) | 0.48 | <0.01 |
| rs1799853 (*CYP2C9*) | −0.10 (−0.05 to −0.14) | 0.50 | <0.01 |

VKA, vitamin K antagonists.

combination explain a total of 64%. Loading plots of *component one versus component two* are shown in fig. 2. The main variation in VKA dose is described by the variables spanning the *x*-axis or *component one* of the model. Furthest apart from VKA dose in this direction are the two correlated *VKORC1* polymorphisms rs9934438 and rs9923231 and age. Age is clustered closely with variables such as retirement, hypertension, polypharmacy, betablockers and numbers of medications taken. In the opposite direction, on the *x*-axis is a group of closely related variables including weight, BMI, waist circumference and hip circumference. Nearby, 'drinks 2' is found, reflecting the consumption of diet sodas (fig. 1). *Component two* or the *y*-axis is spanned by SNPs and the variables reflecting body weight in one direction and those reflecting age in the other.

Figure 3 is a loading plot similar to fig. 2, based on a model with only 20 predictor variables. The two loading plots are very similar regarding variables spanning the two components. A notable difference is that age is placed on the opposite side of VKA dose in fig. 3 in both *component one* and *component two*. *Component one* explains 52%, and the first two components combined explain 60% of the total variation in VKA dose.

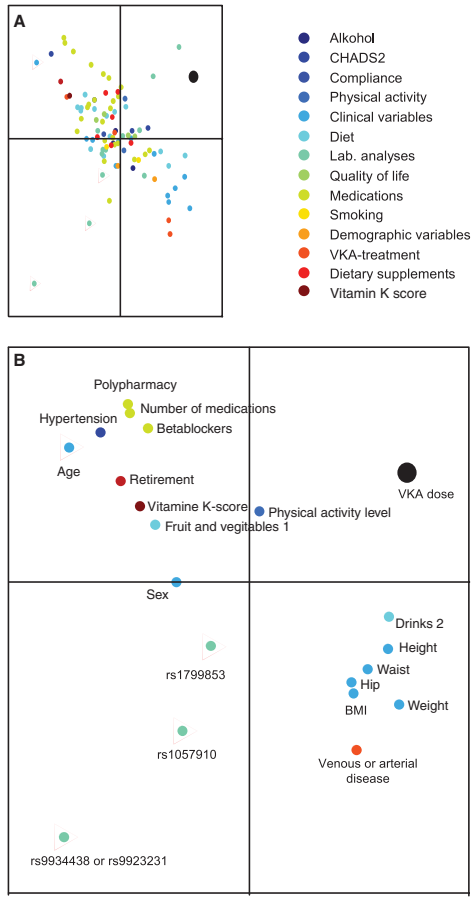



Fig. 2. Loading plot of component one *versus* component two of the partial least square regression model with 94 predictor variables and vitamin K antagonists dose as dependent variable. The different groups of variables are coded by colour and the triangles mark the variables selected by the linear regression model. (A) Illustrates the loadings of all predictor variables, including those making minimal contributions. (B) The variables contributing the most plus some of high interest for interpretation are included, all marked by a legend.

## Discussion

For the present study, a wide range of pharmacogenetic, anthropometric, behavioural and psycho-sociological variables were estimated in patients from a Danish anticoagulant clinic. The influence of these variables on VKA dose was investigated using two different statistical models, linear regression with forward selection and partial least square regression.

In agreement with a large body of similar studies [15,16], we found the rs9923231 or rs9934438 polymorphism in *VKORC1*, age and the *2 and *3 *CYP2C9* genotypes to be the most important predictors of VKA dose in the linear regression model. Possession of variant SNP alleles was negatively associated with VKA dose, so was higher age.

A number of variables were not selected by our linear regression model, even though they have been found to influence VKA dose in previous reports and are included in published dosing algorithms. These include weight (or the related variable BMI or body surface area), sex, smoking status and venous thrombosis as the indication for VKA treatment [18,27]. The quantitatively very small contribution from the two latter variables [27] offers a good explanation for their absence in our model. The PLS model provides a more detailed understanding of the somewhat ambiguous role played by weight. We were also intrigued by the absence of amiodarone in the linear regression model, because this medication is known to strongly influence VKA maintenance dose [15]. When directly comparing amiodarone users and non-users, we found significant differences in VKA dose [26]. Only 21 of the study participants used amiodarone, and this example illuminates that even though amiodarone use will profoundly affect an individual's dose requirement, it is

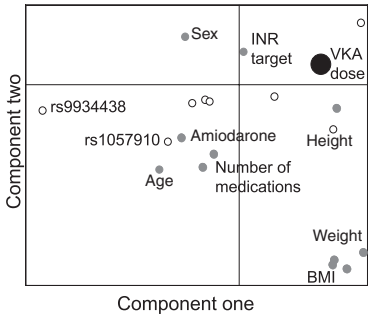



Fig. 3. Loading plot of component one *versus* component two of the 'minimal' partial least square regression model, only including pharmacogenetic and selected clinical variables. The open circles represent genetic polymorphisms, the large black circle vitamin K antagonists dose and the grey circles the clinical variables.

not among the most important sources of variation on the population level.

By employing a multivariable model to explore the variation in VKA dose, we found the same variables to be most influential as we did by the simpler linear regression. The first two components of the full PLS model explained a substantially greater part of the variation in VKA dose (64%) than the linear regression model (50%). The minimal PLS model had fewer predictor variables but was about as good at explaining the variation in VKA dose (60%) as the full model. The small difference in percentage of VKA dose variation explained between the two PLS models is shared by more than 70 variables. Therefore, among the variables describing health-related behavioural and socio-psychological factors, there is not one strong determinant of VKA dose. As the full model with *more* variable groups included was *not* significantly poorer, the added information is only redundant and not necessarily irrelevant. Hence, the full model adds knowledge concerning the entire system but does not contribute to much additional predictive performance.

The visualization of the full model may explain the somewhat surprising role played by weight in determining VKA dose. We expect heavier patients to require a higher VKA dose, which is also clearly observed in *component one* of our PLS models. In *component two*, however, weight is negatively correlated to VKA dose, which is positively correlated to intake of vitamin K-rich vegetables and physical activity. It has previously been suggested that more active patients require a higher VKA dose [20,21]. Intuitively, one would expect a large dietary intake of vitamin K to correlate with a large dose requirement, but our analyses did not identify vitamin K score as an important predictor of VKA dose. Similar weak or non-existing associations have been reported in other studies [17–19]. All in all, it seems that behavioural markers (intake of vegetables, physical activity) and weight are complexly correlated and that these relations 'overrule' the effect of weight in *component two*.

The study is limited by the fact that all data describing diet were estimated from food frequency questionnaires. We are, therefore, unable to make any firm conclusions about overall caloric intake or the distribution of macro- or micronutrients, and the vitamin K score is an arbitrary estimation of vitamin K intake. As the population showed little variation in terms of age and ethnicity, the impact of this error is limited.

In conclusion, the present report suggests that health-related behaviour, socio-demographic factors and subjective health perceptions have very limited influence on VKA dose. For dosing algorithms, pharmacogenetic information is of the absolutely highest importance. Even though our results indicate that anticoagulated patients should not receive any different advice regarding nutrition and physical activities than other individuals, the possible impact of health-related behaviour on INR fluctuations must not be forgotten. The focus of future studies should therefore be directed at the possible impact of behaviour, socio-demographic factors and subjective health perceptions on the intra-individual stability of dose and response to VKA treatment.

Acknowledgements

The authors would like to thank Gunhild Andreasen, Anders Vestergaard Fournaise, Tanja Graff, Anette Larsen, Bodil Leed, Pernille Tandrup Nielsen, Asta Nørregård and Kathrine Overgaard for technical assistance and Jørgen Gram for helpful comments on an earlier draft of the manuscript.

Funding

This project was supported by grant number 09-063088 from the Danish Council for Strategic Research.

## References

1 Pirmohamed M. Warfarin: almost 60 years old and still causing problems. Br J Clin Pharmacol 2006;**62**:509–11.
2 Gage BF, van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode BS *et al.* Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation 2004;**110**:2287–92.
3 Camm AJ, Kirchhof P, Lip GY, Schotten U, Savelieva I, Ernst S *et al.* Guidelines for the management of atrial fibrillation: the Task Force for the Management of Atrial Fibrillation of the European Society of Cardiology (ESC). Eur Heart J 2010;**31**:2369–429.
4 Ansell J, Hirsh J, Hylek E, Jacobson A, Crowther M, Palareti G. Pharmacology and management of the vitamin K antagonists: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest 2008;**133**:160S–98S.
5 Poller L. International Normalized Ratios (INR): the first 20 years. J Thromb Haemost 2004;**2**:849–60.
6 Oake N, Jennings A, Forster AJ, Fergusson D, Doucette S, van Walraven C. Anticoagulation intensity and outcomes among patients prescribed oral anticoagulant therapy: a systematic review and meta-analysis. CMAJ 2008;**179**:235–44.
7 Hylek EM, Chang YC, Skates SJ, Hughes RA, Singer DE. Prospective study of the outcomes of ambulatory patients with excessive warfarin anticoagulation. Arch Intern Med 2000;**160**:1612–7.

8 Merli GJ, Tzanis G. Warfarin: what are the clinical implications of an out-of-range-therapeutic international normalized ratio? J Thromb Thrombolysis 2009;**27**:293–9.
9 Manotti C, Moia M, Palareti G, Pengo V, Ria L, Dettori AG. Effect of computer-aided management on the quality of treatment in anticoagulated patients: a prospective, randomized, multicenter trial of APROAT (Automated PRogram for Oral Anticoagulant Treatment). Haematologica 2001;**86**:1060–70.
10 Poller L, Shiach CR, MacCallum PK, Johansen AM, Munster AM, Magalhaes A *et al.* Multicentre randomised study of computerised anticoagulant dosage. European Concerted Action on Anticoagulation. Lancet 1998;**352**:1505–9.
11 Poller L, Keown M, Ibrahim S, Lowe G, Moia M, Turpie AG *et al.* An international multicenter randomized study of computer-assisted oral anticoagulant dosage vs. medical staff dosage. J Thromb Haemost 2008;**6**:935–43.
12 Njaastad AM, Abildgaard U, Lassen JF. Gains and losses of warfarin therapy as performed in an anticoagulation clinic. J Intern Med 2006;**259**:296–304.
13 Geisen C, Watzka M, Sittinger K, Steffens M, Daugela L, Seifried E *et al.* VKORC1 haplotypes and their impact on the inter-individual and inter-ethnical variability of oral anticoagulation. Thromb Haemost 2005;**94**:773–9.
14 Moridani M, Fu L, Selby R, Yun F, Sukovic T, Wong B *et al.* Frequency of CYP2C9 polymorphisms affecting warfarin metabolism in a large anticoagulant clinic cohort. Clin Biochem 2006;**39**:606–12.
15 Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT *et al.* Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 2009;**360**:753–64.
16 Lubitz SA, Scott SA, Rothlauf EB, Agarwal A, Peter I, Doheny D *et al.* Comparative performance of gene-based warfarin dosing algorithms in a multiethnic population. J Thromb Haemost 2010;**8**:1018–26.
17 Gage BF, Eby C, Milligan PE, Banet GA, Duncan JR, McLeod HL. Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. Thromb Haemost 2004;**91**:87–94.
18 Aquilante CL, Langaee TY, Lopez LM, Yarandi HN, Tromberg JS, Mohuczy D *et al.* Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. Clin Pharmacol Ther 2006;**79**:291–302.
19 Khan T, Wynne H, Wood P, Torrance A, Hankey C, Avery P *et al.* Dietary vitamin K influences intra-individual variability in anticoagulant response to warfarin. Br J Haematol 2004;**124**:348–54.
20 Shibata Y, Hashimoto H, Kurata C, Ohno R, Kazui T, Takinami M. Influence of physical activity on warfarin therapy. Thromb Haemost 1998;**80**:203–4.
21 Wells PS, Majeed H, Kassem S, Langlois N, Gin B, Clermont J *et al.* A regression model to predict warfarin dose from clinical variables and polymorphisms in CYP2C9, CYP4F2, and VKORC1: derivation in a sample with predominantly a history of venous thromboembolism. Thromb Res 2010;**125**:e259–64.
22 Poller L, Keown M, Ibrahim S, Lowe G, Moia M, Turpie AG *et al.* A multicentre randomised assessment of the DAWN AC computer-assisted oral anticoagulant dosage program. Thromb Haemost 2009;**101**:487–94.
23 Hatch E, Sconce EA, Daly AK, Kamali F. A rapid genotyping method for the vitamin K epoxide reductase complex subunit 1 (VKORC1) gene. J Thromb Haemost 2006;**4**:1158–9.
24 Sipeky C, Csongei V, Jaromi L, Safrany E, Polgar N, Lakner L *et al.* Vitamin K epoxide reductase complex 1 (VKORC1) haplotypes in healthy Hungarian and Roma population samples. Pharmacogenomics 2009;**10**:1025–32.
25 Bladbjerg EM, Gram J, Jespersen J, de Maat MP. Internal quality control of PCR-based genotyping methods in research studies and patient diagnostics. Thromb Haemost 2002;**87**:812–6.
26 Skov J, Bladbjerg EM, Sidelmann JJ, Vamosi M, Jespersen J. Plenty of pills: polypharmacy prevails in patients of a Danish anticoagulant clinic. Eur J Clin Pharmacol 2011. Epub ahead of print. DOI: 10.1007/s00228-011-1045-0
27 Gage BF, Eby C, Johnson JA, Deych E, Rieder MJ, Ridker PM *et al.* Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clin Pharmacol Ther 2008;**84**:326–31.

## Supporting Information

Additional Supporting information may be found in the online version of this article:

**Data S1.** Supplementary methods.

**Table S1.** VKA dose across SNP groups.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

# Paper IV

Socio-demographic characteristics and food habits of organic consumers: A study from the Danish National Birth Cohort

Sesilje B. Petersen, Morten A. Rasmussen, Marin Strøm, Thorhallur I. Halldorsson, Sjurdur F. Olsen

*Submitted for Public Health Nutrition*

# Socio-demographic characteristics and food habits of organic consumers – A study from the Danish National Birth Cohort

***Sesilje B Petersen*[1], *Morten A Rasmussen*[2], *Marin Strøm*[1], *Thorhallur I Halldorsson*[1,3], and *Sjurdur F Olsen*[1,4]**

[1] Centre for Fetal Programming, Dept. of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark.

[2] Department of Food Science/Quality and Technology, Faculty of Life Sciences, Copenhagen, Denmark

[3] Unit for Nutrition Research, Faculty of Food Science and Nutrition, School of Health Sciences, University of Iceland

[4] Department of Nutrition, Harvard School of Public Health. Boston, Massachusetts, USA.

**Corresponding author**: Sesilje Bondo Petersen, Statens Serum Institut, Artillerivej 5, building 206, room 306, DK-2300 Copenhagen S, Denmark, Tel: +45 32 68 83 50, Fax: +45 32 68 31 65, E-mail: ssp@ssi.dk

**Running title**: Characterisation of organic food consumers
**Length of manuscript**: 3450 words (excluding acknowledgement, references, tables and figures)
**Number of tables**: 4
**Number of figures**: 1
Abbreviations: OF: organic food, DNBC: Danish National Birth Cohort, FFQ: food frequency questionnaire

**Keywords**: organic food, maternal diet, epidemiology,

**Conflicts of interest**: S.B. Petersen, M.A. Rasmussen, M. Strøm, T.I. Halldorsson, S.F. Olsen, no conflicts of interest.

**Abstract**

**Objective:** To develop a basis for building models that can examine impact of organic food (OF) choices on maternal and offspring health, including identification of factors associated with OF consumption, and underlying dietary patterns.

**Design:** Dietary intake was collected for the preceding month from a food frequency questionnaire (FFQ) in mid-pregnancy and information on socio-demographic characteristics was collected from telephone interviews during pregnancy. From a question about OF consumption in the FFQ, including six food categories, an OF preference index was calculated. Latent variables that captured the variability in OF choices in relation to dietary intake were defined.

**Setting:** The Danish National Birth Cohort (DNBC), 1996 to 2002.

**Subjects:** 60.773 pregnant women from DNBC.

**Results:** We found that frequent OF use was highly associated with age, occupational status, urbanization, smoking and vegetarianism. By principal component analysis we identified two eating patterns – 'Western dietary pattern' and 'Prudent dietary pattern' that explained 14·2 % of the variability in data. Frequent OF users consumed a more 'prudent' diet compared to non-users and had a significantly higher intake of vegetables (+67 %), fibre (+13 %), n-3 fatty acids (+11 %) and less saturated fat (-8 %).

**Conclusion:** Frequent OF users seemed to have a healthier lifestyle compared to non users. These findings highlight a major challenge in observational studies examining impact of OF consumption on health due to potentially irremediable confounding factors.

**Introduction**

During the past decade the demand for organic food (OF) products has grown considerably and various alternatives to the mainstream conventional food products and distribution have been developed and are increasing, especially in developed countries [1]. It is difficult to identify reasons for this growth as studies show that OF consumption reflects a complex web of determinants including its availability in food stores, socio-demographic and cultural factors, and personal values and attitudes. Studies, however, also reveal that health considerations are a major factor behind the growth in OF consumption [2-12]. This is noteworthy because a causal association between OF consumption and better health remains to be scientifically established [13-18]. Several types of studies have been conducted to address the question, but with few or unclear answers for various reasons: the biological and chemical studies in the field are often incomparable, as chemical composition of crops is easily affected by temperature, soil and sort [14;15;17;18]; few comparable animal studies exist, and some may not be relevant for today's farming practice [13;17]; and very few intervention studies or prospective observational studies have addressed the potential health benefits of OF products [16;19-21].

In relation to health effects of OF consumption the observational study design is complicated by the role of potential confounding factors. Thus, underlying determinants of OF purchase need to be described and included in statistical models in order to minimize potential confounding. The Danish National Birth Cohort (DNBC) [22;23] is suitable for addressing this issue as it is one of the largest prospective cohort studies worldwide to have recorded OF preferences during pregnancy along with a multitude of other dietary-, health- and socio-economic factors. Furthermore, Denmark has a strong system for control and certification of organic farming and manufacturing, a matured organic market, and the highest consumption of OF per capita in Europe [24]. Therefore, the DNBC offers unique opportunities to study the impact of OF in pregnancy on maternal and offspring health.

The aim of this study was to develop a basis for building models that can examine the impact of OF choices on maternal and offspring health among Danish pregnant women. The first step was to identify factors associated with OF consumption. The second step was to employ multivariate methods to identify underlying patterns and to define latent variables that can capture the variability in OF choices among the study population.

**Materials and methods**

***The Danish National Birth Cohort (DNBC)***

The DNBC is a cohort with information from 100.000 pregnancies [22]. Women were recruited between 1996 and 2002 during the first antenatal visit to the general practitioner around weeks 6-10 of gestation. The data collection in the study included four telephone interviews (two prenatal interviews conducted in gestational week 12 and 30 and two postnatal interviews when the child was 6 and 18 months old) and a semi-quantitative food frequency questionnaire (FFQ) mailed to the women in week 25 of gestation [22;23]. It was estimated that during the study period approximately 35 % of all deliveries in Denmark were included in the cohort [22]. Some of the women are registered in the cohort twice or more through subsequent pregnancies during the recruitment period. However, in our study, only the first pregnancy for each participating woman was included to avoid inter-correlation within subjects. Furthermore, multiple pregnancies were excluded from the study. The DNBC complies with the Declaration of Helsinki and was approved by the Danish National Committee on Biomedical Research Ethics.

***Definition of organic food preference index***

Information on OF consumption was collected through the FFQ [23]. About 70 % of the women returned the questionnaire, which was a modified form of a questionnaire used by the Danish Cancer Registry [25]. Of the 100.000 women in DNBC, 60.773 met the inclusion criteria and had answered the question about OF consumption in the FFQ.

The FFQ covered the pregnant woman's diet during the preceding four weeks and included one question about organic food consumption: "How often do you eat organic foods?". The question was divided in six categories: milk products, cereals, egg, vegetables, fruit and meat. The answer categories were never, sometimes, regularly or always. Based on the question we calculated an organic food preference index. Each answer category was given a score (never = 1, sometimes = 2, regularly = 3, always = 4) and summarized across categories to form an organic index (OX). Missing values were characterized as "never eating organic". Due to the fact that vegetarians and vegans in general have no consumption of meat and vegans have no consumption of meat, egg and milk products the meat category was excluded for both groups and the egg and milk categories were excluded for vegans. Based on the constructed index the women were segmented into four groups: non users (OX= 6), low users ($6 < OX \leq 12$), moderate users ($12 < OX \geq 18$) and frequent users ($OX > 18$).

***Assessment of dietary patterns***

In the FFQ the women were asked about frequencies for approximately 360 different items of foods and beverages. To estimate food intake standard portion sizes and standard recipes were applied for

all items in the questionnaire. Standard portion sizes were multiplied with the daily frequencies to estimate intake of each food item in grams. For more complex items standard recipes were made and the intakes of foods present in different items were aggregated. The estimated amount of all items were coupled with the Danish Food Tables [26] and corrected for loss of fat, water, vitamins and minerals. The 360 different items were divided into 35 main food groups and 65 more specific food groups representing the entire diet of the women. All nutrients were energy adjusted by the residual method, as described by Willett et al. [27]. Only women with an energy intake > 4500 kJ and < 20.000 kJ per day were included in the analyses for dietary intakes to avoid unrealistic estimates. The FFQ used in the DNBC has been validated in a group of younger non-pregnant women [28] and in the DNBC for intake of fruit, vegetables and pregnancy relevant nutrients (folate, protein, retinol and n-3 fatty acids) by a 7-day weighed food diary and biomarkers [29;30].

***Assessment of socio-demographic and lifestyle factors***

Socio-demographic and lifestyle variables were gathered from the consent form, the FFQ and the telephone interviews and include age ( < 20, 20-24, 25-29, 30-34, 35-39, ≥ 40), parity (0, 1, 2, 3+), occupational status (high-level proficiencies, medium-level proficiencies, skilled, student, unskilled, unemployed), cohabitation status (single, couple/married), urbanization (capital city, capital suburbs, 100.000+, 10.000 – 99.999, < 10.000 citizens), smoking during pregnancy (non-smoker, occasional smoker, < 15 cigarettes per day, ≥ 15 cigarettes per day), alcohol intake in pregnancy (not at all, yes), energy intake (in quintiles), physical activity (no, light, moderate, high level), intake of dietary supplements in the pregnancy (no, yes), maternal pre-pregnant body mass index (BMI; < 18·5, 18·5-25, 25-30, 30-35, 35+), living area in Denmark (West or East part of the country) and vegetarianism (yes, no). Most of these variables have been described and used in earlier studies based on the DNBC [31-40].

***Statistical analysis***

Univariate and multivariate logistic regression was used to estimate the association between OF consumption and socio-demographic characteristics and linear regression was used to analyze differences in dietary intake between non and frequent OF users. In these analyses the focus was on the differences between non users and frequent users in order to obtain the biggest contrast between OF consumers. All analyses were performed with SAS software, version 9.1 (SAS Institute Inc., Cary, North Carolina).

Principal component analyses (PCA) were used for exploration of the associations between the 65 food groups. PCA is conducted to uncover the systematic correlation structure between variables while excluding the non systematic variation. PCA is a widely used method for compression of

large datasets, often with a large number of variables, into a few underlying latent variables (principal components), describing the systematic variation. The number of variables in the present work is manageable from a univariate point of view, and PCA is hence not applied as a *second to none* alternative compared to the univariate analysis. PCA in combination with visualization reveals the inter variable correlation structure and hence adds a dimension on top of what can be explored from univariate analysis [41]. PCA models were implemented in MATLAB version 7.9.0.529 (R2009b) using PLS toolbox version 5.2.2. (Eigenvector Research inc.) and in-house algorithms for plotting of results.

**Results**

The responses to the OF question are shown in table 1. The frequencies of organic consumption through the six food categories differed substantially. The consumption of organic eggs and milk was common among the women, whereas the intake of organic cereals, vegetables, fruit and meat was low. According to the constructed OF preference-index, 12 % of the study population were classified as 'non users', 44 % were 'low users', 37 % 'moderate users' and 7 % were 'frequent users'.

Table 2 shows associations between OF preferences and socio-demographic and lifestyle characteristics as odds ratios (OR) for being a frequent user as opposed to a non user. The crude ORs are based on univariate analyses, whereas the adjusted ORs are based on a multivariate analysis showing the associations with OF preferences for each explaining variable independently of all the other variables in the table. The women's age had a strong and independent association with OF preferences and so did vegetarianism. Social group, smoking, BMI, physical activity, living area and urbanization were all associated with OF preferences in the crude analysis; however, adjusting for other factors attenuated their associations with OF preferences, though living area and urbanization still had strong associations. Regarding cohabitation status, adjustment tended to strengthen the association with OF preferences. For alcohol intake, use of dietary supplements, physical activity and occupational status the association with OF preferences was eliminated upon adjustment for the other covariates.

Table 3 shows daily intakes of food items according to OF preferences and crude and adjusted increments in intake for frequent users compared to non users (reference group) of OF. Intakes differed significantly across OF preference for almost all foods and food groups. The most marked differences in intake were observed for vegetables, legumes, fruit and berries, nuts, lamb, seafood, plant oils and tea; all with higher intakes for frequent OF consumers. Adjustment for covariates attenuated the observed differences and reversed the association between OF preferences and the intake of alcohol and desserts (candy, ice-cream and cakes). Compared with non users, frequent users seemed to substitute certain items with others, e.g. margarines with oils, white bread with dark bread, pork with poultry, lamb and fish, coffee with tea and soft drinks with water and juice.

Table 4 shows daily intakes of specified nutrients according to OF preferences. Nearly all comparisons between frequent users and non users were statistically significant. The most marked differences were observed for n-3 fatty acids, fibre, iodine, beta-carotene, folate and vitamin D, K, and C – which were higher among frequent users – and saturated fatty acids (SFA), monounsaturated fatty acids (MUFA), n-6 fatty and trans fatty acids, cholesterol and retinol – which

were lower among frequent users. Adjustment attenuated the differences, however significantly higher intakes of certain nutrients were still observed.

In figure 1 results from the principal component analysis are shown. Information concerning distribution of samples and correlation structure of variables are highlighted. The sample/women distribution in relation to degree of organic consumption is exploited as four ellipsoids, one for each organic consumption group. The centre and half axis correspond to the mean, and standard deviation of the group, with respect to principal component one (PC1) and two (PC2), and cover approximately 47 % of the data in each group. Though highly significant differences for almost every food group (see table 3) a fairly large overlap between distributions is observed. This indicates that the statistical significant differences are driven by large number of persons and not large deviations between groups. Inter food group correlation is shown as a scatter plot of the first two principal components for the food groups. The food groups are colored according to common food classes. When two variables are positioned close to each other they are correlated with respect to the variance explained by the two components.

From the PCA (figure 1) two distinct eating patterns, describing 14·2 % of the total variation in data, can be derived. PC1 is associated with a dietary pattern comprising more vegetables, cabbage, root, legumes, fish etc. as these food groups obtain high positive values in PC1. We named this component the “prudent dietary pattern”. PC2 is characterized by a high intake of pork, mixed/processed meat, white bread, margarine, French fries etc, as these food groups obtain high positive values in PC2. This component we named the “Western dietary pattern”. The positions of the four distributions in relation to organic consumption (the ellipsoids in figure 1) indicate that the “prudent dietary pattern” is positively correlated with OF consumption, whereas the “Western dietary pattern” is negatively correlated to OF consumption. Nevertheless, these two components are not correlated but orthogonal, implying that Western dietary pattern is not the opposite of prudent dietary pattern.

**Discussion**

In the present study we found that OF use was an eating habit related to higher social classes and healthier lifestyle and diet – all characteristics that predispose OF users to lower risks of chronic diseases, that may affect foetal health during pregnancy. Thus, the study illustrates the major confounder problem that faces researchers who are seeking to tease out the relationship between OF consumption and health outcomes.

Very few studies have compared the diet of non and frequent OF users and in general they are of poor quality. However, there seems to be an overall tendency towards higher intake of fruit and vegetables and lower intake of meat among OF users [42-44], consistent with our findings. It has been argued that the healthier diet observed among frequent OF users can be explained by differences in food supply and prices [45]. Frequent users have a higher propensity to purchase OF products from speciality shops, but also from direct sales channels such as farm gates, box schemes, street stalls in urban areas, etc, that may affect OF product availability and consumption [46]. In the beginning of 2000 the availability of organic foods in supermarkets and discount stores was lower than today [47], especially for organic meat, which can explain the lower OF intake for that product category.

A higher number of vegetarians among frequent OF users might in turn explain the lower intake of meat among frequent OF users in our study. However, stratification by vegetarianism showed a significantly lower adjusted intake of total meat (data not shown) among non-vegetarian OF users compared to non-vegetarian non OF users. This finding is supported by a recent survey among 515 Danish consumers which showed that the highest quartile in relation to organic preferences consumed 50 % less meat compared to non-users [51].

Several efforts have been made to describe OF consumers through descriptive, socio-economic and behavioural factors. However, comparisons between studies are complicated by different market conditions between countries and different study methods. In general, reviews across countries show little consistency and no clear differences or patterns between organic and conventional users [7;48]. Nevertheless, higher OF consumption tends to be related to vegetarianism [8;43;44;49;50], educational level [2;45;46] and urbanization [2;45;46] which is supported by our findings.

In previous studies frequent OF users have been described as 'intellectuals' from urban areas [2], and the most common reason not to purchase OF products was lack of knowledge or awareness [7]. It has been hypothesized that higher educational level provides the consumer more information and experiences to believe that personal behaviour, including OF purchasing behaviour, and personal decisions affect other people [45]. We found occupational status to be strongly associated with OF

preferences, which supports these previous studies. In relation to the higher OF use in or near the capital city our findings support the description of 'intellectuals' and the tendency to higher purchase in urban areas.

In general, income does not seem to explain differences in OF purchasing behaviour [7;48], and income is a weak determinant in highly industrialized countries such as Denmark [45]. Other findings suggest that OF users in some cases may have lower food expenditures than conventional households, despite the fact that OF products are more expensive and this can be due to differences in dietary habits of the households [44]. Thus, higher price for OF products appears irrelevant in relation to other incentives underlying OF preferences.

There seems to be a good agreement concerning incentives for OF use among countries [7;45]. Several studies have found that health considerations are one of the most important incentives for organic preferences followed by concern for the environment [2;7;12;48] and concern for pesticide residues [10]. In fact, it has been argued that frequent OF users consider the concern for health and environment to be one and the same thing [45]. Since health apparently is a serious concern for OF users, it can be assumed that they follow recommendations about health and exercise to a higher extent than non users. This is also reflected in our results, even after adjustment for occupational status.

The strength of our study is the large sample size as we have been able to include more than 60,000 pregnant women. It can be argued that self-reported dietary intake may be prone to bias, such as over- or underestimation, but a FFQ is a valid method for classifying individuals according to high or low intake, which was the main interest with respect to OF consumption. The FFQ has been validated against a 7-day weighed food record and the validation showed that the FFQ was useful in separating high and low intake [29]. The dietary calculations were based on assumptions of average portions, sizes and standard recipes for complex dishes, which may have introduced bias in the estimates. In this study we do not focuse on the accuracy of specific nutrient estimates, but instead one the differences between estimates. Thus, we find the dietary intake between non and frequent users to be valid.

Until today, very few etiological studies about organic consumption have been published. The major explanation underlying this may be found in the impact of surrounding multiple factors, unbalance in data, low compliance, and lack of knowledge about dietary components and pesticides' impact on human health. Our findings add one more parameter because OF consumers in general seem to have a healthier lifestyle and diet. In relation to the previous findings of OF users

being more conscious regarding health observational studies aiming at examining the impact of OF use on human health are complicated.

It is relevant to consider whether OF consumption is part of a specific organic lifestyle including healthy diet, physical activity and health and environmental awareness. If this is the case it may be of no significance to estimate the relationship between OF use and health outcomes in observational studies, because the risk of chronic diseases already is lowered by the diet and exercise. However, it is still important to investigate whether OF products can contribute to lower risk of diseases. Therefore careful epidemiological modelling that can control for confounding factors is needed.

In theory, a randomised controlled trial would be the optimal study design for investigating health effects of human OF consumption. However, in many cases such a trial would require a long intervention period and strict control of foods consumed and would be affected of long study period, high costs and low compliance. Measurements of biomarkers in blood, e.g. pesticide residues and fatty acids composition, would be desirable; however, in a study including 60,000 women this would be financially unfeasible. Moreover, possible health effects can be related to other factors that are undetecable in blood.

The DNBC gives us an opportunity to examine associations in observational studies; however, statistical models used to analyze these associations must be designed to manage residual confounding and several covariates. Our approach is to devise a stratification strategy for selecting exchangeable groups of women for low and high organic food consumption based on relevant confounders and our basis for this will be principal component analysis. This will restrict the study population and hence reduce the statistical power, but in return produce conservative estimates with reduced bias for effects under the assumption of perfect exchangeability.

In conclusion, frequent OF users in the DNBC had a healthier lifestyle and consumed a more prudent diet with higher intake of fruit and vegetables, fibre, vitamins, minerals, n-3 fatty acids and less saturated fat. Furthermore, they had a higher occupational status and were living in urban areas, which together indicates an impact of a social gradient on OF purchasing behaviour. Our findings point to a major challenge in examining the impact of OF consumption on health in observational studies due to potentially irremediable confounding by generally healthier food choices among frequent users. Thus, in future studies it is crucial to manage this particular confounder problem. Our detailed analyses constitute a strong basis for such later advancement of strategies for analyses that can allow for unbalance in data, when we compare maternal organic and non-organic food consumers and their offspring in relation to health outcomes.

**Literature Cited**


(1) Lea E (2005) Food, health, the environment and consumers' dietary choices. *Nutr Diet* **62**, 21-5.

(2) Torjusen H, Sangstad L, O'Doherty Jensen K et al. (2004) European Consumers' Conceptions of Organic Food: A Review of Available Research. National Institute for Consumer Resarch, Oslo Norway.

(3) Midmore P, Ayres N, Lund TB et al. (2008) Understanding the Organic Consumer through Narratives: an International Comparison. *IFOAM Organic World Congress, Modena Italy, June 16-20,* (http://orgprints.org/12574/).

(4) Wier M, Andersen LM (2003) Consumer demand for organic foods - attitudes, values and purchasing behaviour. *DARCOFenews* **3**, June.

(5) Lund TB, O'Doherty Jensen K (2008) Consumption of organic foods from a life history perspective: An explorative study among Danish consumers. Sociology of Food Research Group, Dept.of Human Nutrition, University of Copenhagen.

(6) Ayres N, Midmore P (2009) Consumption of organic foods from a life history perspective: An exploratory study of British consumers. School of Management and Buisness, Aberystwyth University.

(7) Yiridoe EK, Bonti-Ankomah S, Martin RC (2005) Comparison of consumer perceptions and preference toward organic versus conventionally produced foods: A review and update of the literature. *Renewable Agriculture and Food Systems* **20**, 193-205.

(8) Schifferstein HNJ, Oude Pphuis PAM (1998) Health-related determinants of organic food consumption in the Netherlands. *Food Qual Prefer* **9**, 119-33.

(9) Shepherd R, Magnusson M, Sjoden PO (2005) Determinants of consumer behavior related to organic foods. *Ambio* **34**, 352-9.

(10) Byrne PJ, Bacon JR, Toensmeyer UC (1994) Pesticide residue concerns and shopping location likelihood. *Agribusiness* **10**, 491-501.

(11) Huang CL (1996) Consumer preferences and attitudes towards organically grown produce. *Eur Rev Agric Econ* **23**, 331-42.

(12) Magnusson MK, Arvola A, Hursti UK et al. (2003) Choice of organic foods is related to perceived consequences for human health and to environmentally friendly behaviour. *Appetite* **40**, 109-17.

(13) Velimirov A, Huber M, Lauridsen C et al. (2009) Feeding trials in organic food quality and health research. *J Sci Food Agric* **90**, 175-82.

(14) Dangour AD, Dodhia SK, Hayter A et al. (2009) Nutritional quality of organic foods: a systematic review. *Am J Clin Nutr* **90**, 680-5.

(15) Williams CM (2002) Nutritional quality of organic food: shades of grey or shades of green? *Proc Nutr Soc* **61**, 19-24.

(16) Dangour AD, Lock K, Hayter A et al. (2010) Nutrition-related health effects of organic foods: a systematic review. *Am J Clin Nutr* **92**, 203-10.

(17) Magkos F, Arvaniti F, Zampelas A (2003) Organic food: nutritious food or food for thought? A review of the evidence. *Int J Food Sci Nutr* **54**, 357-71.

(18) Woese K, Lange D, Boess C et al. (1997) A Comparison of Organically and Conventionally Grown Foods - Results of a Review of the Relevant Literature. *J Sci Food Agric* **74**, 281-93.

(19) Grinder-Pedersen L, Rasmussen SE, Bugel S et al. (2003) Effect of diets based on foods from conventional versus organic production on intake and excretion of flavonoids and markers of antioxidative defense in humans. *J Agric Food Chem* **51**, 5671-6.

(20) Kummeling I, Thijs C, Huber M et al. (2008) Consumption of organic foods and risk of atopic disease during the first 2 years of life in the Netherlands. *Br J Nutr* **99**, 598-605.

(21) Rist L, Mueller A, Barthel C et al. (2007) Influence of organic diet on the amount of conjugated linoleic acids in breast milk of lactating women in the Netherlands. *Br J Nutr* **97**, 735-43.

(22) Olsen J, Melbye M, Olsen SF et al. (2001) The Danish National Birth Cohort - its background, structure and aim. *Scand J Public Health* **29**, 300-7.

(23) Olsen SF, Mikkelsen TB, Knudsen VK et al. (2007) Data collected on maternal dietary exposures in the Danish National Birth Cohort. *Paediatr Perinat Epidemiol* **21**, 76-86.

(24) The world of organic agriculture - Statistics & emerging trends 2011 (2011) : IFOAM, Bonn, & FiBL.

(25) Overvad K, Tjonneland A, Haraldsdottir J et al. (1991) Development of a Semiquantitative Food Frequency Questionnaire to Assess Food, Energy and Nutrient Intake in Denmark. *Int J Epidemiol* **20**, 900-5.

(26) The official Danish food composition database (www.foodcomp.dk) (2009) National Food Institute - Technical University of Denmark (DTU).

(27) Willett WC, Howe GR, Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* **65**, 1220S-8S.

(28) Friis S, Kruger Kjaer S, Stripp C et al. (1997) Reproducibility and relative validity of a self-administered semiquantitative food frequency questionnaire applied to younger women. *J Clin Epidemiol* **50**, 303-11.

(29) Mikkelsen TB, Osler M, Olsen SF (2006) Validity of protein, retinol, folic acid and n-3 fatty acid intakes estimated from the food-frequency questionnaire used in the Danish National Birth Cohort. *Public Health Nutr* **9**, 771-8.

(30) Mikkelsen TB, Olsen SF, Rasmussen SE et al. (2007) Relative validity of fruit and vegetable intake estimated by the food frequency questionnaire used in the Danish National Birth Cohort. *Scand J Public Health* **35**, 172-9.

(31) Halldorsson TI, Thorsdottir I, Meltzer HM et al. (2009) Dioxin-like activity in plasma among Danish pregnant women: dietary predictors, birth weight and infant development. *Environ Res* **109**, 22-8.

(32) Halldorsson TI, Meltzer HM, Thorsdottir I et al. (2007) Is high consumption of fatty fish during pregnancy a risk factor for fetal growth retardation? A study of 44,824 Danish pregnant women. *Am J Epidemiol* **166**, 687-96.

(33) Halldorsson TI, Thorsdottir I, Meltzer HM et al. (2008) Linking Exposure to Polychlorinated Biphenyls With Fatty Fish Consumption and Reduced Fetal Growth Among Danish Pregnant Women: A Cause for Concern? *Am J Epidemiol* **168**, 958-65.

(34) Klemmensen AK, Tabor A, Østerdal ML et al. (2009) Intake of Vitamin C and E in pregnancy and risk of pre-eclampsia: prospective study among 57,346 women. *BJOG* **116**, 964-74.

(35) Knudsen VK, Hansen HS, Ovesen L et al. (2007) Iron supplement use among Danish pregnant women. *Public Health Nutr* **10**, 1104-10.

(36) Mikkelsen TB, Osler M, Orozova-Bekkevold I et al. (2006) Association between fruit and vegetable consumption and birth weight: a prospective study among 43,585 Danish women. *Scand J Public Health* **34**, 616-22.

(37) Mikkelsen TB, Osterdal ML, Knudsen VK et al. (2008) Association between a Mediterranean-type diet and risk of preterm birth among Danish women: a prospective cohort study. *Acta Obstet Gynecol Scand* **87**, 325-30.

(38) Oken E, Osterdal ML, Gillman MW et al. (2008) Associations of maternal fish intake during pregnancy and breastfeeding duration with attainment of developmental milestones in early childhood: a study from the Danish National Birth Cohort. *Am J Clin Nutr* **88**, 789-96.

(39) Osterdal ML, Strom M, Klemmensen AK et al. (2009) Does leisure time physical activity in early pregnancy protect against pre-eclampsia? Prospective cohort in Danish women. *BJOG* **116**, 98-107.

(40) Strom M, Mortensen EL, Halldorsson TI et al. (2009) Fish and long-chain n-3 polyunsaturated fatty acid intakes during pregnancy and risk of postpartum depression: a prospective study based on a large national birth cohort. *Am J Clin Nutr* **90**, 149-55.

(41) Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemometr Intell Lab* **2**, 37-52.

(42) Denver S, Christensen T, Krarup S (2007) Får økologiske forbrugere oftere 6 om dagen? (Consume organic users more often 6 servings per day?). *Tidsskrift for Landøkonomi* **2**, 109-18.

(43) Holt G (1992) Investigating the diet of "organic eaters". *Nutr Food Sci* **6**, 13-6.

(44) Brombacher J, Hamm U (1990) Expenses for nutrition with food from organic agriculture. *Ecology and Farming* **1**, 13-6.

(45) O'Doherty Jensen K, Larsen HN, Mølgaard JP et al. (2001) Økologiske fødevarer og menneskets sundhed - Rapport fra vidensyntese udført i regi af Forskningsinstitut for Human Ernæring, KVL. FØJO-rapport nr. 14. Forskningscenter for Økologisk Jordbrug, FØJO.

(46) Wier M, O'Doherty Jensen K, Andersen LM et al. (2008) The character of demand in mature organic food markets: Great Britain and Denmark compared. *Food Policy* **33**, 406-21.

(47) Statistics Denmark (2011) StatBank Denmark. (http://www.statbank.dk/statbank5a/default.asp?w=1920).

(48) Bonti-Ankomah S, Yiridoe EK (2006) Organic and conventional food: A litterature review of the economics of consumer perceptions and preferences - Final Report. Organic Agriculture Centre of Canada, Nova Scotia Agricultural College

(49) Torjusen H, Brantsaeter AL, Haugen M et al. (2010) Characteristics associated with organic food consumption during pregnancy; data from a large cohort of pregnant women in Norway. *BMC Public Health* **10**, 775.

(50) Onyango BM (2007) Purchasing organic food in US food systems - A study of attitudes and practice. *Br Food J* **109**, 399-411.

(51) FDB Analyse (2010) Økologiske forbrugere belaster klimaet mindre (Organic consumers affect the environment to a lesser extent). (http://fdb.dk/nyhed/%C3%B8kologiske-forbrugere-belaster-klimaet-mindre).

**Table 1**: Distribution of answers in the six food categories (N = 60.773).

| Organic food | Never | | Sometimes | | Regularly | | Always | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| Milk | 15865 | 26 | 20679 | 34 | 16366 | 27 | 7862 | 13 |
| Cereals | 18666 | 31 | 26482 | 43 | 12664 | 21 | 2961 | 5 |
| Egg | 15825 | 26 | 17418 | 29 | 11570 | 19 | 15958 | 26 |
| Vegetables | 15162 | 25 | 31082 | 51 | 13065 | 22 | 1723 | 2 |
| Fruit | 24994 | 36 | 35089 | 50 | 8630 | 12 | 1464 | 2 |
| Meat | 31107 | 52 | 23468 | 39 | 4675 | 8 | 820 | 1 |

**Table 2:** Association between organic food consumption and different socio-demographic factors.

| Variable | Non user | | Frequent user | | Odds ratio (95 % CI) | | | | P # |
|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | | Crude | | Adjusted* | Trend adjusted |
| **Age** | | | | | | | | | < 0·0001 |
| Age < 20 | 148 | 2 | 29 | 1 | 0·43 | (0·29-0·65) | 0·33 | (0·16-0·70) | |
| 20 < Age < 25 | 1476 | 21 | 296 | 7 | 0·44 | (0·39-0·51) | 0·48 | (0·39-0·60) | |
| 25 < Age < 30 | 3102 | 43 | 1402 | 34 | | REF | | REF | |
| 30 < Age < 35 | 1901 | 26 | 1646 | 40 | 1·92 | (1·75-2·10) | 1·88 | (1·64-2·16) | |
| 35 < Age < 40 | 491 | 7 | 692 | 17 | 3·12 | (2·73-3·56) | 3·56 | (2·90-4·37) | |
| Age > 40 | 47 | 1 | 81 | 2 | 3·81 | (2·65-5·49) | 4·49 | (2·71-7·45) | |
| **Occupational status** | | | | | | | | | < 0·0001 |
| High | 266 | 4 | 519 | 14 | | REF | | REF | |
| Medium | 1107 | 17 | 1264 | 33 | 0·58 | (0·49-0·69) | 0·83 | (0·67-1·04) | |
| Skilled | 1534 | 23 | 484 | 13 | 0·16 | (0·13-0·19) | 0·27 | (0·21-0·34) | |
| Student | 543 | 8 | 654 | 17 | 0·62 | (0·51-0·74) | 1·01 | (0·78-1·31) | |
| Unskilled | 2194 | 33 | 447 | 12 | 0·10 | (0·09-0·12) | 0·25 | (0·20-0·32) | |
| Unemployed | 984 | 15 | 432 | 11 | 0·22 | (0·19-0·27) | 0·48 | (0·37-0·62) | |
| **Living area** | | | | | | | | | |
| West Denmark | 5212 | 77 | 1617 | 43 | | REF | | REF | |
| East Denmark | 1567 | 23 | 2100 | 57 | 4·32 | (3·96-4·71) | 2·44 | (2·05-2·90) | < 0·0001 |
| **Urbanization** | | | | | | | | | < 0·0001 |
| Capital city | 216 | 4 | 1036 | 29 | 23·8 | (20·1-28·2) | 8·04 | (6·23-10·4) | |
| Capital suburbs | 333 | 5 | 480 | 13 | 7·15 | (6·07-8·44) | 2·70 | (2·10-3·48) | |
| 100·000 + | 590 | 9 | 673 | 19 | 5·66 | (4·92-6·52) | 5·04 | (4·22-6·02) | |
| 10·000 – 99·999 | 2124 | 34 | 801 | 22 | 1·87 | (1·66-2·11) | 1·45 | (1·25-1·69) | |
| < 10·000 | 3988 | 48 | 602 | 17 | | REF | | REF | |
| **Cohabitation status** | | | | | | | | | |
| Couple/married | 6731 | 98 | 3831 | 97 | | REF | | REF | |
| Single | 125 | 2 | 103 | 3 | 1·45 | (1·11-1·88) | 2·25 | (1·43-3·55) | 0·0001 |
| **Parity** | | | | | | | | | 0·66 |
| Nulliparous | 3649 | 53 | 1789 | 45 | | REF | | REF | |
| 1 child | 1990 | 29 | 1520 | 39 | 1·56 | (1·43-1·70) | 1·54 | (1·34-1·76) | |
| 2 children | 941 | 14 | 537 | 14 | 1·16 | (1·03-1·31) | 1·15 | (0·95-1·39) | |
| 3+ children | 278 | 4 | 89 | 2 | 0·65 | (0·51-0·83) | 0·43 | (0·29-0·64) | |
| **Smoking** | | | | | | | | | < 0·0001 |
| Non-smoker | 5014 | 71 | 3210 | 78 | | REF | | REF | |
| Occasional smoker | 741 | 10 | 559 | 14 | 1·18 | (1·05-1·33) | 1·34 | (1·12-1·61) | |
| < 15 cigarettes per day | 1091 | 15 | 313 | 8 | 0·45 | (0·39-0·51) | 0·62 | (0·50-0·75) | |
| > 15 cigarettes per day | 264 | 4 | 33 | 1 | 0·20 | (0·14-0·28) | 0·28 | (0·16-0·48) | |
| **Alcohol in pregnancy** | | | | | | | | | |
| Not at all | 3506 | 49 | 1803 | 44 | | REF | | REF | |
| Yes | 3600 | 51 | 2309 | 56 | 1·25 | (1·15-1·35) | 0·90 | (0·80-1·01) | 0·21 |
| **Dietary habits** | | | | | | | | | |
| Eating meat | 7172 | 99·7 | 3909 | 94 | | REF | | REF | |
| Vegetarian/vegan | 23 | 0·3 | 237 | 6 | 18·8 | (12·2-28·9) | 18·7 | (9·77-35·7) | < 0·0001 |
| **Prepregnant BMI** | | | | | | | | | < 0·0001 |
| BMI < 18,5 | 274 | 4 | 229 | 6 | 1·15 | (0·96-1·38) | 1·26 | (0·96-1·66) | |
| 18,5 ≤ BMI < 25 | 4016 | 60 | 2911 | 75 | | REF | | REF | |
| 25 ≤ BMI < 30 | 1604 | 24 | 561 | 15 | 0·48 | (0·43-0·54) | 0·68 | (0·58-0·79) | |
| 30 ≤ BMI < 35 | 594 | 9 | 127 | 3 | 0·29 | (0·24-0·36) | 0·51 | (0·39-0·67) | |
| BMI ≥ 35 | 252 | 4 | 39 | 1 | 0·21 | (0·15-0·30) | 0·31 | (0·19-0·51) | |
| **Energy intake** | | | | | | | | | < 0·0001 |
| Quintile 1 | 1573 | 22 | 668 | 16 | | REF | | REF | |
| Quintile 2 | 1290 | 19 | 762 | 19 | 1·39 | (1·22-1·58) | 1·23 | (1·02-1·49) | |
| Quintile 3 | 1342 | 19 | 790 | 19 | 1·38 | (1·22-1·57) | 1·24 | (1·03-1·50) | |
| Quintile 4 | 1270 | 18 | 889 | 22 | 1·65 | (1·45-1·87) | 1·59 | (1·32-1·91) | |
| Quintile 5 | 1509 | 22 | 980 | 24 | 1·53 | (1·35; 1·73) | 1·80 | (1·50-2·15) | |
| **Physical activity** | | | | | | | | | < 0·0001 |
| Non | 4673 | 68 | 2162 | 55 | | REF | | REF | |
| Light | 1358 | 20 | 910 | 23 | 1·45 | (1·31-1·60) | 1·30 | (1·12-1·49) | |
| Moderate | 718 | 11 | 734 | 19 | 2·21 | (1·97-2·48) | 1·72 | (1·45-2·03) | |
| High | 92 | 1 | 110 | 3 | 2·58 | (1·95-3·42) | 1·24 | (0·80-1·93) | |
| **Dietary supplements** | | | | | | | | | |
| No | 671 | 10 | 211 | 5 | | REF | | REF | |
| Yes | 6334 | 90 | 3891 | 95 | 1·95 | (1·66-2·29) | 1·28 | (1·00-1·64) | 0·073 |

* Mutually adjusted.

# P value for trend for exposures with >2 categories.

**Table 3:** Associations between intake of different food items and organic consumption in the Danish National Birth Cohort (non-users is the referent).

| Food item | Non users | | Frequent users Crude | | | Frequent users Adjusted* | | |
|---|---|---|---|---|---|---|---|---|
| | MEAN | SD | INCREASE IN INTAKE # | | | INCREASE IN INTAKE # | | |
| | | | (g/day) | 95 % CI | % | (g/day) | 95 % CI | % |
| Vegetables | 92·4 | 82·7 | 81·6 | 78·2; 85·1 | 88 | 61·9 | 58·1; 65·8 | 67 |
| Legumes | 9·5 | 17·6 | 8·5 | 7·6; 9·4 | 90 | 6·2 | 5·3; 7·2 | 65 |
| Fruit and berries | 128 | 104 | 64·4 | 60·3; 68·5 | 50 | 42·0 | 37·3; 46·7 | 33 |
| Nuts | 1·2 | 2·6 | 2·1 | 1·9; 2·9 | 175 | 1·5 | 1·3; 1·7 | 125 |
| Potatoes | 144 | 103 | -13·8 | -17·1; -10·5 | -10 | -3·7 | -7·1; -2·4 | -3 |
| French fries | 9·4 | 10·9 | -4·4 | -4·7; -4·1 | -47 | -2·8 | -3·1; -2·4 | -30 |
| Rice | 10·1 | 8·2 | 2·2 | 1·9; 2·5 | 22 | 1·4 | 1·1; 1·8 | 14 |
| Pasta | 13·0 | 9·5 | 2·1 | 1·7; 2·4 | 16 | 1·3 | 0·9; 1·7 | 10 |
| Whole grain bread/flour | 135 | 84 | 28·8 | 25·6; 32·0 | 21 | 13·0 | 9·8; 16·3 | 10 |
| White bread/flour | 93·8 | 58·8 | -17·2 | -19·2; -15·2 | -18 | -14·1 | -16·2; -12·1 | -15 |
| Breakfast cereals | 25·6 | 28·3 | 9·0 | 7·9; 10·0 | 35 | 6·2 | 5·0; 7·5 | 24 |
| Poultry | 19·7 | 17·8 | 4·7 | 4·0; 5·4 | 24 | 2·6 | 1·8; 3·5 | 13 |
| Pork | 29·9 | 19·9 | -12·8 | -13·5; -12·2 | -43 | -8·9 | -9·6; -8·3 | -30 |
| Beef/veal | 41·7 | 30·7 | -0·8 | -1·8; 0·3 | -2 | 1·1 | 0·1; 2·2 | 3 |
| Lamb | 0·8 | 4·1 | 2·9 | 2·8; 3·1 | 363 | 2·1 | 1·9; 2·3 | 263 |
| Processed meat | 19·5 | 15·4 | -7·6 | -8·1; -7·1 | -39 | -5·3 | -5·8; -4·8 | -27 |
| Seafood | 22·7 | 28·0 | 11·7 | 10·8; 12·6 | 52 | 9·3 | 8·4; 10·2 | 41 |
| Egg | 14·7 | 14·1 | 2·1 | 1·7; 2·6 | 14 | 1·5 | 1·1; 2·0 | 10 |
| Whole fat milk products | 67·2 | 175 | -8·6 | -13·7; -3·5 | -13 | -8·8 | -14·4; -3·1 | -13 |
| Light milk products | 478 | 407 | -6·3 | -21·0; 8·5 | -1 | 18·9 | 2·2; 35·7 | 4 |
| Yoghurt | 43·2 | 56·0 | 13·0 | 10·9; 15·1 | 30 | 8·1 | 5·7; 10·6 | 19 |
| Butter | 7·7 | 9·3 | 2·0 | 1·6; 2·3 | 26 | 1·5 | 1·1; 1·9 | 20 |
| Cheese | 28·2 | 24·2 | 8·3 | 7·3; 9·2 | 29 | 4·5 | 3·5; 5·6 | 16 |
| Oils | 0·9 | 2·1 | 1·9 | 1·8; 2·0 | 211 | 1·2 | 1·1; 1·3 | 133 |
| Margarine | 23·2 | 22·0 | -5·4 | -6·2; -4·7 | -23 | -5·1 | -5·9; -4·4 | -22 |
| Dressing/sauce | 5·6 | 7·9 | -1·8 | -2·0; -1·5 | -32 | -1·2 | -1·5; -1·0 | -21 |
| Tea | 122 | 202 | 69·9 | 62·0; 77·9 | 57 | 33·0 | 23·6; 42·4 | 27 |
| Coffee | 165 | 266 | -33·7 | -42·3; -25·2 | -20 | -32·4 | -41·8; -23·0 | -20 |
| Drink, sweated | 203 | 300 | -27·1 | -38·5; -15·7 | -13 | -29·0 | -41·8; -16·2 | -14 |
| Drink, light | 251 | 333 | -42·2 | -47·7; -36·7 | -17 | -30·0 | -36·5;-23·4 | -12 |
| Juice | 167 | 268 | 21·2 | 12·2; 30·2 | 13 | 20·8 | 10·5; 31·0 | 12 |
| Water | 985 | 553 | 172 | 153; 192 | 18 | 118 | 94·9; 141 | 12 |
| Alcohol | 20·3 | 37·5 | 4·5 | 2·8; 6·3 | 22 | -1·6 | -3·4; 0·3 | -8 |
| Snack | 4·8 | 5·3 | -1·8 | -2·0; -1·6 | -38 | -1·4 | -1·6; -1·2 | -29 |
| Dessert | 43·3 | 31·2 | 0·7 | -0·4; 1·8 | 2 | -2·8 | -3·9; -1·6 | -6 |

* Covariates: cohabitation status, age, smoking habits, parity, prepregnant BMI, occupational status, physical activity, energy intake, urbanization and living area.
# Δ compared to non users

**Table 4:** Associations between different nutrients from the diet and organic consumption in the Danish National Birth Cohort.

| **Nutrient** Energy adjusted | **Non users** | | **Frequent users** Crude | | | **Frequent users** Adjusted | | |
|---|---|---|---|---|---|---|---|---|
| | MEAN | SD | INCREASE IN INTAKE (per day) | 95 % CI | % | INCREASE IN INTAKE (per day) | 95 % CI | % |
| **Energy (kJ)** | 10058 | 2756 | 320 | 222; 418 | 3 | 442 | 327; 557 | 4 |
| **Fat (g)** | 83·3 | 16·5 | -6·5 | -7·1; -5·9 | -8,0 | -4·9 | -5·5; -4·2 | -6 |
| SFA (g) | 34·9 | 9·0 | -4·1 | -4·5; -3·8 | -12 | -2·9 | -3·3; -2·6 | -8 |
| MUFA (g) | 26·2 | 5·7 | -2·0 | -2·3; -1·8 | -8 | -1·5 | -1·7; -1·2 | -6 |
| PUFA (g) | 11·8 | 2·3 | 0·25 | 0·16; 0·34 | 2 | 0·09 | -0·01; 0·2 | 1 |
| N-3 (g) | 0·7 | 0·2 | 0·10 | 0·09; 0·11 | 15 | 0·07 | 0·06; 0·08 | 11 |
| N-6 (g) | 2·7 | 0·7 | -0·36 | -0·38; -0·33 | -13 | -0·19 | -0·22; -0·16 | -7 |
| Trans fatty acids (g) | 1·6 | 0·7 | -0·15 | -0·18; -0·13 | -9 | -0·12 | -0·15; -0·09 | -7 |
| Cholesterol (mg) | 326 | 96·7 | -17·4 | -20·8; -14·0 | -5 | -10·5 | -14·5; -6·4 | -3 |
| **Protein (g)** | 87·0 | 14·0 | 2·7 | 2·2; 3·2 | 3 | 2·8 | 2·2; 3·4 | 3 |
| **Carbohydrate (g)** | 312 | 36·9 | 12·8 | 11·4; 14·1 | 4 | 9·1 | 7·5; 10·6 | 3 |
| Starch (g) | 107 | 31·2 | 1·8 | 0·7; 2·9 | 2 | 1·0 | -0·3; 2·3 | 1 |
| Sugar (g) | 118 | 37·8 | 2·0 | 0·7; 3·4 | 8 | 2·0 | 0·5; 3·6 | 2 |
| Fibre (g) | 24·7 | 6·8 | 4·4 | 4·2; 4·7 | 18 | 3·1 | 2·8; 3·4 | 13 |
| **Vitamins** | | | | | | | | |
| Vitamin A (RE) | 889 | 392 | 45·6 | 31·1; 60·0 | 5 | 45·6 | 25·6; 62·6 | 5 |
| Retinol (µg) | 687 | 336 | -106 | -117; -94·3 | -15 | -77·3 | -90·7; -63·9 | -11 |
| Beta-carotene (µg) | 2338 | 2666 | 1810 | 1690; 1930 | 77 | 1465 | 1325; 1604 | 63 |
| Vitamin D (µg) | 3·0 | 1·6 | 1·08 | 1·00; 1·15 | 36 | 0·67 | 0·58; 0·75 | 22 |
| Vitamin E (α-TE) | 7·6 | 1·9 | 0·85 | 0·78; 0·92 | 11 | 0·56 | 0·48; 0·64 | 7 |
| Vitamin K (µg) | 79·9 | 42·4 | 39·7 | 37·8; 41·5 | 50 | 31·4 | 29·3; 33·6 | 39 |
| Vitamin $B_6$ (mg) | 1·5 | 0·3 | 0·16 | 0·15; 0·17 | 11 | 0·14 | 0·13; 0·15 | 9 |
| Vitamin $B_{12}$ (µg) | 6·2 | 2·3 | 0·12 | 0·04; 0·20 | 2 | 0·27 | 0·18; 0·37 | 4 |
| Folates (µg) | 330 | 65·7 | 49·2 | 46·6; 51·8 | 15 | 36·9 | 33·9; 39·8 | 11 |
| Vitamin C (mg) | 123 | 77·9 | 27·8 | 24·9; 30·7 | 23 | 23·8 | 20·4; 27·2 | 19 |
| **Minerals** | | | | | | | | |
| Calcium (mg) | 1347 | 453 | 57·1 | 40·4; 73·7 | 4 | 69·4 | 49·9; 88·9 | 5 |
| Magnesium (mg) | 375 | 66·1 | 37·5 | 35·0; 39·9 | 10 | 29·6 | 26·8; 32·4 | 8 |
| Iron (mg) | 10·9 | 1·6 | 0·80 | 0·74; 0·86 | 7 | 0·63 | 0·56; 0·70 | 6 |
| Zink (mg) | 12·2 | 1·9 | 0·31 | 0·24; 0·38 | 3 | 0·33 | 0·25; 0·41 | 3 |
| Iodine (µg) | 260 | 81·9 | 24·4 | 21·1; 27·6 | 9 | 27·2 | 23·4; 31·0 | 10 |
| Selenium (µg) | 41·2 | 9·5 | 4·1 | 3·7; 4·4 | 10 | 3·6 | 3·2; 4·0 | 9 |
| Cobber (mg) | 4·5 | 1·8 | 0·56 | 0·49; 0·62 | 13 | 0·37 | 0·30; 0·45 | 8 |

*Covariates: cohabitation status, age, smoking habits, parity, prepregnant BMI, occupational status, physical activity, urbanization and living area.
[#] Δ compared to non users

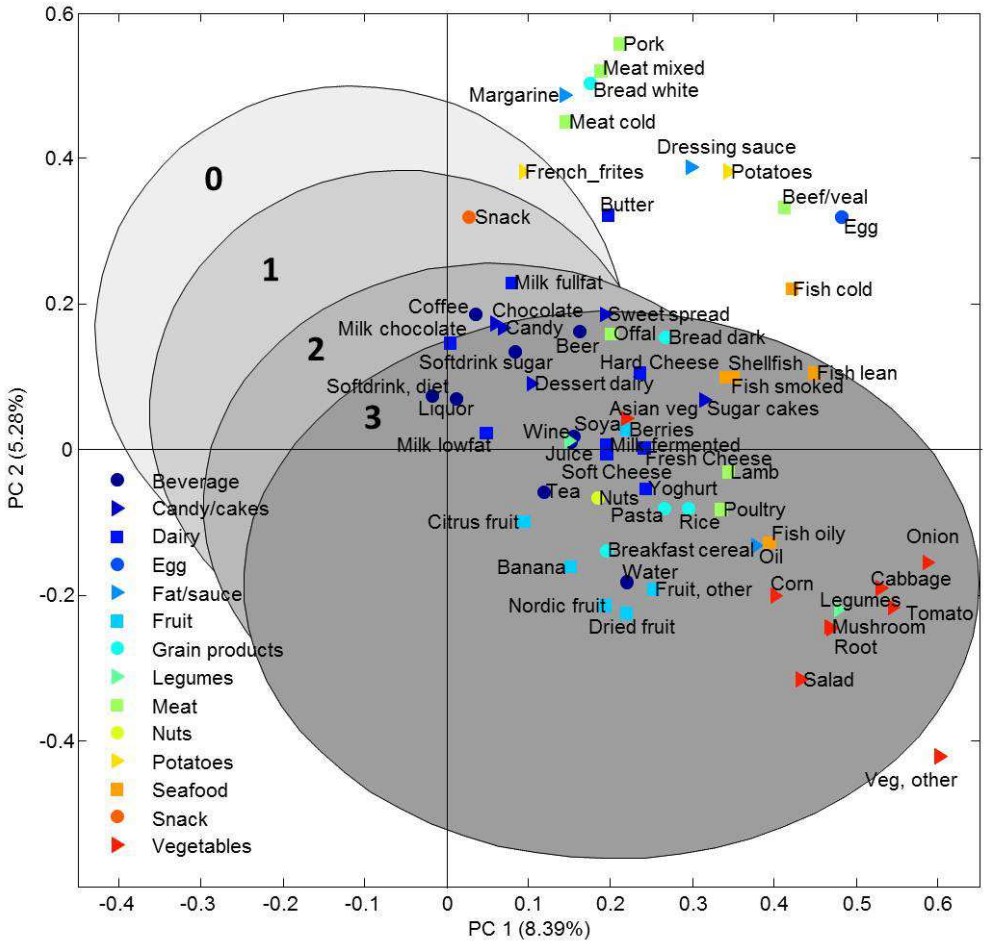



Figure 1 Principal component analysis results for the first two principal components reflecting the sample distribution and the associations between different food groups. The distribution of the 60.773 women is reflected as four ellipsoids associated with degree of organic food consumption (0 = non users, 1 = light user, 2 = middle user, 3 = frequent user). The food groups are imposed on this graph and are colored according to food classes. The axis correspond to factor correlations with the individual food groups.

# Paper V

Neonatal Cytokine Profile in the Airway Mucosal Lining Fluid is skewed by Maternal Atopy

Nilo Følsgaard, Bo Chawes, **Morten A. Rasmussen**, Anne L Bischoff, Charlotte G Carson, Jakob Stokholm, Louise Pedersen, Trevor T Hansel, Klaus Bønnelykke, Susanne Brix, Hans Bisgaard

*American Journal of Respiratory and Critical Care Medicine,* ***185*** *(2012), 275-280*

# Neonatal Cytokine Profile in the Airway Mucosal Lining Fluid Is Skewed by Maternal Atopy

**Nilofar V. Følsgaard[1], Bo L. Chawes[1], Morten A. Rasmussen[2], Anne L. Bischoff[1], Charlotte G. Carson[1], Jakob Stokholm[1], Louise Pedersen[1], Trevor T. Hansel[3], Klaus Bønnelykke[1], Susanne Brix[4], and Hans Bisgaard[1]**

[1]Copenhagen Prospective Studies on Asthma in Childhood, Health Sciences, University of Copenhagen, Copenhagen University Hospital, Gentofte, Denmark; [2]Faculty of Life Sciences, University of Copenhagen, Frederiksberg, Denmark; [3]Imperial Clinical Respiratory Research Unit, National Heart and Lung Institute, Imperial College, London, United Kingdom; and [4]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Søltofts Plads, Lyngby, Denmark

***Rationale*: Heredity from mother or father may impact differently in complex diseases, such as atopy. Maternal atopy is a stronger risk factor than paternal atopy for the development of atopy in the offspring. We hypothesized that mother's and father's atopy would have a differential imprinting on the cytokines and chemokines in the upper airway mucosal lining fluid of healthy neonates.**

***Objectives*: To study parental atopic imprinting on the cytokines and chemokines in the upper airway mucosal lining fluid of healthy neonates.**

***Methods*: Eighteen cytokines and chemokines were quantified in nasal mucosal lining fluid in 309 neonates from the novel unselected Copenhagen Prospective Study on Asthma in Childhood (COPSAC) birth cohort.**

***Measurements and Main Results*: Maternal, but not paternal, atopic status (asthma, hay fever, or eczema with or without sensitization) was associated with general down-regulation of all 18 mediators assessed by principal component analysis (overall $P = 0.015$).**

***Conclusions*: Maternal atopy, but not paternal atopy, showed a strong linkage with a suppressed mucosal cytokine and chemokine signature in asymptomatic neonates, suggesting imprinting by the maternal milieu *in utero* or perinatal life.**

**Keywords:** neonatal airways; parental atopic disease; skewed immunology

**AT A GLANCE COMMENTARY**

**Scientific Knowledge on the Subject**

Maternal atopy is a stronger risk factor than paternal atopy for the development of atopy in the offspring.

**What This Study Adds to the Field**

Maternal atopy, but not paternal atopy, showed a strong linkage with a suppressed mucosal cytokine signature in asymptomatic neonates, suggesting imprinting by the maternal milieu *in utero* or perinatal life.

Heredity from mothers or fathers may impact disease development in the child differently in complex diseases, such as diabetes, inflammatory bowel disease, and atopy (1–4). Thus, atopic hereditary disease linkage in the offspring is stronger for maternal than paternal atopy (3–5). This suggests imprinting (i.e., an inheritance process independent of the classical mendelian inheritance) of the maternal immunologic status or milieu on the child during pregnancy or perinatal life, which is a period of immune immaturity.

The airway mucosa is exposed to the mother's milieu before birth and to allergens, pollutants, and the microbiome after birth. The ability of the newborn to mount a balanced and appropriate local immune response at the mucosal surfaces in response to the exposome (i.e., the totality of exposures received by a person during life [6]) is essential for maintaining healthy airways because atopic diseases presumably develop from an inappropriate immune response caused by complex interactions between genes, the exposome, and the host immune system. Indeed, a recent large-scale, genome-wide study of asthma showed strong associations to genes controlling innate and adaptive immune components (7).

The cytokines and chemokines involved in the pathogenesis of asthma and allergic rhinitis have been studied extensively *in vitro* in stimulated cord blood or peripheral blood from neonates (8–11), whereas there is no available *in vivo* evidence from the target organ, the airway mucosa. It is not known how parental atopy may affect early immunity in the target organ, the respiratory mucosa.

We hypothesized that atopic heredity from mother and father would have a differential imprinting on the cytokines and chemokines in the upper airway mucosa lining fluid of healthy neonates. To investigate this hypothesis, we studied subjects enrolled in the novel unselected Copenhagen Prospective Study on Asthma in Childhood ($COPSAC_{2010}$) birth cohort. Some of the results of this study have been previously reported in the form of an abstract (12).

(*Received in original form August 15, 2011; accepted in final form October 26, 2011*)

Supported by the Lundbeck Foundation, the Pharmacy Foundation of 1991, the Augustinus Foundation, the Danish Medical Research Council, and the Danish Pediatric Asthma Centre. COPSAC is funded by private and public research funds all listed on www.copsac.com. The funding agencies did not have any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

*Author Contributions*: H.B. is the guarantor of the study and is responsible for the integrity of the work as a whole, from conception and design to acquisition of data, analysis and interpretation of data, and writing of the manuscript. N.V.F. contributed to acquisition of data, data analyses and interpretation, and writing of the manuscript. S.B. performed analysis of cytokines and chemokines, data interpretation, and writing of the manuscript. B.L.C. and K.B. contributed to data analyses, interpretation, and writing of the manuscript. T.T.H. contributed with data interpretation and writing of the manuscript. M.A.R. performed statistical analyses. A.L.B., C.G.C., J.S., B.L.C., and L.P. contributed to data collection. All authors made important intellectual contributions and critical final revision of the manuscript.

Correspondence and requests for reprints should be addressed to Hans Bisgaard, M.D., D.M.Sc., Copenhagen Prospective Studies on Asthma in Childhood, Health Sciences, University of Copenhagen, Copenhagen University Hospital, Gentofte, Ledreborg Allé 34, DK-2820 Gentofte, Copenhagen, Denmark. E-mail: bisgaard@copsac.com

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

**Am J Respir Crit Care Med Vol 185, Iss. 3, pp 275–280, Feb 1, 2012**

**Originally Published in Press as DOI: 10.1164/rccm.201108-1471OC on November 10, 2011**
**Internet address: www.atsjournals.org**

## METHODS

### The COPSAC$_{2010}$ Birth Cohort

The neonates were part of the novel COPSAC$_{2010}$ cohort, an ongoing unselected prospective clinical study of a birth cohort of 700 children recruited in Zealand, Denmark, during 2009–2010. Written invitations were sent to pregnant women in Zealand. Women who were interested in participating were interviewed by telephone and subsequently attended the COPSAC clinical research unit for enrollment at gestational week 24, excluding women with chronic endocrinologic, nephrologic, or cardiac diseases. The neonates were enrolled at 1 week of age excluding any baby with severe congenital abnormality.

### Measurements of Airway Inflammatory Mediators

The neonates visited the clinical research unit at 1 month of age. Upper airway mucosal lining fluid was collected with a pair of 3- × 15-mm strips of filter paper (Accuwik Ultra; fibrous hydroxylated polyester sheets, cat no. SPR0730, Pall Life Sciences, Portsmouth, Hampshire, UK; this product is no longer manufactured, but Leukosorb from Pall Life Sciences is an alternative) inserted onto the anterior part of the inferior nasal turbinate of both nostrils (*see* online video of the sampling procedure; www.copsac.com). The strips were left for 2 minutes (13). After removal, the filter papers were frozen at −80°C and stored until analysis.

The levels of IFN-γ, IL-1β, IL-2, IL-4, IL-5, IL-10, IL-12p70, IL-13, tumor necrosis factor (TNF)-α, CXCL8 (IL-8), CCL11 (eotaxin-1), CCL26 (eotaxin-3), CXCL10 (IP-10), CCL2 (MCP-1), CCL13 (MCP-4), CCL22 (MDC), CCL4 (MIP-1β), and TARC (CCL17) in the extracted upper airway mucosal lining fluid were analyzed in duplicate using the Ultrasensitive Meso Scale Discovery Multi-spot Human TH1/TH2 10-Plex cytokine assay and 9-plex chemokine assay (Meso Scale Discovery, Gaithersburg, MD). Samples were read using the Sector Imager 6000 (Meso Scale Discovery). The lower limit of detection was set as the mean signal from blanks +3 SD (*see* online supplement). For validation, IL-8 was analyzed with both cytokine and chemokine assay showing complete correlation between the two (rho = 0.98). IL-8 chemokine values are presented in this article. As for IL-4, 55% of the measurements were under the set detection level (1.05 pg/ml), and 20 were undetectable or missing (i.e., 0 pg/ml) (*see* Table E2 in the online supplement). IL-4 has been included in the analysis because of its central role in Th2 skewing, but interpretation of these results must be done with caution.

### Atopic Predisposition

Parents' atopic status was defined from a history of doctor-diagnosed asthma, hay fever, or atopic dermatitis independent of sensitization. This was determined by a structured clinical interview performed by the research doctors at the parents' first visit to the COPSAC research clinic during gestational week 24.

### Statistical Analyses

Analyses were done by conventional statistics and by pattern recognition using principal component analysis (PCA). Mediator levels were log-transformed before analysis. In the conventional statistics, differences in cytokine and chemokine concentrations between the offspring of parents who are atopic versus parents who are not atopic were tested by Student unpaired *t* test. Our primary outcomes were mother who is atopic (irrespective of paternal atopic status) and father who is atopic (irrespective of maternal atopic status). Estimates were expressed as relative differences with the nonatopic group as reference and with corresponding 95% confidence intervals.

For the pattern recognition analyses, PCA was used to extract underlying latent variables, called principal components, which describe the systematic part of the variation across the 18 mediators in a few uncorrelated new and latent variables. The latent variables (principal components) reflect the distribution of samples across all variables. In this way it can be seen how the different variables contribute to the principal components and thereby understand what variables dominate the dataset in biologic patterns. Scatter plots of the principal components were used to investigate differences in overall cytokine–chemokine response between the children with or without mothers who are atopic.

The data processing was conducted using MATLAB R2009b v. 7.9.0.529 (MathWorks, Natick, MA) with the statistical and PLS toolbox (*t* test, PCA, visualization). In PCA, the NIPALS algorithm (nonlinear iterative partial least squares) was used to estimate the model while dealing with missing data (14).

Additional methodologic details are given in the online supplement.

## RESULTS

### Baseline Characteristics

The study included the first 309 consecutive neonates from the novel COPSAC$_{2010}$ birth cohort at the mean age of 31 days (SD, 5 d). Sampling was completed in all neonates without any complications or adverse events.

Atopic disease had been diagnosed in 173 of the mothers (56% of the study group) and 142 (47%) of the fathers, and 241 (78%) of the neonates had either a mother or a father with a history of atopy. Baseline characteristics of the neonates of mothers who are atopic (with or without fathers who are atopic) did not differ significantly from neonates of mothers who are not atopic (with or without fathers who are atopic) with respect to fathers' atopy, ethnicity, household income, antibiotics in third trimester, smoking, alcohol intake in third trimester, furred animals in the home during pregnancy, gestational age, birth weight, caesarean section, Apgar score less than 7 (15), sex, older siblings, or exclusive breastfeeding (Table 1). Neonates of fathers who are atopic (with or without a mother who is atopic) versus fathers who are not atopic (with or without mothers who are atopic) and neonates of mothers who are atopic (without fathers who are atopic) versus fathers who are atopic (without mothers who are atopic) were also comparable with respect to the previously mentioned characteristics (*see* Table E1A and E1B).

### Cytokine and Chemokine Levels in Neonates with and without Parents Who Are Atopic

Mean levels and variation in cytokine and chemokine levels are shown in Table E2.

Neonates of mothers who are atopic (with or without fathers who are atopic) had lower levels of mediators compared with neonates of mothers who are not atopic (with or without fathers who are atopic) (Figure 1A). Univariate statistical analyses showed *P* less than or equal to 0.01 for eotaxin-3 and IP-10, and *P* less than or equal to 0.05 for IFN-γ, IL-13, IL-1β, IL-2, IL-4, IL-8, TNF-α, eotaxin-1, MCP-1, MCP-4, and MDC. The mediators IL-10, IL-12p70, IL-5, MIP-1β, and TARC exhibited a nonsignificant trend of down-regulation (*see* Table E3). The conventional statistical approach was verified in a data-driven unsupervised PCA. The loading plot for all mediators and maternal atopic status shows that all variables are strongly correlated, because the first component (PC1) describes 66.5% of the total variation (*see* Figure E1). Therefore, the significant univariate associations revealed by conventional statistics mainly reflect one general underlying inhibitory effect in the neonates of mothers who are atopic. This is confirmed by the separation plot (Figure 2) showing significantly different distribution for individuals with and without mothers who are atopic ($P = 0.01$).

Neonates with fathers who are atopic (with or without mothers who are atopic) showed similar mediator levels as fathers who are not atopic (with or without mothers who are atopic) (Figure 1B; *see* Table E3). All univariate statistical comparisons were nonsignificant except MCP-1 (up-regulated in neonates with fathers who are atopic; $P = 0.02$). This pattern was confirmed in the data-driven PCA (*see* Figure E2).

Neonates of mothers who are atopic and fathers who are not atopic exhibited significantly reduced levels of IL-10, IL-12p70, IL-2, IL-4, TNF-α, eotaxin-3, MCP-1, MCP-4, and TARC compared

**TABLE 1. BASELINE DEMOGRAPHICS**

| | All % (*N*) | Mothers Who Are Atopic % (*N*) | Mothers Who are Not Atopic % (*N*) | *P* Value |
|---|---|---|---|---|
| Baseline | 309 | 56 (173) | 44 (136) | — |
| Prenatal exposure | | | | |
| Father who is atopic | 47 (141) | 43 (73) | 52 (68) | 0.16 |
| Race—white | 95 (293) | 96 (165) | 94 (128) | 0.32 |
| Household income—high* | 35 (108) | 31 (53) | 40 (55) | 0.37 |
| Antibiotics in third trimester | 18 (56) | 19 (32) | 18 (24) | 0.85 |
| Furred animals during pregnancy | 34 (104) | 35 (60) | 32 (44) | 0.77 |
| Maternal smoking during pregnancy | 5 (15) | 3 (6) | 7 (9) | 0.20 |
| Alcohol >1 unit/wk during third trimester | 6 (18) | 6 (10) | 6 (8) | 0.97 |
| Postnatal exposure | | | | |
| Gestational age† | 5 (16) | 5 (8) | 6 (8) | 0.62 |
| Birth weight (<2,500 g) | 1 (3) | 2 (2) | 1 (1) | 0.74 |
| Male | 50 (154) | 48 (83) | 52 (71) | 0.46 |
| Apgar score <7 at 1 min | 4 (13) | 5 (9) | 3 (4) | 0.53 |
| Caesarian section | 12 (37) | 12 (20) | 12 (17) | 0.26 |
| Older sibling | 48 (148) | 47 (82) | 49 (66) | 0.84 |
| Exclusive breastfeeding until study day | 90 (279) | 90 (156) | 90 (123) | 0.45 |

*High; above €110,000 a year.
† Before 37 weeks.

with neonates of fathers who are atopic and mothers who are not atopic. IL-8, IP-10, and MDC exhibited a nonsignificant trend of down-regulation (Figure 1C; *see* Table E3). The PCA confirmed the significant general down-regulation of mediators (*see* Figure E3).

## DISCUSSION

### Main Findings

Healthy neonates born to mothers who are atopic have down-regulated cytokines and chemokines in the upper airway mucosal lining fluid, whereas there is no association to father's atopic status. This suggests a maternal imprinting on the child's immune response developing during pregnancy or the first weeks of life.

### Strengths and Limitations

We developed this noninvasive method of sampling upper airway mucosal lining fluid by filter paper with efficient absorption (wicking) properties, but with minimal protein binding capacity, so that mediators can be efficiently eluted (13, 16). This simple and noninvasive method makes it possible to assess the baseline innate and adaptive immune responses in the airway mucosa as reflected by a panel of cytokines and chemokines. The novel method was feasible in all neonates and without inconvenience to the child.

It is an advantage of this method that mediators could be measured in the upper airway mucosal lining fluid reducing the problem of dilution inherent to nasal lavage where the dilution is a significant confounder and may reduce the mediator concentration below the detection limit of the assay (17). It is a strength that the biologic sampling of mucosal lining fluid is from one of the target organs of allergic disease. A further strength of the method is that these mediators could be assessed in asymptomatic healthy neonates at baseline without stimulation, hence avoiding the excessive response accompanying challenge models (18).

It seems probable that the *in vivo* mediator concentrations are the product of networks of interactions between different host mucosal cell types reacting with the exposome, and hence this *in vivo* measure is a more complete assessment of the final common pathway of the immune response than a response to stimulated purified cells or cell lines determined *in vitro*.

The validity of the study is increased by the comprehensive prospective data collection from personal interviews in the clinic on potential confounders, including the pregnant mother's use of antibiotics, alcohol, and tobacco, mode of delivery, breastfeeding, and presence of siblings or animals at home.

The statistical evidence is strong with conventional statistics confirmed in the unsupervised, data-driven PCA (i.e., the potential issue of multiple testing by univariate analyses of each of the 18 mediators is mitigated by the overall significance in the first principal component). This PCA analysis simultaneously analyzes variation in all cytokines and chemokines, and allows testing for difference between groups by one single test. This revealed significantly different cytokine and chemokine levels for neonates with and without mothers who are atopic.

It is a limitation of the study that the definition of parents who are atopic did not include an objective measure, such as specific-IgE or skin prick test.

It may be a limitation to our conclusions that, despite being an unselected birth cohort, 78% have at least one parent who is atopic compared with the Danish population where 54% of children have been reported to have at least one parent who is atopic, reflecting a bias toward the afflicted population even in unselected cohort studies (19). It is well-known that parents with atopic disease tend to avoid certain behaviors or environments (20). Therefore, the home environment, maternal dietary habits, and the exposure to allergens in these families might differ from families without atopic predisposition, which may have affected the maternal imprinting on the child's immune response.

### Interpretation

The *in vivo* airway cytokine and chemokine concentrations in neonates with mothers who are atopic presented a general down-regulation compared with mothers who are not atopic and compared with fathers who are atopic, whereas there was no influence from the fathers' atopic status compared with fathers who are not atopic. This general suppression in all mediators may be interpreted as a delayed maturation of the immune response.

The immune response is generally immature in early life as recently demonstrated in the experiment by nature, where the H1N1v pandemic caused more severe disease in neonates and young infants than in older children, all unprotected from passive

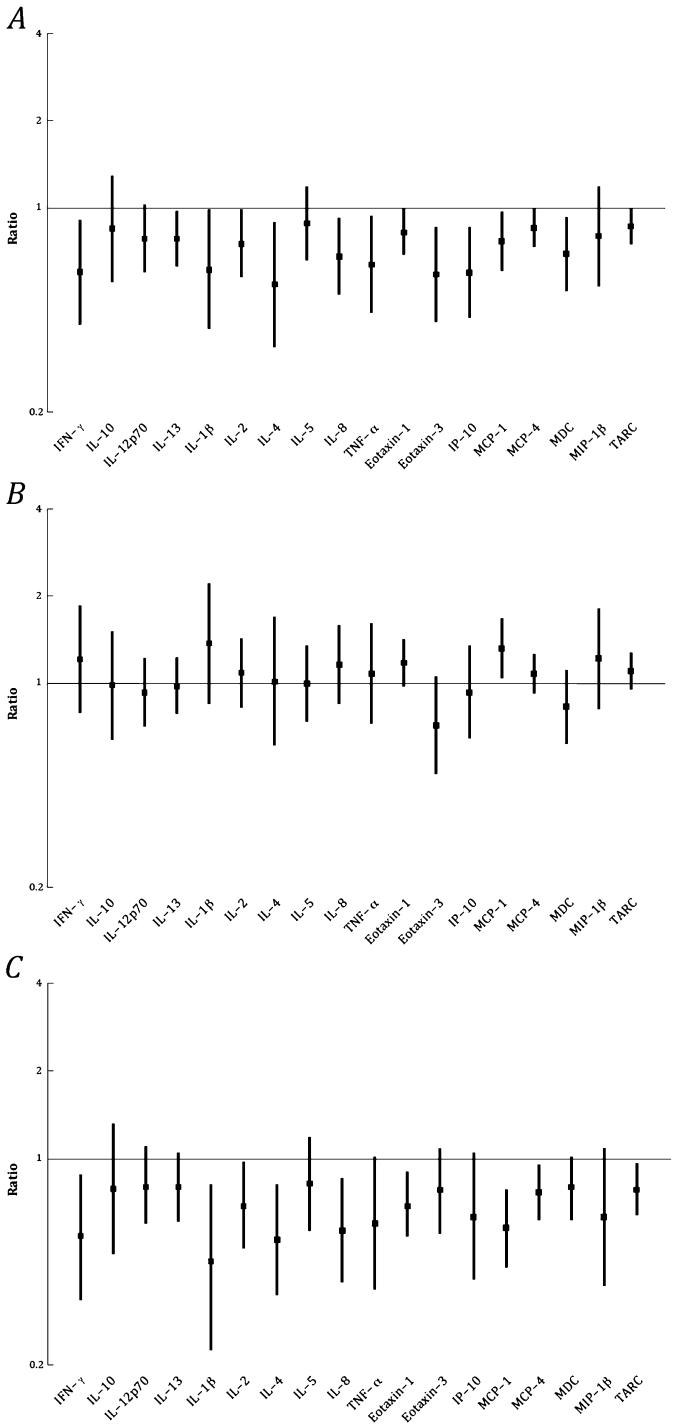



***Figure 1.*** Ratios of cytokines and chemokines in the upper airway mucosal lining fluid in healthy neonates. Means with 95% confidence interval. (*A*) Ratio: Neonates with mothers who are atopic versus mothers who are not atopic. (*B*) Ratio: Neonates with fathers who are atopic versus fathers who are not atopic. (C) Ratio: Neonates with mothers who are atopic (fathers who are not atopic) versus neonates with fathers who are atopic (mothers who are not atopic). eotaxin-1 = CCL11; eotaxin-3 = CCL26; IFN-γ = interferon-γ; IL = interleukin; IL-8 = CXCL8; IP-10 = CXCL10; MCP-1 = CCL2; MCP-4 = CCL13; MDC = CCL22; MIP-1β = CCL4; TARC = CCL17; TNF-α = tumor necrosis factor-α.

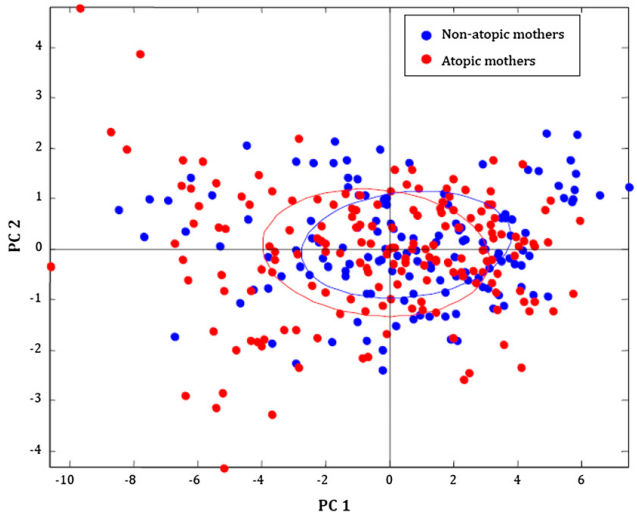



***Figure 2.*** Separation plot for neonates with mothers who are atopic (with or without fathers who are atopic) versus mothers who are not atopic (with or without fathers who are atopic). For separation in PC1 ($P = 0.035$, Mann-Whitney *U* test; $P = 0.015$, *t* test) and PC2 ($P = 0.38$, Mann-Whitney *U* test; $P = 0.24$, *t* test). Ellipsoids are centered at the mean and rotated according to correlation with latent variables (PC1 and PC2) using standard deviation as semiaxis.

immunity to this novel virus (21). This illustrates the neonatal period as a period of immune immaturity and it is our interpretation that the milieu of the atopic mothers causes a further delay of the child's immune maturation.

Previous reports on the effect of maternal atopic status on neonatal immune profiles are ambiguous probably because of the limitation of *in vitro* studies including use of different stimulation regimens, problems with the harvesting of cells, use of single-cell lines with absence of epithelial–immune cell cross-talk, and different laboratory practices. In agreement with our findings, *in vitro* studies stimulating the innate immune cells from cord blood reported a reduced response of the Th1-type cytokine IFN-γ in children of mothers who are atopic (8, 9, 22–24). Likewise, regulatory T cells and their suppressive function and associated cytokines (IL-10) were found suppressed (24), together with reduced levels of Toll-like receptor (TLR) 2, TLR4/CD14, and TLR9 expression in cord blood from high-risk infants (10, 25). A linkage between maternal atopy and reduced Th1 (IFN-γ) response toward fetal histocompatibility complex antigens was previously reported (26), thus indicating that maternal atopy may impact fetal immune programming by immune regulatory processes at the maternofetal interface. Together these *in vitro* studies support our *in vivo* data suggesting that that maternal heredity delays the immune maturation in the offspring. One previous *in vivo* study assessed cytokines directly in the cord blood from 20 high-risk infants and 36 control subjects reporting detectable levels of the chemokines IP-10, I-TAC, MDC, and TARC with no significant correlations to the atopic status of the mother, whereas other Th1- and Th2-associated cytokines were undetectable (11). Other studies found no association between atopic heredity and the *in vitro* immune response to nonspecific stimulation of cord blood cells (27, 28). Studies have even reported enhanced levels of IFN-γ and IL-12 in cord blood from high-risk infants on stimulation of the innate immune system (29), and enhanced IL-13 response in stimulated cord blood (28).

Maternal specific imprinting has also been reported for other complex diseases, such as diabetes (1) and Crohn's disease (2). The mechanism behind maternal imprinting cannot be determined by our data. It may be speculated to arise from the maternal immune system signaling to the developing fetus during pregnancy or may involve epigenetic mechanisms (30). Alternatively, exposure to environmental immune activators, such as maternal nutrition during pregnancy (e.g., omega-3 fatty acids and vitamin D intake) or the microbiome of the mother, is imprinting the fetal immune response toward the atopic status. We recently reported an association between abnormal bacterial colonization of the upper airways of neonates and later development of asthma suggesting that, *post or propter*, the human microbiome is associated with the skewing of the immune balance in people with asthma (31).

## CONCLUSIONS

Maternal atopic disease was associated with a general down-regulation of cytokine and chemokine levels in the upper airway mucosal lining fluid in healthy neonates. There was no paternal linkage to the mucosal immune response pattern, suggesting maternal programming of the fetus or neonate is causing an aberrant local airway immune profile in the newborn child.

**Author disclosures** are available with the text of this article at www.atsjournals.org.

***Acknowledgment***: The authors thank the children and parents participating in the $COPSAC_{2010}$ cohort and the COPSAC study team. The study was conducted in accordance with the guiding principles of the Declaration of Helsinki and approved by the Ethics Committee for Copenhagen (H-B-2008–093) and the Danish Data Protection Agency (j.nr. 2008–41–2599). Both parents gave their informed consent before enrollment of the children.

### References

1. Cerf ME. Parental high-fat programming of offspring development, health and beta-cells. *Islets.* 2011;3:118–120.
2. Akolkar PN, Gulwani-Akolkar B, Heresbach D, Lin XY, Fisher S, Katz S, Silver J. Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. *Am J Gastroenterol* 1997;92:2241–2244.
3. Lim RH, Kobzik L, Dahl M. Risk for asthma in offspring of asthmatic mothers versus fathers: a meta-analysis. *PLoS ONE* 2010;5:e10134.
4. Ruiz RG, Kemeny DM, Price JF. Higher risk of infantile atopic dermatitis from maternal atopy than from paternal atopy. *Clin Exp Allergy* 1992;22:762–766.

5. Liu CA, Wang CL, Chuang H, Ou CY, Hsu TY, Yang KD. Prenatal prediction of infant atopy by maternal but not paternal total IgE levels. *J Allergy Clin Immunol* 2003;112:899–904.
6. Rappaport SM. Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol* 2011;21:5–9.
7. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, Mutius EV, Farrall M, Lathrop M, Cookson WO. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211–1221.
8. Prescott SL, King B, Strong TL, Holt PG. The value of perinatal immune responses in predicting allergic disease at 6 years of age. *Allergy* 2003; 58:1187–1194.
9. Rinas U, Horneff G, Wahn V. Interferon-gamma production by cord-blood mononuclear cells is reduced in newborns with a family history of atopic disease and is independent from cord blood IgE-levels. *Pediatr Allergy Immunol* 1993;4:60–64.
10. Reece P, Thanendran A, Crawford L, Tulic MK, Thabane L, Prescott SL, Sehmi R, Denburg JA. Maternal allergy modulates cord blood hematopoietic progenitor Toll-like receptor expression and function. *J Allergy Clin Immunol* 2011;127:447–453.
11. Sandberg M, Frykman A, Ernerudh J, Berg G, Matthiesen L, Ekerfelt C, Nilsson LJ, Jenmalm MC. Cord blood cytokines and chemokines and development of allergic disease. *Pediatr Allergy Immunol* 2009;20: 519–527.
12. Folsgaard NV, Chawes BL, Rasmussen MA, Bonnelykke K, Brix S, Bisgaard H. Cytokine and chemokines in the airway epithelial lining fluid are down-regulated in newborns of atopic mothers [abstract]. Amsterdam: European Respiratory Society; 2011.
13. Chawes BL, Edwards MJ, Shamji B, Walker C, Nicholson GC, Tan AJ, Folsgaard NV, Bonnelykke K, Bisgaard H, Hansel TT. A novel method for assessing unchallenged levels of mediators in nasal epithelial lining fluid. *J Allergy Clin Immunol* 2010;125:1387–1389.
14. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; 2:37–52.
15. Kliegman RM. Nelson textbook of pediatrics, 18th ed. Amsterdam: Saunders Elsevier; 2007.
16. Nicholson GC, Kariyawasam H, Tan AJ, Hohlfeld J, Quinn D, Walker C, Rodman D, Westwick J, Jurcevic S, Kon O, *et al*. The effects of an anti-IL-13 mAb on cytokine levels and nasal symptoms following nasal allergen challenge. *J Allergy Clin Immunol* 2011;128:800–807.e9.
17. Bisgaard H, Krogsgaard OW, Mygind N. Measurement of secretion in nasal lavage. *Clin Sci (Lond)* 1987;73:217–222.
18. Pipkorn U, Proud D, Lichtenstein LM, Kagey-Sobotka A, Norman PS, Naclerio RM. Inhibition of mediator release in allergic rhinitis by pretreatment with topical glucocorticosteroids. *N Engl J Med* 1987; 316:1506–1510.
19. Benn CS, Wohlfahrt J, Aaby P, Westergaard T, Benfeldt E, Michaelsen KF, Bjorksten B, Melbye M. Breastfeeding and risk of atopic dermatitis, by parental history of allergy, during the first 18 months of life. *Am J Epidemiol* 2004;160:217–223.
20. Kummeling I, Thijs C, Stelma F, Huber M, Brandt PA, Dagnelie PC. Do parents with an atopic family history adopt a 'prudent' lifestyle for their infant? (KOALA Study). *Clin Exp Allergy* 2006;36:489–494.
21. Libster R, Bugna J, Coviello S, Hijano DR, Dunaiewsky M, Reynoso N, Cavalieri ML, Guglielmo MC, Areso MS, Gilligan T, *et al*. Pediatric hospitalizations associated with 2009 pandemic influenza A (H1N1) in Argentina. *N Engl J Med* 2010;362:45–55.
22. Liao SY, Liao TN, Chiang BL, Huang MS, Chen CC, Chou CC, Hsieh KH. Decreased production of IFN gamma and increased production of IL-6 by cord blood mononuclear cells of newborns with a high risk of allergy. *Clin Exp Allergy* 1996;26:397–405.
23. Gabrielsson S, Soderlund A, Nilsson C, Lilja G, Nordlund M, Troye-Blomberg M. Influence of atopic heredity on IL-4-, IL-12- and IFN-gamma-producing cells in *in vitro* activated cord blood mononuclear cells. *Clin Exp Immunol* 2001;126:390–396.
24. Schaub B, Liu J, Hoppler S, Haug S, Sattler C, Lluis A, Illi S, Mutius EV. Impairment of T-regulatory cells in cord blood of atopic mothers. *J Allergy Clin Immunol* 2008;121:1491–1499.
25. Krauss-Etschmann S, Hartl D, Heinrich J, Thaqi A, Prell C, Campoy C, Molina FS, Hector A, Decsi T, Schendel DJ, *et al*. Association between levels of Toll-like receptors 2 and 4 and CD14 mRNA and allergy in pregnant women and their offspring. *Clin Immunol* 2006; 118:292–299.
26. Prescott SL, Breckler LA, Witt CS, Smith L, Dunstan JA, Christiansen FT. Allergic women show reduced T helper type 1 alloresponses to fetal human leucocyte antigen mismatch during pregnancy. *Clin Exp Immunol* 2010;159:65–72.
27. Martinez FD, Stern DA, Wright AL, Holberg CJ, Taussig LM, Halonen M. Association of interleukin-2 and interferon-gamma production by blood mononuclear cells in infancy with parental allergy skin tests and with subsequent development of atopy. *J Allergy Clin Immunol* 1995; 96:652–660.
28. Kopp MV, Zehle C, Pichler J, Szepfalusi Z, Moseler M, Deichmann K, Forster J, Kuehr J. Allergen-specific T cell reactivity in cord blood: the influence of maternal cytokine production. *Clin Exp Allergy* 2001; 31:1536–1543.
29. Prescott SL, Noakes P, Chow BW, Breckler L, Thornton CA, Hollams EM, Ali M, van den Biggelaar AH, Tulic MK. Presymptomatic differences in Toll-like receptor function in infants who have allergy. *J Allergy Clin Immunol* 2008;122:391–399.
30. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A, *et al*. Parental origin of sequence variants associated with complex diseases. *Nature* 2009;462:868–874.
31. Bisgaard H, Hermansen MN, Buchvald F, Loland L, Halkjaer LB, Bonnelykke K, Brasholt M, Heltberg A, Vissing NH, Thorsen SV, *et al*. Childhood asthma after bacterial colonization of the airway in neonates. *N Engl J Med* 2007;357:1487–1495.

# Paper VI

Endotyping Atopic Dermatitis Children by Filaggrin Gene Mutation Status in a Prospective Cohort Study.

Charlotte G. Carson, Morten A. Rasmussen, Jonas P. Thyssen, Torkil Menné, Hans Bisgaard

*Submitted for British Journal of Dermatology*

**Endotyping Atopic Dermatitis Children by Filaggrin Gene Mutation Status in a Prospective Cohort Study.**

Charlotte Giwercman Carson[1], Morten Arendt Rasmussen[2], Jacob P. Thyssen[3], Torkil Menné[3], Hans Bisgaard[1]

[1]Copenhagen Prospective Studies on Asthma in Childhood; COPSAC, Health Sciences, University of Copenhagen, Copenhagen University Hospital, Gentofte, Ledreborg Alle 34,Gentofte 2820,Copenhagen, Denmark

[2]Faculty of Life Sciences, University of Copenhagen, DK-1870 Frederiksberg, Denmark

[3]National Allergy Research Centre, Department of Dermato-Allergology, Copenhagen University Hospital Gentofte, DK-2900 Hellerup, Denmark

Address for correspondence:

Copenhagen Prospective Studies on Asthma in Childhood

Danish Pediatric Asthma Center

Copenhagen University Hospital, Gentofte

Ledreborg Alle 34

DK-2820 Gentofte

Copenhagen

Denmark

Tel: +45 39 77 73 60

Fax: +45 39 77 71 29

E-mail: bisgaard@copsac.dk

Website: www.copsac.com

**Sources of funding**

COPSAC is funded by private and public research funds. Grants above 100.000 Euro were donated by The Lundbeck Foundation; the Pharmacy Foundation of 1991; Augustinus Foundation; the Danish Medical Research Council; The Danish Pediatric Asthma Centre.

The funding agencies did not have any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Disclosure**

The guarantor of the study is HB who has been responsible for the integrity of the work as a whole, from conception and design to conduct of the study and acquisition of data in COPSAC, analysis and interpretation of data and writing of the manuscript. The contributors of the study include CGC, MRA, JPT and TM who contributed to the analyses and interpretation of the data and writing of the manuscript. All contributors have contributed important intellectual input and approval of final version of the manuscript.

Data from this manuscript has not been presented previously except in abstract form.

Short title: Endotyping atopic dermatitis by filaggrin

Article Type: Original article

Word count: 2842 (abstract 200)

Figure count: 6

Tables: 0

Supplementary data, text: 206 words

Supplementary data, figures: 3

Supplementary data, tables: 1

**Abbreviations**

AD: Atopic dermatitis

FLG: Filaggrin

COPSAC: The Copenhagen Prospective Studies on Asthma in Childhood

**Abstract**

Filaggrin null mutations result in impaired skin barrier functions and increase the risk of atopic dermatitis. We wanted to characterize the clinical presentation and course of atopic dermatitis associated with filaggrin null mutations within the first 7 years of life in our prospective, clinical birth cohort study (COPSAC) of 411 children born to mothers with asthma followed with scheduled visits every 6 months as well as visits for acute exacerbations of dermatitis. Atopic dermatitis was defined in accordance with international guidelines and described at every visit using 35 predefined localizations and 10 different characteristics. A total of 170 (43%) of 397 Caucasian children suffered from atopic dermatitis. R501X and/or 2282del4 filaggrin null mutations were present in 26 (15 %) children and were associated with an early age at onset of dermatitis (128 vs. 299 days, $p<0.0001$), a higher number of unscheduled visits (3.6 vs. 2.7; $p=0.04$), more severe (moderate-severe SCORAD 44 % vs. 31%; $p=0.14$) and widespread dermatitis (10 % vs. 6 % of the body area, $p<0.001$), with predilection to exposed areas (hands, feet, extensor areas, cheeks), and lesions characterized by an up regulation in both acute and chronic morphologically. This indicates a filaggrin specific endotype of atopic dermatitis.

**Introduction**

9-10% of the general population carry at least one null mutation in the filaggrin (FLG) gene (Irvine, 2007). Normal gene expression results in intracellular FLG proteins which aggregate keratin filaments leading to keratinocyte compaction and formation of the stratum corneum. The cornified cell envelope is crucial for the skin barrier function as it prevents epidermal water loss and penetration of microbes, toxic chemicals and allergens (Candi *et al*., 2005). Heterozygous, and especially homozygous or compound heterozygous, carriers of FLG null variants such as R501X and 2282del4, may experience dry, scaly and fissured skin (Thyssen and et al, 2011a;Sergeant *et al*., 2009). Furthermore, it has repeatedly been shown that FLG null mutations are major predisposing factors for atopic dermatitis (AD), a condition that affects up to 20% of children (Palmer *et al*., 2006;Marenholz *et al*., 2006;Weidinger *et al*., 2006;Irvine, 2007;1998). FLG null mutations are also associated with early onset of AD, persistence of AD into adulthood and asthma and allergic sensitization (Palmer *et al*., 2006;Marenholz *et al*., 2006;Barker *et al*., 2007;Brown *et al*., 2007;Bonnelykke *et al*., 2010). Clinical studies have shown that most homozygous individuals develop dermatitis very early in life whereas heterozygous individuals may experience a milder course or no symptoms at all (Thyssen and et al, 2011b;Smith *et al*., 2006).

Phenotyping based on FLG null mutations represent a novel endotype with known underlying molecular causes and presumably distinct clinical features and treatment responses. Endotyping of AD is important for segmentation of patients and future investigations of individualized treatment possibilities. Also, it may help us to improve disease prediction and perform research into novel preventive approaches. Phenotypical characterization based on prospective data collection has not been reported previously for the FLG null mutations. Therefore, we meticulously characterized the pattern of atopic dermatitis in the Copenhagen Prospective Birth Cohort during the first 7 years of life and performed stratification by FLG mutation status.

**Results**

Twenty six (15.3 %) of 170 AD children and 15 (7.1 %) of 211 non-AD children were FLG null mutation carriers. The lifetime prevalence of AD was 63 % in children with the FLG null genotype and 43 % in children with the wild type (p= 0.011).

*Severity*

The skin was examined at 905 separate visits of which 276 (30.5 %) were unscheduled visits for acute skin symptoms.

Children with the FLG null mutations had more unscheduled visits than wild type children (mean 3.6 vs. 2.7 visits, p=0.036) (figure 1) and FLG null children presented with higher SCORAD score than wild type children, albeit not statistically significant (44 % vs. 31 % with moderate-severe SCORAD, p=0.14). Similar, non-significant trends were found in SCORAD values at unscheduled and scheduled visits separately (results not shown).

FLG null carriers had earlier onset of dermatitis (median age 128 vs. age 299 days, p<0.0001), but FLG null children did not have more chronic dermatitis judged from the yearly point prevalence ratio (figure 2) and the prevalence by age 7 did not differ significantly (dermatitis in 33% of children with FLG null genotype compared with 24 % of wild type children; p=0.36).

*Anatomical dermatitis localizations*

FLG null children had more widespread dermatitis at each visit compared to wild type children (median 10% vs. 6% of the body, p<0.001). A similar difference was seen at both scheduled and unscheduled visits (results not shown).

The anatomical localization of dermatitis stratified by FLG mutation status (affected per visit) is presented in figure 3. Here, 35 predefined areas were grouped into 11 (for details, please see table S1 in the Supplementary data). The analysis showed that involvement of the palm and back of the

hands, the flexor and extensor extremities, the feet and the cheeks was associated with the FLG null genotype. PLSDA analysis of the original 35 regions confirmed this positive association with a tendency to higher number of different localizations for the FLG null children, as all score values were positive in the first component, mainly driven by affection of the cheeks and the back of the hands, whereas no differences were seen for the flexural area ($p<0.01$) (figure S1 in the Supplementary data). Figure 4 summarizes the localizations that were more often affected in FLG null children (red and blue areas), with the red areas illustrating the localizations significantly selected by the PLSDA (cheeks and back of the hand). To investigate whether the anatomical localizations associated with the FLG null genotype were dependent on age further stratification was performed (0-3 and 3-7 years). However, we found no change in predilection of dermatitis (data not shown).

*Morphology:*

First, three regions were excluded from the analyses due to a low number of registrations (eye area, foot (sole) and nose). Second, two components describing 58% of the total variation were extracted by PCA (component 1 and 2). The PCA loading plot separated chronic signs of dermatitis such as fissures, lichenification and crusting (x-axis) from acute signs of dermatitis such as edema and erythema (y-axis). Characteristics closely positioned exhibited similar pattern, i.e. high scores for erythema track with high scores for edema and vesicles etc (Figure 5a). Third, this information was combined with information about the morphology at the separate skin localizations to make a score plot (figure 5b). At the score plot, the positions in truncated principal component space of the dermatitis lesions are shown. Each region is shown as two points (● FLG wild type, ● FLG null) connected with an arrow from FLG wild type to FLG null. This plot showed that the majority of arrows pointed up or towards the right, i.e. pointing the same directions as the acute and chronic

markers at the loading plot, which means that a general up-regulation of both acute and chronic markers was observed in FLG null children when compared to wild type children. Interpretation of single region differences between the two groups revealed a pattern with certain regions up-regulated in acute markers (extensor areas, extremities flex areas, truncus, hand (palm/back)) and in chronic markers (feet area, wrist (front/back), hand back). At figure 5b the localizations are grouped according to table S1 (Supplementary data). For original plot see figure S2 in the Supplementary data.

**Discussion**

*Main finding*

This study showed that FLG null children with atopic dermatitis represents an endotype characterized by 1) high frequency of dermatitis episodes (as illustrated by more unscheduled visits to our clinic due to dermatitis, 2) more generalized, and 3) severe dermatitis (higher SCORAD) when compared to FLG wild type children with atopic dermatitis. In addition, FLG null children more often had dermatitis at anatomical localizations that tend to be exposed to drying conditions (e.g. wind, cold, sun and radiations from indoor heating), especially the back of the hands and the cheeks (figure 4). Finally, dermatitis lesions were characterized by an up regulation in both acute and chronic markers in FLG null children when compared to wild type children.

*Strengths and limitations*

The COPSAC cohort consists of prospective collection of data including detailed description of dermatitis morphology, anatomical localization, area in percent of the total body surface, and the severity of the dermatitis episode defined by the SCORAD severity score; all collected during the first 7 years of life. The diagnosis, detailed phenotype and management of skin lesions have been controlled solely by the clinical research unit physicians from standard operating procedures and treatment algorithms, and not by the general practitioner or others. Hence, the specificity of the AD diagnosis is high and the risk of misclassification expected to be low. This is of particular importance in the clinical evaluation of AD where inter-observer variation may be a problem(Williams *et al*., 1994). The children were followed every 6 months and in case of acute flare-up of dermatitis. Such prospective data collection reduces the risk of recall bias.

Our study is limited by all children being of Caucasian descent and being high-risk children with mothers suffering from asthma, thereby risking interaction with asthma genetics. Therefore, our

results need replication in an unselected population. Because of the project protocol with the children being followed closely by trained physicians when having acute flare ups, these children are likely to be treated better than the general population. This optimized care can be a confounder diluting the differences between FLG null and wild type children with AD.

We used classic bar charts to illustrate our main findings and confirmed them by multivariate pattern analysis, i.e. PLSDA and PCA. This approach facilitates the handling of many variables in the same analysis, which is especially important for complex phenotypic data. Multivariate pattern analysis makes it possible to study the systematic variation, while filtering out uncorrelated random variation and hence observe patterns not otherwise recognized by traditional univariate statistical analysis. We were able to extract plausible acute and chronic components from the morphological data and characterize different skin regions. In this way, we use a data analytical approach more in line with the clinicians approach to clinical problems, often characterized by pattern recognition rather than any single markers.

*Interpretation*

AD has traditionally been classified as acute vs. chronic, intrinsic vs. extrinsic, associated with ichtyosis vulgaris, based on morphology (nummular, atopic prurigo, lichen planus-like, pityriasis alba) or based on localization (hand, juvenile plantar and palmar, eyelid, cheilitis, nipple, periorificial). However, none of these classifications seem to be satisfying because patients often experience dynamic changes between the suggested categories and with affection of different sites.(Pugliarello *et al*., 2011) We showed that FLG null children had more severe dermatitis characterized by a higher number of visits in the clinical research unit, a more generalized dermatitis and a non-significant trend of higher SCORAD value. Also, FLG null mutations were associated with an earlier onset of disease. These observations are in agreement with previous

cross-sectional reports (Brown *et al*., 2009;Pugliarello *et al*., 2011;Barker *et al*., 2007;Brown *et al*., 2007) and suggest that FLG genotyping should be considered in the initial diagnostic work of patients suspected with AD and that it may be useful for classification.

Our study suggests an association between FLG null carrier status and dermatitis on the hands. In line with this, a recent general population study showed that the FLG null genotype significantly increased the risk of hand dermatitis in individuals with atopic dermatitis and that it was associated with early onset before age 6 years and persistence into adulthood (Thyssen *et al*., 2010b). Furthermore, the clinical observation was recently made that adult patients with the FLG null genotype had a distinct phenotype of hand dermatitis characterized by fissured eruptions on the back of the hands and wrists with only sparse involvement of the palmar aspects, similar to the clinical description of hand dermatitis among atopic individuals previously described by other groups (Thyssen *et al*., 2010a;Simpson *et al*., 2006). It was interesting that we could partly confirm this finding in our cohort of children as FLG null carriers mainly had dermatitis location on the back of the hands. Furthermore we found that skin fissures, but also the other morphological characteristics, were up regulated in FLG null children compared to wild type. Finally, our data suggests that FLG null mutations are associated with dermatitis at other "exposed areas", besides the back of the hands, such as the feet, extensor areas and cheeks. These sites are typically exposed to drying and irritant factors such as water, soap, sun, wind, children playing, changes in temperatures and radiation from indoor heating. It is possible that these may act as a local triggers resulting in dry skin and then dermatitis. Supporting this interpretation, FLG deficient flaky tail mice bred in cages with contact to the environment had a higher level of dermatitis and skin inflammation compared to mice bred in cages with no environmental contact (Fahy *et al*., 2011). In contrast, dermatitis in wild type children was not located on these rather exposed sites. Another

mechanism must be responsible for their dermatitis, e.g. changes in immunomodulatory factors, heat or bacterial growth.

*Conclusion*

Children with AD and FLG null mutation status have a genetically defined endotype, characterized by early onset of dermatitis, a more severe course with more generalized dermatitis resulting in more frequent medical consultations, and with dermatitis often located to exposed areas of the body, in particular hands and cheeks. These findings will hopefully help us to segment patients and promise individualized treatment in the future as well as improved disease prediction and research into novel preventive approaches. However, future studies including replication in an unselected population and bigger cohorts with more FLG homozygotes/compound heterozygotes represented are needed to confirm our phenotype findings.

**Materials and methods**

*Participants*

The Copenhagen Study on Asthma in Childhood (COPSAC) is a prospective, clinical, birth cohort study including 411 children born to mothers with asthma. The study was conducted in accordance with the guiding principles of the Declaration of Helsinki, and approved by The Danish Data Protection Agency (2002-41-2434) and the Ethics Committee for Copenhagen (KF 01-289/96). Data validity was assured by compliance with "Good Clinical Practice" (GCP) guidelines and quality control procedures. Data were collected on-line and locked after external monitoring and an audit trail was run routinely. Informed written consent was obtained from parents. The children were enrolled at one month of age and visited the clinical research unit at scheduled visits every six months as well as for any acute skin symptoms. The main recruiting area of the cohort was greater Copenhagen, Denmark and all children were born between August 1998 and December 2001. The study was previously detailed (Bisgaard *et al*., 2007;Bisgaard, 2004;Bisgaard *et al*., 2006). In this study, we included data from Caucasians only, i.e. 397 of 411 enrolled children. Among these, 172 (43 %) were diagnosed with AD before age 7 years. For 2 children, information about FLG mutation status and registration of dermatitis were missing leaving 170 children for analyses. Two homozygous/compound heterozygous children were grouped with the heterozygous children. Skin examinations, diagnoses and treatment of dermatitis were handled by medical doctors employed for this purpose in the clinical research unit.

*Risk assessments*

AD was stratified by FLG mutation status and the groups were analyzed for differences during the first 7 years of life. AD was defined based on the criteria of Hanifin and Rajka(Hanifin JM and Rajka G, 1980) and FLG genotyping was performed for R501X and 2282del4 (Palmer *et al*., 2006).

At each visit, the following observations were registered:

- **Anatomical localization** of dermatitis lesions were divided into 35 predefined areas: abdomen, ankle (back), ankle (front), back (lower), back (upper), cheek, chest, chin, ear, elbow (back), elbow (front), eye area, foot (back), foot (sole), forearm (back), forearm (front), forehead, hand (back), hand (palm), knee (back), knee (front), lower leg (back), lower leg (front), nappy region, neck (back), neck (front), nose, perioral, scalp, upper arm (back), upper arm (front), upper leg (back), upper leg (front), wrist (back) and wrist (front).
- **Morphology** of the single dermatitis lesion was based on the following characteristics: area (in percent of the total body surface area), erythema, lichenification, crusts, dryness, vesicles, squamation, fissures, edema and excoriations (graduated from 0 (none) to 3 (severe)).
- **Severity** of the dermatitis episode was assessed using the Scoring Atopic Dermatitis index (SCORAD)(1993).

*Statistical analysis*

Associations between FLG mutation status (null vs. wild type) and different variables were investigated using the following analyses: chi-square test (PROC FREQ) (AD diagnosis & age at end of diagnosis), log-rank test (PROC LIFETEST) (age at onset of AD), the non-parametric Wilcoxon Rank-Sum Test (PROC NPAR1WAY) (number of visits in the clinical research unit), and GEE-model (PROC GENMOD) (SCORAD & total area of the body involved pr visit). All analyses were done in SAS version 9.1(SAS Institute Inc, Cary, NC). The overall significance level used was 0.05.

A multivariate approach was employed to detect clinical patterns related to FLG null mutation status. We applied Partial Least Squares Discriminant Analysis (PLSDA) (Barker and Rayens, 2003) to describe the anatomical location of dermatitis and Principal Component Analysis (PCA) to describe dermatitis morphology, in both cases stratifying by FLG mutation status.

*1) Anatomical location:* The number of registrations (continuous) was used as predictors. The pattern of 35 different regions was visualized including how they were associated with the FLG genotype both individually and compared to the other regions. The component consists of a primary *score matrix* with samples distribution followed by a *loading matrix* where the relation between the predictors (regions) was created and displayed.

*2) Dermatitis morphology:* For each region (n=35) and each FLG mutation type (wild type and null), the morphology parameters (erythema, lichenification, crusts, dryness, vesicles, squamation, fissures, edema and excoriations) were represented by a weighted average across all registrations taking the different registration frequencies into account. This resulted in a 70 by 9 matrix. PCA with two components and varimax rotation was then conducted on this matrix to make a *loading* and *score plot,* respectively, visualized by a scatter plot (Kaiser Henry F, 1958).

PLSDA and PCA were conducted using the PLStoolbox ver. 6.0.1 (Eigenvector Inc, Manson, Washington, USA.). In addition inhouse algorithms were used for visualization. All analysis where conducted in Matlab® R2010b version 7.11.0.584.

Additional methodological details are given in the Supplementary Materials and Methods.

**Conflict of interest**

The authors state no conflict of interest.

**Acknowledgements**

The authors wish to thank the children and parent participating in the COPSAC cohorts as well as the COPSAC study teams.

**Reference List**

Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. Dermatology 186:23-31 (1993).

Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. Lancet 351:1225-1232 (1998).

Barker JN, Palmer CN, Zhao Y, Liao H, Hull PR, Lee SP, Allen MH, Meggitt SJ, Reynolds NJ, Trembath RC, McLean WH: Null mutations in the filaggrin gene (FLG) determine major susceptibility to early-onset atopic dermatitis that persists into adulthood. J Invest Dermatol 127:564-567 (2007).

Barker M, Rayens W: Partial least squares for discrimination. J Chemometrics 17:166-173 (2003).

Bisgaard H: The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. Ann Allergy Asthma Immunol 93:381-389 (2004).

Bisgaard H, Hermansen MN, Buchvald F, Loland L, Halkjaer LB, Bonnelykke K, Brasholt M, Heltberg A, Vissing NH, Thorsen SV, Stage M, Pipper CB: Childhood asthma after bacterial colonization of the airway in neonates. N Engl J Med 357:1487-1495 (2007).

Bisgaard H, Hermansen MN, Loland L, Halkjaer LB, Buchvald F: Intermittent inhaled corticosteroids in infants with episodic wheezing. N Engl J Med 354:1998-2005 (2006).

Bonnelykke K, Pipper CB, Tavendale R, Palmer CN, Bisgaard H: Filaggrin gene variants and atopic diseases in early childhood assessed longitudinally from birth. Pediatr Allergy Immunol 21:954-961 (2010).

Brown SJ, Relton CL, Liao H, Zhao Y, Sandilands A, McLean WH, Cordell HJ, Reynolds NJ: Filaggrin haploinsufficiency is highly penetrant and is associated with increased severity of eczema: further delineation of the skin phenotype in a prospective epidemiological study of 792 school children. Br J Dermatol 161:884-889 (2009).

Brown SJ, Sandilands A, Zhao Y, Liao H, Relton CL, Meggitt SJ, Trembath RC, Barker JN, Reynolds NJ, Cordell HJ, McLean WH: Prevalent and Low-Frequency Null Mutations in the Filaggrin Gene Are Associated with Early-Onset and Persistent Atopic Eczema. J Invest Dermatol (2007).

Candi E, Schmidt R, Melino G: The cornified envelope: a model of cell death in the skin. Nat Rev Mol Cell Biol 6:328-340 (2005).

Fahy CMR, McLean WHI, Irvine AD: Variation in the development of phenotype in filaggrin-deficiency: a murine model of atopic dermatitis 2011).

Hanifin JM, Rajka G: Diagnostic features of atopic dermatitis. Acta Derm Venereol 92:44-47 (1980).

Irvine AD: Fleshing out filaggrin phenotypes. J Invest Dermatol 127:504-507 (2007).

Kaiser Henry F: The varimax criterion for analytic rotation in factor analysis. Psychometrica 23:187-200 (1958).

Marenholz I, Nickel R, Ruschendorf F, Schulz F, Esparza-Gordillo J, Kerscher T, Gruber C, Lau S, Worm M, Keil T, Kurek M, Zaluga E, Wahn U, Lee YA: Filaggrin loss-of-function mutations predispose to phenotypes involved in the atopic march. J Allergy Clin Immunol 118:866-871 (2006).

Palmer CN, Irvine AD, Terron-Kwiatkowski A, Zhao Y, Liao H, Lee SP, Goudie DR, Sandilands A, Campbell LE, Smith FJ, O'Regan GM, Watson RM, Cecil JE, Bale SJ, Compton JG, DiGiovanna JJ, Fleckman P, Lewis-Jones S, Arseculeratne G, Sergeant A, Munro CS, El HB, McElreavey K, Halkjaer LB, Bisgaard H, Mukhopadhyay S, McLean WH: Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. Nat Genet 38:441-446 (2006).

Pugliarello S, Cozzi A, Gisondi P, Girolomoni G: Phenotypes of atopic dermatitis. J Dtsch Dermatol Ges 9:12-20 (2011).

Sergeant A, Campbell LE, Hull PR, Porter M, Palmer CN, Smith FJ, McLean WH, Munro CS: Heterozygous null alleles in filaggrin contribute to clinical dry skin in young adults and the elderly. J Invest Dermatol 129:1042-1045 (2009).

Simpson EL, Thompson MM, Hanifin JM: Prevalence and morphology of hand eczema in patients with atopic dermatitis. Dermatitis 17:123-127 (2006).

Smith FJ, Irvine AD, Terron-Kwiatkowski A, Sandilands A, Campbell LE, Zhao Y, Liao H, Evans AT, Goudie DR, Lewis-Jones S, Arseculeratne G, Munro CS, Sergeant A, O'Regan G, Bale SJ, Compton JG, DiGiovanna JJ, Presland RB, Fleckman P, McLean WH: Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. Nat Genet 38:337-342 (2006).

Thyssen JP, Carlsen BC, Johansen JD, Meldgaard M, Szecsi PB, Stender S, Menne T: Filaggrin null-mutations may be associated with a distinct subtype of atopic hand eczema. Acta Derm Venereol 90:528 (2010a).

Thyssen JP, Carlsen BC, Menne T, Linneberg A, Nielsen NH, Meldgaard M, Szecsi PB, Stender S, Johansen JD: Filaggrin null mutations increase the risk and persistence of hand eczema in subjects with atopic dermatitis: results from a general population study. Br J Dermatol 163:115-120 (2010b).

Thyssen JP, et al: Filaggrin mutation R501X and 2282del4 carrier status is associated with fissured skin on the hands: results from a cross-sectional population study. Submitted to Br J Dermatol (2011a).

Thyssen JP, et al: Individuals who are homozygous for the 2282del4 and R501X filaggrin null mutations do not always develop dermatitis and complete long-term remission is possible. Accepted for publication in Journal of the European Academy of Dermatology & Venereology (2011b).

Weidinger S, Illig T, Baurecht H, Irvine AD, Rodriguez E, az-Lacava A, Klopp N, Wagenpfeil S, Zhao Y, Liao H, Lee SP, Palmer CN, Jenneck C, Maintz L, Hagemann T, Behrendt H, Ring J, Nothen MM, McLean WH, Novak N: Loss-of-function variations within the filaggrin gene predispose for atopic dermatitis with allergic sensitizations. J Allergy Clin Immunol 118:214-219 (2006).

Williams HC, Burney PG, Strachan D, Hay RJ: The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. II. Observer variation of clinical diagnosis and signs of atopic dermatitis. Br J Dermatol 131:397-405 (1994).

**Figure legends**

**Figure 1:** Frequency of unscheduled visits in the clinical research unit in relation to the FLG null mutations.

**Figure 2:** The yearly point prevalence ratio of dermatitis cases observed at children having the FLG null mutations compared to the number of dermatitis children with the FLG wildtype.

**Figure 3:** Frequency of localizations in relation to the number of visits at the clinical research unit, 0-7y, grouped.

**Figure 4:** The figure summarizes the localizations more often affected in children with FLG null mutations compared to wild type children (red and blue areas), with red areas illustrating the localizations specifically selected by PLSDA as driving sites.

**Figure 5a:** PCA loading plot for the first two varimax rotated principal components. Morphological characteristics such as fissures, lichenification and crust were grouped at the x-axis (component 1, chronic markers), whereas edema, erythema and vesicles were grouped at the y-axis (component 2, acute markers).

**Figure 5b:** PCA score plot showing the morphology of the individual localizations. Each region is shown as two points (● FLG wild type, ● FLG null) connected with an arrow from FLG wild type to FLG null. E.g. hand (back) obtained higher values with respect to chronically inflammatory markers than e.g. forearm (front). A general up-regulation of both acute and chronic markers was observed for FLG null children across almost all regions.

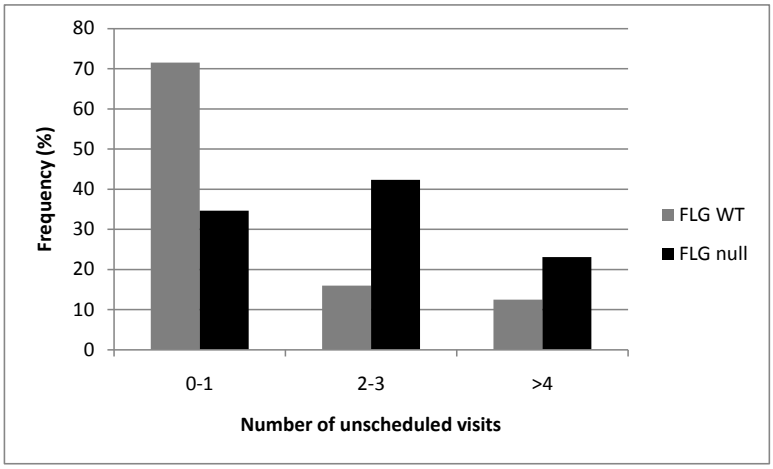

80
70
60
50
40
30
20
10
0
Frequency (%)
0-1
2-3
>4
Number of unscheduled visits
FLG WT
FLG null

Figure 1

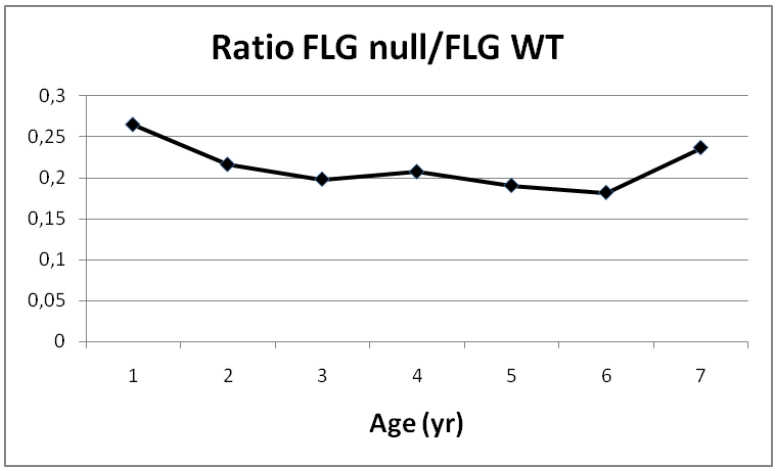



**Figure 2**

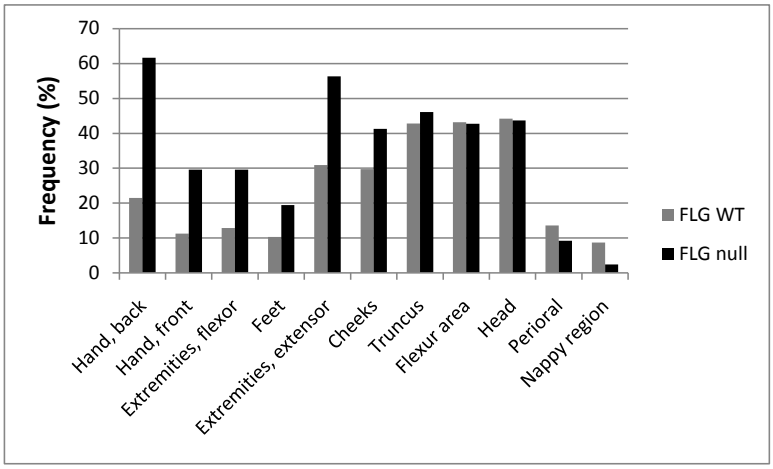
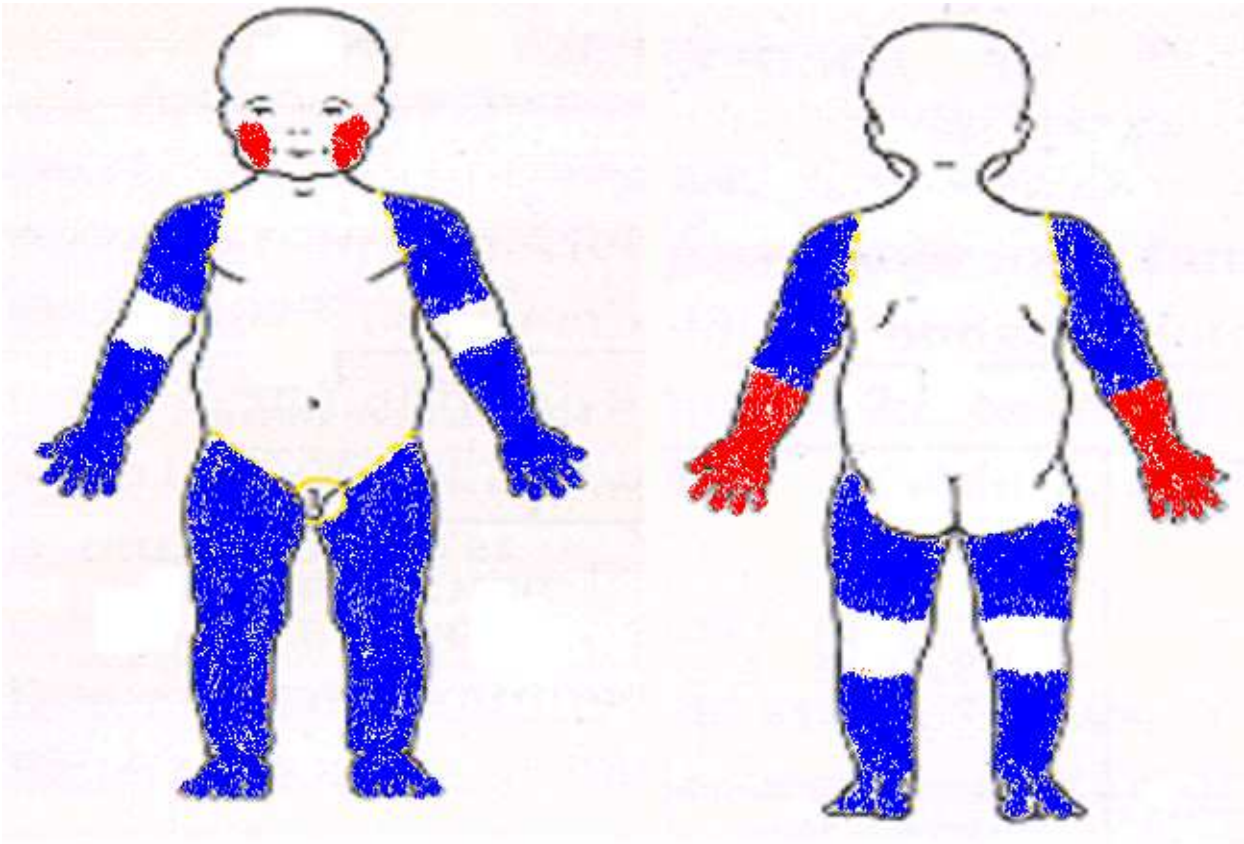

Frequency (%)
70
60
50
40
30
20
10
0
Hand, back
Hand, front
Extremities, flexor
Feet
Extremities, extensor
Cheeks
Truncus
Flexur area
Head
Perioral
Nappy region
FLG WT
FLG null

**Figure 3**



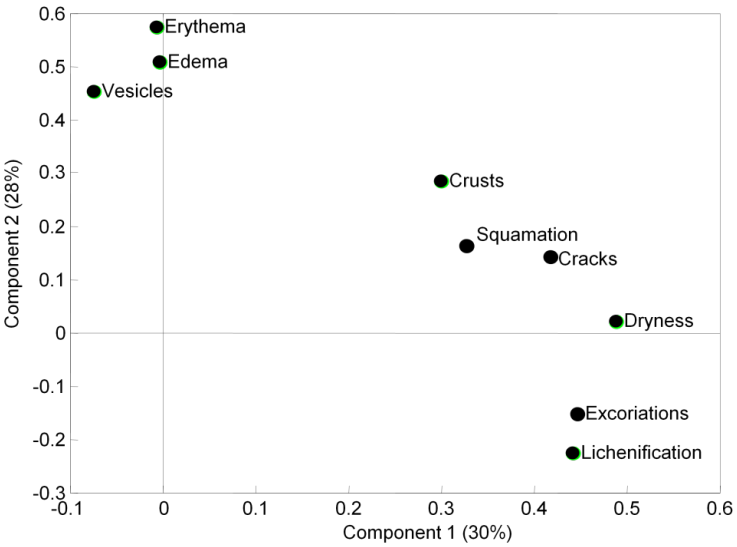

Erythema
Edema
Vesicles
Crusts
Squamation
Cracks
Dryness
Excoriations
Lichenification
Component 2 (28%)
Component 1 (30%)
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 5a

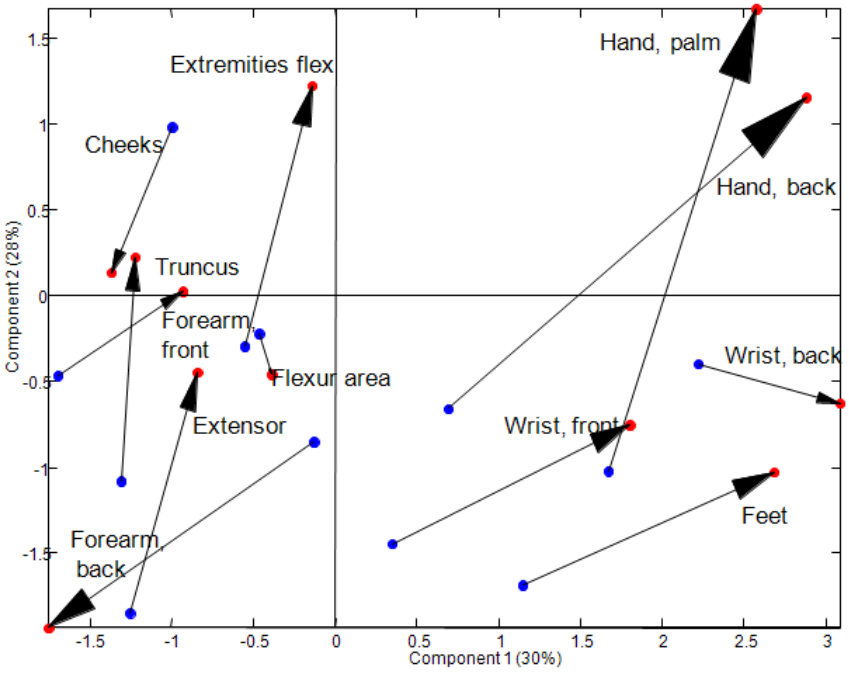

Extremities flex
Cheeks
Truncus
Forearm, front
Flexur area
Extensor
Forearm, back
Hand, palm
Hand, back
Wrist, back
Wrist, front
Feet
Component 2 (28%)
Component 1 (30%)
1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3

**Figure 5b**

**Endotyping Atopic Dermatitis Children by Filaggrin Gene Mutation Status in a Prospective Cohort Study.**

**"Supplemental material"**

Charlotte Giwercman Carson[1], Morten Arendt Rasmussen[2], Jacob P. Thyssen[3], Torkil Menné[3], Hans Bisgaard[1]

[1]Copenhagen Prospective Studies on Asthma in Childhood; COPSAC, Health Sciences, University of Copenhagen, Copenhagen University Hospital, Gentofte, Ledreborg Alle 34,Gentofte 2820,Copenhagen, Denmark

[2]Faculty of Life Sciences, University of Copenhagen, DK-1870 Frederiksberg, Denmark

[3]National Allergy Research Centre, Department of Dermato-Allergology, Copenhagen University Hospital Gentofte, DK-2900 Hellerup, Denmark

Address for correspondence:

Copenhagen Prospective Studies on Asthma in Childhood

Danish Pediatric Asthma Center

Copenhagen University Hospital, Gentofte

Ledreborg Alle 34

DK-2820 Gentofte

Copenhagen

Denmark

Tel: +45 39 77 73 60

Fax: +45 39 77 71 29

E-mail: bisgaard@copsac.dk

Website: www.copsac.com

## Materials and methods

*Statistical analyses*

We applied Partial Least Squares Discriminat Analysis (PLSDA) and principal component analysis (PCA) which may be regarded as extensions of the multiple linear regression model (y=ax+b). In the latter model, one can make reasonable estimates (predictions) for new observations based on the linear relationship between variables (e.g. between height and weight). The PLS is to some degree related to the linear regression model as it also aims to identify a linear model based on observed and expected values, but in contrast, these variables are projected into a new space. The PLS model can be regarded as a dimension reduction approach that is coupled with a regression model. It is typically used in situations where the matrix of predictors has more variables than observations or when there is multicollinearity among the x-values (this is when two or more predictor variables are highly correlated). The PLS model has for these reasons also been used for analysis of high-dimensional genomic data where expression of thousands of genes is evaluated. In general, PLS can be regarded as a robust statistical model suited for a large number of variables and which also takes into account the interdependencies that may exist among the variables.

**Tables and Figures**

| **Truncus** | **Flexur area** | **Hand area, back** | **Hand area, front** | **Head** | **Cheeks** | **Extremities, extensor** | **Nappy region** | **Perioral** | **Feet** | **Extremities flexor** |
|---|---|---|---|---|---|---|---|---|---|---|
| Abdomen | Elbow, front | Hand, back | Hand, palm | Chin | Cheek | Elbow, back | Nappy region | Perioral | Foot, back | Lower leg, back |
| Back, lower | Knee, back | Wrist, back | Wrist, front | Ear | | Knee, front | | | Foot, sole | Upper arm, front |
| Back, upper | | Forearm, back | Forearm, front | Eye area | | Lower leg, front | | | Ankle, back | Upper leg, back |
| Chest | | | | Forehead | | Upper arm, back | | | Ankle, front | |
| | | | | Nose | | Upper leg, front | | | | |
| | | | | Scalp | | | | | | |
| | | | | Neck, back | | | | | | |

**Table S1:** Grouping the 35 predefined localizations into 11 groups.

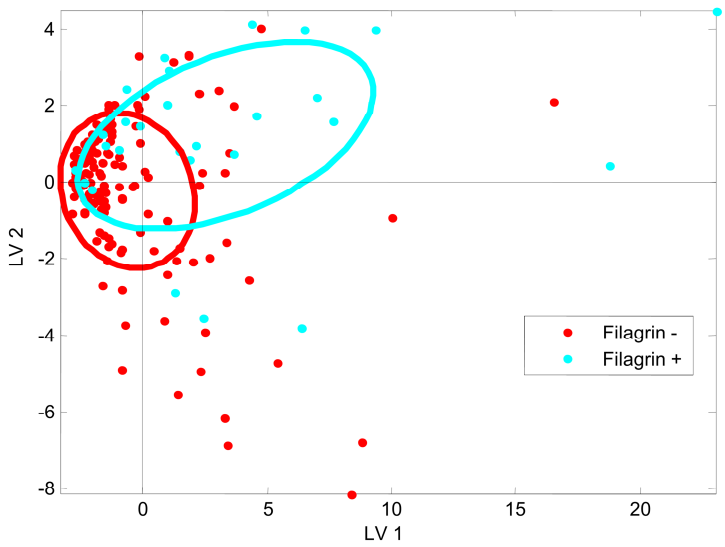



**Figure S1A: PLSDA score plot**

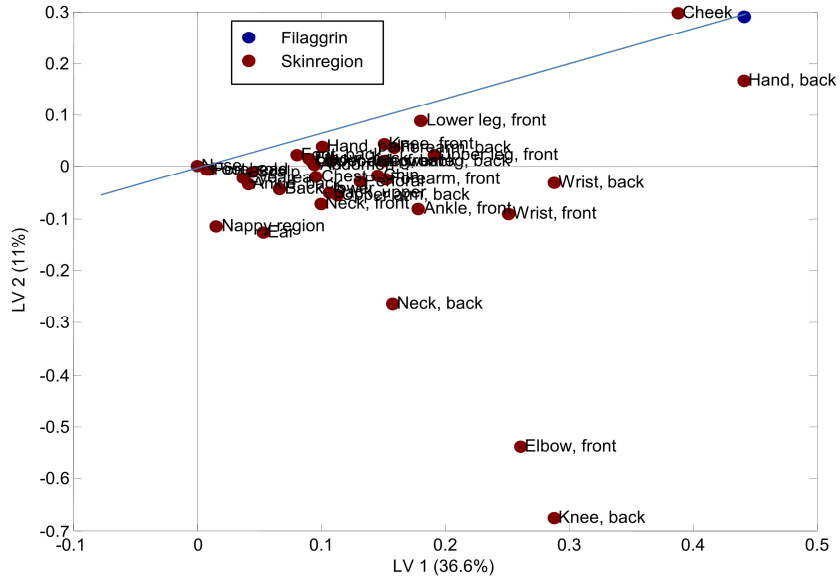



**Figure S1B: PLSDA loading plot**

PLSDA score- (**A**) and loading (**B**) plot for having dermatitis on a given localization in relation to number of visits in the clinic. Ellipsoids are centered at population mean with half axis corresponding to the standard deviation and under normality assumption hence cover ~50% of data.

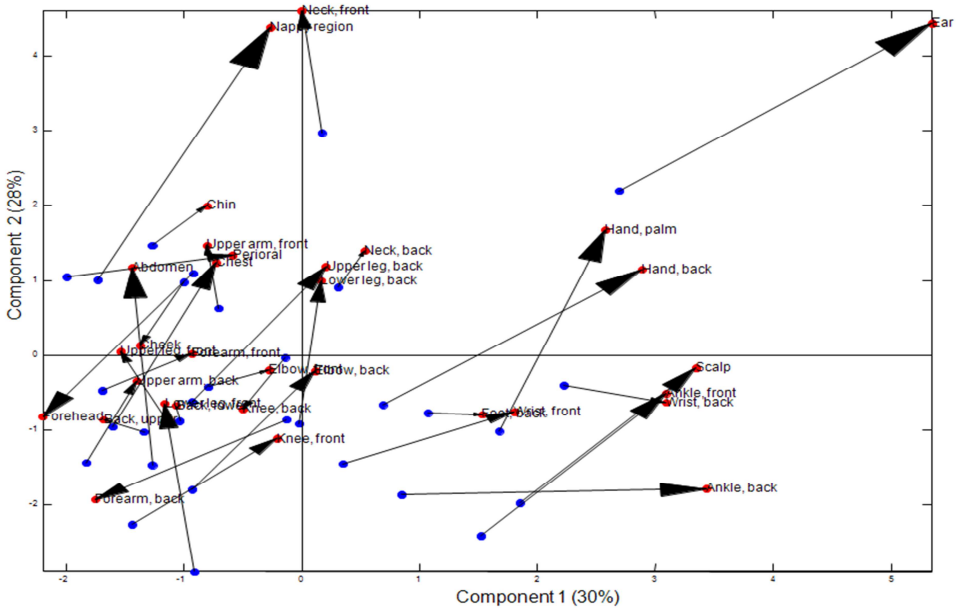



**Figure S2:** Original PCA score plot showing the morphology of the individual localizations. Each region is shown as two points (● FLG wild type, ● FLG null) connected with an arrow from FLG wild type to FLG null. E.g. hand (back) obtained higher values with respect to chronically inflammatory markers than e.g. forearm (front). A general up-regulation of both acute and chronic markers was observed for FLG null children.

# Paper VII

No genetic footprints of the fat mass and obesity associated (FTO) gene in human plasma $^{1}$H CPMG NMR metabolic profiles

Karin Kjeldahl, Morten A. Rasmussen, Annelouise Hasselbalch, Kirsten O. Kyvik, Lene Christiansen, Emma Per-Trepat, Serge Rezzi, Sunil Kochhar, Torkild I. A. Sørensen, Rasmus Bro

*in preparation*

**Metabolomics manuscript No.**
(will be inserted by the editor)

# No genetic footprints of the fat mass and obesity associated (FTO) gene in human plasma $^1$H CPMG NMR metabolic profiles.

**K. Kjeldahl · M. A. Rasmussen · A. L. Hasselbalch · K. O. Kyvik · L. Christiansen · E. Per-Trepat · S. Rezzi · S. Kochhar · T. I. A. Sørensen · R. Bro**

Received: date / Accepted: date

**Abstract** In this paper it was investigated if any genotypic footprints from the fat mass and obesity associated (FTO) SNP could be found in 600 MHz $^1$H CPMG NMR profiles of around 1000 human plasma samples from healthy Danish twins. The problem was addressed with a combination of univariate and multivariate methods. The NMR data was substantially compressed using Principal Component Analysis (PCA) or Multivariate Curve Resolution (MCR) with focus on chemically meaningful feature selection reflecting the nature of chemical signals in an NMR spectrum. The possible existence of an FTO signature in the plasma samples was investigated at the subject level using supervised multivariate classification in the form of Extended Canonical Variate Analysis (ECVA), classification tree modeling (CART) and Lasso (L1) regularized linear logistic regression model (GLMNET). Univariate hypothesis testing of peak intensities was used to explore the genotypic effect on the plasma at the population level. The multivariate classification approaches indicated poor discriminative power of the metabolic profiles whereas univariate hypothesis testing provided seven spectral regions with $p < 0.05$. Applying false discovery rate (FDR) control, no reliable markers could be identified, which was confirmed by test set validation. We conclude that it is very unlikely that an FTO-correlated signal can be identified in these $^1$H CPMG NMR plasma metabolic profiles and speculate that high-throughput un-targeted genotype-metabolic correlations will in general be a difficult path to follow.

**Keywords** FTO · CPMG · Data compression · ECVA · Lasso · CART

K. Kjeldahl · M. A. Rasmussen · R. Bro
Univ Copenhagen, Dept Food Sci, Fac Sci, DK-1958 Frederiksberg C, Denmark
Tel.: +45-35333197
Fax: +45-35333245
E-mail: mortenr@life.ku.dk

A. L. Hasselbalch · T. I. A. Sørensen
Inst of Preventive Medicine, Øster Søgade 20, DK-1357 Copenhagen, Denmark

## 1 Introduction

The quantitative genetic contribution to body mass index variation and hence to obesity is well established [32,34] and has been confirmed in the population used for the present study [26]. Several obesity candidate genes have been discovered, among which the most consistent associations have been found between body weight and single nucleotide polymorphisms (SNP) in the fat mass and obesity associated (FTO) gene [8,20,17,**?**]

The FTO gene is located on chromosome 16. To date, the locus with the strongest association with obesity is the rs9939609 [32], which is a cluster of 10 SNP located in the first intron of the gene. The verification of the association between the FTO gene and body mass strongly supports the suggestion that this gene has common variants (the AA and AT genotypes) that predispose to obesity, relative to the wild (TT) genotype. Varying effect sizes of the FTO locus have been reported [19] but a number of studies states an effect size in the area of 0.35 kg/m2 (0.1 z-score units for BMI) per susceptibility allele.

Expression studies indicate that *FTO* is widely expressed in many tissues, but has its highest expression in the brain, particularly the arcuate nucleus of the hypothalamus [8], where it is believed to be involved in energy uptake rather than energy expenditure [14,**?**]. The findings of Wardle [33] that the *FTO* risk allele was associated with reduced satiety responsiveness in children support this putative functional

role, whereas another study [13] found no association between *FTO* and increased energy intake or food preferences. The study of Wahlen et al [31] indicate a role of *FTO* in fat cell lipolysis.

Gerken *et al*. [11] suggested that *FTO* catalyzes demethylation of 3-methylthymine in DNA, with concomitant production of succinate, formaldehyde, and carbon dioxide, but a direct functional role of the *FTO* gene in obesity development remains unsolved. It is likely that the *FTO* variants are in linkage disequilibrium with the true causative variant [24].

As part of the Danish nationwide GEMINAKAR study which took part 1997-2000 fasting blood samples were collected from healthy Danish twins and analyzed for a number of constituents. Genotyping with respect to the *FTO* locus rs9939609 was also conducted for a subset of the twins, and at a later stage, Nestlé (Lausanne, Switzerland) subjected the plasma to $^1$H NMR analysis for metabolomics studies.

The purpose of the present study was to combine these two datasets and investigate whether the *FTO* (rs9939609) genotype is reflected in the blood composition as measured with 600 MHz $^1$H NMR.

In this paper, we attempted to identify metabolic associations between *FTO* polymorphism and metabolic signatures through the interrogation of 1H CPMG NMR generated metabolic profiles with a combination of multivariate and univariate statistics.

## 2 Materials and Methods

The GEMINAKAR study regarded the relative influence of environmental and genetic factors on especially the metabolic syndrome, and was based on data from 756 healthy Danish twin pairs. The details of the GEMINAKAR study are described elsewhere [26,27,5,12]. The blood samples were collected during the years 1997-2000, added NaF as an anticoagulant and preservative and stored at $-80\,^{\circ}$C until 2005 when the plasma was subjected to NMR analysis, whereby individual fasting metabolic profiles were obtained. SNP data was obtained from the same blood samples. Combined NMR and SNP data was available for 1116 individuals.

Data handling and analysis was performed using the commercial software package MATLAB®, ver. 7.10.0 including PLS_Toolbox ver. 5.5.1.

### 2.1 NMR profiles

*NMR Acquisition*

For each plasma sample, a Free Induction Decay (FID) was acquired in 1.98 seconds on a Bruker DRX 600 MHz NMR instrument using a 5 mm probe (Bruker Biospin, Rheinstetten, Germany). The spin-echo pulse sequence was standard Carr-Purcell-Meiboom-Gill (CPMG) [21]. With CPMG the fast relaxation of protons in macromolecules (short T2) is used to filter particularly those signals out and leave peaks from small molecules or signals from molecules with significant segmental motion [3].

*NMR preprocessing*

Prior to Fourier transformation, NMR FIDs were multiplied by an exponential weighting function corresponding to a line broadening of 1.0 Hz followed by manual correction for phase and baseline distortions. These actions were performed using the Topspin software (ver. 2.1, Bruker Biospin, Rheinstetten, Germany).

The spectra were calibrated against the alpha-glucose doublet peak at 5.23 ppm and interpolated to a common ppm axis. The spectral range 0.1-8.25 ppm was kept for further analysis with exclusion of the water band at 4.55-5.17 ppm. No normalization was applied. Each spectrum then consisted of 29027 data points. 85 spectra were excluded due to poor data quality (poor water suppression or shimming) or due to obvious signals from ethanol or drugs.

The very high number of variables represents a computational challenge, but may potentially also hamper the data analysis adversely by substantial power reduction and by presence of spurious correlations in data. It is therefore desirable to reduce the data in a reasonable way. An integration approach was applied in the following way:

1. Define $K$ spectral regions, if possible so that each region contains only one peak. Omit peaks which one would not trust if they turn out to be significant, i.e. peaks with very weak intensities or regions with many overlapping small signals. Obvious spin-spin coupling splits (i.e. doublets, triplets etc), with no interfering peaks should be put in same region, but emphasis is rather on representing all informative signals than taking care of structural information. For each peak region, $j$:
   (a) Align peaks using the *I*coshift tool [25]. *I*coshift shifts the peaks of each region individually and preserves the peak shape. Possibly apply Savitsky-Golay differentiation to the spectra prior to alignment for improved alignment and apply the resulting spectral adjustment to the original spectra.
   (b) If the peak is a small signal on a shoulder of large broad parent peak, model the parent peak as baseline and subtract this from the small peak.
   (c) Decompose region by Principal Component Analysis (PCA) [16] or Multivariate Curve resolution (MCR) [28] using a reasonable number of components $l$, $l \in 1,2$ determined manually for each region. For

**Table 1** Distribution of FTO genotypes in this study

| Genotype | number of subjects | % |
|---|---|---|
| TT | 347 | 34% |
| AT | 479 | 47% |
| AA | 197 | 19% |

improved robustness, include only the approx. 80% most normal samples as determined by initial PCA submodels for each class (AA, AT, TT) for the modeling, but eventually project all samples onto the regional PCA or MCR model. The choice of method (PCA or MCR) was made by comparison of the spectra with the obtained loadings; the shape of the loadings should ideally only model the peak of interest in the region and not artefacts arising from interfering signals. After this step, each region is represented by $l$ score variables.

2. Collect scores to form the new matrix $\mathbf{X}$.

Eventually, three extreme outliers were removed by use of PCA diagnostics from a model on $\mathbf{X}$, resulting in a dataset consisting of 1028 samples and 171 variables originating from 164 peak regions. The distribution of the three FTO genotypes is shown in Table 1.

## 2.2 SNP data

We genotyped the FTO SNP (rs9939609) by conducting allelic discrimination using pre-designed Taqman® SNP genotyping assays (Applied Biosystems). The conditions described by the manufacturer were applied. We performed PCR in the ABI Prism 7700 and we analyzed PCR using the Sequence Detection System software (Applied Biosystems).

## 2.3 Data analysis

Immediately after the preprocessing, the dataset was split in a training set of 800 samples and a test set of 228 samples by Kennard-Stone subset selection [18,7] with the constraint that siblings should be put in the same set. The two datasets had comparable composition with respect to FTO classes.

## 2.4 Distribution of variation

The distribution of variation sources was estimated for each variable using simple ANalysis Of VAriance (ANOVA), with gender as categorical and age as linear effects. For the total variable space simple means of the contributions to the individual variables were used.

*Data quality*

An initial confirmation of data quality of both full spectra and the compressed data was performed by (a) modeling of age and (b) visual verification that multivariate gender differences were present. Age was modeled by Partial Least Squares (PLS) regression and gender differences were visualized via PCA.

The relation between the metabolomic and the genotypic profiles was investigated both uni- and multivariately:

1. Multivariate classification on the basis of all peak variables.
2. Univariate significance testing of spectral features.

*Multivariate modeling*

Multivariate classification was performed using Extended Canonical Variate Analysis (ECVA) [22]. Furthermore a classification tree model (CART) [6] and a Lasso (L1) regularized linear logistic regression model (GLMNET) [9,10] were fit, but since neither of these two produced meaningful models for these data, they are not treated further here.

The ECVA modeling aims to separate the two classes by seeking new directions which separate the variation between the classes relative to the variation within the classes. This is a supervised method, and consequently there is a risk of over-fitting in the sense that optimistically good separation may be found. For the ECVA modeling, only AA and TT samples were included to simplify the problem to a two-class problem with only the most extreme groups.

A major advantage of the applied data reduction is that data is essentially noise free, and thus up-weighting of small peaks is not associated with noise boosting. Consequently, data for the ECVA models was autoscaled, i.e. each variable was mean-centered and scaled to unit variance. Prior to this, the compressed spectral data were adjusted for age and gender variation as described below.

The performance of the ECVA modeling was assessed by a combination of cross-validation and permutation test. Specificity and sensitivity were first assessed by cross-validation where the samples were split randomly into 13 segments where each segment conserved the class distribution. This was repeated 200 times, whereby 200 (specificity, sensitivity)-pairs were obtained. To assesses if the obtained performance was better than random, modeling and cross-validation were repeated 1000 times with a shuffled class membership.

*Univariate significance*

Each of the 164 individual spectral regions was tested in parallel. A test statistic, $p_j$ $(j = 1,..,164)$, was obtained for each spectral region by comparing two nested generalized

linear models predicting the occurrence of A in the FTO locus via logistical regression.

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{ij}\beta + \varepsilon_{ij} \quad (1)$$

with a pure intercept spectral independent model:

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \varepsilon_i \quad (2)$$

where $p_i$ refers to the probability of locus A for the $i$'th person ($i = 1,...,n$).

The deviance ($-2*log$ likelihood) between the two models $dev_{0j} - dev_{1j}$ is assumed $\chi^2(df_j)$ with $df_j = df_0 - df_{1j}$. $x_{ij}$ is the derived scores for person $i$ spectral region $j$ corrected for age and gender information (see handling of covariates below).

Thus a significance test was produced for each spectral region ($p_1,..,p_{164}$), such that the minimum p-value refers to the most interesting spectral region. The 164 spectral regions were split into a set of significant regions and a set of non significant regions using the method of Benjamini-Hochberg [4] for control of false discovery rate (FDR).

The selected spectral regions suffer from selection bias, that is; the most significant regions are selected due to a combination of true effect size, but also by chance. This bias is known as Winner's curse [35]. The spectral regions were investigated for selection consistency by applying a non parametric bootstrap procedure evaluating the frequency of selection for individual regions at different false discovery rate (FDR) settings.

*Handling of covariates*

The plasma profiles systematically reflect gender and age, hence a priori correction for those was appropriate. Let **D** be a design matrix corresponding to gender and age, such that the first and second column are dummy vectors for male and female respectively and the third column is age. Linear correction for both age and gender can be done by orthogonalization with respect to **D**:

$$\mathbf{X}_{corr} = \left(\mathbf{I} - \mathbf{D}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\right)\mathbf{X} \quad (3)$$

It is assumed that the FTO is independent of gender and age. Nevertheless the collected data might exhibit a small partial confounding with either gender or age. Under such circumstances (3) is an overcorrection, and subsequent models will suffer from that. In order to remove information *only* related to gender and age, while retaining all information related to FTO, the design matrix **D** is projected onto the null space of FTO. Let **F** be an FTO design matrix with three columns corresponding to TT, AT and AA. For the ECVA models only TT and AA columns were included.

$$\mathbf{D}^* = (\mathbf{I} - \mathbf{F}(\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}})\mathbf{D} \quad (4)$$

$\mathbf{D}^*$ is then used for correction of $\mathbf{X}$ in equation (3).

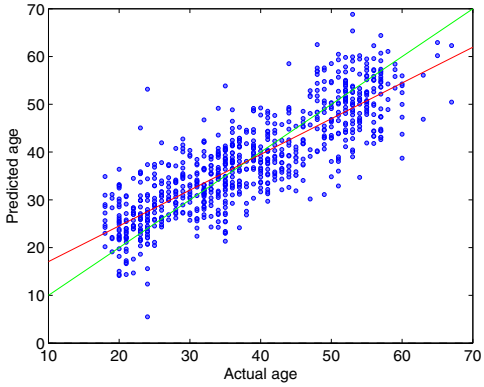



**Fig. 1** Prediction of subject age using compressed data. Green line: ideal prediction fit, red line: obtained cross-validation prediction fit. Root mean square error of test set prediction ($RMSEP = 6.8$ years)

## 3 Results

### 3.1 Data quality

PLS models of subject age based on (a) raw spectra (without water peak) and (b) compressed data showed good performance in both cases. Fig. 1 shows the performance of the model with the compressed data, similar results were found for the full-spectrum model. Model complexity was estimated using 6-fold cross-validation on the training set in both models, and performance was assessed using the test set. The two models showed comparable root mean squared error of prediction (RMSEP) at 6.5 and 6.8 years respectively.

Gender differences could also be recognized using PCA for both data sets. Fig. 2 shows a scores plot for the compressed data set, where it is obvious that gender differences are found in this data set.

Examination of the distribution of variation sources across all variables reveals that both age and gender at most contribute with 1% variance for single variables, and 0.1% across all variables (results not shown).

### 3.2 Multivariate classification

ECVA classification models were built to separate the two extreme genotypes (TT, AA). Cross-validation estimated a mean sensitivity of 0.42 and specificity of 0.70. These values are quite poor, and it is not obvious whether they are better than random. The permutation test (Fig. 3) showed that the obtained performance appear to be slightly better than random, but in no way convincing.

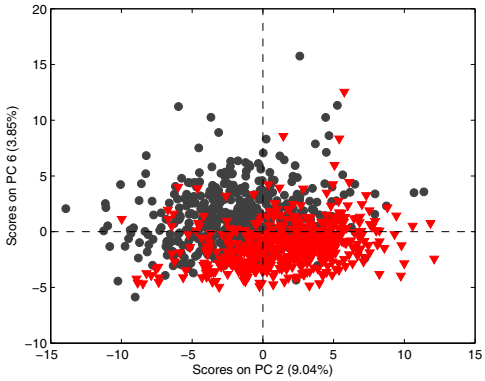



**Fig. 2** Gender differences are present in compressed data. Visualized by PCA score plot

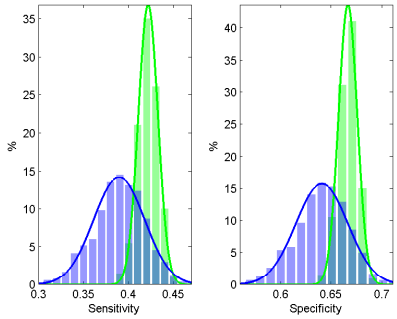



**Fig. 3** Sensitivity (left) and specificity (right) obtained by multivariate ECVA models (green), compared to permutation test (blue)

### 3.3 Univariate significance

Each of the 164 spectral regions was tested one by one, resulting in p-values ranging from 0.028 up to 0.991. Regions were found to be significant with an FDR threshold $q$ of 0.2, i.e. allowing 20% of the positive findings to be false discoveries. Ignoring the multiple testing correction, seven regions were found to have $p \leq 0.05$. The positions of these in the NMR spectrum are shown in Fig. 4. The p-values and the FDR control were bootstrapped and by this procedure, the frequency of "winning" (i.e. being selected as significant) was assessed for each of the variables. The seven variables were selected as significant at a frequency between 44% and 56%, which is not overwhelming, but considerably higher than the average of 20%.

The validity of the seven regions was evaluated with the test set of 223 samples. The mean values of each of the seven regions are shown in Fig. 5, and it is obvious that the results

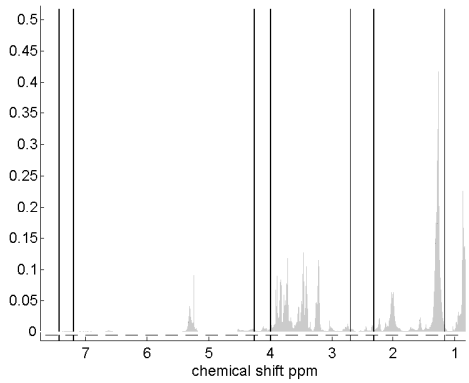



**Fig. 4** NMR regions with $p \leq 0.05$

from the training set are not valid for the test set. The result of region 44 is the only one where effect size and pattern across genotypes are comparable between the training and the test set. In fact region 65, 151 and 161 display completely opposite effect for the two sets. Closer investigation of the distribution of intensities of region 44 showed that differences were so small that they were of no value for any purpose (not shown).

## 4 Discussion

It is interesting to investigate possible markers of the FTO gene in NMR based plasma profiles. Such markers would be interesting candidates for further investigations. The results show that we have not been able to identify reliable plasma markers of the FTO genotype when we took care of multiple testing by Benjamini-Hochberg FDR control. The use of a test set illustrated that candidate markers with $p < 0.05$ had very little validity. This underlines that multiple testing situations require careful handling of p-values and that the use of test sets is recommendable.

Does that mean that the FTO gene does not leave any footprint in a blood sample? There may be several reasons why we did not make any positive findings in this study. We shall here address (a) the quality of the data, (b) the data analysis and (c) the overall research questions.

The data quality check indicated that the substantial data compression from 29027 to 171 variables did not result in massive information loss since the information in the compressed data was adequate to model a highly complex attribute such as age, which was found to contribute with less than 1% of the variation in the 171 variables. One could consider to include more than 164 spectral regions as many

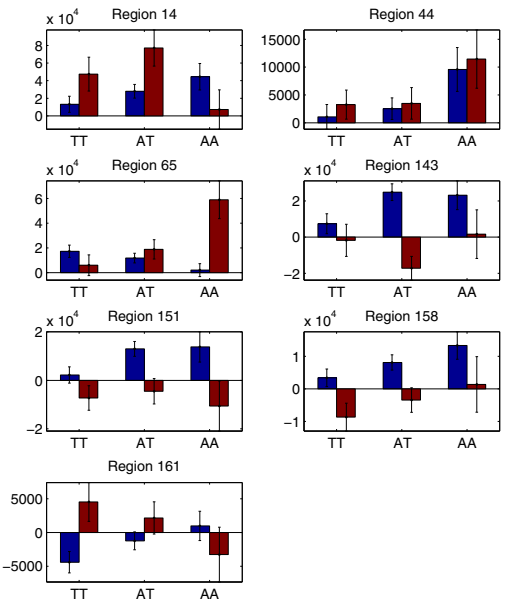



**Fig. 5** Mean values (with standard errors of the mean) of the seven regions with $p \leq 0.05$ for adjusted training (blue) and test set (red) data.

more are definitely present, but visual inspection of the remaining spectral regions made us very uncertain about concluding anything about these as most of them were close to the noise limit or weak shoulder peaks difficult to align etc.

Low sensitivity is an intrinsic problem in NMR and as a result only around hundred metabolites give signal in the NMR spectrum, which is about 10% of the total metabolome [29]. On top of this the CPMG pulse sequence filters the signals by suppressing signals from macromolecules. Lipoproteins are thus suppressed in these spectra, which potentially could be an important loss of information. It might thus have been useful to combine the CPMG recordings with e.g Overhauser enhancement spectroscopy (NOESY) recordings which provide a good overview of all the types of molecules present in the sample matrix [3]. LC-MS is an alternative platform with much higher sensitivty but this technique certainly represents other challenges. In any case recording of more information consequently lowers the statistical power.

In the data analysis, four (three multivariate and one univariate) classification paths were investigated. The ECVA modeling investigates multivariate solutions where all variables are active. The LASSO search path recovers sparse solutions and it is hence the assumption that a combination of a subset of the variables has discriminatory power. Classification trees are scale invariant and superiority of such a model relies on non linearity compared to the linear ECVA and LASSO models. If the different regions are independent, multiple univariate tests sorts the regions in terms of association with FTO. The different models reveal different representations of the biological system and this exhaustive search confirms lack of consistent information oppose to wrong modeling choices.

In this study we searched for any kind of metabolic response in a plasma sample correlated with the FTO genotype. The metabolic profile is a result of a very complex network of interactions between genetic and environmental factors and at system level there is a long path from genotype to metabolic profile. Epistatic effects, either at the genomic or phenotypic level may mask the signal, interactions with the environment and the fact that the whole homeostatic system is very robust shrink the correlation between genotype and metabolic profile. Physical activity, for example, is one factor which has been found to modify the effect of FTO [30,2,23]. Despite the reported approximate effect size of 0.1 z-score units [15] for BMI per susceptibility allele which is actually quite substantial, it is a very difficult task we have given ourself here. The variation related to age and gender is only $\approx 0.1\%$ across all 171 variables, and at most 1% for single variables, revealing a high degree of unexplained variation. Targeted analysis opposed to un-targeted metabolic profiling examining a narrow range of biological relevant metabolites in connection with FTO expression and genetic and environmental confounders would increase the probability of discovery. This is supported by a range of recent works [1] which suggest data driven analysis in combination with mechanistic understanding as a strategic path to uncovering associations from high throughput methods. In this way, narrowing the model range by a priori parameter restriction leads to more statistical power, which is critical for data with large degree of unexplained variation.

## 5 Conclusion

This study shows that $^1$H NMR CPMG plasma profiles do not contain signals which can be strongly associated with FTO genotype and we speculate that high-throughput untargeted genotype-metabolic correlations will in general be a difficult path to follow due to correlation shrinkage by multiple interacting factors and inadequate power. The study further underlines the importance of careful handling of significance in cases with multiple testing and advocates for the sound use of chemical and biological knowledge in data analysis.

**Acknowledgements** The GEMINAKAR study was supported by grants from the Danish Medical Research Fund, the Danish Diabetes Association, the NOVO Foundation, the Danish Heart Foundation, and Apotekerfonden. The present study was supported by the Diogenes study which is the acronym for "Diet, Obesity and Genes" supported by the European Community (Contract no. FP6-513946), http://www.diogenes-eu.org/.

## References

1. Gary An, John Bartels, and Yoram Vodovotz. In silico augmentation of the drug development pipeline: examples from the study of acute inflammation. *Drug Development Research*, 2010.
2. C.H. Andreasen, K.L. Stender-Petersen, M.S. Mogensen, S.S. Torekov, L. Wegner, G. Andersen, A.L. Nielsen, A. Albrechtsen, K. Borch-Johnsen, S.S. Rasmussen, et al. Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes*, 57(1):95, 2008.
3. O. Beckonert, H.C. Keun, T.M.D. Ebbels, J. Bundy, E. Holmes, J.C. Lindon, and J.K. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, 2(11):2692–2703, 2007.
4. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
5. B. Benyamin, T. Sørensen, K. Schousboe, M. Fenger, P. Visscher, and K. Kyvik. Are there common genetic and environmental factors behind the endophenotypes associated with the metabolic syndrome? *Diabetologia*, 50:1880–1888, 2007. 10.1007/s00125-007-0758-1.
6. L Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software. Monterey, CA., 1984.
7. M. Daszykowski, B. Walczak, and DL Massart. Representative subset selection. *Analytica Chimica Acta*, 468(1):91–103, 2002.
8. T.M. Frayling, N.J. Timpson, M.N. Weedon, E. Zeggini, R.M. Freathy, C.M. Lindgren, J.R.B. Perry, K.S. Elliott, H. Lango, N.W. Rayner, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889, 2007.
9. J. Friedman, T. Hastie, H. H "ofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals*, 1(2):302–332, 2007.
10. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
11. T. Gerken, C.A. Girard, Y.C.L. Tung, C.J. Webby, V. Saudek, K.S. Hewitson, G.S.H. Yeo, M.A. McDonough, S. Cunliffe, L.A. McNeill, et al. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science*, 318(5855):1469, 2007.
12. Pia Skov Hansen, Thomas Heiberg Brix, Thorkild I. A. Sørensen, Kirsten Ohm Kyvik, and Laszlo Hegedus. Major genetic influence on the regulation of the pituitary-thyroid axis: A study of healthy danish twins. *J Clin Endocrinol Metab*, 89(3):1181–1187, 2004.
13. A.L. Hasselbalch, L. Angquist, L. Christiansen, B.L. Heitmann, K.O. Kyvik, and T.I.A. Sørensen. A Variant in the Fat Mass and Obesity-Associated Gene (FTO) and Variants near the Melanocortin-4 Receptor Gene (MC4R) Do Not Influence Dietary Intake. *Journal of Nutrition*, 140(4):831, 2010.
14. A. Haupt, C. Thamer, H. Staiger, O. Tschritter, K. Kirchhoff, F. Machicao, HU Haring, N. Stefan, and A. Fritsche. Variation in the FTO gene influences food intake but not energy expenditure. *Exp Clin Endocrinol Diabetes*, 117:194–197, 2009.
15. Branwen Hennig, Anthony Fulford, Giorgio Sirugo, Pura Rayco-Solon, Andrew Hattersley, Timothy Frayling, and Andrew Prentice. Fto gene variation and measures of body mass in an african population. *BMC Medical Genetics*, 10(1):21, 2009.
16. H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417–441, 1933.
17. T. Jess, E. Zimmermann, SII Kring, T. Berentzen, C. Holst, S. Toubro, A. Astrup, T. Hansen, O. Pedersen, and T.I.A. Sørensen. Impact on weight dynamics and general growth of the common fto rs9939609: a longitudinal danish cohort study. *International Journal of Obesity*, 32(9):1388–1394, 2008.
18. R.W. Kennard and L.A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.
19. R.J.F. Loos. Recent progress in the genetics of common obesity. *British Journal of Clinical Pharmacology*, 68(6):811–829, 2009.
20. R.J.F. Loos, C.M. Lindgren, S. Li, E. Wheeler, J.H. Zhao, I. Prokopenko, M. Inouye, R.M. Freathy, A.P. Attwood, J.S. Beckmann, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics*, 40(6):768–775, 2008.
21. S. Meiboom. Carr-Purcell-Meiboom-Gill sequence (CPMG). *Rev Sci Instrum*, 29:688–691, 1959.
22. L. Nørgaard, R. Bro, F. Westad, and S.B. Engelsen. A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics*, 20:425 – 435, 2006.
23. E. Rampersaud, B.D. Mitchell, T.I. Pollin, M. Fu, H. Shen, J.R. O'Connell, J.L. Ducharme, S. Hines, P. Sack, R. Naglieri, et al. Physical activity and the association of common FTO gene variants with body mass index and obesity. *Archives of internal medicine*, 168(16):1791, 2008.
24. Catherine L. Saunders, Benedetta D. Chiodini, Pak Sham, Cathryn M. Lewis, Victor Abkevich, Adebowale A. Adeyemo, Mariza de Andrade, Rector Arya, Gerald S. Berenson, John Blangero, Michael Boehnke, Ingrid B. Borecki, Yvon C. Chagnon, Wei Chen, Anthony G. Comuzzie, Hong-Wen Deng, Ravindranath Duggirala, Mary F. Feitosa, Philippe Froguel, Robert L. Hanson, Johannes Hebebrand, Patricia Huezo-Dias, Ahmed H. Kissebah, Weidong Li, Amy Luke, Lisa J. Martin, Matthew Nash, Miina Ohman, Lyle J. Palmer, Leena Peltonen, Markus Perola, R. Arlen Price, Susan Redline, Sathanur R. Srinivasan, Michael P. Stern, Steven Stone, Heather Stringham, Stephen Turner, Cisca Wijmenga, and David A.Collier. Meta-analysis of genome-wide linkage studies in bmi and obesity[ast]. *Obesity*, 15(9):2263–2275, September 2007.
25. F. Savorani, G. Tomasi, and SB Engelsen. *i*coshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.
26. K. Schousboe, PM Visscher, B. Erbas, KO Kyvik, JL Hopper, JE Henriksen, BL Heitmann, and TIA Sørensen. Twin study of genetic and environmental influences on adult body size, shape, and composition. *International Journal of Obesity*, 28(1):39–48, 2003.
27. A. Skytthe, K. Kyvik, N.V. Holm, J.W. Vaupel, and K. Christensen. The Danish Twin Registry: 127 birth cohorts of twins. *Twin Research*, 5(5):352–357, 2002.
28. R. Tauler and D. Barceló. Multivariate curve resolution applied to liquid chromatography–diode array detection. *TrAC Trends in Analytical Chemistry*, 12(8):319–327, 1993.
29. M.R. Viant, C. Ludwig, and U.L. Gunther. 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling. *Metabolomics, metabonomics and metabolite profiling*, page 44, 2008.
30. Karani S Vimaleswaran, Shengxu Li, Jing Hua Zhao, Jian'an Luan, Sheila A Bingham, Kay-Tee Khaw, Ulf Ekelund, Nicholas J Wareham, and Ruth JF Loos. Physical activity attenuates the body mass indexincreasing influence of genetic variation in the fto gene.

*The American Journal of Clinical Nutrition*, 90(2):425–428, August 2009.

31. K. Wahlen, E. Sjolin, and J. Hoffstedt. The common rs9939609 gene variant of the fat mass-and obesity-associated gene FTO is related to fat cell lipolysis. *The Journal of Lipid Research*, 49(3):607, 2008.
32. Andrew J. Walley, Julian E. Asher, and Philippe Froguel. The genetic contribution to non-syndromic human obesity. *Nat Rev Genet*, 10(7):431–442, July 2009.
33. Jane Wardle, Susan Carnell, Claire M A Haworth, I. Sadaf Farooqi, Stephen O'Rahilly, and Robert Plomin. Obesity associated genetic variation in fto is associated with diminished satiety. *J Clin Endocrinol Metab*, 93(9):3640–3643, Sep 2008.
34. W. Yang, T. Kelly, and J. He. Genetic epidemiology of obesity. *Epidemiologic reviews*, 29(1):49, 2007.
35. S. Z
    "ollner and J.K. Pritchard. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615, 2007.

DEPARTMENT OF FOOD SCIENCE
PHD THESIS 2012 · ISBN 978-87-7611-499-2

MORTEN ARENDT RASMUSSEN

**Medicometrics**

## Medicometrics

Medicometrics is the science of integrating different sources of measurements related to a pathological system. It is an interfacial discipline utilizing elements from applied mathematics, multivariate statistics, chemometrics, pharmacometrics, medicine, biology, biochemistry, etc. The philosophy of Medicometrics is highly exploratory and holistic, aiming at multivariate patterns oppose to single factor associations.

SL grafik