Investigation of the machine learning method Random Survival Forest as an exploratory analysis tool for the identification of variables associated with disease risks in complex survival data.

vorgelegt von

Dipl.-Biologe, MSc. Bioinformatik, MSc. Epidemiologie,

Stefan Dietrich

geb. in Berlin

von der Fakultät VII – Wirtschaft und Management
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Gesundheitswissenschaften/Public Health
– Dr. P.H. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Elke Schäffner

Gutachter: Prof. Dr. Heiner Boeing

Gutachter: Prof Dr. Reinhard Busse

Tag der wissenschaftlichen Aussprache: 23. Juni 2016

Berlin 2016

Meinen Eltern Renate und Uwe Dietrich gewidmet.

„Wir müssen unbedingt Raum für Zweifel lassen, sonst gibt es keinen Fortschritt, kein Dazulernen. Man kann nichts Neues herausfinden, wenn man nicht vorher eine Frage stellt. Und um zu fragen, bedarf es des Zweifelns.“

Richard P. Feynman

Contents

List of tables V
List of figures VI
List of abbreviations VII
Summary VIII
Zusammenfassung X
1 Introduction 1
1.1 Background 1
1.2 Public health relevance and epidemiology 2
1.3 Variable selection 3
1.4 Statistical challenge of variable selection 5
1.5 Decision tree learning 7
1.6 Random Survival Forest 8
1.7 Objective 10
2 Methods 12
2.1 Random Survival Forest 12
2.1.1 Minimal depth measurement 13
2.1.2 Prediction error rate 13
2.1.3 Random Survival Forest backward algorithm 14
2.1.4 Partial plots 15
2.2 Simulation study 15
2.2.1 Generation of Simulation Data 15
2.2.2 Simulation Analyses 17
2.3 European Prospective Investigation into Cancer and Nutrition-Potsdam study 18
2.3.1 Blood sample collection at baseline 19
2.3.2 Assessment of serum metabolites 20
2.3.3 Assessment of dietary intake 20
2.3.4 Assessment of covariates 21
2.3.5 Ascertainment of endpoints 22
2.4 Analytical study sample 22
2.5 Statistical Analysis 24
2.5.1 Descriptive statistics 25
2.5.2 Choice of covariates 25
2.5.3 Variable selection using the Random Survival Forest backward algorithm 26
2.5.4 Determination of associations using partial plots 26
2.5.5 Analysis of correlation structures using Gaussian Graphic Models 27

3 Results 29
3.1 Simulation study 29
3.1.1 Sensitivity of the minimal depth measurement to *ntree* and *nsplit* 29
3.1.2 Sensitivity of the minimal depth measurements to correlated variables 32
3.1.3 Sensitivity of the minimal depth measurement to covariates 33
3.1.4 Sensitivity of the Random Survival Forest backward selection procedure 35
3.2 Identification of metabolites associated with incident type 2 diabetes [1] 37
3.3 Identification of food groups associated with incident hypertension 45
3.3.1 Selected food groups associated with incident hypertension in men 45
3.3.2 Selected food groups associated with incident hypertension in women 50
3.3.3 Comparison of food group selection in data of men and women 56
4 Discussion 61
4.1 Discussion of methodical findings 61
4.1.1 The suitability of the minimal depth ranking measurement 62
4.1.2 Random Survival Forest and the preference of node splits on continuous variables 63
4.1.3 Random Survival Forest and multicollinearity 64
4.1.4 Random Survival Forest and confounding 65
4.1.5 Variable selection using the Random Survival Forest backward algorithm 67
4.1.6 Application of the Random Survival Forest backward algorithm to observational data 68
4.1.7 Advantages and Disadvantages of Random survival forest 71
4.2 Discussion of biological plausibility of results 73
4.2.1 Serum metabolites associated with incident type 2 diabetes 73
4.2.2 Food groups associated with incident hypertension 78
4.3 Strengths and Limitations 85
4.4 Conclusion and implications for public health 88
5 Literature 90
Appendix ..... **XII**
Danksagung ..... **XVIII**

List of tables

TABLE 1: SELECTION STEPS OF THE RANDOM SURVIVAL FOREST BACKWARD ALGORITHM WHEN APPLIED TO SIMULATION DATA. 36

TABLE 2: BASELINE CHARACTERISTICS OF THE ANALYSED EPIC-POTSDAM STUDY POPULATION. 37

TABLE 3: COMPUTED PREDICTION ERROR RATES OF DIFFERENT RANDOM SURVIVAL FOREST MODELS REGARDING THE INCIDENT OF TYPE 2 DIABETES. 38

TABLE 4: COMPARISON BETWEEN MULTIVARIATE COX PROPORTIONAL HAZARDS REGRESSION AND THE RANDOM SURVIVAL FOREST BACKWARD ALGORITHM REGARDING SELECTED METABOLITES. 42

TABLE 5: BASELINE CHARACTERISTICS OF MALE AND FEMALE PARTICIPANTS OF THE ANALYSED EPIC-POTSDAM SAMPLE POPULATION. 45

TABLE 6: COMPUTED PREDICTION ERROR RATE OF DIFFERENT RANDOM SURVIVAL FOREST MODELS FOR MEN REGARDING INCIDENT HYPERTENSION. 46

TABLE 7: COMPUTED PREDICTION ERROR RATE OF DIFFERENT RANDOM SURVIVAL FOREST MODELS FOR WOMEN REGARDING INCIDENT HYPERTENSION. 51

TABLE 8: COMPARISON OF FOOD GROUPS SELECTED BY THE RANDOM SURVIVAL FOREST BACKWARD ALGORITHM IN MEN AND WOMEN. 58

Table S1: SUMMARY OF SERUM METABOLITES XII

TABLE S2: COMPOSITION OF FOOD GROUPS XV

TABLE S3: SUMMARY OF FOOD GROUPS XVII

List of figures

FIGURE 1: ILLUSTRATION OF DECISION TREE LEARNING. 8
FIGURE 2: ILLUSTRATION OF THE RANDOM SURVIVAL FOREST METHOD. 9
FIGURE 3: ILLUSTRATION OF MINIMAL DEPTH. THE MINIMAL DEPTH IS INDICATED BY AN INTEGER VALUE INSIDE THE NODES. 13
FIGURE 4: ILLUSTRATION OF THE THREE SCENARIOS USED IN THE SIMULATION STUDY TO EVALUATE THE RSF METHOD. 18
FIGURE 5: DEFINITION AND EXCLUSION CRITERIA OF THE TWO ANALYSED EPIC POTSDAM STUDY SAMPLES. 24
FIGURE 6: BOXPLOTS FOR MINIMAL DEPTH VALUES OF PREDICTIVE AND NOISE VARIABLES FOR DIFFERENT VALUES OF *NTREE* AND *NSPLIT*. 31
FIGURE 7: BOXPLOTS FOR MINIMAL DEPTH VALUES OF PREDICTIVE, CORRELATED AND NOISE VARIABLES. 33
FIGURE 8: BOXPLOTS FOR MINIMAL DEPTH VALUES OF PREDICTIVE AND NOISE VARIABLES TAKING INTO ACCOUNT A COVARIATE. 34
FIGURE 9: SELECTED METABOLITES THAT ARE MOST PREDICTIVE FOR INCIDENT TYPE 2 DIABETES. . 39
FIGURE 10: PARTIAL PLOTS OF THE SELECTED METABOLITES MOST INFORMATIVE REGARDING INCIDENT TYPE 2 DIABETES. 40
FIGURE 11: CORRELATION STRUCTURE FOR SELECTED ACYL-ALKYL PHOSPHATIDYLCHOLINES. 43
FIGURE 12: SERUM METABOLITE NETWORK OF THE EPIC-POTSDAM SUBCOHORT. 44
FIGURE 13: IDENTIFIED FOOD GROUPS AND COVARIATES INCLUDED IN THE FINAL RANDOM SURVIVAL FOREST MODEL WHICH WERE MOST INFORMATIVE FOR INCIDENT HYPERTENSION IN MEN. 47
FIGURE 14: PARTIAL PLOT OF IDENTIFIED FOOD GROUPS MOST PREDICTIVE FOR INCIDENT HYPERTENSION IN MEN. 49
FIGURE 15: IDENTIFIED FOOD GROUPS AND COVARIATES INCLUDED IN THE FINAL RANDOM SURVIVAL FOREST MODEL WHICH WERE MOST INFORMATIVE FOR INCIDENT HYPERTENSION IN WOMEN.. 52
FIGURE 16: PARTIAL PLOT OF IDENTIFIED FOOD GROUPS MOST PREDICTIVE FOR INCIDENT HYPERTENSION IN WOMEN. 54
FIGURE 17: FOOD GROUP NETWORK OF MALE PARTICIPANTS OF EPIC-POTSDAM. 59
FIGURE 18: FOOD GROUP NETWORK OF FEMALE PARTICIPANTS OF EPIC-POTSDAM. 60
FIGURE 19: ILLUSTRATION OF SPLIT POINT SELECTION FOR BINOMINAL, CATEGORICAL AND CONTINUOUS VARIABLES. 64

List of abbreviations

a, acyl

aa, diacyl

AC Acylcarnithine

ae, acyl-alkyl

BMI, body mass index

EPIC, European Prospective Investigation into Cancer and Nutrition

FFQ, food frequency questionnaire

IPAI, Improved physical activity index

GGM, gaussian graphic model

HDL high density lipoprotein

HR, hazards ratio

LDL low density lipoprotein

OOB, out-of-bag

RSF, Random Survival Forest

RF, Random Forest

PC, phoshatidylcholines

T2D, type two diabetes mellitus

VLDL, very low density lipoprotein

Summary

The containment of the global epidemic increase of chronic diseases represents a major objective of health care systems worldwide. However, the fulfillment of this objective is complicated by the multifactorial origin of many frequent chronic diseases. Comprehensive investigations are necessary to grasp the complexity of the pathophysiological mechanisms of chronic diseases. However, this frequently results in the acquisition of complex data with numerous highly correlated variables. The statistical analysis of such complex data to identify disease associated markers is a daunting challenge. In general the application of regression methods to complex data is accompanied by problems of multiple testing and of multicollinearity. A promising approach for the survival time analysis of complex data represents the machine learning method Random Survival Forest (RSF).

Against this background, the present thesis aimed to evaluate the applicability of RSF for survival analysis of complex data in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. A RSF backward selection algorithm was developed for the purpose of variable selection. A simulation study was then performed to evaluate the RSF method and the RSF backward algorithm. Subsequently, the RSF backward algorithm was applied to prospective observational data of the EPIC-Potsdam study to identify metabolites associated with incident T2D and to identify food groups associated with incident hypertension.

The conducted simulation study confirmed the suitability of the RSF method and the implemented RSF backward algorithm as a tool for variable selection. It was demonstrated that the RSF method is able to identify predictive variables while taking into account possible confounders and can handle also the problem of multicollinearity. The subsequent application of the RSF backward algorithm to data of the EPIC-Potsdam study resulted in the

successful identification of several metabolites and food groups which were associated with incident T2D and incident hypertension, respectively. Beside hexose, the metabolite diacyl-phosphatidylcholine (PC) C38:3, acyl-alkyl-PC C34:4, the amino acids valine, tyrosine, and glycine and a correlation pattern of five acyl-alkyl-PC and two diacyl-PC were associated with the incidence of T2D. Regarding the incidence of hypertension, a lunch and dinner pattern was most informative in women. In addition, a pattern reflecting dairy fat and cheese consumption and the consumption of spirits were also associated with incident hypertension in women and men. By using partial plots the direction of non-linear associations between identified variables and incident T2D and hypertension were visualised which enhanced the interpretability of the findings.

In conclusion, the findings of the present thesis demonstrated that the RSF method and the implemented RSF backward algorithm represent a sensible complement to existing survival analysis methods. The RSF backward algorithm is particularly useful for exploratory analysis of complex survival data to identify unknown biomarkers associated with time until event of interest. However, the verification of the implemented RSF backward algorithm and of the present findings in external cohorts as well as the translation of the present findings for clinical diagnosis, prevention strategies and dietary recommendations should be a matter for future research.

Zusammenfassung

Die Eindämmung der globalen epidemischen Zunahme chronischer Krankheiten stellt weltweit eine Hauptaufgabe für Gesundheitssysteme dar. Diese Aufgabe wird erschwert durch den multifaktoriellen Ursprung vieler chronischer Krankheiten. Umfangreiche Forschungen sind notwendig, um die Komplexität der pathophysiologischen Mechanismen chronischer Krankheiten zu erfassen. Dies ist häufig verbunden mit der Erfassung komplexer Daten mit einer Vielzahl von hoch korrelierten Variablen. Die statistische Analyse dieser Daten mit dem Ziel krankheitsauslösende Faktoren innerhalb der Daten zu identifizieren, stellt eine große Herausforderung dar. So müssen Probleme aufgrund von multiplen Testen oder von Multikollinearität beachtet werden, wenn Regressionsmethoden angewendet werden. Eine vielversprechende Methode für die Überlebenszeitanalyse von komplexen Daten stellt die maschinelle Lernmethode Random Survival Forest (RSF) dar.

Vor diesem Hintergrund war das Ziel dieser Dissertation die Anwendbarkeit von RSF für die Überlebenszeitanalyse von komplexen Daten in der European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Studie zu evaluieren. Für den Zweck der Variablenselektion wurde ein RSF backward Algorithmus entwickelt. Eine Simulationsstudie wurde durchgeführt, um die RSF Methode und den RSF backward Algorithmus zu evaluieren. Anschließend wurde der RSF backward Algorithmus auf prospektive Beobachtungsdaten der EPIC-Potsdam Studie angewendet, um Metabolite zu identifizieren, die mit inzidentem Typ 2 Diabetes mellitus (T2D) und Lebensmittelgruppen die mit inzidenter Hypertonie assoziiert sind.

Die durchgeführte Simulationsstudie bestätigte die Eignung der RSF Methode und des implementierten RSF backward Algorithmus für die Variablenselektion. Es wurde demonstriert, dass die RSF Methode prädiktive Variablen identifiziert unter Berücksichtigung

möglicher Confounder und zudem das Problem von Multikollinearität handhaben kann. Die Anschließende Anwendung des RSF backward Algorithmus auf Daten der EPIC-Potsdam Studie resultierte in der erfolgreichen Identifizierung verschiedener Metabolite, die mit inzidentem T2D assoziiert waren und von Lebensmittelgruppen die mit inzidenter Hypertonie assoziiert waren. Neben Hexose waren die Metabolite diacyl-Phosphatidylcholin (PC) C38:3, acyl-alkyl-PC C34:4, die Aminosäuren Valin, Tyrosin, Glycin und ein Korrelationsmuster aus fünf acyl-alkyl-PC und zwei diacyl-PC mit inzidenter Hypertonie assoziiert. Bezogen auf die Inzidenz der Hypertonie von Frauen war ein Korrelationsmuster, welches Mittag und Abendmahlzeiten widerspiegelte, am informativsten. Zusätzlich war ein Korrelationsmuster, welches den täglichen Fett- und Käsekonsum widerspiegelt, sowie Alkoholkonsum bei Frauen und Männern mit inzidenter Hypertonie assoziiert. Durch die Verwendung von Partial Plots konnte die Richtung von nicht-linearen Assoziationen zwischen den identifizierten Variablen und inzidentem T2D und inzidenter Hypertonie visualisiert werden, welches die Interpretierbarkeit der Ergebnisse erhöhte.

In der Schlussfolgerung zeigten die Ergebnisse der vorliegenden Studie, dass die RSF Methode und der implementierte RSF backward Algorithmus eine adäquate Ergänzung von existierenden Methoden der Überlebenszeitanalysen darstellt. Der RSF backward Algorithmus ist insbesondere für die explorative Analyse von komplexen Überlebensdaten geeignet, um unbekannte Biomarker zu identifizieren, die mit der Ereigniszeit von Interesse assoziiert sind. Jedoch sollte in zukünftigen Studien eine Verifizierung des implementierten RSF backward Algorithmus und der gezeigten Ergebnisse in externen Kohorten stattfinden, sowie eine Übertragbarkeit der gezeigten Ergebnisse für die klinische Diagnostik, Präventionsstrategien und Ernährungsempfehlungen untersucht werden.

1 Introduction

1.1 Background

Nowadays, societies worldwide are confronted with an epidemic increase of incidence and prevalence of chronic diseases with dramatic consequences for affected individuals and considerable expenditure for health care systems. Many of the frequent chronic diseases, e.g. cardiovascular diseases, diabetes mellitus type two (T2D) and dementia, are caused and promoted by a variety of cellular, molecular and metabolic conditions as well as environmental factors (1-7). Modern technologies and methods enable the exploratory acquisition of possible disease triggers by the generation of complex data which capture the entire spectrum of genes, RNAs, proteins as well as metabolites (8-11). Furthermore, in recent epidemiological studies in depth detailed information on anthropometric markers, diet, and environmental determinants are recorded which in addition increases the analysable data volume (12-15).

However, the statistical analysis of such complex data to identify disease associated markers is a daunting challenge. In general, complex data consist of a variety of highly correlated variables causing problems of multiple testing and multicollinearity when using regression methods to identify disease markers (16-19). In order to address these well-known problems of statistical testing, several new variable selection methods have been developed (20-23). A promising method with respect to the statistical analysis of right censored survival data represents the machine learning method Random Survival Forest (RSF) (23). However, applications of RSF in epidemiological studies with complex data to identify biological markers promoting the development of chronic diseases are still a rarity.

Against this background, the present thesis aimed to examine the applicability of RSF for survival analysis of complex data in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. A RSF backward selection algorithm was developed for the

purpose of variable selection and applied to identify metabolites associated with incident T2D and to identify food groups associated with incident hypertension in the EPIC-Potsdam study.

1.2 Public health relevance and epidemiology

The identification of disease triggers associated with the incidence of hypertension and T2D is of particular relevance for public health systems, as one third of adults worldwide, suffer from high blood pressure and one-tenth of adults has diabetes mellitus (24). From 1980 to 2008 the number of people with hypertension increased globally from 600 million to nearly 1 billion (24). During the same period, also the number of people with diabetes mellitus increased globally from 153 million in 1980 to 347 million in 2008 (25). Of all diabetes mellitus diagnoses, 90% are attributable to T2D (26). According to the latest survey study from the Robert Koch Institute, it is estimated that 20 million of German adults between 18 and 79 suffer from hypertension and 4.6 million from T2D (27, 28). The high prevalence rates are particularly alarming given the fact that both diseases can cause further health complications. Hypertension, the chronic increase of blood pressure, which causes high stress on heart and blood vessels, is a major risk factor for stroke (29, 30), myocardial infarction (31, 32), heart failure (33), peripheral artery disease, and kidney disease (34). For T2D, the increased blood sugar concentration caused by insulin resistance or deficiency, can result in diabetic retinopathy, diabetic nephropathy, diabetic foot syndrome, heart attack, and stroke (35-38). In addition, both conditions may cause a reduced quality of life and life expectancy among those individuals affected and also result in high financial burden on health care systems. Actually, in 2008 the treatment costs for the German health care system amounted to 15.3 billion euros for both chronic diseases (39).

The epidemic increase of hypertension and T2D is supposed to be a consequence of a rising world population, coupled with prolonged life expectancy and dissemination of Western lifestyle (24). Furthermore, hypertension and T2D are associated with obesity, physical inactivity, and unhealthy diet (40-42). Thus, weight reduction, promotion of physical activity, and consequent change in diet are important elements of treatment and prevention strategies (40-42). In addition to these modifiable risk factors, a number of non-modifiable risk factors exists, such as increasing age, gender, and genetic predisposition affecting the occurrence and development of hypertension and T2D (40-42).

Despite great efforts in research and science, the underlying biological mechanisms of these two diseases are not yet completely understood. Hence, hypertension and T2D are often diagnosed when consequential health damages have already occurred. In this context, the usage of scientific opportunities established by latest technologies of data-mining is of great interest to identify disease associated markers and triggers allowing early diagnosis, effective therapy and prevention activities. The concepts to identify disease associated markers and triggers, and the related statistical challenges are introduced in the next chapters.

1.3 Variable selection

An important aspect of data-mining is variable selection, also known as feature selection. Variable selection is applied to complex data, in order to identify a variable or a subset of variables which are associated with the outcome of interest. In general, the variable selection process is accompanied by data reduction which is being sought for several reasons (43-45):

1.) Redundant predictors will add noise to statistical analysis causing bias, misleading estimators and reduced prediction accuracy.

2.) Overfitting, problems of multiple testing and multicollinear effects can be reduced or avoided if simple models are available.
3.) Statistical interpretability is increased with a reduced number of variables.
4.) In the context of clinical diagnosis and prediction of diseases, usage of a few, highly informative variables avoid misinterpretations and reduce cost.

Selection of informative variables can in general be achieved in three different ways which are categorized into top-down (backward selection), bottom-up (forward selection) and stepwise approaches (43-45).

Top-down approaches start with the full model including all variables to be analysed (43-45). Step by step noise (non-informative) variables are eliminated then. For this purpose, the variables are often classified based on a statistical measure and the variable with the worst value is eliminated. No further improvement of model quality can be used as stop criterion. Contrary, bottom-up approaches start with a simple model including one or more informative variables (43-45). Subsequently, other variables are added step by step to the model until the stop criterion (e.g. improvement of model quality) is fulfilled. Thereby, identification of an optimal, informative start variable is necessary and an appropriate stop criterion must be defined. Stepwise selection approaches presents a mixture of forward and backward selection (43-45). They form final model by stepwise adding of variables or use shrink processes by stepwise elimination of variables. However, at each step it is verified whether the model performance is improved. If not, the variable is replaced by another variable until the model performance is further improved. Although stepwise selection leads to more exact models they tend to be very computationally intensive resulting in excessive computation times.

Application of a backward selection procedure is often applied in data-mining when complex data are analysed where no-prior knowledge exists, such like genomic or metabolomic data (43-45). Thus, there is no need to define an initial variable and it is ensured that all measured informative variables are considered (43-45). A reduced computation time is an advantage compared to stepwise selection approaches, albeit at the cost of accuracy. However, when applying variable selection procedures on complex-data some statistical challenges have to be considered which are discussed next.

1.4 Statistical challenge of variable selection

The application of variable selection approaches on complex data to identify disease associated markers poses challenging statistical issues. The challenges originate from the structure of analysed complex data, from the nature of applied statistical method, but also from prerequisites that have to be fulfilled for applied statistical methods. Complex data often consist of a large number of variables, e.g. hundreds or thousands of variables in case of metabolomics or genomics. In contrast, the number of observations used for sampling often depends on financial resources, workload, material cost and available examination time. In the consequence, this results in a small number of observations leading to a loss of statistical power (46). To counteract the problem of reduced statistical power, resampling methods like bootstraps, permutation tests, and cross validation are frequently applied in variable selection approaches (47, 48).

Another challenge regarding variable selection is mullticollinearity within complex data. The term multicollinearity refers to the case that variables are highly correlated with each other, as it occurs frequently in biological data e.g. due to tight co-regulation (19, 49). When applying variable selection based on regression approaches to highly correlated data,

inflated standard errors are the consequence which affects confidence intervals and calculations of p-values of individual correlated predictors (19, 49). This increases the risk of arbitrary predictor choices in variable selection processes which hampers the identification of disease-related variables or pathways in regression models (50, 51).

Cox regression approaches are commonly used for variable selection in survival data to assess the association between variables to survival time individually (52). However, testing each metabolite individually at a time – which is a frequently used approach in exploratory data analysis – increases the probability of type I error. Several methods to adjust for multiple testing are available (17). Some of them have been successfully applied to identify biomarkers in data sets derived from metabolomic approaches (52, 53). Yet, correction for multiple testing may substantially decrease statistical power in datasets containing a large number of noise variables, e.g. in untargeted metabolomic studies.

In addition, the usage of regression approaches is often accompanied by statistical prerequisites to be fulfilled which hamper the data analysis. Just to name a few: Modeling of Cox proportional hazards models requires the fulfillment of the proportional hazards assumption (54). Furthermore, normal distribution of underlying data and linearity of associations have to be test for multiple regression approaches (55). Consequently, extensive and time consuming statistical analysis must be performed to fulfill prerequisites of regression approaches and to model non-linear effects.

In view of these statistical challenging issues, various statistical methods have been developed including Cox-regression under lasso penalization (20, 56), partial least squares (21, 22), *L*2-boosting (57), or componentwise Cox-likelihood based boosting (58). One of the most promising approaches for the identification of disease-associated markers in complex

survival data represents the RSF method which was introduced in 2008 (23) and is described in the following paragraph.

1.5 Decision tree learning

To provide a better understanding of the basic methodical concept of the RSF method, first, decision tree learning is introduced. Decision trees are ordered, directed trees using hierarchically structured decision rules to classify data objects, similar to the process of diagnosis by physicians (59). As visualized in **Figure 1**, a decision tree consists of nodes representing split variables and edges connecting parent nodes with the daughter nodes.

The decision tree learning - or in other words, the classifications of objects - begins with the root node and branches into many nodes (59). At each node of the tree a classification variable (e.g. age, cholesterol concentration) is determined which best classifies analysed objects into subsets regarding the outcome of interest (e.g. health status or time until event). For the determination of the best classification variable, a statistical measure like the gini index or the entropy criterion can be used (60). The classified objects are then assigned to the respective subsequent daughter nodes. Thus, for example, one daughter node contains several objects older than 60 years with shorter follow-up time and the other daughter node contains objects equal to or younger than 60 years with a longer follow-up time. The splitting process is repeated on each derived subset of objects in a recursive proceeding until the number of objects in each terminal node becomes too small for further splits.

The concept of decision tree learning is commonly applied in data-mining procedures, as it may deliver high classification accuracy, is easily understandable, illustrates complex decision problems in a simple way, and allows the detection of key feature (61). However,

the usage of single decision tree learning is accompanied by some drawbacks. Particularly noteworthy are the instability of decision trees in response to small changes in data used, susceptibility to noise and overfitting to the training data (59, 62). Against this background and to counteract the drawbacks, ensemble decision tree learning methods like Random Forest (RF) or RSF have been designed (23, 63).

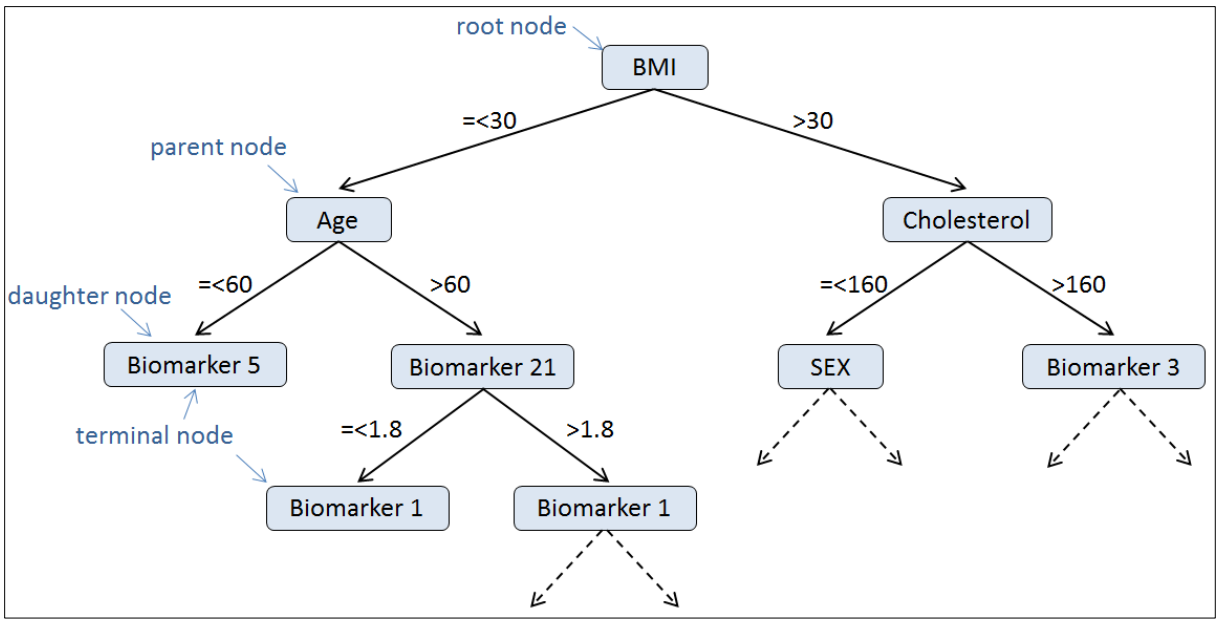

Figure 1: Illustration of decision tree learning.

1.6 Random Survival Forest

RSF is a multivariate machine learning method (23) and computes an ensemble of decision trees, which can be used for selecting the most important variables that are associated with event time of interest (e.g. time from baseline until the diagnosis of hypertension or T2D). The computation of an ensemble of decision trees with the RSF method is based on two random processes comprising bootstrapping and random node splitting (23). As described by Ishwaran et al. (23), several independent bootstrap samples are randomly drawn from the study population and each of the individual bootstrap samples is used to compute one

decision tree **(Figure 2)**. The purpose of using bootstrap samples is to reduce the problem of overfitting to the training data (64).

The second random process is introduced at the level of node splitting. At each node of the decision tree a predefined number of candidate variables is chosen from the pool of available variables. Within the candidate variables the one variable with a cut point that maximizes the survival difference between daughter nodes is selected to split the respective node. Commonly the log-rank statistic is used to determine the maximisation of the survival differences (65, 66). Thereby, per each candidate variable several split points can be tested. However, those splitting processes are known to favour continuous variables if continuous and categorical variables are used together. Nevertheless, the introduced two random processes altogether counteract bias and variance which affects single decision trees.

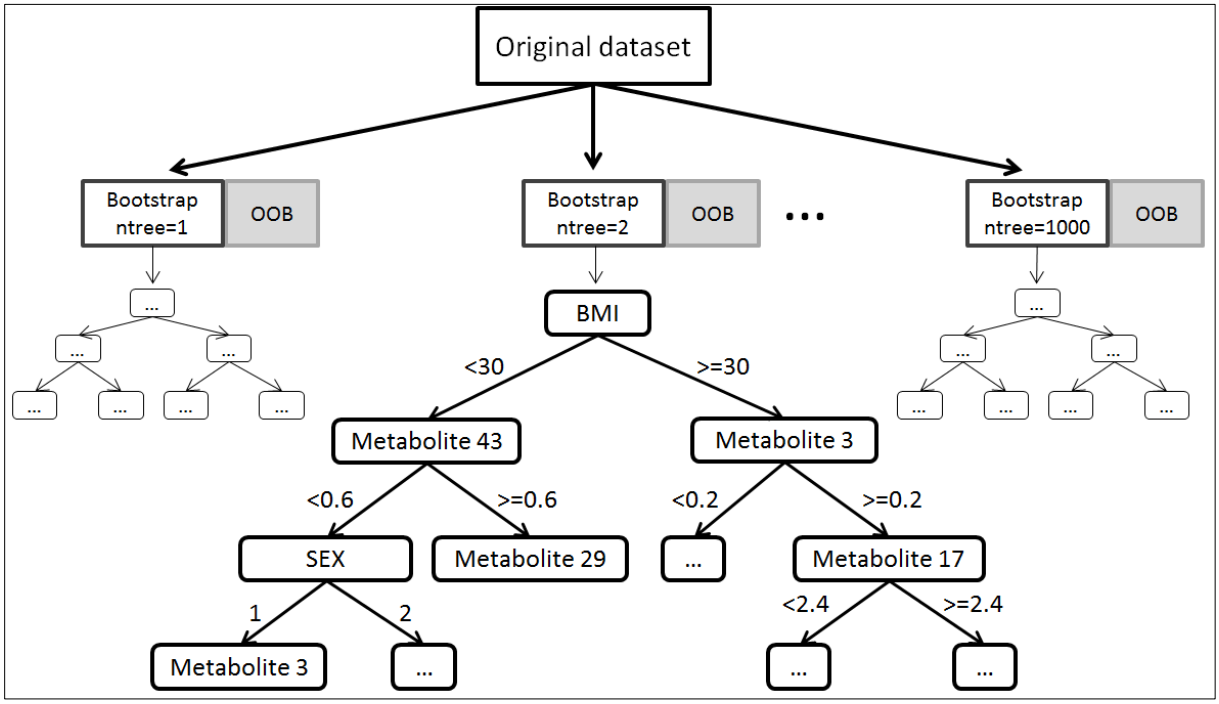

Figure 2: Illustration of the RSF method. Reference: Dietrich et al. (92).

Once a RSF is computed, it can be determined which variables are most predictive regarding time until event. For this purpose, the distance from the root node to the node where a variable splits first is determined (67). According to this approach, the RSF ranking procedure

is named ‘minimal depth’ (67). Please note that a second RSF ranking procedure is available which is named variable importance. However, this method has been criticized because it is prone to bias and challenging to regulate in theoretical studies (67). Hence, this method is not considered in the present thesis.

The application of RSF is recommended and seems reasonable because it is a completely data-driven method, does not rely on constraints that have to be fulfilled for regression approaches, and can automatically deal with high-level interactions due to random node splitting (23). Accordingly, the RSF method has been successfully applied to identify risk factors of different diseases in some recent studies (68-70). In this context, it was also demonstrated that the RSF method results in prediction models with comparable accuracy to other available survival analysis methods (68, 70). Although, the RSF method generates accurate prediction models, deep tree growing may induce loss of prediction accuracy. Data-reduction by application of a RSF variable selection algorithm seems appropriate to solve this issue and to achieve more accurate prediction models with a low number of predictive variables.

1.7 Objective

Against this background the present thesis aimed to evaluate the underlying statistical properties of the RSF method and to apply a RSF variable selection approach in the EPIC-Potsdam study for identification of disease-associated factors. Thereby the following objectives were of interest:

1.) Development of a RSF backward algorithm appropriate for variable selection.

2.) Evaluation of the RSF method and of the RSF backward algorithm regarding identification of predictive variables.

3.) Identification of metabolites associated with incident T2D in the EPIC-Potsdam study by using the implemented RSF backward algorithm.

4.) Identification of food groups associated with incident hypertension in the EPIC-Potsdam study by using the implemented RSF backward algorithm.

For the purpose of evaluation, right censored survival data were simulated including predictive, correlated, and noise variable as well as a confounding variable. For Identification of metabolites associated with incident T2D the implemented RSF backward algorithm was applied to a dataset of the EPIC-Potsdam study which was recently used by Floegel et al. (52). In this previous study, Cox proportional hazards regression and principal component analysis (PCA) were applied to analyze the association between metabolites and risk of T2D (52). The present thesis extend the previous work and discusses the achieved results in comparison with those from Floegel et al. (52). In the last step the RSF backward algorithm was applied to food-group data that have already been extensively examined in the EPIC-Potsdam study, however, not in regard to incident hypertension.

2 Methods

2.1 Random Survival Forest

To compute a RSF numerous random bootstrap sample are drawn (23, 71). The number of bootstrap samples can be defined in the package randomForestSRC of the statistical software R by a parameter which is named *ntree* (72). In the present thesis this nomenclature is maintained to refer to the number of bootstrap samples. Each bootstrap sample included on average two thirds of the original data. The remaining one third of the data is excluded and called out-of-bag (OOB) data. A single decision tree is computed based on each bootstrap sample. To grow a decision tree, at each node a random set of candidate variables is chosen for random node splitting. The number of candidate variables corresponds with the square root of the total number of variables. As splitting rule, for node splitting, the log-rank statistic was used in the present thesis (66, 71). The log-rank statistic maximized the survival differences between nodes by the following formula (66, 71):

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}}$$

d - number of deaths
Y - individuals at risk at time t
i - number of observations

Equation 1

To calculate the log-rank statistics for each candidate variable several random split points can be randomly chosen. The number of random split points per candidate variable, can be pre-defined before computation of a RSF model when using the package randomForestSRC in the statistical software R. In the RSF package this RSF parameter is named *nsplit* and this nomenclature is maintained in the present thesis.

2.1.1 Minimal depth measurement

How informative a variable is regarding time until event can be assessed from a computed RSF by the minimal depth measurement (67). As in detail described by Ishwaran et al. (67), the minimal depth of a variable is determined by the distance from the root node to the closest node were the respective variable splits first. This concept is illustrated in **Figure 3**. The metabolite 5 in **Figure 3** splits at a node with a minimal depth of 1 and also of 2. The minimal depth of 1 represents the closest split to the root node and is thus assigned to the variable. The value is recorded for the variable and in the following repeated in any further computed decision tree and finally averaged over the whole RSF (67).

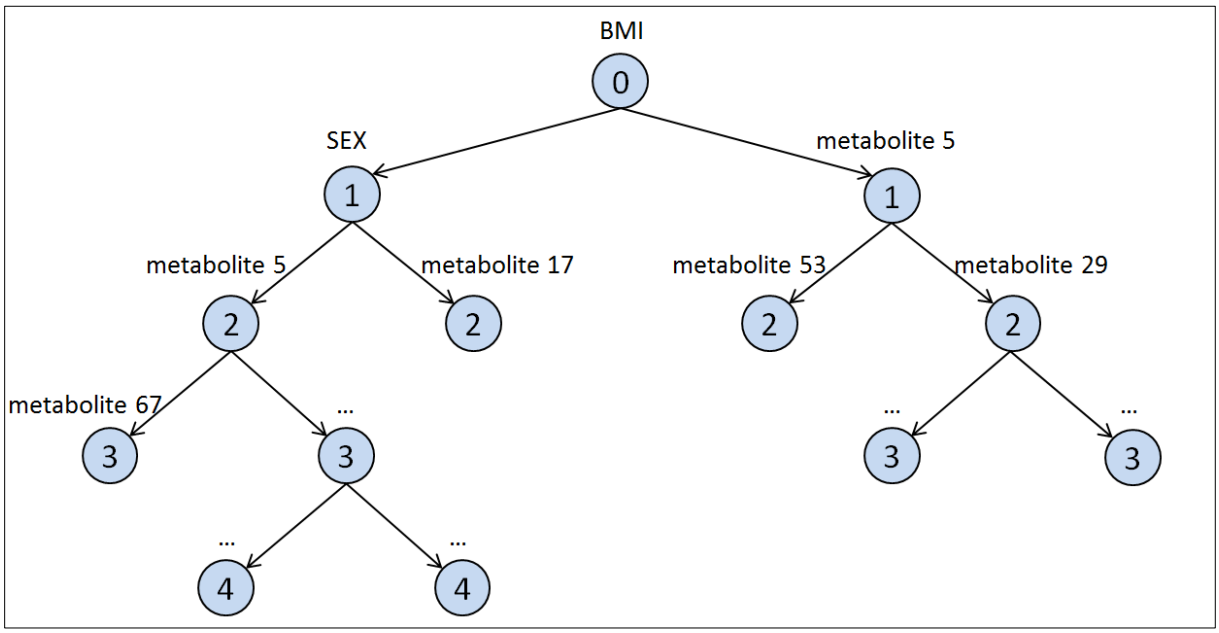

Figure 3: Illustration of minimal depth. The minimal depth is indicated by an integer value inside the nodes.

2.1.2 Prediction error rate

To determine the prediction accuracy of a RSF model, the RSF prediction error rate can be calculated based on Harrell's concordance index (C-index) (23, 73). For this purpose, the OOB samples of each decision tree are used and dropped down the respective decision tree According to Harrell's C-index, the probability is then estimated, that within a randomly

selected pair of OOB samples with an event, the OOB sample with the shorter follow-up time has the worst predictive outcome (23, 73). The RSF prediction error rate is conform to 1-C-index with values between 0 and 1, where lower RSF prediction error rate correspond to RSF models with more precise prediction accuracy (23). RSF prediction error rates of 0.5 refer to RSF models which based on chance (23).

2.1.3 Random Survival Forest backward algorithm

For the purpose of variable selection a RSF backward selection algorithm was implemented. The algorithm works in the following consecutive steps:

(1) Compute a RSF model with the full data set including covariates and all the variables (metabolites and food groups) to be tested.

(2) Rank the variables by the minimal depth values derived from the computed RSF model.

(3) Remove the one variable with the "worst" (highest) minimal depth value from the data. Covariates are not considered.

(4) Save the prediction error rate of the computed RSF model.

(5) Compute a new RSF model based on the reduced data set including the covariates and the remaining variables.

(6) Repeat Step 2 to 6 until only a RSF model can be computed based on covariates.

(7) The final set of variables, which is chosen, is the RSF model with the lowest predicted error rate.

The suitability of the RSF backward algorithm was examined in a simulation study. Subsequently, the RSF backward algorithm was applied to observational data of the EPIC-Potsdam study.

2.1.4 Partial plots

In order to graphically illustrate associations between concentrations of selected variables and predicted event-free survival, partial plots can be drawn based on a computed RSF model (71, 74). Partial plots represent the effect of each selected variable on predicted event-free survival after accounting for the average effects of the remaining variables of the respective RSF model (71, 74). To calculate a partial plot of a respective variable, several replications are performed. At each replication the values of the respective variable in the data are replaced by another constant value. The constant value is derived from the values of the respective variable and represents a so-called partial value. A RSF is computed based on the new data including the respective variable with constant values, the other selected variables, and the covariates. The difference between the predicted event-free survival of the new computed RSF and the original RSF is recorded and used to draw the partial plots of the respective variable which is thus also adjusted for all other included variables. The computation of partial plots is conducted by an automatically function of the R-package randomForestSRC (72).

2.2 Simulation study

To evaluate the RSF method and the implemented RSF backward algorithm a simulation study was performed. The analyses of the simulation study were conducted with the statistical software R (version 3.0.0) and the R-package randomForestSRC (Version 1.2) (72).

2.2.1 Generation of Simulation Data

For the simulation study, data with a sample size of 1000 observations and the following variables were generated:

- As outcome variable the censoring time until event with a follow-up time of approximately ten years on average
- Three predictive variables with normal distribution $X_i \sim N(\mu, \sigma^2), i = 1,2,3$
- One categorical predictive variable with three uniformly distributed categories $X_4 \in \{1, 2, 3\}$
- Four correlated variables XC, for each predictive variable one correlated variable
- One confounding variable $C \sim N(\mu, \sigma^2)$
- 20 random (noise) variables R:
  - Five random variables with normal distribution $R_i \sim N(\mu, \sigma^2), i = 1, \dots, 5$
  - Five random variables with lognormal distribution $R_i \sim LN(\mu, \sigma^2)$, $i = 6, \dots, 10$
  - Three random variables with binominal distribution $R_i \sim B(1000, p), i = 11, 12, 13$
  - Two random categorical variables with three categories which were uniformly distributed $R_i \in \{1, 2, 3\}, i = 14, 15$
  - Five random variables with right skewed distribution $R_i \sim Exp(\lambda),\ i = 16, \dots, 20$

The censoring time (T) was calculated based on the Gompertz distribution with the following formula (75):

$$T = \left(\frac{1}{\alpha}\right) * log\left(1 - \left(\frac{\alpha * \log(U)}{\lambda * \exp(\sum_{i=1}^{k} \beta_i * X_i)}\right)\right)$$

$\alpha = \{x \in R | 0 \leq x \leq 1\}$
$U = \{x \in R | 0 \leq x \leq 1\}$
$\lambda = \{x \in R | 10^{-1} \leq x \leq 10^{-7}\}$
X_i = dependent variables
β_i = beta estimators
k = number of dependent variables

Equation 2

To achieve a simulated censoring time of approximately 10 years, the parameter α and λ were varied within the specified ranges (Equation 2). The event rate was fixed at 10%. The beta parameters of the three continuously predictive variables were chosen so that the variables reflect hazards ratios (HR) of 0.7, 1.3 and 2.0, respectively. For the categorical predictive variable the beta parameter was chosen to reflect a HR of 2.0.

2.2.2 Simulation Analyses

To evaluate the RSF method, three simulation scenarios (**Figure 4**) were performed

(1) Data included four predictive and 20 noise variables.

(2) Data included four predictive, four correlated and 20 noise variables.

(3) Data included four predictive, one confounding variable and 20 noise variables.

50 independent replicated simulation data sets per scenario were created and the RSF method was then applied on each data set. For each resulting RSF model, the minimal depth values of the used variables were recorded. After the 50 computation runs, a boxplot for each variable was calculated based on the minimal depth values.

Moreover, the data of the first scenario were used to determine optimal values for the two RSF parameter *ntree* (number of bootstrap samples) and *nsplit* (number of split per candidate variable). For this purpose, the parameter *ntree* was fixed at 100 and 1000, respectively. For the second parameter *nsplit* values of 1, 2, 5 and 10 were tested. After the evaluation of the parameter *ntree* and *nsplit* in the first scenario, scenarios 2 and 3 were performed with a fixed parameter *ntree* at 1000 bootstrap samples and with the parameter *nsplit* of 1 and 10.

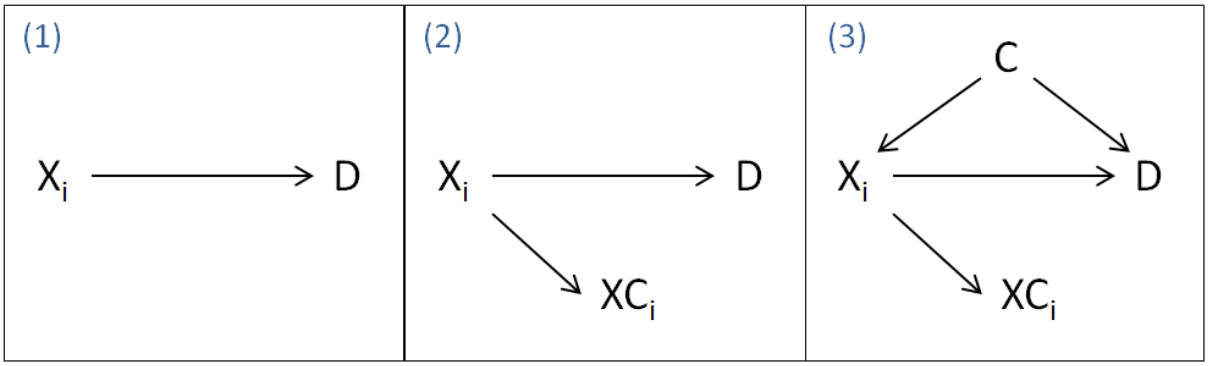

Figure 4: Illustration of the three scenarios used in the simulation study to evaluate the RSF method.
Scenario (1): X_i is associated with D, scenario (2): X_i is associated with D and correlated with XC_i, scenario (3): X_i and XC_i are associated with D and X_i is correlated with XC_i and C is a confounding variable regarding X_i and D. X_i, predictive variable; XC_i variable correlated with X_i; D, outcome; C, confounding variable; i=1-4

To evaluate whether the implemented RSF backward algorithm is able to identify predictive variables regarding time until event, the RSF backward algorithm was applied to the simulated data of scenario 3. For this purpose the parameters *ntree* and *nsplit* were used with a value of 1000 and 1, respectively.

2.3 European Prospective Investigation into Cancer and Nutrition-Potsdam study

In the present thesis, data of the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study were analysed. EPIC-Potsdam is part of the ongoing multicenter prospective EPIC study comprising more than 500,000 participants (mostly aged 35-70 years) from 23 study centers located in ten European countries. The primary aim of the EPIC study is to investigate the association between nutrition and cancer, with the potential for studying other risk factors and chronic diseases (e.g. T2D, cardiovascular disease) as well (76, 77). The EPIC study is coordinated by the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO).

For the EPIC-Potsdam study 27,548 participants that were aged mainly between 35 and 64, were recruited in Potsdam and the surrounding area of Potsdam based on general

population registries (78). The recruitment and baseline examinations were carried out between 1994 and 1998 and most of the participants were followed up to the present (78). The EPIC-Potsdam study procedures were approved by the ethics committee of the medical association of the State of Brandenburg (Germany) and all participants provided written informed consent. At baseline, participants underwent examinations including anthropometric and blood pressure measurements, filled in self-administered questionnaires on diet and lifestyle, and answered personal computer-assisted interviews (76). Blood samples (30 mL) were collected at baseline and immediately fractionated, aliquoted into straws, and stored at -196°C until measurement of serum metabolites.

2.3.1 Blood sample collection at baseline

At baseline, one blood sample of 30 mL venous blood was collected by qualified medical staff from 95.7% of the EPIC-Potsdam participants (76). Prior to the blood withdrawal 28% of the participants had consumed beverages and the other participants had fasted overnight (76). Of the 30 ml blood, 20 mL were inserted into monovettes containing citrate, and 10 mL into monovettes without anticoagulant (76). The blood samples were immediately fractionated into serum, plasma, buffy coat and erythrocytes, and aliquoted into a total of 28 straws of 0.5 mL according to a standardized protocol (76). At the end 12 straws of plasma, 8 straws of serum, 4 straws of buffy coat, and 4 straws of erythrocytes were filled (76). The straws were stored in tanks of liquid nitrogen at -196°C until measurement of serum metabolites (76).

2.3.2 Assessment of serum metabolites

Baseline blood serum samples were analysed to determine metabolite concentrations by using AbsoluteIDQ p150 Kits (Biocrates Life Scienes AG, Innsbruck Austria) which is based on flow injection analysis tandem mass spectrometry (FIA-MS/MS) technique (79). Analysis was done by the Genome Analysis Center at the Helmholtz Zentrum München. For analytical details of serum metabolite concentrations see Römisch-Margl et al. (79).

36 metabolites with concentrations below the detection limit or very high analytical variance were excluded, leaving 127 quantified metabolites for statistical analyses (80). The following metabolite classes were used for statistical analyses (**Table S1**): one hexose (sum of six-carbon monosaccharides without distinction of isomers), 14 amino acids, 14 spingomyelins (SM), 17 acylcarnitines (Cx:y; with x indicating the number of carbon atoms and y indicating the number of double bonds), and 81 glycerophospholipids (37 acyl-alkyl-, 34 diacyl-, and 10 lyso-phosphatidylcholines (PC)). The biological variation and reliability of the metabolites was evaluated in a previous study by Floegel et al. (80).

2.3.3 Assessment of dietary intake

Assessment of dietary intake was ascertained at the time of enrollment by means of a validated, self-administered, semi-quantitative food-frequency questionnaire (FFQ) (81-83). The FFQs has been sent before baseline examinations by post to the participants. At baseline examination the FFQs were collected, optically read and missing or implausible information were requested from the participants on site (84).

The FFQs assessed frequencies and portion sizes of 148 food items (food and beverage consumption) during the 12 month before the baseline examination. The frequencies of consumption were scaled from ‘never’, ‘one time per month or less’ to ‘five times per day or

more'. The estimates of the quantity of portion sizes were visualized by photographs. In addition, fat content of dairy products, fat quality for food preparation and dietary supplement were also recorded. The food items of each participant were converted into grams or milliliters per day. For statistical analysis, the determined 148 food items were than composed into 49 food groups including also beverages (85). In a subset of 104 EPIC-Potsdam participants the validity and reproducibility of the FFQs were evaluated by comparison with repeated 24-hour dietary recalls as described previously (81, 82). Overall, the validity was reported to be moderate or high for estimates of food group intake (85). However, low validity was reported for food groups that were consumed rarely, such as legumes, nuts, and fish (85). Detailed information about how the food groups are composed and how much of the respective food groups were consumed is given in the supplement (**Table S2 and Table S3**).

2.3.4 Assessment of covariates

For anthropometric measurements, the participants were lightly dressed and measured without shoes by qualified personal (84). Body mass index (BMI) was calculated as the ratio of weight (kg) to height squared (m^2). Education at attainment, physical activity and smoking were acquired by a standardized interview. For analyses, education at attainment was categorized as no degree/vocational training, trade/technical school, university degree. Smoking behavior was categorized as never smoker, former smoker, and current smoker (≤ 20 cigarettes/day, and > 20 cigarettes/day). To account for physical activity, the improved physical activity index (IPAI) adjusted for sex and age was calculated as described by Wientzek et al. (86). For identification of metabolites associated with T2D alcohol intake from beverages was categorized into non-consumer, and consumer (women >0-6, 6-12, and

>12 g/day; and men >0-12, 12-24, and >24 g/day). Cases of prevalent T2D, hypertension, myocardial infarction and stroke were assessed during a standardized interview at baseline and verified by the treating physician. Participants with (i) systolic BP≥140 mmHg and/or, diastolic BP≥90 mmHg as defined by the mean of the second and third measurements, (ii) self-reported hypertension diagnosis, or (iii) use of antihypertensive medication at baseline were classified as cases of prevalent hypertension (87).

2.3.5 Ascertainment of endpoints

During the follow-up period, a follow-up questionnaire was send by mail every two to three years to the EPIC-Potsdam participants. The follow-up questionnaire contained items to identify potential incident cases of chronic diseases, including T2D and hypertension (88). Items on a new diagnosis of disease, disease-relevant medication use, or change in diet were requested (88). For all self-reports, the treating physician or clinic was contacted for verification to obtain the exact date and type of the diagnosis (88). In addition, to provide death certificates for confirmation of the diagnosis, local health offices were contacted. This procedure contributed to high quality of follow-up data (88) and enabled the acquisition of only physician confirmed diagnosed cases of incident T2D (ICD-10: E11) and incident hypertension (ICD-10: I10).

2.4 Analytical study sample

The present analysis included two EPIC-Potsdam data sets: One to identify metabolites associated with T2D and the other to identify food groups associated with incident hypertension. The first analysed data set based on a case-cohort study which is nested within the EPIC-Potsdam cohort and was previously used to identify metabolites which were associated with T2D (52). The same data as previously used by Floegel et al. were used in

the present thesis (52). The case-cohort study consists of 849 cases of incident T2D recorded until August 2005 and 2,500 randomly drawn study participants (52). After excluding participants with prevalent T2D at baseline (n=110), with missing or non-verified data on incident or prevalent T2D (n=13), with missing blood samples or biomarker measurement (n=64), and with missing covariate information (n=80) the analytical study population included 800 cases of incident T2D and 2,197 non-cases of incident T2D (**Figure 5**) (52).

The second analysed data set based on the full EPIC-Potsdam cohort. For analysis participants with prevalent hypertension at baseline (n=13,470), with missing or non-verified data on incident hypertension (n=521), cases of prevalent T2D, myocardial infarction or stroke (n=178), with the use of antihypertensive medication (n=3), with missing blood pressure measurements (n=87), with missing covariate information (n=331), with age younger than 35 years or older than 65 years (n=304), with pregnant status at baseline (n=15), with censoring at baseline (n=212), or a follow up time shorter than 1 year (n=23) were excluded (**Figure 5**). The recorded food groups included no missings. After exclusion, the study sample included 1,215 cases of incident hypertension and 10,270 non-cases of incident hypertension. For the identification of food groups associated with incident hypertension gender-differentiated analyses were performed. Thus, the final study sample of men included 389 cases and 3,103 non-cases of incident hypertension and the final study sample of women included 826 cases and 7,167 non-cases of incident hypertension.

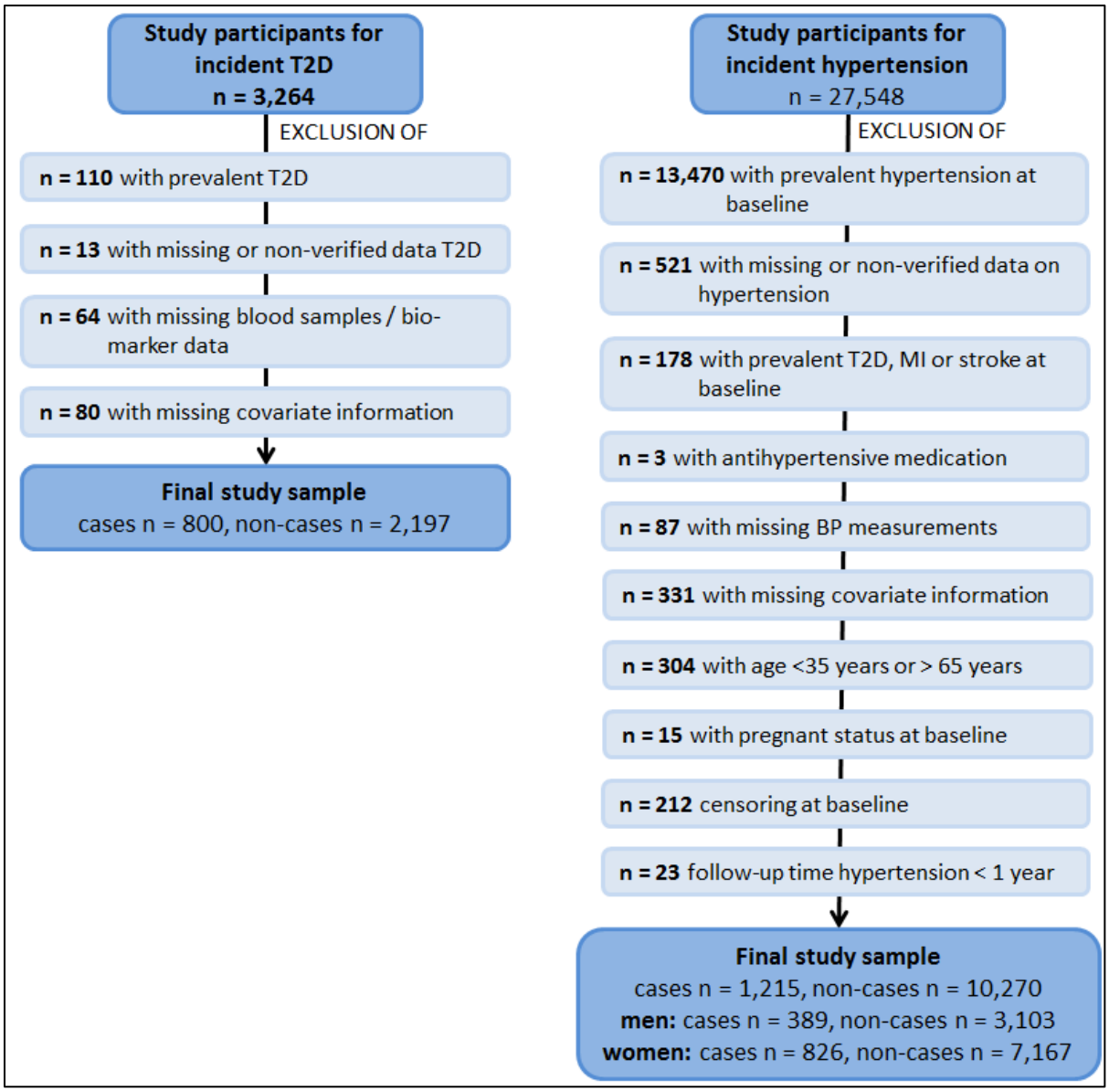

Figure 5: Definition and exclusion criteria of the two analysed EPIC Potsdam study samples.
Cases presented are incident cases. Abbreviations: T2D, type 2 diabetes; Mi, myocardial infarction; BP, blood pressure.

2.5 Statistical Analysis

Statistical analyses were performed with the statistic software R (version 3.0.0), the R-package randomForestSRC (Version 1.2) (72) and SAS software package, release 9.4 (SAS Institute, Cary, NC).

2.5.1 Descriptive statistics

Baseline characteristics of participants of the EPIC-Potsdam study sample - analysed regarding incident T2D - were calculated as age- and sex-adjusted mean and standard error (SE) for continuous variables, or percentages for categorical variables. Baseline characteristics of participants of the EPIC-Potsdam study sample - analysed regarding incident hypertension - were calculated as age-adjusted mean and standard error (SE) for continuous variables, or percentages for categorical variables. Thereby, depending on the analysed study population, properties describing lifestyle, socioeconomic status, as well as obesity measures and medical conditions were considered.

2.5.2 Choice of covariates

For the identification of metabolites associated with incident T2D and food groups associated with incident hypertension two different sets of covariates that may be possible confounding variables were used. For metabolite selection, the same set of covariates was used as previously used by Floegel et al. (52). This set included age, sex, BMI (kg/m²), waist circumference (cm), alcohol intake from beverages (non-consumer, women >0-6, 6-12, and >12 g/day; men >0-12, 12-24, and >24 g/day), smoking (never smoker, former, current ≤20 cigarettes/day, current >20 cigarettes/day), cycling and sports (h/week), level of education (no degree/vocational training, trade/technical school, university degree), coffee intake (cups/day), red meat intake (g/day), whole-grain bread intake (g/day), and prevalent hypertension. The usage of the same set of covariates allowed comparability of selected metabolites by the RSF backward algorithm and the previously applied multivariate Cox proportional hazards regression approach (52).

For food group selection, covariates were chosen based on literature search. Subsequently a minimal sufficient adjustment covariate set was selected according to the Directed Acyclic Graph (DAG) theory using the software DAG program v0.21 (89). The final set of covariates for food group analyses included age, BMI (kg/m²), smoking (never smoker, former, current ≤20 cigarettes/day, current >20 cigarettes/day), level of education (no degree/vocational training, trade/technical school, university degree), IPAI, total energy intake, and prevalent T2D (yes/no).

2.5.3 Variable selection using the Random Survival Forest backward algorithm

The implemented RSF backward selection algorithm was applied to the respective EPIC-Potsdam study sample data, in order to identify metabolites associated with incident T2D and food groups associated with incident hypertension. For metabolite selection, the RSF backward algorithm was applied using the parameter value 1000 for the number bootstrap samples (*ntree*) and the parameter value 10 for random node splitting (*nsplit*). For food group selection the parameter value 1000 was chosen for the number of bootstrap samples (*ntree*) and for random node splitting the parameter *nsplit* was chosen equal to 1. The final set of selected metabolites and food groups was the RSF model with the lowest RSF prediction error rate.

2.5.4 Determination of associations using partial plots

To visualise the associations between concentrations of selected variables and event-free survival, partial plots were computed. The partial plots were derived from a RSF model that was computed based on the final set of selected variables and covariates. For the selected

metabolites partial plots were computed representing associations between metabolite concentrations and 5-year predicted T2D-free survival. For the selected food groups the gender-specific partial plots represent associations between food group intake and 10-year predicted hypertension-free survival.

2.5.5 Analysis of correlation structures using Gaussian Graphic Models

To analyze the correlation structure of the investigated metabolite and food group data and to identify possible patterns Gaussian graphic models (GGMs) were used. A GGM represents undirected probabilistic graphs useful to analyse and visualised the dependency structure of variables of complex data (90). A GGM graph consists of a set of nodes, representing variables, and a set of edges (90). Edges of the GGM represent pairwise correlation between two variables controlled for the correlation effects of all other variables. Thereby, missing edges between two nodes refer to conditional independence between the two nodes and thus between variables.

For serum metabolites an existing GGM was used that was previously calculated as described by Floegel et al. (91). The GGM of serum metabolites was colour-coded with the intention to visualise and identify metabolite patterns. Different colouring of metabolite nodes illustrate metabolites that were selected by the RSF method, a previously applied stepwise Cox regression method, or by both methods.

To identify dietary pattern associated with incident hypertension sex-specific GGMs were derived from the food group data. The GGM analyses were conducted using the statistic software R with the R-packages “corpcor” and “glasso”. To construct the two GGMs, the food group data were, first, log-transformed to fulfil the required normal distribution. The normalised food group data were used to calculate a covariance matrix of the 49 food

groups. In a subsequent step, a spearman partial correlation matrix was estimated from the covariance matrix. To get sparse partial correlation matrices, graphical least absolute shrinkage was applied to the spearman partial correlation matrix. The software yEd graph editor (yWorks GmbH, Tübingen, www.yworks.com) was used to visualise the resulting spearman partial correlation matrix as a GGM. Colour-coding of the GGM was used to illustrate directions of associations between selected food group concentrations and 10-year predicted hypertension-free survival derived from the partial plots.

3 Results

3.1 Simulation study

3.1.1 Sensitivity of the minimal depth measurement to *ntree* and *nsplit*

In the following chapter, results are presented that showed how the minimal depth values of the respective variables is altered regarding modification in the number of bootstrap samples (*ntree*) and numbers of splits (*nsplit*). The analyses were conducted with the data of the first scenario.

The present results showed - independent of the chosen parameter for *ntree* and *nsplit* – that predictive variables had lower minimal depth values than noise variables (**Figure 6**). In addition, different minimal depth values were also observed for the predictive variables. Indeed, the two predictive variables 03 and 04 with simulated HR of 2.0 had lower minimal depth values than the predictive variables 01 and 02 with hazard ratios of 0.7 and 1.3, respectively. However, the results also demonstrated that the choice of the parameter values of *ntree* and *nsplit* affected the minimal depth values of predictive and noise variables.

Modification of the parameter *ntree* from 100 to 1000 bootstrap samples resulted in lower interquartile ranges and lower distance between the minimum and maximum values of the boxplots of the respective simulated variables when using higher number of bootstrap samples (**Figure 6**). This indicates improved precision of the minimal depth measurement and thus greater stability of the computed RSF models when using 1000 bootstrap samples instead of 100 bootstrap samples.

Tree based methods are known to favour node splits on continuous variables when a mixture of continuous and categorical variables is used. This bias was also demonstrated in the results obtained in the present thesis. A modification of the RSF parameter *nsplit* affected especially the computed minimal depth values of simulated categorical variables

(**Figure 6** predictive04 and r11 to r15). If the value of the parameter *nsplit* was 1, the simulated categorical variables were ranked similar as comparable simulated continuous variables regarding the minimal depth measurement. However, when the value of *nsplit* increased from 1 to 2, 5 and 10, then the minimal depth values of the categorical variables worsened, whereas the minimal depth values of the simulated continuous variables did not vary due to increasing values of *nsplit*.

The predictive variable 03 and 04 had the same hazard ratio of 2.0. However the predictive variable 04 was a categorical variable. With increasing values of the parameter *nsplit* the minimal depth value of the predictive categorical variable increased and thus worsened. In contrast, the minimal depth value of the predictive variable 03 remained relatively constant with increased values of the parameter *nsplit*. However, this increase of minimal depth values for the predictive categorical variable 04 was only small compared with the alteration of the minimal depth values of the noise categorical variables r11-r15. For the noise categorical variables the minimal depth values worsened stronger with increasing values of the parameter *nsplit*, whereas the minimal depth values of noise continuous variables remains even constant. Moreover, the noise categorical variables r11 to r13 presented binominal categorical variables, whereas the noise categorical variables r14 and r15 presented variables with three categories. As shown an increase of the parameter value *nsplit* resulted in a greater increase of the minimal depth values of the two binominal noise variables than for the minimal depth values of the three categorical noise variables with three categories.

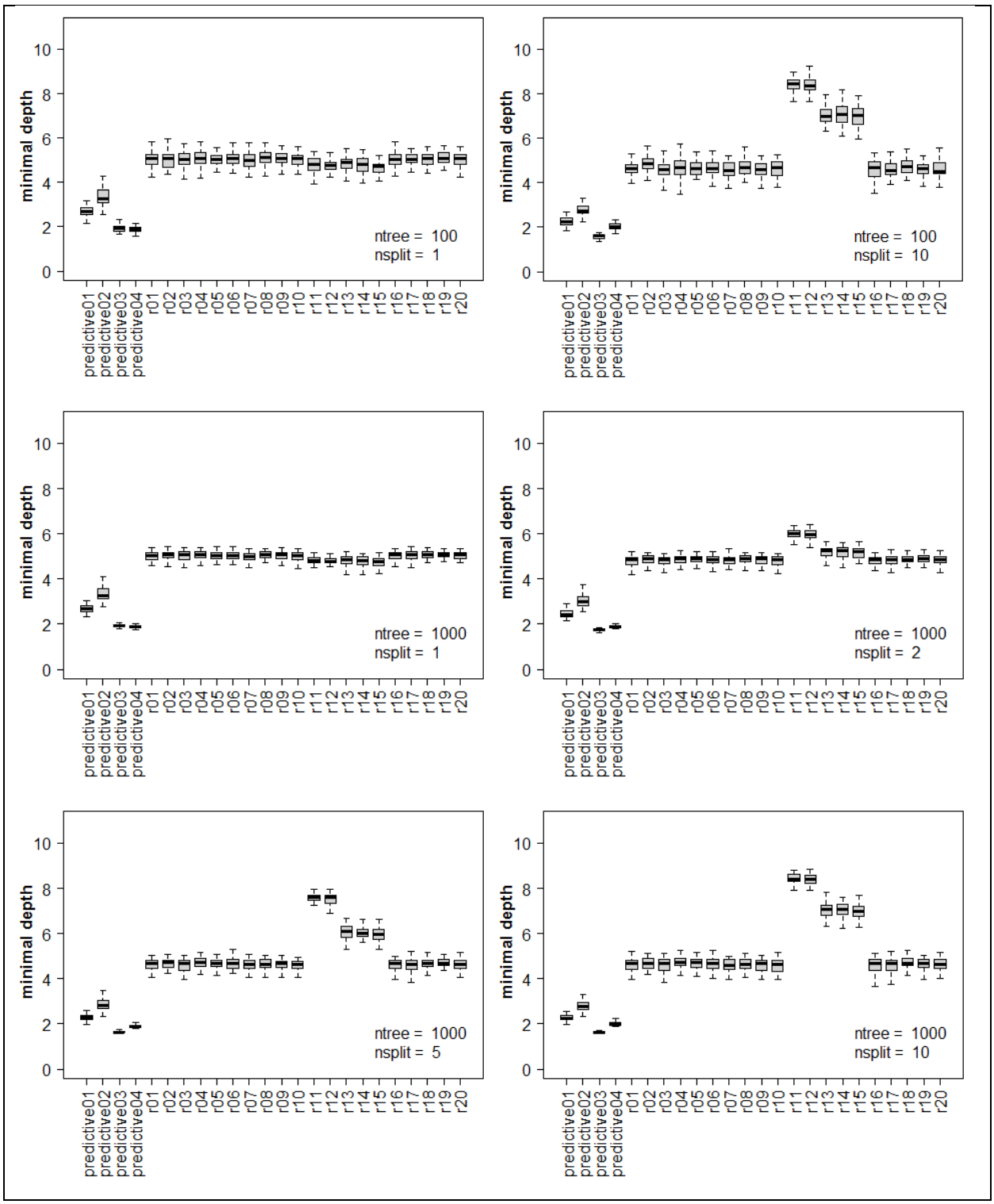

Figure 6: Boxplots for minimal depth values of predictive and noise variables for different values of *ntree* and *nsplit*.

A boxplot represents the minimal depth values of 50 computed random survival forests for the respective variable. r01-r04 = noise normal distributed variables simulated Hazard ratios: predictive01 = 0.7, predictive02 = 1.3, predictive03 and 04 = 2.0.

In summary, higher values of the parameter *ntree* enhanced the precision of the minimal depth values. If the value of the parameter *nsplit* was equal to 1 categorical variables had

comparable minimal depth values as continuous variables. Moreover, the results also demonstrated - independent of modifications of the parameter *ntree* and *nsplit* – that all predictive variables had lower minimal depth values than the noise variables indicating that RSF is able to distinguish between predictive and noise variables.

3.1.2 Sensitivity of the minimal depth measurements to correlated variables

In this chapter results for scenario 2 are presented. In scenario 2 for each predictive variable an associated correlated variable was added to the simulated data (**Figure 7**, assignment of variables: predictive 01 and correlated01; predictive02 and correlated02; and so forth). The inclusion of the correlated data led to a general increase in the minimal depth values (**Figure 7**). Thus, the minimal depth values of noise variables changed from approximately 5 to 6 when comparing data with and without correlated variables. However, the addition of correlated variables did not affect the differentiation between predictive and noise variables.

The correlated variables had lower minimal depth values than the noise variables (**Figure 7**). Compared to the minimal depth values of the respective predictive variables, the correlated variables showed higher minimal depth values. Moreover, the two correlated variables 03 and 04 of the predictive variables 03 and 04 with HR of 2.0 had lower minimal depth values than the correlated variables 01 and 02 of the predictive variables with HR of 0.7 and 1.3. The minimal depth values of the two correlated variables 03 and 04 were minimal lower than or similar to the minimal depth values of the predictive variables 01 and 02. Alterations in the values of the parameter *nsplit* from 1 to 10 affected especially the correlated variable04, resulting in a worsened minimal depth value for correlated variable04.

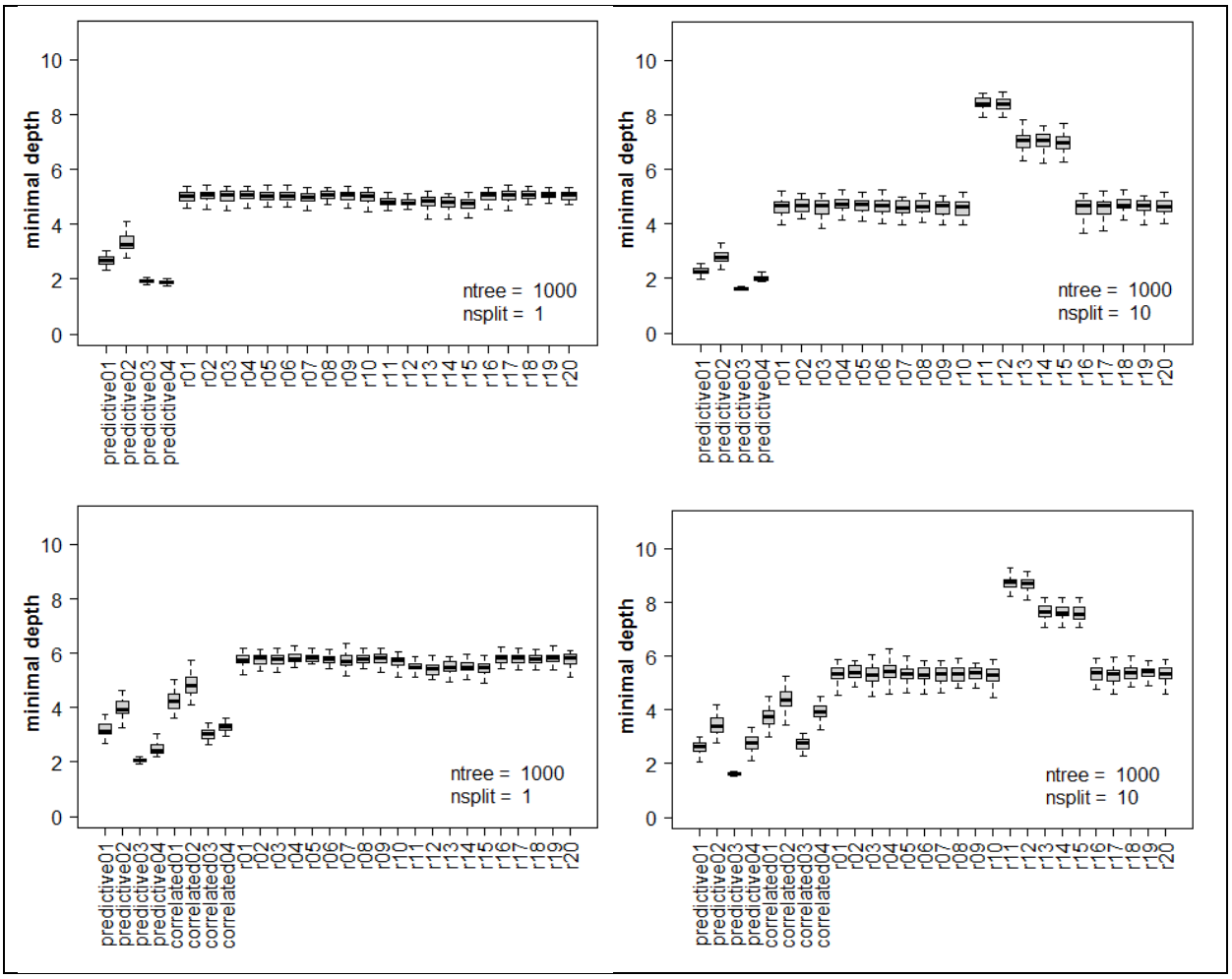

Figure 7: Boxplots for minimal depth values of predictive, correlated and noise variables.
In the top row, as reference, the results without correlated variables and in the bottom row with correlated variables. A boxplot represents the minimal depth values of 50 computed random survival forests for the respective variable. Results are shown for *ntree* values (number of bootstrap samples) of 1000 and *nsplit* values (number of splits per variable) of 1 and 10, respectively.

These results indicate that predictive as well as their correlated variables had lower minimal depth values than the noise variables. Thus, existing correlation structures with a true signal regarding time until event may be identified in RSF models when using the minimal depth measurement.

3.1.3 Sensitivity of the minimal depth measurement to covariates

In this chapter results for scenario 3 are presented. In scenario 3 a confounding factor (covariate) was added to the simulation data of scenario 2. When comparing the boxplot of minimal depth values of RSF models with and without the covariate no great differences

were observed (**Figure 8**)Figure 8: Boxplots for minimal depth values of predictive and noise variables taking into account a covariate.. However the predicted variable 01 and their correlated variable 01 showed decreased minimal depth values when the covariate was considered for computation of the respective RSF model. However, this shift was moderate and the predictive variable 01 still had minimal depth values which were lower than the minimal depth values of the noise variables. In addition, in the RSF model where the covariate was considered, the covariate had lower minimal depth values than the noise variables.

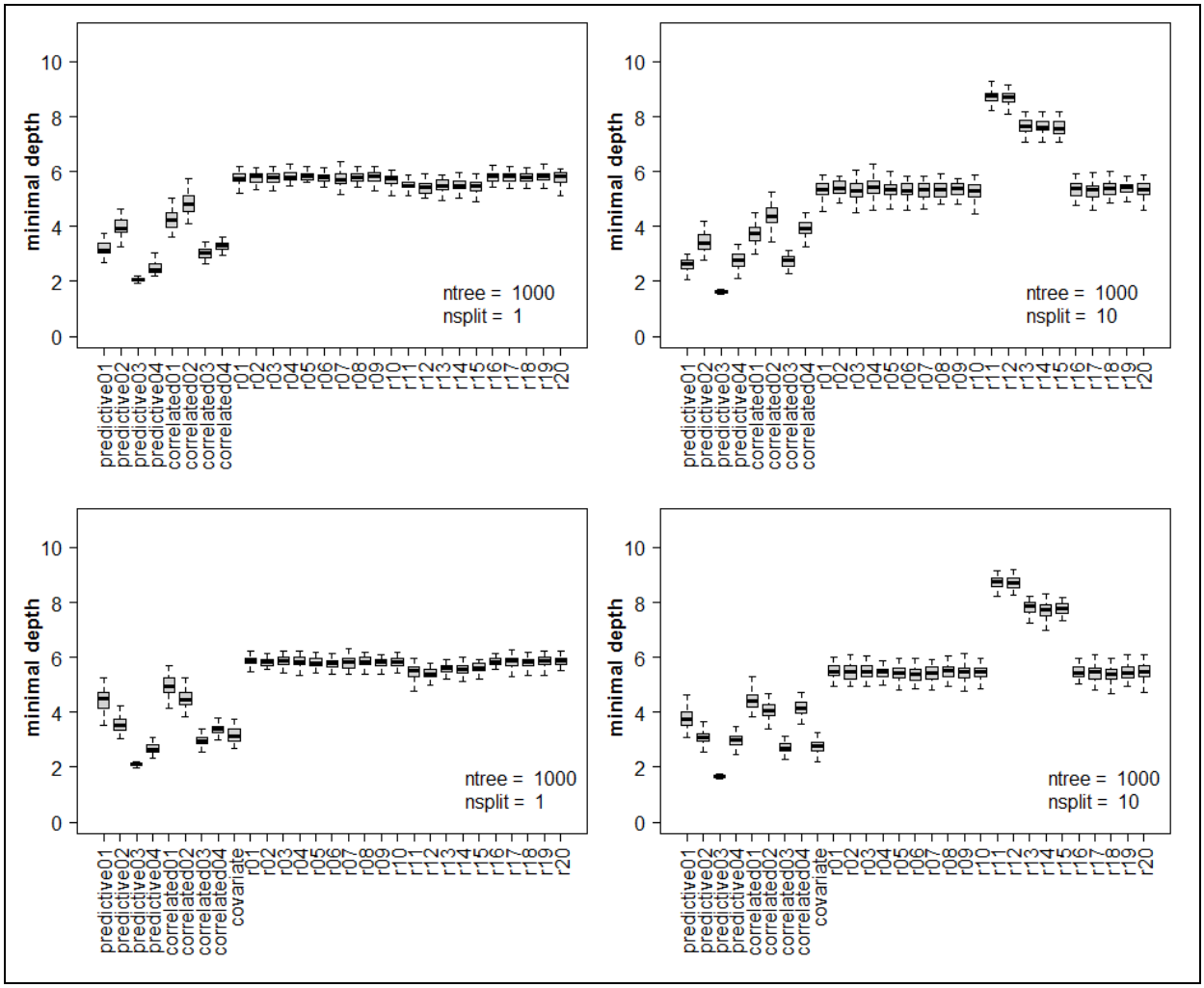

Figure 8: Boxplots for minimal depth values of predictive and noise variables taking into account a covariate.

In the top row, as reference, the averaged results of 50 RSF models that were computed without a covariate and in the bottom row the averaged results of 50 RSF models that were computed with a covariate. A boxplot represents the minimal depth values of 50 computed random survival forests for the respective variable. Results are shown for *ntree* values (number of bootstrap samples) of 1000 and *nsplit* values (number of splits per variable) of 1 and 10, respectively.

These results indicate that RSF models computed without possible confounding factors can result in misleading minimal depth values. The inclusion of possible confounding factors can contribute to the avoidance of misleading minimal depth values.

3.1.4 Sensitivity of the Random Survival Forest backward selection procedure

In the next step, the RSF backward selection procedure was applied to the simulation data consisting of predictive, correlated and noise variables as well as one confounding variable with the purpose to identify variables associated with time until event. **Table 1** illustrates the results of the stepwise backward selection procedure.

In the first 20 steps the noise variables were removed which was accompanied by an improvement of the RSF error rate from values of 0.0380 to 0.0262. Moreover, the RSF error rate of 0.0262 in the selection step 20 represents the smallest RSF error rate during the whole selection process. No further improvement of the RSF error rate was achieved by removal of other variables. Per definition of the RSF backward algorithm the final model to be chosen is the model with the minimal RSF error rate. Derived from **Table 1** the final model which most accurate predict time until event includes all predictive variables and all correlated variables.

Table 1: Selection steps of the Random Survival Forest backward algorithm when applied to simulation data.

selection step	stepwise removed variable	RSF error rate
0	full model	0.0380
1	random variable 07	0.0367
2	random variable 16	0.0368
3	random variable 06	0.0362
4	random variable 19	0.0385
5	random variable 02	0.0361
6	random variable 04	0.0357
7	random variable 10	0.0325
8	random variable 18	0.0343
9	random variable 09	0.0329
10	random variable 14	0.0328
11	random variable 20	0.0318
12	random variable 15	0.0308
13	random variable 05	0.0320
14	random variable 03	0.0324
15	random variable 08	0.0310
16	random variable 17	0.0296
17	random variable 01	0.0296
18	random variable 13	0.0276
19	random variable 12	0.0265
20	random variable 11	**0.0262**
21	correlated variable 01	0.0283
22	predictive variable 01	0.0375
23	correlated variable 02	0.0372
24	correlated variable 04	0.0360
25	predictive variable 02	0.0465
26	correlated variable 03	0.0445
27	predictive categorical variable 04	0.1012
28	predictive variable 03	1.0000

Shown are the removed variables at each selection step and the corresponding RSF error rate of the respective RSF model at the selection step. Variables of the final RSF model were blue coloured. Note: when a variable was removed from the data the resulting RSF error rate was calculated and recorded for the respective selection step.

In summary, the present results confirm the usefulness of the RSF backward algorithm as a variable selection tool for survival analysis. The application of the RSF backward algorithm to the simulated data resulted in a final RSF model that included all predictive variables and correlated variables but no noise variables.

3.2 Identification of metabolites associated with incident type 2 diabetes [1]

In the next step the RSF backward algorithm was applied to observational data of the EPIC-Potsdam study in order to identify metabolites associated with incident T2D. The baseline characteristics of the analysed study population are presented in **Table 2**. In general, non-cases of incident T2D included a greater proportion of women, had lower prevalence rates of hypertension and were younger than cases of incident T2D. Moreover, non-cases of incident T2D had a lower BMI and waist circumference, a higher level of education, a higher intake of whole grain bread and a lower intake of red meat than cases of incident T2D. Furthermore, non-cases of incident T2D tend to smoke less than cases of incident T2D.

Table 2: Baseline Characteristics of the analysed EPIC-Potsdam study population.

	Non-cases (n=2179)	Incident Type 2 Diabetes Cases (n=800)
Age (years) †	49.3 (8.9)	54.7 (7.3)
Women (%) †	57.8	42.2
BMI (kg/m²)	25.9 (0.09)	30.1 (0.15)
Waist circumference, men (cm) ‡	93.3 (0.34)	103.6 (0.46)
Waist circumference, women (cm) ‡	80.1 (0.30)	93.4 (0.62)
Prevalent Hypertension (%)	48.3	70.8
Education		
No degree/vocational training (%)	36.6	45.6
Trade/technical school (%)	23.9	25.4
University degree (%)	39.5	29.0
Smoking status		
Never (%)	47.4	36.2
Former (%)	32.3	42.3
Current (%)	20.3	21.5
Among smokers: number of cigarettes/d	12.5 (0.44)	16.0 (0.74)
Physical activity (h/week) §	2.9 (0.08)	2.2 (0.13)
Alcohol intake from beverages (g/d)	14.8 (0.42)	14.5 (0.71)
Coffee consumption (cups/d)	2.8 (0.05)	2.7 (0.08)
Whole grain bread intake (g/d)	46.4(1.12)	38.2 (1.91)
Red meat intake (g/d)	43.0 (0.62)	48.9 (1.06)

Presented are age- and sex-adjusted mean (standard error) for continuous variables or percentages for categorical variables. †Unadjusted mean (standard deviation) or percent. ‡Age-adjusted mean (standard error). §Average of cycling and sports during summer and winter season. Abbreviation: BMI, body mass index

[1] In the framework of this thesis, parts of the results are from a manuscript which is in revision at the International Journal of Epidemiology: (92). Dietrich S., Floegel A, Boeing H., Schulze M.B., Illig T., Pischon T., Knüppel S., Drogan D. Random Survival Forest in practice – a method for modelling high-dimensional metabolomics data in time to event analysis. In revision at International Journal of Epidemiology

The application of the RSF backward algorithm on the data of 127 metabolites resulted in a reduced set of 16 metabolites (**Figure 9**). This set of metabolites had the smallest RSF prediction error rate (0.165) during selection process suggesting high relevance of this subset of metabolites for incident T2D (**Table 3**). A RSF model based on data consisting of covariates only or on data consisting of all metabolites and the covariates resulted in higher prediction error rates of 0.217 or 0.172, respectively. This suggests that the selected metabolites may improve the prediction of incident T2D when the selected metabolites are used in addition to known risk factors (covariates) in a RSF model.

Table 3: Computed prediction error rates of different Random Survival Forest models regarding the incident of type 2 diabetes.

RSF model	RSF prediction error rate
Only covariates	0.217
All metabolites + covariates	0.172
Selected metabolites + covariates	0.165

Among the selected metabolites, hexose appeared to have the strongest influence on T2D risk according to the minimal depth measurement. Beside hexose, the metabolites acyl-alkyl-PC C34:3 and diacyl-PC C38:3 had also low minimal depth values. Furthermore several acyl-alkyl-PC (C42:4, C42:5, C44:4, C44:5, C44:6), diacyl-PC (C32:0, C42:0, C42:1), aminoacids (valine, tyrosine, glycine), lyso-PC C18:2 and acylcarnitine C16 were selected.

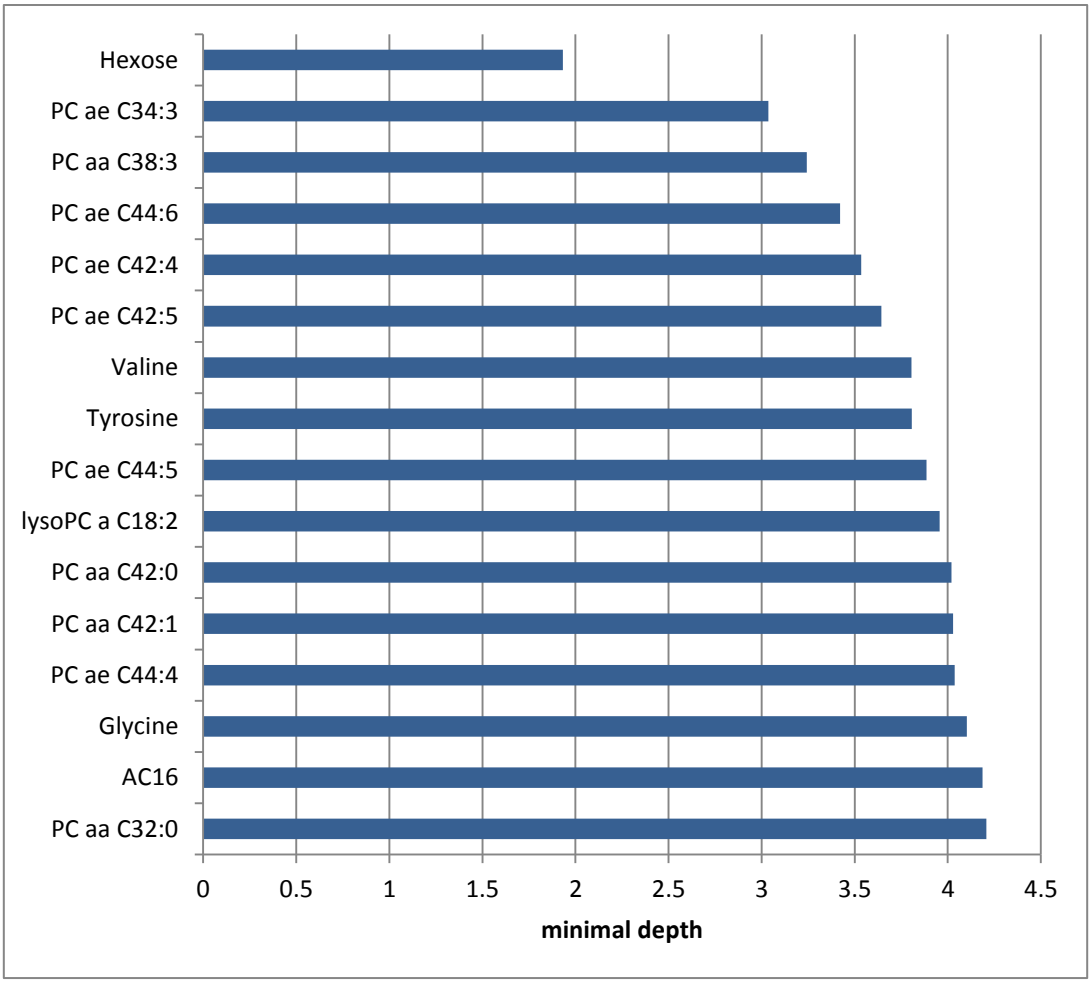

Figure 9: Selected metabolites that are most predictive for incident type 2 diabetes. Metabolites are ranked by the minimal depth measurement. a, acyl; aa, diacyl; ae, acyl-alkyl; PC, phosphatidylcholine; AC, acylcarnitine. Reference: Dietrich et al. (92).

Direction and non-linearity between selected metabolites and predicted 5-year predicted T2D-free survival was assessed visually in partial plots (**Figure 10**). The 5-year predicted T2D-free survival decreased noticeably as values of hexose, diacyl-PC C38:3, valine, tyrosine, and acylcarnitine C16 increased. Threshold values were approximately 5000 µmol/L of hexose and 50 µmol/L of diacyl-PC C38:3. Individuals with the lowest values of hexose had approximately 25% higher 5-year predicted T2D-free survival compared to individuals with highest values. In contrast, increasing values of all selected acyl-alkyl-PC, lyso-PC C18:2, glycine as well as diacyl-PC C42:0 and C42:1 were associated with an increase of 5-year predicted T2D-free survival. Most of the partial plots indicate a non-linear relationship between the respective metabolites and 5-year predicted T2D-free survival.

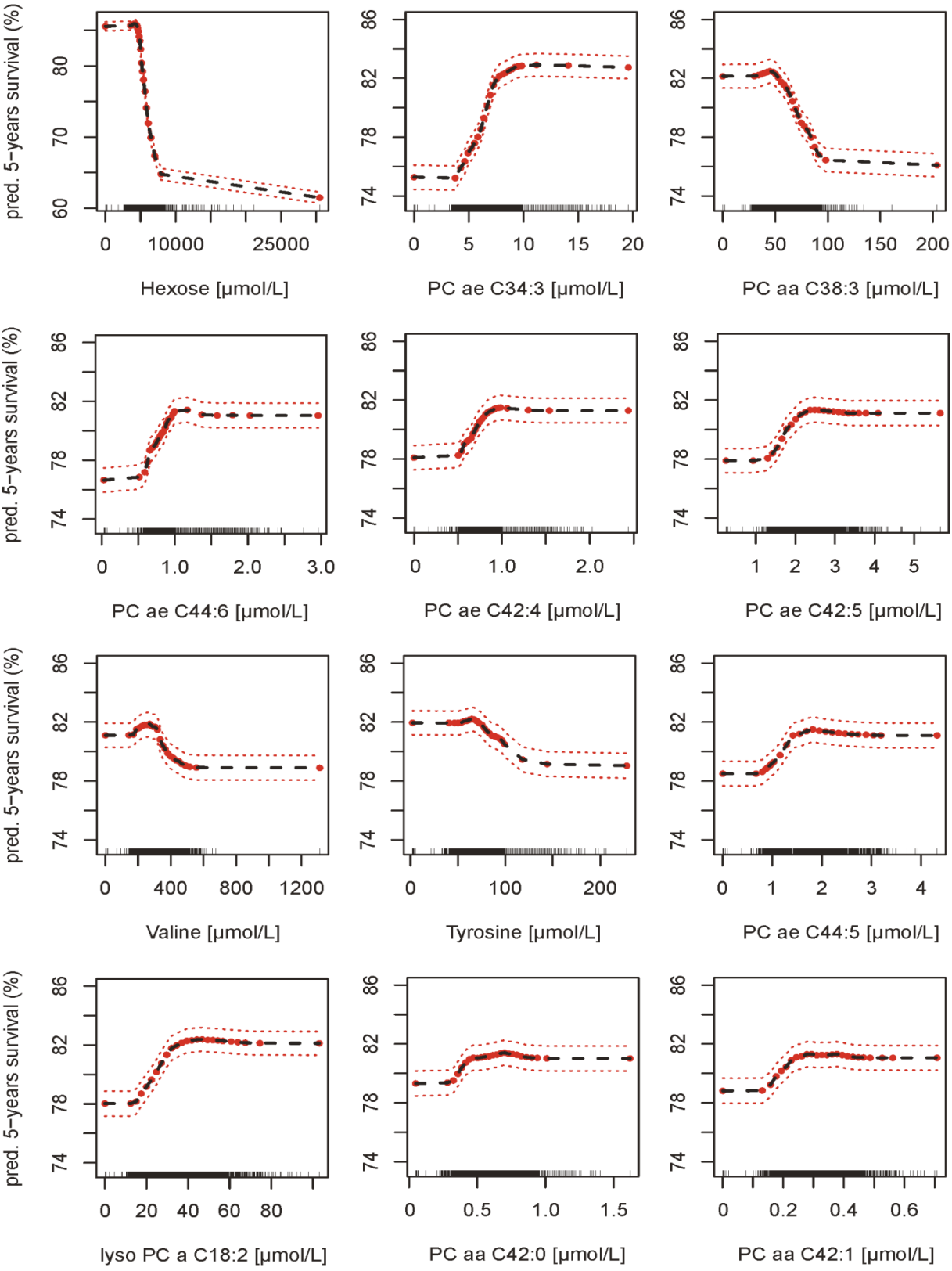

Figure 10: Partial plots of the selected metabolites most informative regarding incident T2D.
The plots including the partial values (red points) ± 2 SE (dashed red lines). Values on the vertical axis represent predicted 5-year type 2 diabetes-free survival for a given variable after adjusting for all other variables (covariates and selected metabolites). a, acyl; aa, diacyl; ae, acyl-alkyl, PC, phosphatidylcholine; AC, acylcarnitine. Reference: Dietrich et al. (92).

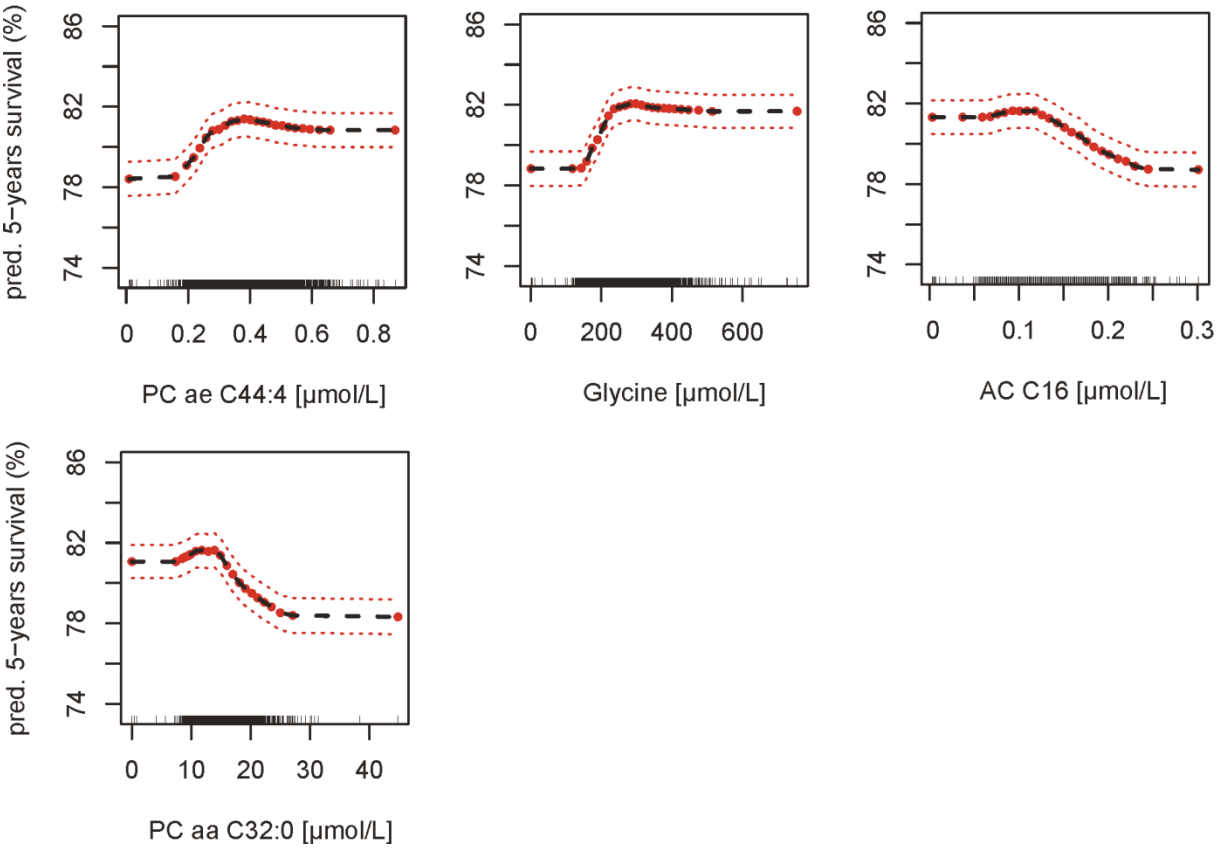

Figure 10 continued

In a previous study, Floegel et al. (52) applied a multivariate Cox proportional hazards regression approach to the same data and equally identified several metabolites associated with incident T2D. The by the two approaches identified metabolites are listed in Table 4. Partly, different metabolites were identified. However, with both approaches the most predictive metabolites were hexose, acyl-alkyl-PC C34:3 and diacyl-PC C38:3. The amino acids tyrosine and valine were selected only by the RSF backward algorithm, whereas phenylalanine was selected by the previously applied multivariate Cox proportional hazards regression approach (52).

In a further step, the identified metabolites were colour-coded in a GGM network (**Figure 12**). A great part of the identified metabolites are distributed throughout the network with no connection to each other, regardless of the selection approach used. However, one metabolite cluster was observed containing five acyl-alkyl-PCs and two diacyl-PC. All seven

metabolites were selected by the RSF backward algorithm, but only three of the acyl-alkyl-PCs were selected by the stepwise Cox proportional hazards regression. The metabolites of this metabolite cluster were connected by edges that based on high partial correlation coefficients.

In **Figure 11** the partial correlation coefficient of all acyl-alkyl-PCs which were selected by the RSF backward algorithm and those selected previously with the stepwise Cox proportional hazards regression were illustrated. RSF selected, in particular, acyl-alkyl-PCs which were in general higher correlated with each other than acyl-alkyl-PCs which were selected by the stepwise Cox proportional hazards regression.

Table 4: Comparison between multivariate Cox proportional hazards regression and the Random Survival Forest backward algorithm regarding selected metabolites.

Metabolites	Both methods	Cox PH Regression	Random Survival Forest
Hexose	+	+	+
Phenylalanine	-	+	-
Glycine	+	+	+
Valine	-	-	+
Tyrosine	-	-	+
SM C16:1	-	+	-
AC C16:0	-	-	+
PC ae C34:3	+	+	+
PC ae C40:6	-	+	-
PC ae C42:4	-	-	+
PC ae C42:5	+	+	+
PC ae C44:4	+	+	+
PC ae C44:5	+	+	+
PC ae C44:6	-	-	+
PC aa C32:0	-	-	+
PC aa C32:1	-	+	-
PC aa C36:1	-	+	-
PC aa C38:3	+	+	+
PC aa C40:5	-	+	-
PC aa C42:0	-	-	+
PC aa C42:1	-	-	+
LysoPC a C18:2	+	+	+
Number of selected variables	8	14	16

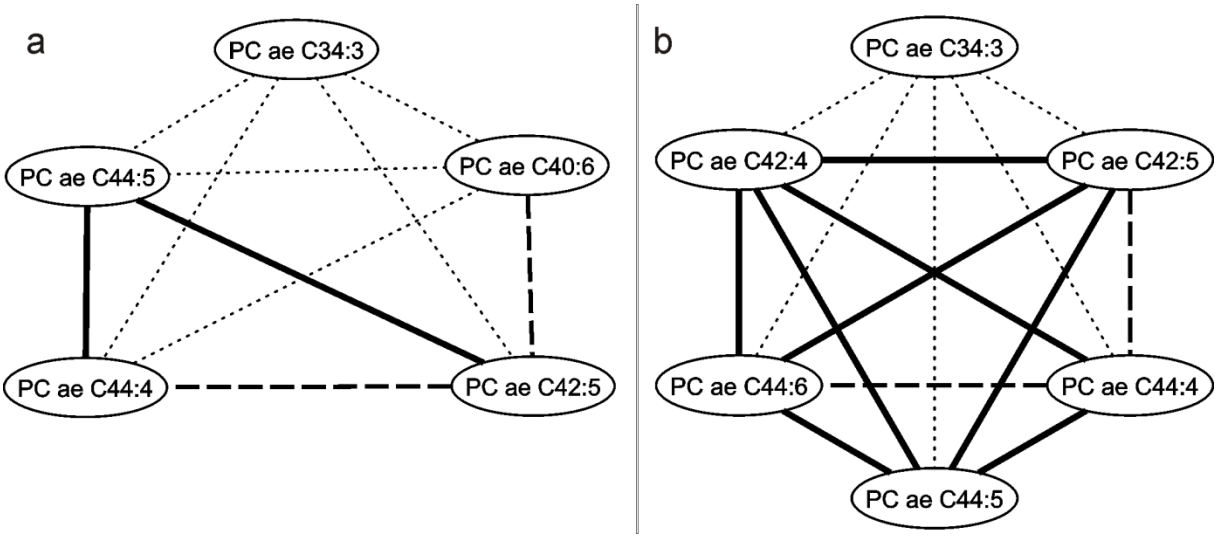

Figure 11: Correlation structure for selected acyl-alkyl phosphatidylcholines.
The presented acyl-alkyl phosphatidylcholines were selected by (a) Cox proportional hazards regression analysis by Floegel et al. (52) and (b) Random Survival Forest backward algorithm. Lines represent spearman correlation coefficients adjusted for used covariates. Dotted lines rs = 0 – 0.5, Thin dashed lines rs = 0.5 – 0.75, thick lines rs > 0.75. ae = acyl-alkyl; PC = phosphatidylcholine. Reference: Dietrich et al. (92).

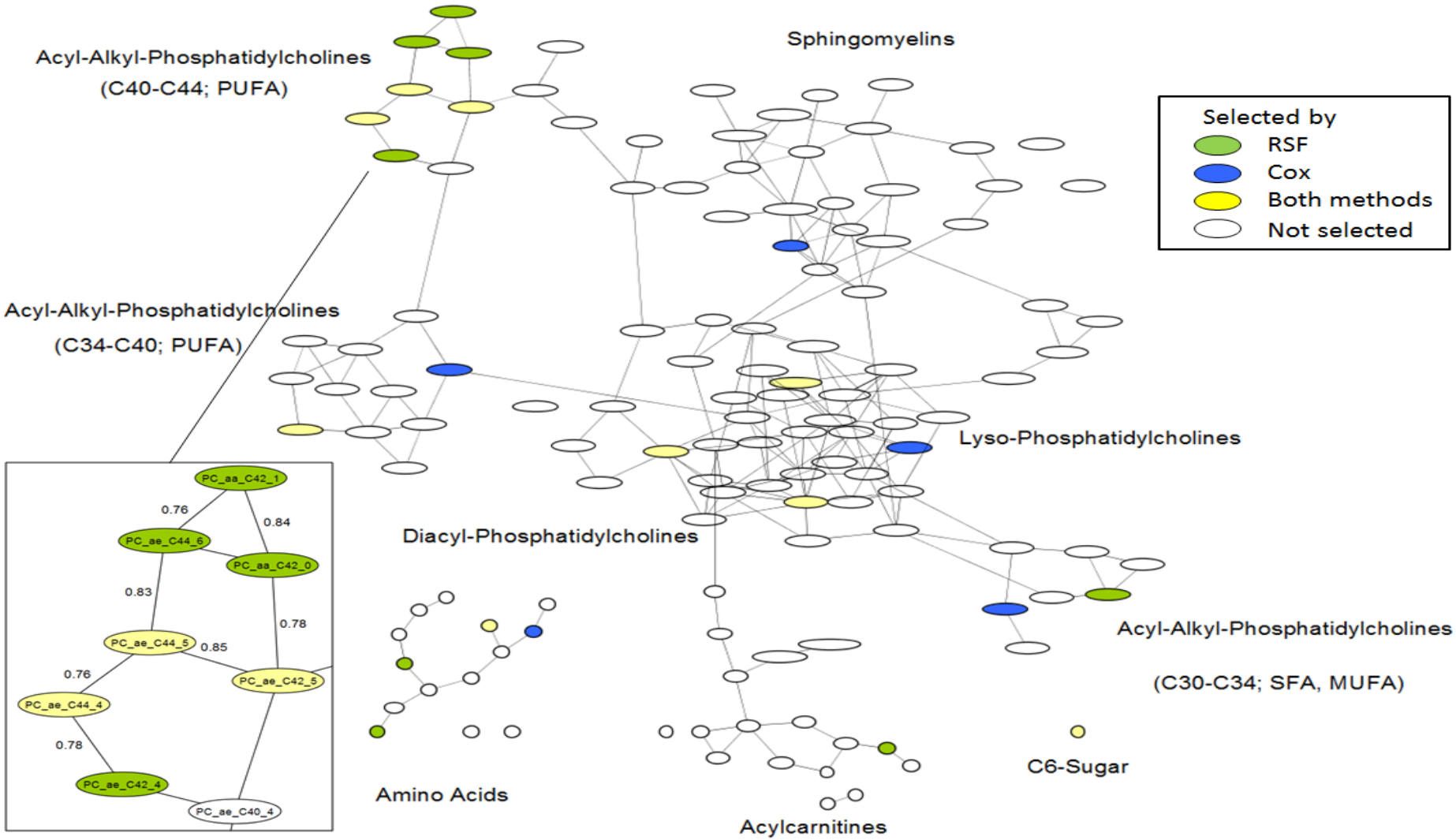

Figure 12: Serum metabolite network of the EPIC-Potsdam subcohort. Each node represents one metabolite and each edge between two nodes represents the partial correlation between two metabolites mutually adjusted for the other metabolites. Green nodes represent metabolites selected by random survival forest, blue nodes metabolites selected by Cox proportional hazards regression and yellow node metabolites selected by both methods. The most important pattern was resized and filled with metabolite names and partial correlation coefficients. Reference: Dietrich et al. (92).

3.3 Identification of food groups associated with incident hypertension

The baseline characteristics of men and women, whose EPIC-Potsdam food group data were analysed in the present thesis, are summarised in **Table 5**. For both sexes cases of incident hypertension tend to be older, had a higher BMI and a higher blood pressure at baseline than non-cases of incident hypertension. Moreover, for both sexes cases of incident hypertension tend to be less educated and had higher prevalence rates of T2D than non-cases of incident hypertension. However, women were on average younger, smoked less, consumed less alcohol and were less educated than men.

Table 5: Baseline characteristics of male and female participants of the analysed EPIC-Potsdam sample population.

	Men		Women	
	Non-cases (n=3103)	cases (n=389)	Non-cases (n=7167)	cases (n=826)
Age (years) †	49.6 (7.5)	52.0 (7.5)	46.0 (8.5)	49.9 (8.7)
BMI (kg/m²)	25.4 (0.05)	26.7 (0.15)	24.2 (0.04)	25.8 (0.12)
Systolic BP (mmHg)	122.5 (0.16)	127.7 (0.45)	115.3 (0.11)	122.5 (0.33)
Diastolic BP (mmHg)	79.6 (0.11)	82.2 (0.31)	76.2 (0.08)	80.5 (0.23)
Education				
No degree/vocational training (%)	30.7	31.9	37.1	40.8
Trade/technical school (%)	14.1	14.7	28.7	30.5
University degree (%)	55.2	53.5	34.2	28.7
Smoking status				
Never (%)	35.3	30.6	54.1	58.5
Former (%)	37.5	38.8	25.5	21.2
Current (%)	27.3	30.6	20.4	20.3
Among smokers: number of cigarettes/day	15.5 (0.34)	16.2 (0.90)	10.9 (0.20)	10.6 (0.59)
Improved Physical Activity Index (IPAI)	35.9 (4.2)	35.5 (3.9)	37.2 (0.04)	37.0 (0.11)
Alcohol intake from beverages (g/day)	22.0 (0.42)	22.7 (1.20)	8.7 (0.12)	8.6 (0.38)
Prevalent T2D (%)	3.2	7.5	1.4	4.8
Total energy intake	10642.6 (55.6)	10338.6 (157.7)	8065.5 (28.4)	8260.0 (84.2)
Follow-up time (years) †	10.7 (2.1)	6.4 (2.6)	10.8 (1.8)	6.3 (2.7)

Presented are age- and sex-adjusted mean (standard error) for continuous variables or percentages for categorical variables. †Unadjusted mean (standard deviation). Abbreviation: BMI, body mass index

3.3.1 Selected food groups associated with incident hypertension in men

The application of the RSF backward algorithm on EPIC-Potsdam food group data for men resulted in a reduced set of 17 selected food groups (**Figure 13**). The set of 17 food groups -

together with used covariates - presents the final RSF model with the smallest RSF prediction error rate during the selection process (**Table 6**). The RSF error rate of the final model was 0.363. A RSF model that was computed from data that includes all food groups and the covariates or only the covariates without food groups showed prediction error rates of 0.396 or 0.393, respectively. This indicates that the selected food groups improve the prediction of incident hypertension in men when used together with the covariates in a RSF model.

Table 6: Computed prediction error rate of different Random Survival Forest models for men regarding incident hypertension.

RSF model	RSF prediction error rate
Only covariates	0.393
All food groups + covariates	0.396
Selected food groups + covariates	0.363

Within the 17 selected food groups, pizza, spirits, vegetarian dishes, high fat dairy products and low fat dairy products showed lowest minimal depth values compared to the other selected food groups (**Figure 13**). This indicates that these food groups are more predictive regarding incident hypertension than the other selected food groups. However, the covariates BMI, age, prevalent T2D and IPAI showed lower minimal depth values than these four selected food groups. Further selected food groups, ranked by the minimal depth values, were soup, mushrooms, pasta and rice, other vegetable fat, wine, cabbage, high fat cheese, tea, sauce, fruit juice, raw vegetables and whole grain bread (**Figure 13**). However, the differences in the minimal depth values were low. The best-rated food groups (pizza and spirits) had minimal depth values of 3.7, whereas the food groups with the worst rates (raw vegetables and whole grain bread) had minimal depth values of approximately 4.2 (**Figure 13**).

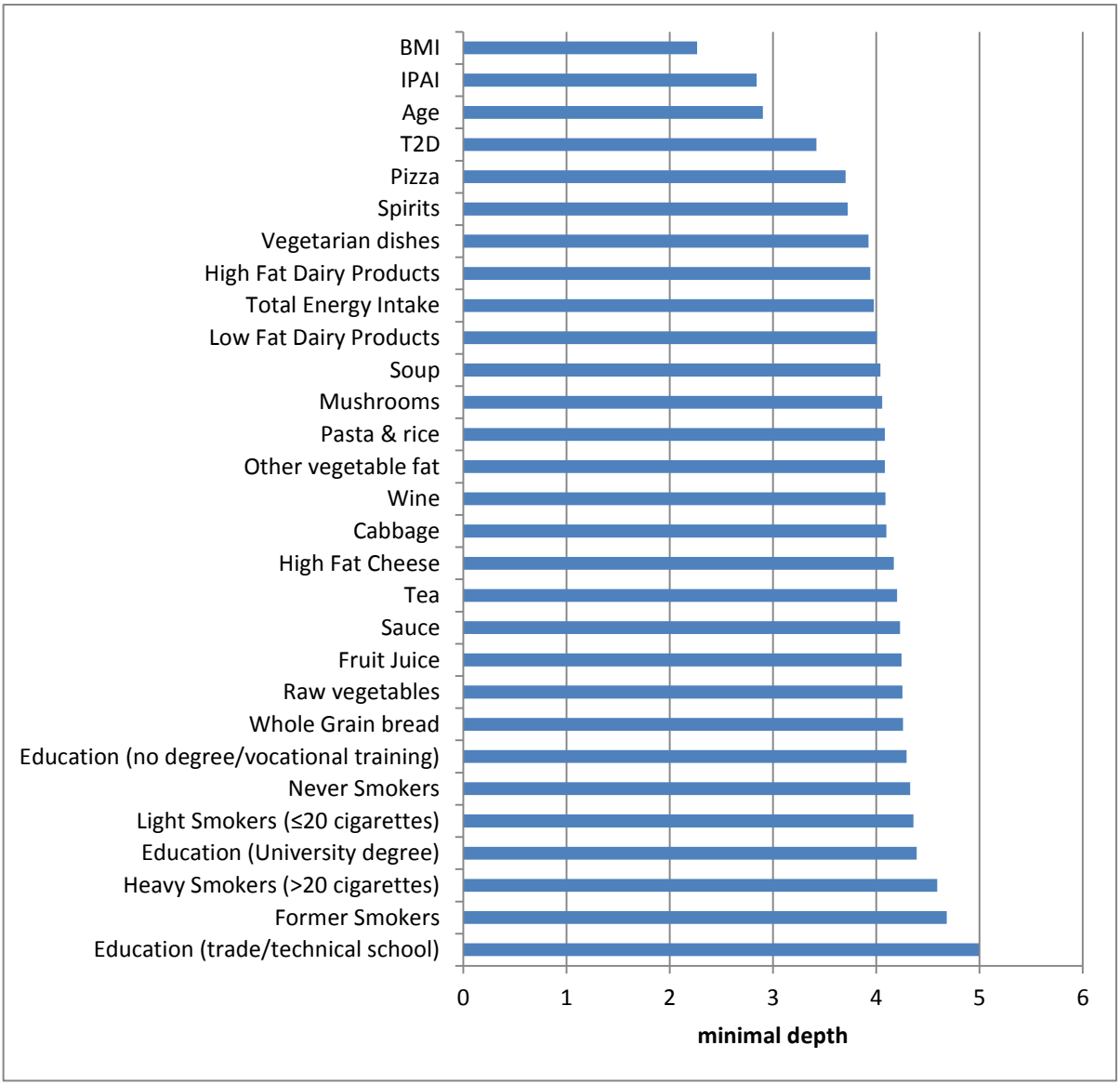

Figure 13: Identified food groups and covariates included in the final RSF which were most informative for incident hypertension in men. The identified food groups and covariates are ranked by the minimal depth measurement.

The visualisation by the partial plots revealed nonlinear associations between selected food groups and 10-year predicted hypertension-free survival (**Figure 14**). For low to medium amounts of consumption of Pizza, vegetarian dishes, high fat dairy products, mushrooms, pasta and rice, cabbage, high fat cheese, and fruit juice an increased 10-year predicted hypertension-free survival was observed. However, for higher consumption levels of these foods, the association between food groups and 10-year predicted hypertension-free survival remained constant. This indicates that a moderate consumption of these food

groups is protective regarding incident hypertension. Contrary, increased consumption of spirits and low fat dairy products were associated with decreased 10-year predicted hypertension-free survival. For higher consumption levels of these foods no further decrease of 10-year predicted hypertension-free survival was observed. The consumption of small quantities of soup resulted in an increase of the 10-year predicted hypertension-free survival. A further consumption of soup that exceeded 25g/day resulted in a decline of the 10-year predicted hypertension-free survival. For other food groups like vegetable fat, wine, tea, sauce and raw vegetables a peak of the 10-year predicted hypertension-free survival was observed for small quantities of consumption, whereas increasing quantities of consumption had only a small or no effect on the 10-year predicted hypertension-free survival. Different consumption of whole grain bread had hardly any effect on the 10-year predicted hypertension-free survival.

A different consumption behavior of spirits had the strongest effect on 10-year predicted hypertension-free survival (**Figure 14**). For spirits the maximal decline in 10-year predicted hypertension-free survival was 5% from lowest consumption values of 0 g/day to values greater than 75 g/day. However, the proportion of male participants consuming more than 10 g spirits per day was low. Different consumer behaviour of the other selected food groups resulted in approximately 2% changes of 10-year predicted hypertension-free survival.

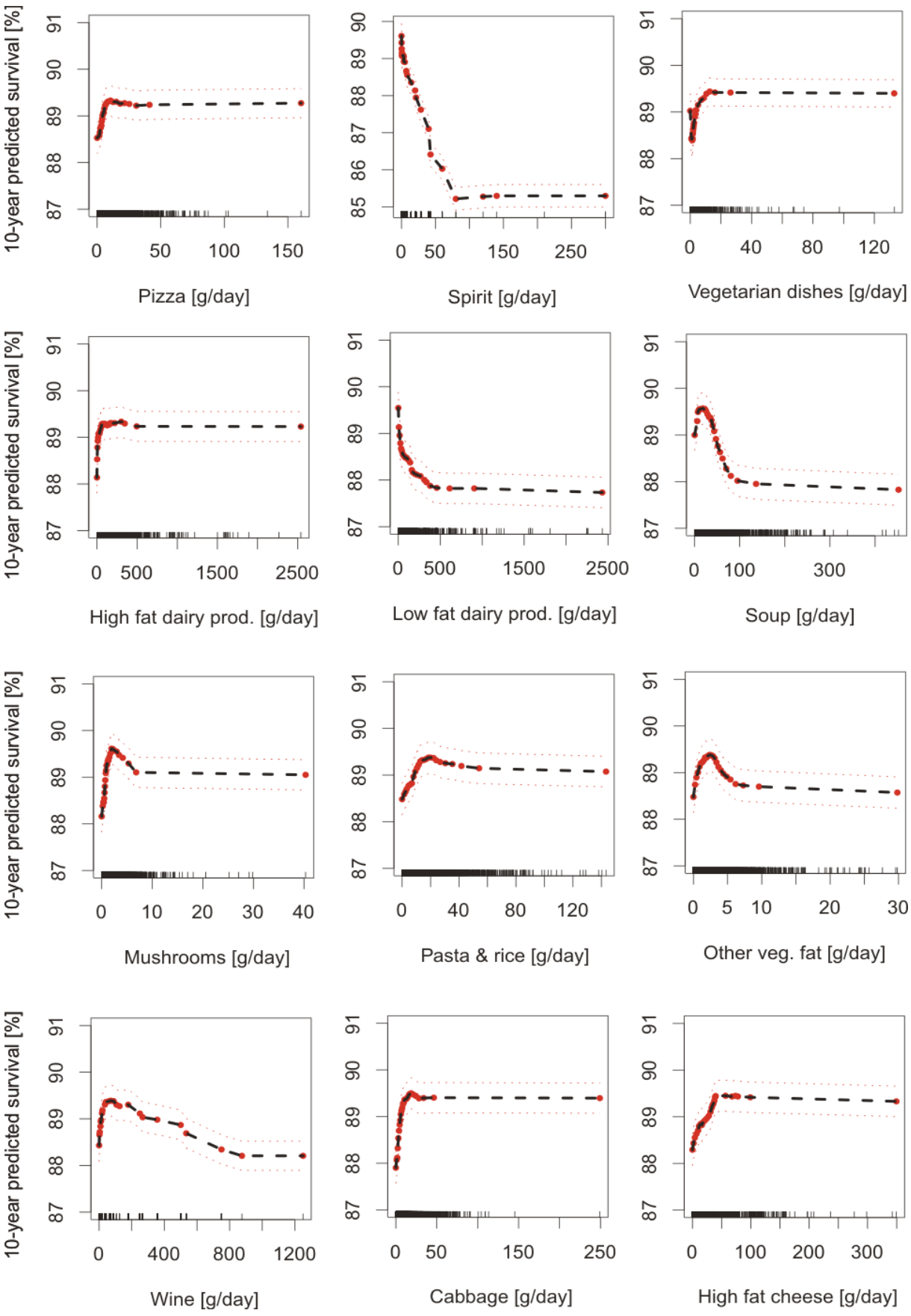

Figure 14: Partial plot of identified food groups most predictive for incident hypertension in men.
The plots including the partial values (red points) ± 2 SE (dashed red lines). Values on the vertical axis represent 10-year predicted hypertension-free survival for a given variable after adjusting for all other variables (covariates and selected metabolites).

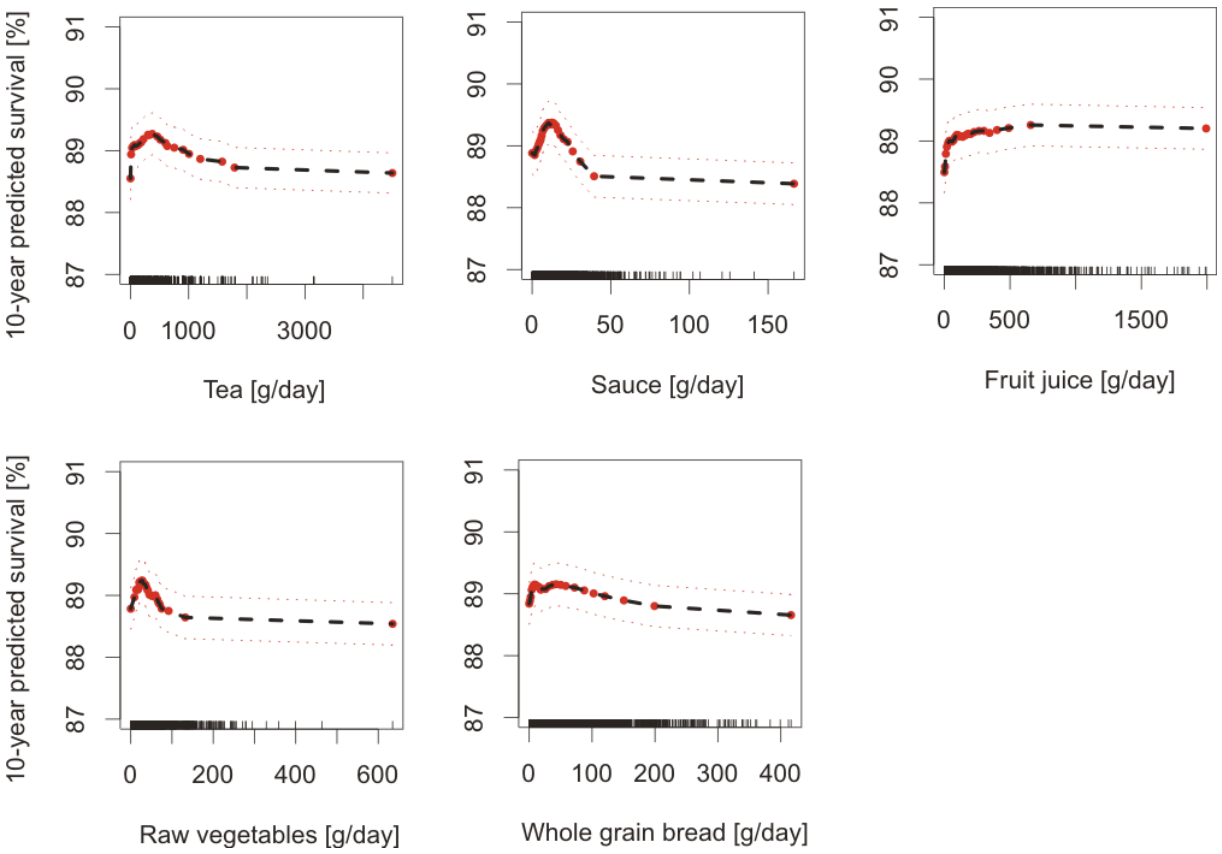

Figure 14 continued

3.3.2 Selected food groups associated with incident hypertension in women

The application of the RSF backward algorithm on EPIC-Potsdam food group data for women resulted in a reduced set of 25 selected food groups. This set of selected food groups and covariates presents the final RSF model which is most predictive for incident hypertension with a RSF prediction error rate of 0.332 (**Table 7**). A RSF model including only covariates or a RSF model including the covariates and all food groups resulted in RSF prediction error rates of 0.350 and 0.340, respectively. This suggests an improved prediction of the final RSF model regarding incident hypertension compared to the other RSF models.

Table 7: Computed prediction error rate of different RSF models for women regarding incident hypertension.

RSF model	RSF prediction error rate
Only covariates	0.350
All food groups + covariates	0.340
Selected food groups + covariates	0.332

Selected food groups most predictive for incident hypertension were potatoes, pizza, red meat and processed meat with minimal depth values of approximately 4 (**Figure 15)**. However, similar to the analysis in men, the covariates BMI, age, prevalent T2D and IPAI showed lower minimal depth values than any food group. Further selected food groups were legumes, mushrooms, low fat cheese, canned fruits, low fat dairy products, fruit juice, high fat dairy products, spirits, poultry, muesli, water, vegetarian dishes cakes and cookies, high energy drinks, other bread, cooked vegetables, sweet bread spreads, whole grain bread, other vegetable fat, pasta and rice, and margarine. These selected food groups had minimal depth values between 4.3 and 4.7.

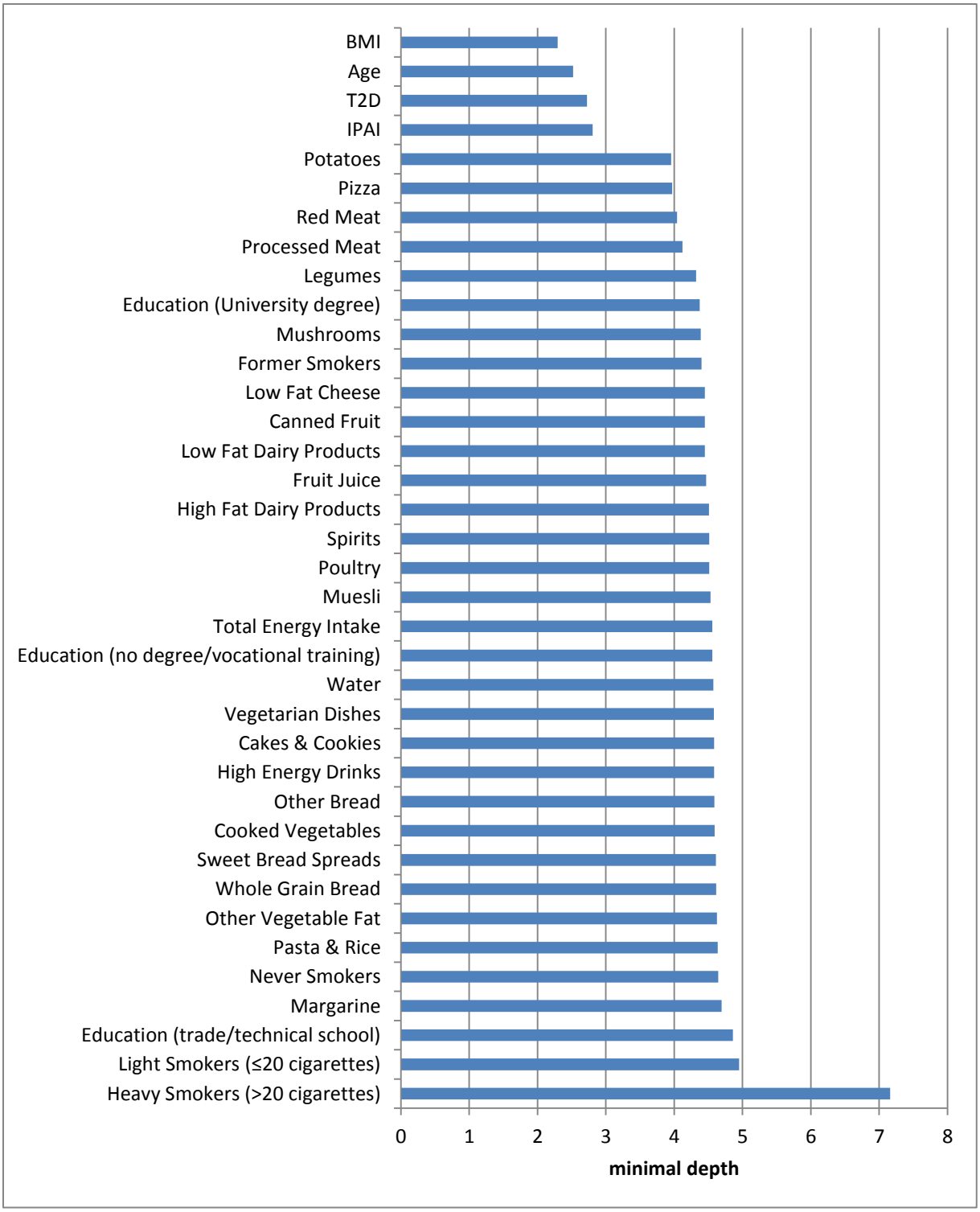

Figure 15: Identified food groups and covariates included in the final RSF which were most informative for incident hypertension in women. The identified food groups and covariates are ranked by the minimal depth measurement.

Similar to identified groups for men, the associations between consumption of food groups and 10-year predicted hypertension-free survival were in women also non-linear as visualised by the partial plots in **Figure 16**. Low and medium consumption of potatoes, red

meat, processed meat, legumes, low fat cheese, low fat dairy products, spirits, and poultry were associated with a decrease of 10-year predicted hypertension-free survival. High consumption levels of these foods resulted in constant lower 10-year predicted hypertension-free survival. Small peaks with increased 10-year predicted hypertension-free survival were observed for low to medium consumption levels of mushrooms, fruit juice, high fat dairy products, cakes and cookies, other bread, cooked vegetables, sweet bread products, whole grain bread, other vegetable fat, and pasta and rice. Compared to other food groups, spirits showed the largest decrease in 10-year predicted hypertension-free survival with values of 3% from lower to higher intake of spirits. For other selected food groups the decrease or increase of 10-year predicted hypertension-free survival was approximately between 1% and 2%.

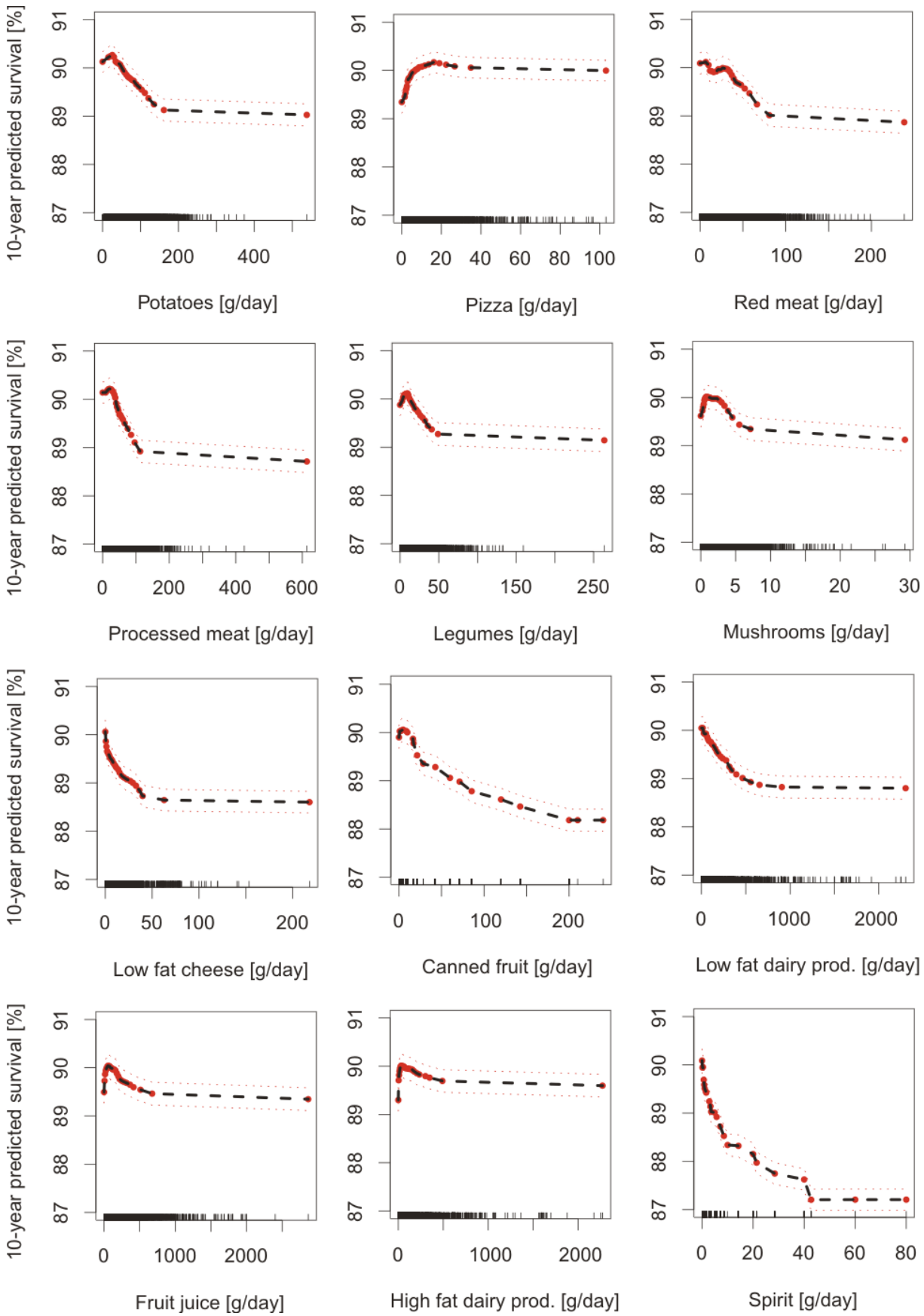

Figure 16: Partial plot of identified food groups most predictive for incident hypertension in women.
The plots including the partial values (red points) ± 2 SE (dashed red lines). Values on the vertical axis represent 10-year predicted hypertension-free survival for a given variable after adjusting for all other variables (covariates and selected metabolites).

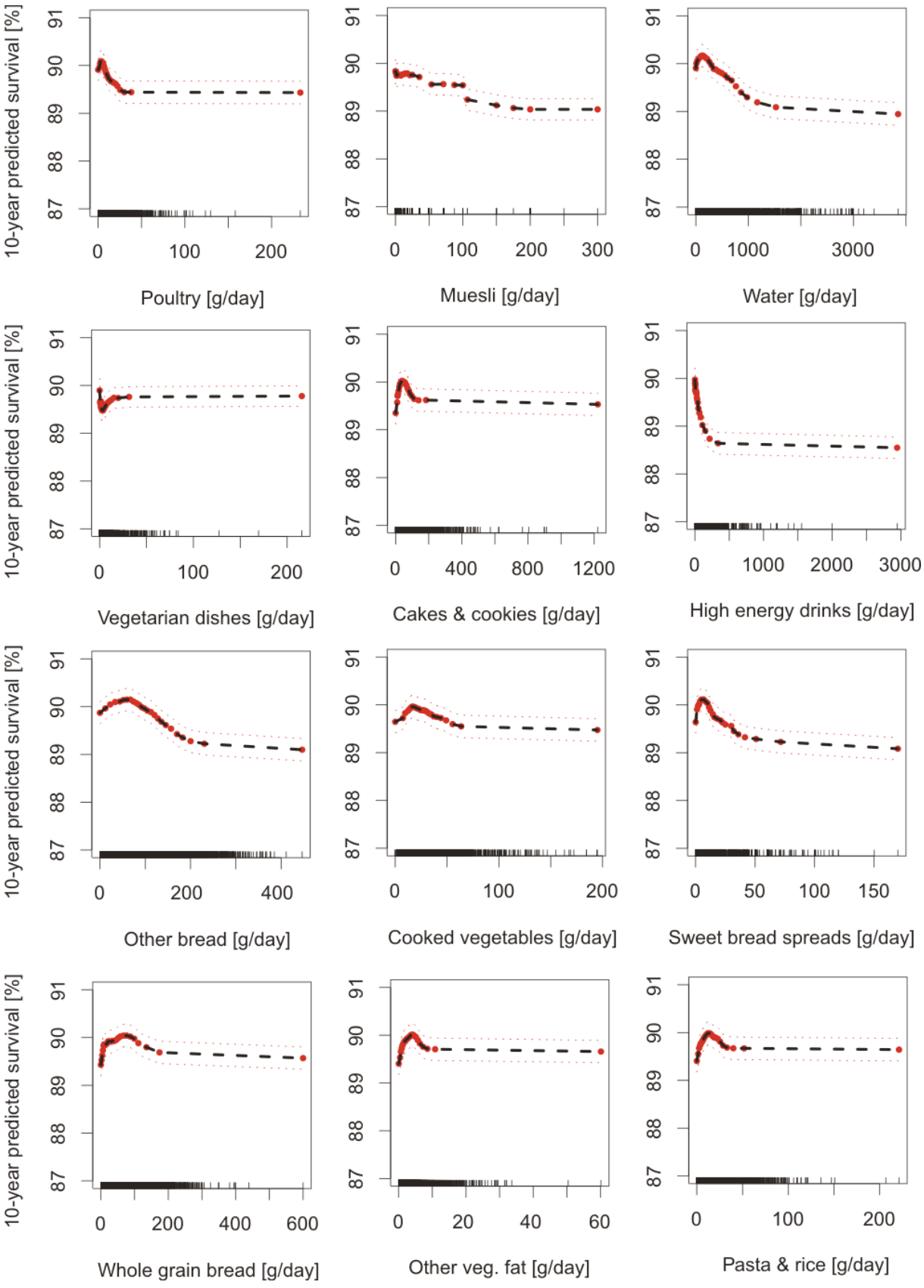

Figure 16 continued

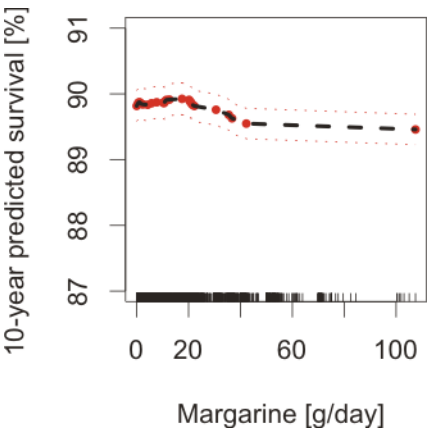

Figure 16 continued

3.3.3 Comparison of food group selection in data of men and women

For men and women partly different food groups were selected regarding the association with incident hypertension (**Table 8**). The final RSF model for women contained eight more food groups than the final RSF model for men. As shown by GGM network for both sexes, the underlying food patterns were the same for men and women (**Figure 17** and **Figure 18**). However, the colour-coding of the direction of the associations between food groups and 10-year predicted hypertension-free survival in the GGM models according to the partial plots illustrated differences between men and women.

In women a food pattern consisting of red meat, poultry, potatoes, legumes, processed meat and whole grain bread was mainly negatively associated with 10-year predicted hypertension-free survival (**Figure 18**). The connecting edges between the food groups showed positive correlations with partial correlation coefficients between 0.30 and 0.41. This suggests that these food groups were frequently consumed together. However, the food groups other bread and whole grain bread were connected by an edge with a negative correlation. This suggests that either other bread or whole grain bread was consumed. In men the food groups of the lunch and dinner food group pattern, with the exception of whole grain bread, were not selected (**Figure 17**). However, the partial plot of whole grain

bread showed only a slight non-linear association with 10-year predicted hypertension-free survival in both sexes (**Figure 14** and **Figure 16**). A positive or negative association cannot be derived from the two whole grain partial plots for both sexes.

By contrast, a food pattern representing fat dairy and cheese products was of importance for incident hypertension for both sexes. Low fat dairy products were negatively and high fat dairy products were positively associated with 10-year predicted hypertension-free survival in both sexes. The low and high fat dairy products were negatively correlated in both sexes suggesting that if one of the two food groups was consumed the other food group was not or less consumed. In women, high fat cheese was also selected to be positively associated with 10-year predicted hypertension-free survival and for men low fat cheese was identified to be negatively associated with 10-year predicted hypertension-free survival. In addition, for women the food group canned fruits and the food group representing cakes and cookies were also selected. The cakes and cookies food group was more positively associated with 10-year predicted hypertension-free survival, whereas the canned fruits food group was more negatively associated with 10-year predicted hypertension-free survival. Other food groups were also selected for both sexes but did not reflect any food patterns.

In both sexes the consumption of spirits showed a strong negative association with 10-year predicted hypertension-free survival. Pizza, however, was positively associated with 10-year predicted hypertension-free survival in both sexes. Especially for women, many further food groups (high energy soft drinks, muesli, sweet bread spreads, canned fruit and water) were selected to be negatively associated with 10-year predicted hypertension-free survival. In contrast, for men other food groups like cabbage, fruit juice, pasta and rice mushrooms, and vegetarian dishes were selected, which were more positively associated with 10-year predicted hypertension-free survival.

Table 8: Comparison of food groups selected by the RSF backward algorithm in men and women.

	men	women
Pizza	+	+
Spirits	+	+
Vegetarian Dishes	+	+
High Fat Dairy Products	+	+
Low Fat Dairy Products	+	+
Soup	+	-
Mushrooms	+	+
Pasta & Rice	+	+
Other Vegetable Fat	+	+
Wine	+	-
Cabbage	+	-
High Fat Cheese	+	-
Tea	+	-
Sauce	+	-
Fruit Juice	+	+
Raw Vegetables	+	-
Whole Grain Bread	+	+
Potatoes	-	+
Red Meat	-	+
Processed Meat	-	+
Legumes	-	+
Low Fat Cheese	-	+
Canned Fruit	-	+
Poultry	-	+
Muesli	-	+
Water	-	+
Cakes & Cookies	-	+
Other Bread	-	+
Cooked Vegetables	-	+
Sweet Bread Spreads	-	+
Magarine	-	+

The plus (+) and minus (-) symbol indicate that the respective food group was or was not selected.

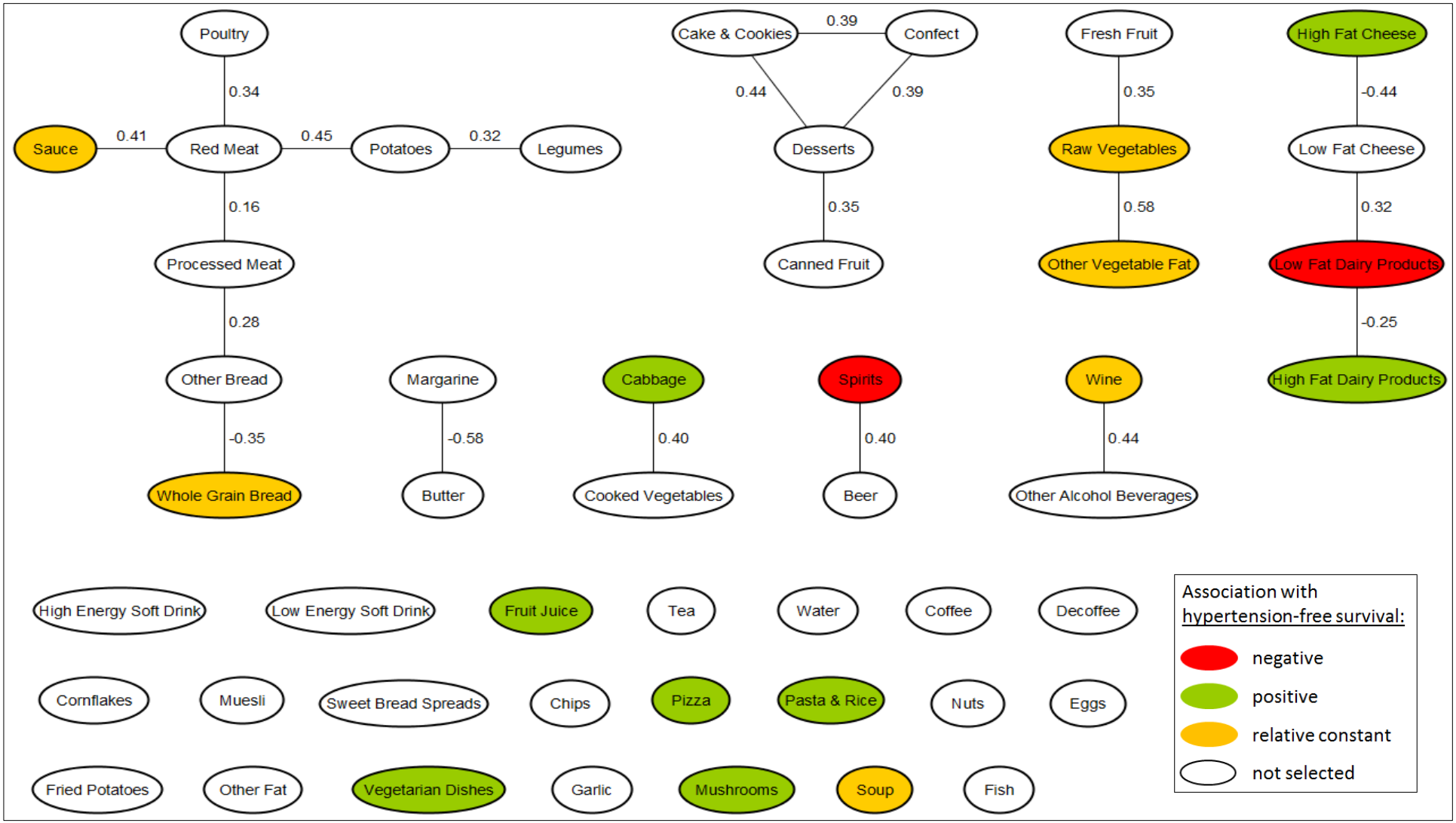

Figure 17: Food group network of male participants of EPIC-Potsdam (n=3492). Each node represents one food group and each edge between two nodes represents The partial correlation between two food groups mutually adjusted for the other food groups. The food group network is colour-coded. The colours represent the association of the food group intakes with 10-year predicted hypertension-free as derived from the partial plots.

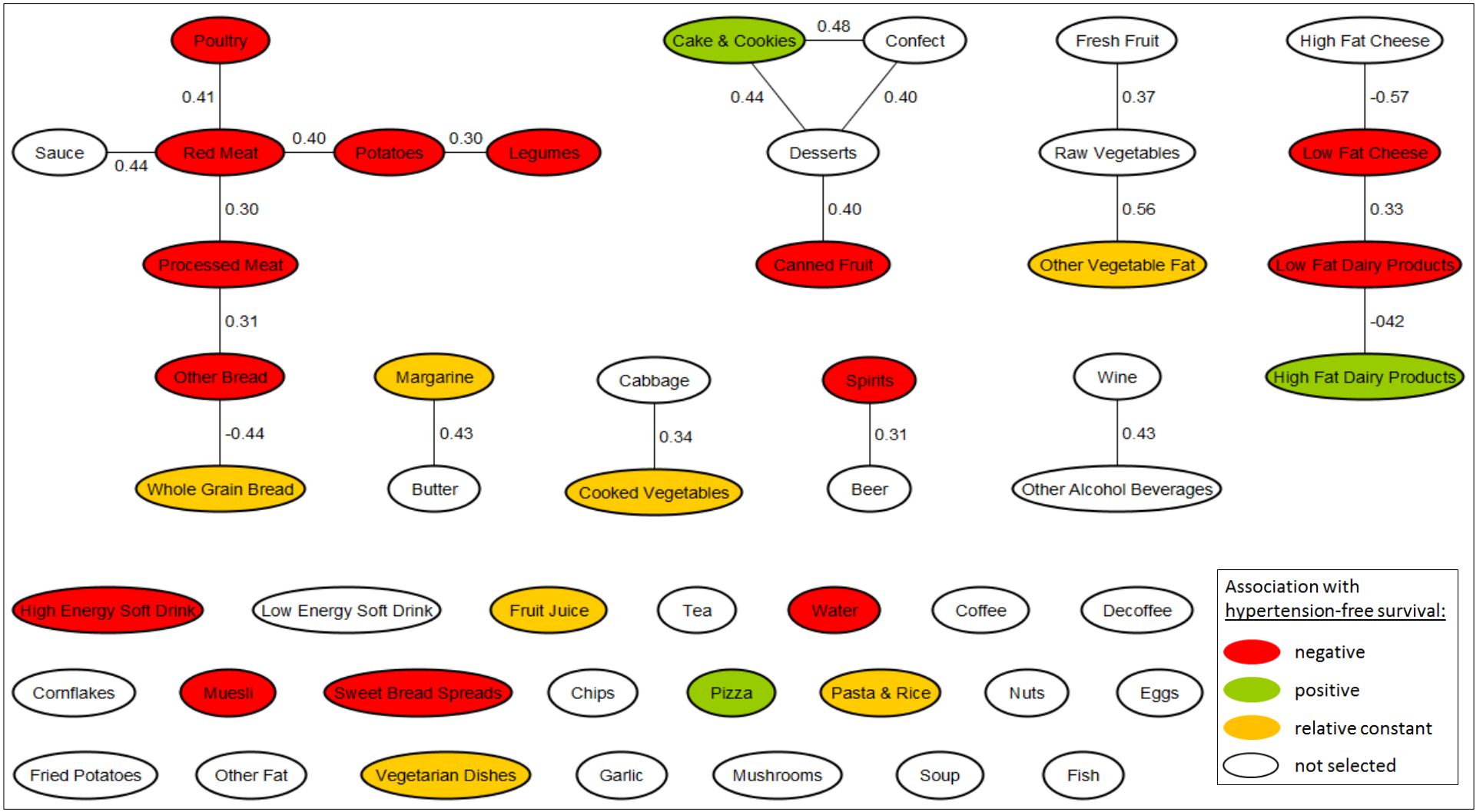

Figure 18: Food group network of female participants of EPIC-Potsdam (n=7993). Each node represents one food group and each edge between two nodes represents the partial correlation between two food groups mutually adjusted for the other food groups. The food group network is colour-coded. The colours represent the association of the food group intakes with 10-year predicted hypertension-free as derived from the partial plots.

4 Discussion[2]

The present thesis investigated the applicability of the machine learning method RSF as a variable selection tool in the scope of exploratory analysis of complex survival data. The conducted simulation study confirmed the suitability of the RSF method and the implemented RSF backward algorithm as a variable selection tool. Furthermore, disease associated correlation patterns can be uncovered and confounding factors can be adequately considered when using the RSF backward algorithm. The application of the RSF backward algorithm to prospective observational data of the EPIC-Potsdam study resulted in the successful identification of metabolites which were associated with incident T2D and of food groups which were associated with incident hypertension. However, a comparison with a previous study by Floegel et al. (52) revealed that the RSF approach resulted in partly different selected variables than the established Cox proportional hazards regression. Furthermore, nonlinear associations between concentrations of identified metabolites or food groups and the predicted event-free survival were revealed by partial plots, an additional RSF feature. Based on these findings, subsequent correlation network analyses with GGM uncovered a T2D associated metabolite pattern and hypertension associated food group patterns.

4.1 Discussion of methodical findings

Simulation studies are an excellent statistical approach to assess the properties, performance and quality of novel statistical methods in certain clearly specified data situations. In contrast, when using observational data for the assessment of novel statistical

[2] In the framework of this thesis, parts of the discussion are from a manuscript which is in revision at the International Journal of Epidemiology: (92) Dietrich S., Floegel A, Boeing H., Schulze M.B., Illig T., Pischon T., Knüppel S., Drogan D. Random Survival Forest in practice – a method for modelling high-dimensional metabolomics data in time to event analysis. In revision at International Journal of Epidemiology

methods, existing associations are often unknown, biased or confounded which altogether hampered the evaluation process. With this in mind, the present thesis generated simulation data by using an approach that has already been successfully applied to simulate survival times for Cox proportional hazards models (75). The approach enabled the generation of predictive variables with defined known hazard ratios regarding time until event. Hence, an appropriate evaluation of the RSF method as well as of the RSF backward algorithm regarding variable selection in simulated survival data was ensured.

4.1.1 The suitability of the minimal depth ranking measurement

The RSF method was specially designed to identify variables associated with event time of interest (23). Accordingly, the simulation study clearly demonstrated that the RSF method correctly identified predictive variables and separated them from noise variables, which was enabled by the usage of the minimal depth measurements. One key finding was that the minimal depth ranking of the predictive simulated variables reflected the different pre-specified HRs of the simulated predictive variables. Thus, selected variables could be categorized according to the strength of their effect on disease development comparable to the HR ranking by regression methods. Nevertheless, it's a disadvantage that the direction of associations cannot directly be derived by the minimal depth measurements. However, subsequent partial plots can be computed to visualise and thus determine the direction of associations with event-free survival.

When using the RSF minimal depth measurement as a ranking tool for variable selection, it must be ensured that the minimal depth values are precise. As shown by the simulation study and suggested by previous studies (67, 93), the precision of the minimal depth ranking can be increased by an appropriate choice of the bootstrap samples. In case of the

simulation study the precision of the minimal depth values increased when 1000 bootstrap samples were used instead of 100 bootstrap samples. However, a high number of bootstrap samples induces longer computation time for the computation of a RSF. This must be taken into account to avoid unnecessary long computation times when applying a backward selection approach to the RSF method. From the present thesis it can be concluded that the choice of 1000 bootstrap samples represents a suitable compromise between computation time and precision of the minimal depth measurements.

4.1.2 Random Survival Forest and the preference of node splits on continuous variables

One criticism of tree-based methods like RSF is that the node splitting process favours continuous variables over categorical variables (94). Indeed, the simulation study of the present thesis confirmed this point of criticism. As shown, when testing more than one split trial (*nsplit*) per candidate variable to split a node then the minimal depth values of categorical variables worsened, in particular for noise categorical variables. In consequence, incorrect and misleading minimal depth values will be computed when testing more than one split per candidate variable for data consisting of a mixture of continuous and categorical data.

For better understanding, if several splits per candidate variable per node are tested continuous variables have a higher chance to maximize the survival difference than categorical variables. For each continuous candidate variables several spilt points can be tested per node, whereas for categorical variables only a limited number of split points can be tested depending on the number of categories, as illustrated in **Figure 19**. In addition, in the present thesis it was also observed that if a categorical variable is highly predictive regarding the endpoint, then the effect of this bias is attenuated and the minimal depth

value is only minimally altered with increasing number of node split trials per candidate variable. However, to avoid or minimize the preference of continuous variables when using data that include continuous and categorical variables, it seems reasonable from the present findings to choose the RSF parameter *nsplit* equal to one. In the simulation study, this RSF setting resulted in equally rating of noise continuous and noise categorical variables regarding their minimal depth values.

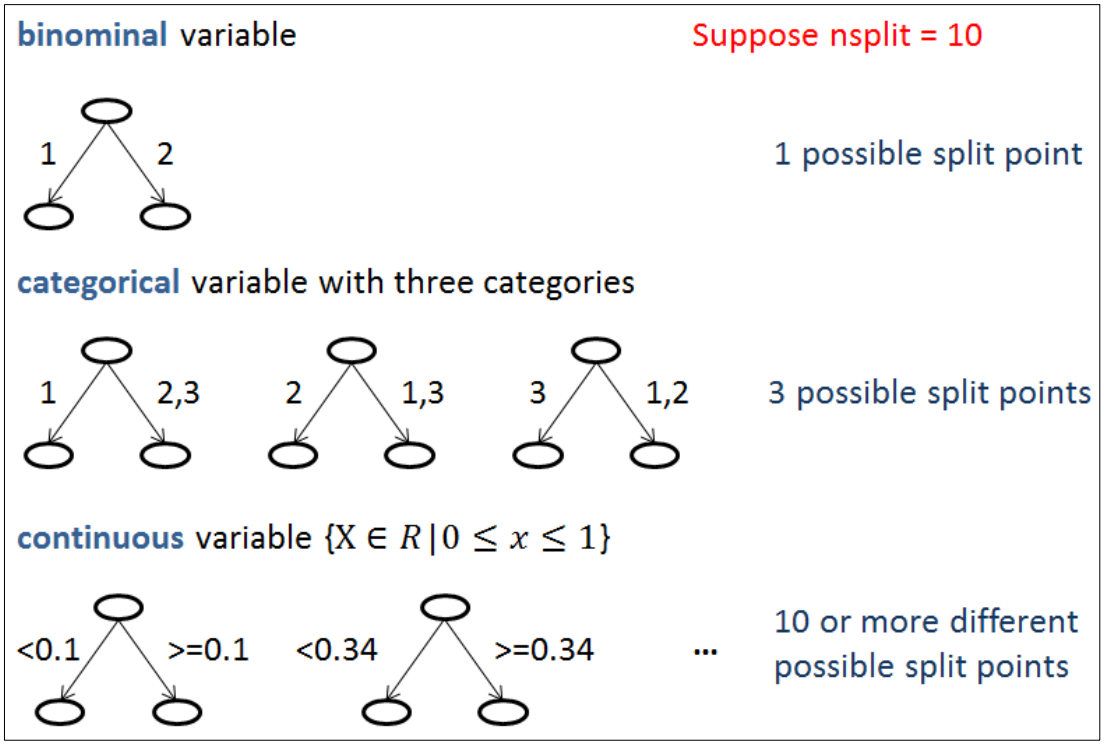

Figure 19: Illustration of split point selection for binominal, categorical and continuous variables.

4.1.3 Random Survival Forest and multicollinearity

The identification of cluster of correlated variables is an important task for scientist as they may point to biological pathways and biological networks associated with disease outcomes. However, in regression analyses multicollinearity of variables introduces misleading risk estimators and thus caused misclassifications (19, 49). For better understanding, suppose a regression model is fit with two correlated variables both predictive for an outcome. The expectation is that both variables entered the regression model. However, if one of the

correlated variables is given in the regression model the other correlated variable may not entered in the regression model as it, for instance, only contributes with redundant information to the outcome. Thus, regression analyses can result in misinterpretation when multicollinearity is present.

In this regard the RSF method can demonstrate its superior methological performance. As a result of the simulation study, it was shown that predictive variables as well as their correlated variables have lower minimal depth values than noise variables which may point in observational data to a true correlation pattern signal. This effect is attributable to the clustering property of decision trees (67). A node split that occurs with the inclusion of a variable derived from a cluster of correlated variables increases the chance that a following node split is determined with another correlated variables of the cluster (67). Moreover, due to random node splitting correlated variables have independently from each other the same chance to be chosen for node split close to the root node. As a consequence, variables of a correlation pattern with a true signal have small minimal depth values. This ensures the detection of correlation patterns and thus of possible biological pathways when using the RSF approach. However, for a correlation pattern it remains unclear which of the correlated variables have directly adverse or protective effects regarding time until event. Subsequently correlation analyses or if necessary further studies are thus required to elucidates the respective correlation pattern in detail.

4.1.4 Random Survival Forest and confounding

For use in observational epidemiology and clinical interpretation of disease associated markers, it is also important that RSF is able to handle the issue of confounding. Aging processes, weight gain or a lack of exercises, just to name a few possible confounders,

influence the risk to develop T2D or hypertension (40, 42). For instance, different cardiovascular disease risks were observed in epidemiological studies for obese people than for normal weight people or for people with higher physical activity compared to people with lower physical activity (95, 96). A statistical method that aims to identify disease markers should, therefore, also be able to take into account confounding factors.

As far as known, no study has so far investigated the effects of confounding for the RSF method. However, for RF a few approaches had been suggested to account for confounding (97, 98). Thus, for example, a multi-dimensional scaling clustering procedure was suggested followed by an adjustment of data which were than used for RF analysis (97). As a result, the consideration of possible confounding factors enhanced goodness of statistical tree based models.

When using the RSF method, confounding factors can be considered by computation of RSF models that based on data that included also possible confounders (covariates), as demonstrated in the present thesis. This ensures that during the tree growing process, possible confounders can be also chosen for random node splitting. Thus, for example, a node split can be performed that separates the observation into overweight and normal weight individuals based on a variable for BMI. In the branch of overweight or normal weight individuals further node splits can then be performed which enables the group specific identification of variables associated with time until event.

As demonstrated in the simulation study non-consideration of possible confounders may result in misleading minimal depth values (**Figure 8**). The observed effect on the minimal depth values was moderate as only one confounder was used together with idealized data consisting of strong predictors. Although, an identification of all predictive variables was given in the simulation analysis with and without the covariate, in observational data a

mixture of different influencing factors can lead to strong shifted minimal depth values. A non-consideration of possible confounders can then results in erroneous RSF models with misleading predictors.

4.1.5 Variable selection using the Random Survival Forest backward algorithm

Data reduction and data mining approaches like variable selection are applied to epidemiological complex data to avoid bias, erroneous estimators and misleading statistical models introduced by non-informative variables (43-45). By application of variable selection procedures statistical models can be established which allow enhanced statistical interpretation and improved clinical diagnostic (43-45).

Against this background and to identify disease associated markers, a RSF backward algorithm was implemented with the purpose to stepwise eliminate non-informative variables regarding time until event from complex survival data. The results of the present thesis were impressive. In the simulation study the RSF backward algorithm resulted in a final RSF model were all noise variables were removed and only the predictive and correlated variables remained. This was achieved through the combination of the minimal depth measurement ranking and the RSF prediction error rate. The minimal depth measurement was used to rank variables, whereby the variable with the worst minimal depth value could be identified and removed from the data. The goodness of the RSF model was then evaluated by the RSF prediction error rate.

In the present thesis, a backward selection approach was used for implementation because such a procedure needs no prior-knowledge of data, as it is required for forward selection (43-45). Hence, the RSF backward algorithm is particular suitable for variable selection when using omics data where often no prior-knowledge exist. For gene selection and classification

of microarray data with the method RF a similar backward selection procedure was suggested by Díaz-Uriarte et al. (99). However this procedure was performed without retaining possible confounders throughout the selection process. To take into account confounding, the implemented RSF backward algorithm was modified such that the variable selection process is run only on the variables to be investigated while forcing at each step a pre-defined set of potential confounders into the model. In conclusion, the RSF backward algorithm allows the identification and interpretation of variables under consideration of potential confounders.

4.1.6 Application of the Random Survival Forest backward algorithm to observational data

Recently, the method RSF was successfully applied to identify risk factors of different diseases (68-70). However, in the previous studies the RSF method was applied only to data with a low number of variables, whereby there was no need for a variable selection procedure. In contrast, in the present thesis the application of the RSF method to EPIC-Potsdam data with numerous highly correlated variables required the implementation of a RSF variable selection procedure.

The application of the implemented RSF backward algorithm on observational data of the EPIC-Potsdam study resulted in final RSF models (consisting of the selected variables and covariates) which showed smaller RSF prediction error rates than the full RSF models (consisting of all variables and covariates or only the covariates). In addition, it was also shown that the RSF prediction error rate of a RSF model which based only on covariates can be improved when adding the selected metabolites or selected food groups to the RSF models. These findings indicate that mainly metabolites or food groups were removed from

the data without information content regarding incident T2D or incident hypertension, respectively.

However, the final RSF model for incident T2D showed lower RSF prediction error rates (0.165) than the two final RSF model for incident hypertension (0.363 and 0.322). The RSF prediction error rate is a measure of the accuracy of the RSF model and thus provides information how accurate the incidence of a disease can be predicted based on the respective RSF model (23). RSF prediction error rates of 0.5 refer to RSF models based on chance and lower RSF prediction error rate correspond to RSF models with more precise prediction accuracy (23). With this in mind, it must be stated that the two RSF models for incident hypertension are inappropriate for usage in clinical diagnosis of incident hypertension. A lot of disease associated factors may not be considered in the present RSF models of incident hypertension because they are unknown or were not measured. Possible factors that may improve the accuracy are assessed stress level or measured clinical marker of cholesterol levels (100, 101). Nevertheless, the objective to identify food groups associated with incident hypertension was successfully fulfilled because the final RSF model that included the selected food groups had an improved prediction error rate compared to a RSF model which based only on covariates. This suggests that food groups without information contents regarding incident hypertension were removed during the selection process. Based on the available used data, the best final RSF model was determined.

To identify variables associated with time until event, variable selection approaches based on Cox proportional hazards regressions are frequently applied. For instance, in a previous study by Floegel et al. (52) a univariate Cox proportional hazards regression followed by stepwise Cox proportional hazards regression selection method was applied to identify incident T2D associated metabolites. For comparison of the RSF approach with the

previously used Cox proportional hazards regression approach the same data were used in the present thesis as by Floegel et al. (52). A comparison of the metabolites selected by both methods revealed that all of the metabolites selected by RSF also showed a significant association with incident T2D in the univariate Cox proportional hazards regression models. However, after adjustment for multiple testing, the metabolites acyl-alkyl-PC C42:3, C42:4, C44:6, diacyl PC C42:0, C42:1, tyrosine, valine and acylcarnitine C8:1 lost their statistical significance in the subsequent stepwise selection, as previously published (52). Moreover, some of the metabolites, which were part of the final Cox proportional hazards regression model, were not selected for the final RSF model.

As demonstrated in the simulation study and shown for the correlation structure of selected acyl-alky-PC in **Figure 11**, RSF is able to handle multicollinearity and selected highly correlated metabolites with a true signal. Moreover, in **Figure 12** the colour coding of the metabolite network revealed that seven of the metabolites identified by the RSF backward algorithm may represent a correlation pattern. Only three of the seven metabolites were identified in the previous study by Floegel et al. (52). Some of these seven metabolites differ only by the number of double bounds. Therefore, it appears that RSF identified a correlation patter with a true signal regarding incident T2D and it is likely that the seven metabolites are members of the same metabolic pathway. It can be assumed that the different selection of metabolites by the RSF and the Cox proportional hazards regression approach is caused by multicollinearity effects which are known to result in misleading estimators in regression models (19, 49). In addition, non-linear associations between metabolites and 5-year predicted T2D-free survival were observed in the partial plots. This may also have caused misleading estimators in the Cox proportional hazards regression models and thus result in the different selection of metabolites. In addition, for metabolite selection the parameter

nsplit was chosen equal to 10. This choice was made because only some of the covariates represent categorical variables and compared to the variety of continuous metabolites, it was assumed that a higher split number was more appropriate to improve the variable selection process. Nevertheless, misclassified minimal depth measurements cannot be excluded.

In conclusion the present findings indicate that the implemented RSF backward algorithm represent an excellent alternative to existing survival analysis methods. The RSF backward algorithm was successfully applied for the purpose of variable selection in observational data. Based on the RSF prediction error rates, it was demonstrated that the selected variables contribute to an improvement of RSF models when used in addition to traditional risk factors (covariates). Furthermore, nonlinear associations between concentrations of identified metabolites or food groups and the predicted event-free survival were revealed by partial plots, which enhanced the biological interpretations of these findings. The biological plausibility of the present findings is discussed in chapter 4.2. However, a validation in external cohorts and comparison with other survival analysis methods is necessary to confirm the present findings and the suitability of the RSF method and the implemented RSF backward algorithm.

4.1.7 Advantages and Disadvantages of Random survival forest

Compared to regression approaches, the RSF method has several advantages. RSF is completely data driven and thus independent of hypothesis testing (23). It does not test the goodness of fit of data to a hypothesis, but seeks a model that best explains the data. For the application of the RSF approach, no assumptions like underlying distribution of individual variables or proportional hazards must be fulfilled. The raw data can directly be used to

compute RSF models. The RSF process is largely automated and only a few parameters must be specified like the number of bootstrap samples or the number of node splits (23). Thus, the RSF approach represents a suitable tool for exploratory analysis of survival data where prior-knowledge is still limited. In addition, as previously demonstrated, the application of the RSF methods results in RSF models with comparable accuracy then derived with standard statistical survival analysis methods (23, 68-70).

For tree growing, RSF uses random subsets of variables per node. Consequently, correlated variables will be selected independently from each other to split nodes leading to interruption of the correlation structure of variables. As a consequence, there is less competition between highly correlated variables and reliable variable selection is possible even in the presence of multicollinearity (102). Furthermore, the problem of overfitting – e.g. when multiple regression models are performed on a high number of variables without internal validation – is largely reduced due to randomization via bootstrap sampling (103). This feature makes RSF very appealing for explorative survival analysis in omics data, where false-positive discoveries due to overfitting are considered to be a major problem (104).

A disadvantage of the RSF approach is that the relative risk, which is an intuitive and meaningful measure of association in epidemiological studies, cannot immediately calculated for the variables considered in a RSF model. Instead, the contribution of each marker to its relative relatedness with the endpoint needs to be assessed by the minimal depth ranking and the direction of associations by subsequent partial plots. However, variables selected by RSF can be analyzed in subsequent regression models to estimate relative risks. Yet, regression models including all selected metabolites may not be appropriate, given the fact that RSF can independently select structurally related metabolites of high correlations. Another disadvantage of tree-based methods is the

preference of continuous variables for node splitting (94), if the data consists of a mix of continuous and categorical variables. However, as shown, this bias can be countered by using a smaller number of split trials per candidate variable for node splitting. Moreover, each RSF based on the computation of numerous single decision trees. Hence, the underlying structure of a RSF represents a kind of black box and thus it difficult to verify the node splitting process for the respective variables.

4.2 Discussion of biological plausibility of results

In the present thesis the application of the RSF backward algorithm to prospective observational data of the EPIC-Potsdam study resulted in the identification of several serum metabolites associated with incident T2D and several food groups associated with incident hypertension. This was accompanied by an improvement of the prediction error rates indicating that mainly noise metabolites and food groups were excluded. Moreover, using partial plots, non-linear associations between the selected metabolites or food groups and the predicted event-free survival were shown, thereby improving the interpretability of RSF results. The biological plausibility of the present findings is discussed next.

4.2.1 Serum metabolites associated with incident type 2 diabetes

Several metabolites and a correlation pattern of selected metabolites were identified to be associated with incident T2D. The most informative metabolite regarding the incidence of T2D was the metabolite hexose representing all 6-carbon monosaccharides. Accordingly, permanently increased blood glucose levels is one of the most important diagnostic indicators for T2D (105). However, it has to be noted that beside glucose the measured metabolite hexose incorporates also other C6-sugars like fructose, mannose and galactose.

Indeed, recent prospective studies revealed that increased plasma glucose levels as well as increased fructose intake are associated with higher risk of T2D (106-109) confirming the present findings. Permanently elevated glucose levels are tightly linked with insulin resistance which ultimately means that glucose molecules cannot transported by glucose transporters into cells resulting in health complications. Fructose, the second abundant C6-sugar, occurs together with glucose in most consumed food products and is converted in the body into glucose but also into dihydroxyacetone phosphate, the basis for the glycerol in triglycerides and phospholipids (110, 111). Moreover, fructose carbons are also metabolized directly to pyruvic acid and then to acetyl-CoA which is an important molecule in metabolism (e.g. energy, biosynthesis) (110, 111). However, the metabolic role of fructose regarding T2D is not yet fully understood and negative health effects are assumed only for high intake of fructose, particularly if combined with a high energy intake in the form of glucose (110, 111). As far as known, no findings exist regarding the association between the incidence or prevalence of T2D and other C6-sugars than glucose and fructose. This raises the question whether a direct measurement only of glucose and fructose would be more appropriate to determine the incidence of T2D with higher accuracy. Nevertheless, the adverse effect of increased hexose concentration on 5-year predicted T2D-free survival, as visualised by the partial plot, is biological reasonable. In this context, the partial plot of hexose thus enables first indications of clinical cut-points which may be useful for further studies, clinical diagnosis or treatment recommendation.

In the present thesis, the amino acids valine, tyrosine and glycine were also associated with incident T2D which is in line with most of the recent metabolite profiling studies ((53, 112-115). A study which was conducted within the Framingham Offspring study and replicated in the Malmö Diet and Cancer study recently reported significant associations with incident

T2D for isoleucine, leucine, valine, tyrosine, and phenylalanine (53). This was confirmed by another study which reported increased level of leucine/isoleucine, valine, tyrosine, phenylalanine, alanine, proline, glutamate/glutamine and ornithine to be correlated with insulin resistance (113). For Tyrosine it was also reported to be a particularly strong predictor of incident diabetes in South Asian individuals (115).

In line with these studies, the present thesis revealed that higher concentrations of valine and tyrosine promote the incidence of T2D and contrary higher concentrations of glycine are protective regarding the incidence of T2D. A recent study of Kawanaka et al. (116) reported that increased tyrosine levels were associated with increasing fibrotic staging and were also correlated with insulin resistance. Kawanaka et al. (116) speculated that insulin resistance increases the levels of α-ketobutyrate, which is involved in methionine degradation (116). Increase levels of α-ketobutyrate may thus cause a subsequent decrease in tyrosine aminotransferase levels which results in an increase in tyrosine levels (116). The protective effects of glycine regarding T2D may be attributable to the role of anti-inflammatory and immunomodulatory agent which has previously been proposed (117). Moreover, glycine is also involved in gluconeogenesis and the formation of glutathione (117). As far as known, no reasonable biological mechanism has been reported which may explain the observed adverse association of increased valine levels with the incidence of T2D.

In addition, in the study of Wang et al. (53) the amino acids phenylalanine, leucine and isoleucine were also associated with incident T2D, but not in the present thesis. Moreover, in the multivariate approach of Floegel et al. (52) which was applied on the same data as used in the present thesis, only the amino acid phenylalanine were selected to be independently associated with incident T2D. Therefore, it may be assumed that the correlation structure and linear or non-linear associations with incident T2D contributed to

the diverse selection of amino acids by RSF. In addition, the usage of other population samples regarding age, obesity or ethnicity in other studies may have also contributed to inconsistent results.

Many of the identified metabolites belong to the acyl-alkyl-PC and diacyl-PC. It is noteworthy that seven of the identified PCs may reflect a common correlation pattern (**Figure 12)**. Interestingly, increased concentrations of all seven metabolites of the correlation pattern were protectively associated with 5-year predicted T2D-free survival as visualise by the partial plots. Two of the seven metabolites were diacyl-PC. In contrast, increasing concentrations of the diacyl-PC C38:3 and C32:0, which were not part of the correlation pattern, were adversely associated with 5-year predicted T2D-free survival. Currently, few studies investigated the association between PCs and prevalence or incidence of T2D. One of these studies found that diacyl-PCs were associated with lower risk of T2D (118). Furthermore, lower levels of acyl-alkyl-PC were reported in obese individual and patients with insulin resistance (119, 120). In recent studies anti-inflammatory properties were assigned to PC (121-124). For instance, it was observed that serum levels of pro-inflammatory cytokines such as TNF-α, and IL-6, were attenuated by PC (121). In obese people overexpression of TNF-α in adipose tissue induces proinflammatory pathways which result in insulin resistance (125). Moreover, PCs which are synthesized in the liver, are one of the most abundant components of cellular membranes and may be involve in cellular signal transduction (126). In particular, for lipoprotein assembly and hepatic secretion of triglyceride-rich very low density lipoprotein (VLDL) particles and high-density lipoprotein (HDL)-particles, diacyl-PC are required (126). In addition, a vinyl-ether bond in the acyl-alkyl-PCs, however, enables them to act as blood antioxidants to protect lipoproteins from oxidation (126). Accordingly, dyslipidaemia has been tightly linked with insulin resistance

characterized by increased plasma concentrations of VLDL-triglyceride and low levels of HDL in individuals with insulin resistance (127). Hence, an association between altered concentration levels of the identified PCs with the incidence of T2D seems biological reasonable.

A further identified PC metabolite was lyso-PC C18:2. In the present thesis, increasing lyso-PC C18:2 concentrations were protectively associated with incident T2D. This is in line with the results observed by Floegel et al. (52) and also with a study of Wang-Sattler et al. (128). In the study of Wang-Sattler et al. this metabolites was found to be decreased in patients with prevalent T2D (128). However, contrasting results were observed in two other study, where patients with prevalent T2D showed an increase or no alterations in lyso-PC C18:2 levels (118, 129).

Given the observational nature of the EPIC-Potsdam study, the discussed biological explanatory approaches allow hypothesis generating rather than to draw any causal conclusion. Nevertheless, the present findings were greatly in line with current scientific knowledge. Thus, the identified metabolites may be useful as novel biomarkers in clinical diagnostic of incident T2D. Moreover new insights regarding T2D development may be gained by the examination of the causal link between identified metabolites and incident T2D. Particularly informative in this regard are the partial plots which allow the identification of possible cut-points. However, the present findings have to be confirmed by external cohorts, first.

4.2.2 Food groups associated with incident hypertension

Recognising the importance of hypertension for cardiovascular morbidity and mortality (40), prevention strategies are sought for the avoidance of the development of hypertension in high risk individuals. To control for hypertension, beside medical treatments, a consequent change in diet and regulation of caloric intake are recommended (130, 131). Indeed, change in dietary behaviour has been shown to reduce blood pressure (132). Although in recent time, the number of studies increased that prospectively investigated the association between food intake and incident hypertension, more research is required to assess the risk to develop hypertension due to an unbalanced diet. In this context, the application of the RSF backward algorithm on reliable and validated EPIC-Potsdam food group data provided the promising opportunity to identify food groups associated with incident hypertension in a German population. To account for different eating behaviour of women and men, gender specific variable selection was performed in the present thesis.

Indeed, in men and women partly different food groups were associated with incident hypertension. Particularly, for women the subsequent network analyses with GGM revealed that some of the selected food groups (red meat, poultry, processed meat, potatoes, legumes, and other bread) can be assigned to a food group pattern which is characteristic for dinner and lunch meals. Traditionally for lunch in Germany, meat dishes are eaten along with differently prepared potatoes and all kinds of vegetables. For dinner, it is quit common to eat bread covered with processed meat or cheese products. This traditional eating behaviour was reflected by a consistent food pattern in the food group networks for men and women, respectively. However, only in women, this food group pattern of lunch and dinner meals was associated with incident hypertension, but not in men. A high consumption of most of the food groups of this pattern decreased the predicted 10-year

predicted hypertension-free survival and thus, differently expressed, increased the chance to develop hypertension for women.

The present thesis showed that higher intake of poultry, red meat and processed meat were associated with a lower 10-year predicted hypertension-free survival in women. The results obtained were confirmed by a recent study which was conducted in three prospective cohort studies (133). In this study, frequent consumption of total meat (processed and unprocessed red meat) and poultry were associated with increased risk of hypertension in women as well as in men. In another study conducted in the Women's Health Study, higher intake of red meat (unprocessed and processed red meat) was positively associated with the risk to develop hypertension, whereas poultry intake was not associated with the risk to develop hypertension (134). In contrast, the results of the present study indicated that higher intakes of poultry have also an adverse effect regarding incident hypertension in women. The adverse consequences of the discussed meat products regarding hypertension may be caused by the composition of the meat products. Meat products, particularly red meat, have a high content of cholesterol and saturated fat. Indeed, it was shown that cholesterol levels and saturated fat are related with blood pressure and the incidence of hypertension (135-137).

It is noteworthy that in the present thesis meat consumption of men was not associated with incident hypertension. As shown in a study within the Health Professionals Follow-Up Study, it was found that meat consumption may increase the risk to develop hypertension also in men (133). In the present thesis, men consumed more meat than women (**Table S3**). However, meat consumption was the same for male cases and non-cases of incident hypertension, whereas for female cases and non-cases of incident hypertension meat consumption was different (**Table S3**). Together with a smaller number of male incident

hypertension cases compared to women this may point to a statistical power problem. However, RSF is able to handle data with small number of observations and high number of variables. Confounding could be an explanation why RSF could not detect any difference in the survival time regarding incident hypertension for meat consumption in men. A confounding due to misreporting or recall bias cannot be excluded.

Surprisingly, higher intakes of potatoes and legumes were also observed to be associated with a lower 10-year predicted hypertension free survival in women. As far as known, no study to date showed an association of potatoes and legumes with incident or prevalent hypertension. However, both food groups are frequently consumed with meat products and thus are highly correlated with them. As demonstrated, RSF also identifies correlation pattern with a true signal regarding time until event of interest. A high correlation with red meat may be the reason why the food groups potatoes and legumes were selected by the RSF backward algorithm. In addition, potatoes are frequently salted when they were consumed. Salt sensitivity has been linked with hypertension and thus provides another possible explanation why potatoes may have an adverse effect regarding incident hypertension (138).

A second food pattern representing cheese and fat dairy products was in women as well in men associated with the incidence of hypertension. Increased intake of cheese and dairy products with low fat content resulted in lower 10-year predicted hypertension free survival, whereas increased intake of cheese and dairy products with high fat content resulted in higher 10-year predicted hypertension-free survival. This contrasts with existing epidemiological findings. A recent dose-response meta-analysis of 9 prospective cohort studies revealed that total-fat dairy, low-fat dairy and milk intake were inversely and linearly associated with a lower risk of hypertension (139). No significant association with incident

hypertension was found for high-fat dairy, total fermented dairy products, yoghurt and cheese (139). This was in line with another preceding pooled meta-analysis and a systematic review (140, 141). Moreover, in the Dietary Approaches to Stop Hypertension Trial, a diet including low saturated-fat and low total-fat as well as low-fat dairy products was linked with a lowered blood pressure. It is supposed, that minerals (calcium, potassium, and magnesium) and vitamins (vitamin D and folate) contained in cheese and dairy products may have protective properties regarding hypertension (139). However, a recent review showed that in 11 of 16 studies, high-fat dairy intake was inversely associated with measures of adiposity (142). Another recent study showed that a high intake of dairy fat was associated with a lower risk of central obesity, whereas a low dairy fat intake was associated with a higher risk of central obesity (143). This supports the present findings, because obesity is a major risk factor for hypertension and a high fat diet which contributes to a reduction of obesity may thus be protective regarding the incidence of hypertension (40).

Nevertheless, the present findings contrast with previous studies that related dairy fat and cheese consumption to incident hypertension. Although several confounding variables were included in the present thesis, further confounding factors may still remain. Male and female cases of incident hypertension consumed lower amounts of high fat cheese and high fat dairy product and higher amounts of low fat cheese and low fat dairy product than non-cases of incident hypertension (**Table S3**). Dietary recommendations may have result in change of diet in cases of incident hypertension before hypertension diagnosis due to risk factors like overweight and physical inactivity. In addition, as revealed by the partial plots (**Figure 14** and **Figure 16)**, alterations in hypertension-free survival time with increasing consumption of the respective food groups were relative small and thus, small changes in

dietary behaviour may have influenced the results obtained by the RSF approach. Misreporting and recall bias can also not be excluded.

In men and women several further food groups, that were not part of a food group pattern, were associated with incident hypertension. Within the food groups that reflect alcohol consumption (spirits, beer, wine and other alcoholic beverages), only spirit was in men and women associated with an increased incidence of hypertension within the 10-year follow-up. Moreover in men small amounts of wine intake were protective, whereas higher amounts of wine intake also increased the chance to develop hypertension within the 10-year follow-up. In women no association was found for wine. Epidemiological and clinical studies reported inconsistent findings. Low to moderate alcohol consumption was found to be associated with an increased as well as a decreased incidence of hypertension (144). In this context, nonlinear relationships for low to moderate alcohol and hypertension were observed, leading to the speculations that smaller amounts of alcohol consumption may reduce blood pressure (145, 146). However, a linear dose-response association between alcohol intake and hypertension have been demonstrated as well, showing that heavy drinking is associated with increased incidence of hypertension (144, 147).

In the present thesis, beer and other alcoholic beverages were not associated with incident hypertension, but spirit. In general, spirits contain a high alcohol concentration which may explain the observed adverse effect for spirit intake on the incidence of hypertension in the present thesis. Even small amounts of spirit consumption reduced the 10-year hypertension-free survival in men as well as in women. However, the underlying molecular mechanism by which alcohol consumption caused the development of hypertension remains elusive (148, 149). Alcohol consumption may have, for instance, an effect on the central nervous system, baroreceptors, or the renin-angiotensin-aldosterone system, and may cause increased

oxidative stress and endothelial dysfunction (148, 149). Nevertheless, in line with existing findings, the present thesis confirms the public health importance of prevention strategies to reduce alcoholism, thus reducing the incidence of hypertension.

The food group pizza was in men as well as in women selected to be most informative and protective regarding incident hypertension. In general, pizza is composed by different food products, includes high quantities of salt and is seen as a fast food with poor nutrient content which may negatively influenced physical health (150, 151). However, scientific studies that examine the consumption of pizza in relation to hypertension or cardiovascular disease are rare. A cross-sectional study from Iran indicate that pizza consumption increase the risk of hypertension (152). In an Italian population it was shown that regular pizza eating is associated with lower risk of acute myocardial infarction (153). Moreover, it was shown that pizza consumption does not increase the stickiness of endothelium. It may be speculate that pizza can be an indicator for Mediterranean diet. It was shown that the Mediterranean diet have favourable effects of cardiovascular diseases and blood pressure (154, 155). However, due to the limited studies, the role of pizza consumption regarding incident hypertension remains elusive and more protective studies are necessary to confirm the present finding.

In women the food groups canned fruit, sweet bread, muesli and high energy soft drinks were adversely associated with 10-year predicted hypertension-free survival, whereas cakes and cookies were protective. All these food groups represent food products that include high quantities of sugar. Recently, it was reviewed that additional sugar in food contributes to increased blood pressure and increases cardiovascular risk by inciting metabolic dysfunction and dyslipidaemia (156). This confirmed the present findings with exception for the association between cake & cookies consumption and incident hypertension. However,

unpublished results of the EPIC-Potsdam study demonstrated that misreporting may have occurred for the assessment of cakes and cookies intake, which may explain the observed contrasting findings for cakes and cookies compared to the other sugar-containing food groups regarding incident hypertension in women.

These food groups with high sugar content were not selected in men. Instead, in men food groups representing fruit juice, vegetarian dishes, cabbage and mushrooms were protectively regarding the incidence of hypertension. In line with these findings several studies have demonstrated the protective role of fruit and vegetable consumption regarding lowering of blood pressure and hypertension (157-160). Moreover it was shown that increased consumption of fruit and vegetables is related to a reduced risk of coronary heart disease (161). The observed protective effect may be attributed to high content of antioxidant vitamins, folate, and other constituents in fruit and vegetables (162, 163).

Several further food groups were selected to be informative for the incidence of hypertension. But most of them showed only a weak association between food consumption and 10-year predicted hypertension-free survival. A lot of these associations showed a small peak for low food consumption followed by no changes (e.g. whole grain bread and cooked vegetables in women). It is therefore not appropriated to draw conclusion for the consumption of these food groups for the incidence of hypertension.

In summary, the present thesis found several food groups to be associated with incident hypertension in men and women. In particular a lunch pattern in women, and a fat dairy and cheese pattern as well as spirit consumption in men and women were most informative regarding incident hypertension. Although some of the present findings are contrary to previous studies, biological plausible explanations were offered confirming the obtained findings. Nevertheless, it is also possible that recall bias and misreporting during assessment

of the FFQs may have confounded the present findings. In addition, most of the observed associations, seen in the partial plots of selected food groups, indicated small alterations in the 10-year predicted hypertension-free survival due to different food group consumptions. However, when these improvements of 10-year predicted hypertension-free survival due to changes in dietary behaviour are extrapolated to the world population, this can contributes to enhanced life quality of many individuals concerned and to significant financial benefits for health care systems. Lastly, due to the observational nature of the present thesis, the obtained findings are not suitable to draw causal conclusions. Hence, validations in external cohorts are highly recommended to confirm the present findings.

4.3 Strengths and Limitations

A major strengths of the present thesis was the application of the novel machine learning method RSF which is particular suitable for survival analysis of complex data with numerous highly correlated variables. In the present thesis a RSF backward algorithm was implemented allowing the identification of most informative variables regarding time until event of interest. Furthermore, prior to the application on observational data, the RSF method and the RSF backward algorithm was evaluated in a simulation study. This enabled to proof how confounding, correlation and usage of data with a mixture of categorical and continuous data influence the variable selection process of the used RSF approach. Thereby, a reliable identification of metabolites associated with incident T2D and food groups associated with incident hypertension was enabled. As shown the identified metabolites and food groups may contribute to an improved prediction of incident T2D and incident hypertension, respectively. In addition, non-linear associations were revealed by the usage of partial plots allowing the identification of possible clinical thresholds for usage in clinical diagnosis or

subsequent studies. A further strength of the present thesis was the connecting of these finding with correlation networks derived by GGM which resulted in the identification of correlation patterns. In particular, the identified metabolite correlation pattern may point to a biological pathway associated with the incidence of T2D. The prospective design and the large sample size of the used EPIC-Potsdam data represent further strengths of the present study. Moreover, the assessment of various factors in the EPIC-Potsdam study allowed it to adjust for a various set of possible confounders.

Nevertheless, the present thesis has also some limitations. Although the RSF variable selection approach was compared with the study of Floegel et al. (52) that applied a multivariate Cox proportional hazards regression approach on the same metabolite data, additional validation studies are necessary to compare the performance of the introduced RSF approach with other survival analysis approaches like Cox-regression under lasso penalization or *L*2-boosting (56, 57). In addition, a few recent studies showed comparable performance of RSF and Cox regression (23, 68-70). In this context, it may be also appropriate to validate the present findings regarding food groups and incident hypertension with another survival method and in an external cohort. A further limitation is that the present thesis cannot prove causality due to the observational nature of the data. Nevertheless, the present findings revealed a deeper insight in the incidence of T2D and hypertension. However, further investigations in randomized control studies are desirable in order to proof the causality of the present findings.

The used food group data in the present thesis based on FFQs. Source of errors during assessment of the FFQs are recall bias and misreporting and may, therefore, represent also a limitation of the present thesis. However, the used FFQ data were previously validated by comparison with 24h dietary recalls. The reported validity of the FFQ was, in particular, high

for alcohol beverages and moderate for red meat, processed meat, cheese and dairy products. In addition, a limitation was also that it was not possible to consider possible dietary change during the follow-up time as well as the distribution and point in time of meals during the day. Moreover, each respective food groups comprise several food items of the FFQs. RSF is a method that can handle data with a high number of variables and therefore It seems more appropriate to apply the RSF backward algorithm in future studies directly on data of food items to get a more detailed picture of the association between dietary products and incident hypertension.

Another limitation concerns the used metabolic data which based on the measurement of thawed baseline blood samples. Less is known, how the long-term storage of metabolites in a frozen state affects the stability of metabolites. Moreover, measured metabolite concentrations reflect only metabolic state at baseline. Hence, alterations within metabolite concentrations during the follow-up time could not be acquired. However, the biological variation and reliability of the metabolites over a 4 months period was evaluated in a previous study by Floegel et al. (80). Based on the intraclass-correlation coefficient high reliability was found for hexose, sphingolipids, amino acids and glycerophospholipids. Moreover, a metabolite pattern of incident T2D was identified in a metabolite network. Statements on the stability of the metabolite network over time, however, are not possible. Nevertheless, in the future it is aimed to further elucidate the impact of long-term storage on metabolite stability and to proof whether and to what extend metabolite concentrations altered in repetitive measurement at different time points. Moreover correlation network analyses will be performed in three German cohorts to validate and compare the stability of metabolite correlation networks between cohorts.

A lot of possible relevant confounders were included in the present study. However, in particular, for incident hypertension the prediction error rate of the final RSF model was relative low. Hence, further disease associated factors must exist which were not included in the present thesis. Stress or familiar disease history, for example, may be possible confounders for the determination of incident hypertension. A further improvement of the final RSF model regarding the incidence of hypertension may also be achieved by inclusion of clinical markers like HDL cholesterol, LDL cholesterol or triglycerides. Furthermore, like all prospective observational studies, misreporting and measurement errors may have introduced bias in used data. However, the application of standardized questionnaires and measurement procedures in the EPIC-Potsdam study have contribute to minimize those biases.

Lastly, cases of incident T2D and incident hypertension were identified by self-reports during the follow-up waves. In EPIC-Potsdam all self-reports have been verified by treating physician or clinic which enabled the acquisition of only physician confirmed diagnosed cases (88). However, late diagnosis or non-detection of existing cases with incidence T2D or incident hypertension cannot be excluded if participants avoid medical consultations. This may have confounded the present findings, but cannot be prevented in any cohort studies.

4.4 Conclusion and implications for public health

Taken together, it was demonstrated that the RSF method and the implemented RSF backward algorithm represent a sensible complement to existing survival analysis methods. The RSF backward algorithm is particularly useful for exploratory analysis of complex survival data to identify unknown biomarkers associated with time until event of interest. It was shown that the RSF method is able to take into account possible confounders. Moreover the

RSF method is also able to handle the problem of multicollinearity and thus provides the opportunity to identify also disease associated correlation patterns. Moreover, the visualisation by partial plots allows the investigation of the direction and potential non-linearity of individual disease associated markers.

In view of the epidemic increase of several chronic diseases more systematic approaches than the study of single risk factors are required to grasp the complex pathophysiology and multifactorial origin of chronic diseases (164, 165). Nowadays, the technological and methological capabilities are available to assess, for instance, the whole metabolome or genome of individuals. However, the identification of disease associated factors from those complex data is statistical challenging. Against this background, the RSF method can contribute to an improvement of our knowledge regarding the pathophysiological mechanisms of chronic diseases.

In the present thesis the RSF method was successfully applied to identify several metabolites and food groups associated with the incidence of T2D and hypertension, respectively. The inclusion of identified metabolites may contribute to an improved diagnosis of incident T2D. Individuals with a high risk could thus diagnosed earlier and targeted intervention strategies can be recommended in time. Thereby, quality of live can be improved and health care expenditures can be reduced. Although, the improvement of existing incident hypertension diagnosis by selected food groups is questionable, the obtained findings provide indications for dietary recommendations regarding incident hypertension. However, the verification and translation of the present findings for clinical diagnosis, prevention strategies and dietary recommendations should be a matter for future research. Furthermore, replications and validations of the present findings in external cohorts are required to confirm the suitability of the RSF method and the implemented RSF backward algorithm.

5 Literature

1. Ban RH, Kamvissi V, Schulte KM, Bornstein SR, Rubino F, Graessler J. Lipidomic profiling at the interface of metabolic surgery and cardiovascular disease. Current atherosclerosis reports. 2014;16(11):455.
2. Schwenk RW, Vogel H, Schurmann A. Genetic and epigenetic control of metabolic health. Mol Metab. 2013;2(4):337-47.
3. Portha B, Fournier A, Kioon MD, Mezger V, Movassat J. Early environmental factors, alteration of epigenetic marks and metabolic disease susceptibility. Biochimie. 2014;97:1-15.
4. Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015;14(4):388-405.
5. Hameed I, Masoodi SR, Mir SA, Nabi M, Ghazanfar K, Ganai BA. Type 2 diabetes mellitus: From a metabolic disorder to an inflammatory condition. World J Diabetes. 2015;6(4):598-612.
6. Guarner V, Rubio-Ruiz ME. Low-grade systemic inflammation connects aging, metabolic syndrome and cardiovascular disease. Interdiscip Top Gerontol. 2015;40:99-106.
7. De Jesus DF, Kulkarni RN. Epigenetic modifiers of islet function and mass. Trends Endocrinol Metab. 2014;25(12):628-36.
8. Shah NJ, Sureshkumar S, Shewade DG. Metabolomics: A Tool Ahead for Understanding Molecular Mechanisms of Drugs and Diseases. Indian J Clin Biochem. 2015;30(3):247-54.
9. Naba A, Clauser KR, Ding H, Whittaker CA, Carr SA, Hynes RO. The extracellular matrix: Tools and insights for the "omics" era. Matrix Biol. 2015.
10. Zhao YY, Miao H, Cheng XL, Wei F. Lipidomics: Novel insight into the biochemical mechanism of lipid metabolism and dysregulation-associated disease. Chem Biol Interact. 2015;240:220-38.
11. Adamski J, Suhre K. Metabolomics platforms for genome wide association studies--linking the genome to the metabolome. Curr Opin Biotechnol. 2013;24(1):39-47.
12. Bamberg F, Kauczor HU, Weckbach S, Schlett CL, Forsting M, Ladd SC, et al. Whole-Body MR Imaging in the German National Cohort: Rationale, Design, and Technical Background. Radiology. 2015:142242.
13. von Eyben FE, Mouritsen E, Holm J, Dimcevski G, Montvilas P, Suciu G. Computed tomography scans of intra-abdominal fat, anthropometric measurements, and 3 nonobese metabolic risk factors. Metabolism. 2006;55(10):1337-43.
14. Moskal A, Pisa PT, Ferrari P, Byrnes G, Freisling H, Boutron-Ruault MC, et al. Nutrient patterns and their food sources in an International Study Setting: report from the EPIC study. PloS one. 2014;9(6):e98647.
15. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. Environ Health Perspect. 2001;109(1):1-5.
16. Benjamini Y. Simultaneous and selective inference: Current successes and future challenges. Biom J. 2010;52(6):708-21.
17. Bender R, Lange S. Adjusting for multiple testing-when and how? Journal of Clinical Epidemiology. 2001;54(4):343-9.
18. Prunier JG, Colyn M, Legendre X, Nimon KF, Flamand MC. Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. Mol Ecol. 2015;24(2):263-83.
19. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. Am J Physiol. 1985;249(1 Pt 2):R1-12.
20. Zhang HH, Lu W. Adaptive Lasso for Cox's Proportional Hazards Model. Biometrika. 2007;94(3):691-703.
21. Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics. 2002;18(12):1625-32.
22. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics. 2004;20(suppl 1):i208-i15.
23. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Annals of Applied Statistics. 2008;2(3):841-60.

24. World Health Organization. Global status report on noncommunicable diseases 2010. World Health Organization: Geneva, Switzerland, 2010.
25. Danaei G, Finucane MM, Lin JK, Singh GM, Paciorek CJ, Cowan MJ, et al. National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5.4 million participants. Lancet. 2011;377(9765):568-77.
26. Scully T. Diabetes in numbers. Nature. 2012;485(7398):S2-3.
27. Neuhauser H, Thamm M, Ellert U. [Blood pressure in Germany 2008-2011: results of the German Health Interview and Examination Survey for Adults (DEGS1)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2013;56(5-6):795-801.
28. Heidemann C, Du Y, Schubert I, Rathmann W, Scheidt-Nave C. [Prevalence and temporal trend of known diabetes mellitus: results of the German Health Interview and Examination Survey for Adults (DEGS1)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2013;56(5-6):668-77.
29. Palmer AJ, Bulpitt CJ, Fletcher AE, Beevers DG, Coles EC, Ledingham JG, et al. Relation between blood pressure and stroke mortality. Hypertension. 1992;20(5):601-5.
30. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. Stroke; a journal of cerebral circulation. 1991;22(3):312-8.
31. Vasan RS, Larson MG, Leip EP, Evans JC, O'Donnell CJ, Kannel WB, et al. Impact of high-normal blood pressure on the risk of cardiovascular disease. The New England journal of medicine. 2001;345(18):1291-7.
32. Kannel WB. Elevated systolic blood pressure as a cardiovascular risk factor. The American journal of cardiology. 2000;85(2):251-5.
33. Levy D, Larson MG, Vasan RS, Kannel WB, Ho KK. The progression from hypertension to congestive heart failure. JAMA : the journal of the American Medical Association. 1996;275(20):1557-62.
34. Tozawa M, Iseki K, Iseki C, Kinjo K, Ikemiya Y, Takishita S. Blood pressure predicts risk of developing end-stage renal disease in men and women. Hypertension. 2003;41(6):1341-5.
35. Gross JL, de Azevedo MJ, Silveiro SP, Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes care. 2005;28(1):164-76.
36. Alexiadou K, Doupis J. Management of Diabetic Foot Ulcers. Diabetes Therapy. 2012;3(1):4.
37. Viswanath K, McGavin DD. Diabetic retinopathy: clinical findings and management. Community Eye Health. 2003;16(46):21-4.
38. Cade WT. Diabetes-Related Microvascular and Macrovascular Diseases in the Physical Therapy Setting. Physical Therapy. 2008;88(11):1322-35.
39. Statistisches Bundesamt Deutschland; Gesundheit: Krankheitskosten. Website: https://www-genesis.destatis.de/genesis/online/data;jsessionid=6704A70BBED8A9319EF77F0B37CFD7C1.tomcat_GO_1_2?operation=abruftabelleBearbeiten&levelindex=2&levelid=1441084909256&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&selectionname=23631-0001&auswahltext=&werteabruf=Werteabruf.
40. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Bohm M, et al. 2013 ESH/ESC Practice Guidelines for the Management of Arterial Hypertension. Blood pressure. 2014;23(1):3-16.
41. Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes care. 2011;34(6):1249-57.
42. American Diabetes Association. Standards of medical care in diabetes-2014. Diabetes care. 2014;37 Suppl 1:S14-80.
43. Guyon I., Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-82.
44. Gentle J. E., Härdle W. K., Mori Y. Handbook of Computational Statistics - Concepts and Methods. Springer Verlag ISBN-13: 978-3540404644 2012.
45. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning - with Applications in R. Springer Verlag ISBN 978-1-4614-7138-7. 2013.

46. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci. 2012;5(1):7-13.
47. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2; Montreal, Quebec, Canada. 1643047: Morgan Kaufmann Publishers Inc.; 1995. p. 1137-43.
48. Ernst MD. Permutation Methods: A Basis for Exact Inference. 2004:676-85.
49. Yoo W, Mayberry R, Bae S, Singh K, Peter He Q, Lillard JW, Jr. A Study of Effects of MultiCollinearity in the Multivariable Analysis. Int J Appl Sci Technol. 2014;4(5):9-19.
50. Leigh JP. Assessing the importance of an independent variable in multiple regression: is stepwise unwise?1988. 669-77 p.
51. Harrell Jr. FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag,. 2001:p. 64.
52. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost HG, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. Diabetes. 2012.
53. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011;17(4):448-53.
54. Xue X, Xie X, Gunter M, Rohan TE, Wassertheil-Smoller S, Ho GY, et al. Testing the proportional hazards assumption in case-cohort analysis. BMC medical research methodology. 2013;13:88.
55. Osborne J, Waters E. Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, 8(2). 2002.
56. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007; 69: 659–677 doi: 101111/j1467-9868200700607x.
57. Hothorn T, Buhlmann P. Model-based boosting in high dimensions. Bioinformatics. 2006;22(22):2828-9.
58. Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics. 2008;9:14.
59. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software (1984) ISBN 978-0-412-04841-8.
60. Breiman L. Technical Note: Some Properties of Splitting Criteria. Machine Learning. 1996;24(1).
61. Rokach L, Maimon O. Data mining with decision trees: theory and applications. World Scientific Pub Co Inc (2008) ISBN 978-9812771711.
62. Strobl C, Malley J, Tutz G. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological methods. 2009;14(4):323-48.
63. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
64. Breiman L. Random Forests. Mach Learn. 2001;45(1):5-32.
65. Leblanc M CJ. Survival trees by goodness of split. J Amer Stat Assoc. 1993;88:457-67.
66. Segal MR. Regression Trees for Censored-Data. Biometrics. 1988;44(1):35-47.
67. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. Journal of the American Statistical Association. 2010;105(489):205-17.
68. Datema FR, Moya A, Krause P, Back T, Willmes L, Langeveld T, et al. Novel Head and Neck Cancer Survival Analysis Approach: Random Survival Forests Versus Cox Proportional Hazards Regression. Head and Neck-Journal for the Sciences and Specialties of the Head and Neck. 2012;34(1):50-8.
69. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying Important Risk Factors for Survival in Patient With Systolic Heart Failure Using Random Survival Forests. Circulation-Cardiovascular Quality and Outcomes. 2011;4(1):39-45.

70. Omurlu IK, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. Expert Systems with Applications. 2009;36(4):8582-8.
71. Ishwaran H, Kogalur UB. Random survival forest for R. R News. 2007;7(2):25-31.
72. Ishwaran H, Kogalur UB. Package 'randomSurvivalForest'. http://ftpuni-bayreuthde/math/statlib/R/CRAN/src/contrib/Descriptions/randomSurvivalForesthtml. 2013.
73. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA : the journal of the American Medical Association. 1982;247(18):2543-6.
74. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of statistics. 2001;29(5):1189-232.
75. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models by Ralf Bender, Thomas Augustin and Maria Blettner, Statistics in Medicine 2005; 24:1713-1723. Statistics in medicine. 2006;25(11):1978-9.
76. Boeing H, Wahrendorf J, Becker N. EPIC-Germany--A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. Ann Nutr Metab. 1999;43(4):195-204.
77. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5(6B):1113-24.
78. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. Ann Nutr Metab. 1999;43(4):205-15.
79. Römisch-Margl W, Prehn C, Bogumil R, Röhring C, Suhre K, Adamski J. Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics : Official journal of the Metabolomic Society. 2012;8(1):133-42.
80. Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, Adamski J, et al. Reliability of Serum Metabolite Concentrations over a 4-Month Period Using a Targeted Metabolomic Approach. PloS one. 2011;6(6).
81. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J. Reproducibility and relative validity of food group intake in a food frequency questionnaire developed for the German part of the EPIC project. European Prospective Investigation into Cancer and Nutrition. International Journal of Epidemiology. 1997;26(suppl 1):S59.
82. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J. Reproducibility and relative validity of energy and macronutrient intake of a food frequency questionnaire developed for the German part of the EPIC project. European Prospective Investigation into Cancer and Nutrition. International Journal of Epidemiology. 1997;26(suppl 1):S71.
83. Kroke A, Klipstein-Grobusch K, Voss S, Moseneder J, Thielecke F, Noack R, et al. Validation of a self-administered food-frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. Am J Clin Nutr. 1999;70(4):439-47.
84. Kroke A, Bergmann MM, Lotze G, Jeckel A, Klipstein-Grobusch K, Boeing H. Measures of quality control in the German component of the EPIC study. Ann Nutr Metab. 1999;43(4):216-24.
85. Schulze MB, Hoffmann K, Kroke A, Boeing H. Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. Br J Nutr. 2001;85(3):363-73.
86. Wientzek A, Vigl M, Steindorf K, Bruhmann B, Bergmann MM, Harttig U, et al. The improved physical activity index for measuring physical activity in EPIC Germany. PloS one. 2014;9(3):e92005.
87. Schulze MB, Kroke A, Bergmann MM, Boeing H. Differences of blood pressure estimates between consecutive measurements on one occasion: Implications for inter-study comparability of epidemiologic studies. Eur J Epidemiol. 2000;16(10):891-8.
88. Bergmann MM, Bussas U, Boeing H. Follow-up procedures in EPIC-Germany - Data quality aspects. Ann Nutr Metab. 1999;43(4):225-34.

89. Knueppel S, Stang A. DAG program: identifying minimal sufficient adjustment sets. Epidemiology. 2010;21(1):159.
90. Kramer N, Schafer J, Boulesteix AL. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics. 2009;10:384.
91. Floegel A, Wientzek A, Bachlechner U, Jacobs S, Drogan D, Prehn C, et al. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. Int J Obes (Lond). 2014;38(11):1388-96.
92. Dietrich S, Floegel A, Boeing H, Schulze MB, Illig T, Pischon T, et al. Random Survival Forest in practice – a method for modelling high-dimensional metabolomics data in time to event analysis. . In revision at International Journal of Epidemiology.
93. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. Statistical Analysis and Data Mining. 2011;4(1):115-32.
94. Loh W.Y. SYS. Split Selection Methods for Classification Trees. Statistica Sinica, Vol 7, pp 815-840. 1997.
95. Lavie CJ, Milani RV, Ventura HO. Obesity and cardiovascular disease: risk factor, paradox, and impact of weight loss. Journal of the American College of Cardiology. 2009;53(21):1925-32.
96. Ertek S, Cicero A. Impact of physical activity on inflammation: effects on cardiovascular disease risk and other inflammatory conditions. Arch Med Sci. 2012;8(5):794-804.
97. Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, et al. Correction for population stratification in random forest analysis. Int J Epidemiol. 2012;41(6):1798-806.
98. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. Nat Commun. 2015;6:7432.
99. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7:3.
100. Rosenthal T, Alter A. Occupational stress and hypertension. J Am Soc Hypertens. 2012;6(1):2-22.
101. Nelsen DA, Jr. 2014 hypertension and cholesterol guidelines. J Ark Med Soc. 2014;111(1):12-3.
102. Siroky DS. Navigating Random Forests and related advances in algorithmic modeling. Statist Surv Vol3, pp 147-163 2009.
103. van der Schaaf A, Xu CJ, van Luijk P, Van't Veld AA, Langendijk JA, Schilstra C. Multivariate modeling of complications with data driven variable selection: guarding against overfitting and effects of data set size. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2012;105(1):115-21.
104. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics : Official journal of the Metabolomic Society. 2006;2(4):171-96.
105. American Diabetes Association. Standards of medical care in diabetes-2015. Diabetes care. 2015;38 Suppl 1.
106. Tabak AG, Jokela M, Akbaraly TN, Brunner EJ, Kivimaki M, Witte DR. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. Lancet. 2009;373(9682):2215-21.
107. Montonen J, Jarvinen R, Knekt P, Heliovaara M, Reunanen A. Consumption of sweetened beverages and intakes of fructose and glucose predict type 2 diabetes occurrence. J Nutr. 2007;137(6):1447-54.
108. Hu FB, Malik VS. Sugar-sweetened beverages and risk of obesity and type 2 diabetes: epidemiologic evidence. Physiol Behav. 2010;100(1):47-54.
109. Malik VS, Popkin BM, Bray GA, Despres JP, Willett WC, Hu FB. Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes: a meta-analysis. Diabetes care. 2010;33(11):2477-83.
110. Kolderup A, Svihus B. Fructose Metabolism and Relation to Atherosclerosis, Type 2 Diabetes, and Obesity. J Nutr Metab. 2015;2015:823081.

111. Laughlin MR. Normal roles for dietary fructose in carbohydrate metabolism. Nutrients. 2014;6(8):3117-29.
112. Magnusson M, Lewis GD, Ericson U, Orho-Melander M, Hedblad B, Engstrom G, et al. A diabetes-predictive amino acid score and future cardiovascular disease. European heart journal. 2013;34(26):1982-9.
113. Tai ES, Tan ML, Stevens RD, Low YL, Muehlbauer MJ, Goh DL, et al. Insulin resistance is associated with a metabolic profile of altered protein metabolism in Chinese and Asian-Indian men. Diabetologia. 2010;53(4):757-67.
114. Huffman KM, Shah SH, Stevens RD, Bain JR, Muehlbauer M, Slentz CA, et al. Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women. Diabetes care. 2009;32(9):1678-83.
115. Tillin T, Hughes AD, Wang Q, Wurtz P, Ala-Korpela M, Sattar N, et al. Diabetes risk and amino acid profiles: cross-sectional and prospective analyses of ethnicity, amino acids and diabetes in a South Asian and European cohort from the SABRE (Southall And Brent REvisited) Study. Diabetologia. 2015;58(5):968-79.
116. Kawanaka M, Nishino K, Oka T, Urata N, Nakamura J, Suehiro M, et al. Tyrosine levels are associated with insulin resistance in patients with nonalcoholic fatty liver disease. Hepat Med. 2015;7:29-35.
117. Wang W, Wu Z, Dai Z, Yang Y, Wang J, Wu G. Glycine metabolism in animals and humans: implications for nutrition and health. Amino acids. 2013;45(3):463-77.
118. Zhu C, Liang QL, Hu P, Wang YM, Luo GA. Phospholipidomic identification of potential plasma biomarkers associated with type 2 diabetes mellitus and diabetic nephropathy. Talanta. 2011;85(4):1711-20.
119. Wallner S, Schmitz G. Plasmalogens the neglected regulatory and scavenging lipid species. Chemistry and physics of lipids. 2011;164(6):573-89.
120. Pietilainen KH, Sysi-Aho M, Rissanen A, Seppanen-Laakso T, Yki-Jarvinen H, Kaprio J, et al. Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects--a monozygotic twin study. PloS one. 2007;2(2):e218.
121. Jung YY, Nam Y, Park YS, Lee HS, Hong SA, Kim BK, et al. Protective effect of phosphatidylcholine on lipopolysaccharide-induced acute inflammation in multiple organ injury. Korean J Physiol Pharmacol. 2013;17(3):209-16.
122. Ghyczy M, Torday C, Kaszaki J, Szabo A, Czobel M, Boros M. Oral phosphatidylcholine pretreatment decreases ischemia-reperfusion-induced methane generation and the inflammatory response in the small intestine. Shock. 2008;30(5):596-602.
123. Treede I, Braun A, Sparla R, Kuhnel M, Giese T, Turner JR, et al. Anti-inflammatory effects of phosphatidylcholine. J Biol Chem. 2007;282(37):27155-64.
124. Eros G, Varga G, Varadi R, Czobel M, Kaszaki J, Ghyczy M, et al. Anti-inflammatory action of a phosphatidylcholine, phosphatidylethanolamine and N-acylphosphatidylethanolamine-enriched diet in carrageenan-induced pleurisy. Eur Surg Res. 2009;42(1):40-8.
125. Nieto-Vazquez I, Fernandez-Veledo S, Kramer DK, Vila-Bedmar R, Garcia-Guerra L, Lorenzo M. Insulin resistance associated to obesity: the link TNF-alpha. Arch Physiol Biochem. 2008;114(3):183-94.
126. Cole LK, Vance JE, Vance DE. Phosphatidylcholine biosynthesis and lipoprotein metabolism. Biochimica et biophysica acta. 2012;1821(5):754-61.
127. Choi SH, Ginsberg HN. Increased very low density lipoprotein (VLDL) secretion, hepatic steatosis, and insulin resistance. Trends Endocrinol Metab. 2011;22(9):353-63.
128. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, et al. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol. 2012;8:615.
129. Ha CY, Kim JY, Paik JK, Kim OY, Paik YH, Lee EJ, et al. The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. Clin Endocrinol (Oxf). 2012;76(5):674-82.
130. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the

panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA : the journal of the American Medical Association. 2014;311(5):507-20.
131. Appel LJ, Brands MW, Daniels SR, Karanja N, Elmer PJ, Sacks FM. Dietary approaches to prevent and treat hypertension: a scientific statement from the American Heart Association. Hypertension. 2006;47(2):296-308.
132. O'Shaughnessy KM. Role of diet in hypertension management. Curr Hypertens Rep. 2006;8(4):292-7.
133. Borgi L, Curhan GC, Willett WC, Hu FB, Satija A, Forman JP. Long-term intake of animal flesh and risk of developing hypertension in three prospective cohort studies. Journal of hypertension. 2015.
134. Wang L, Manson JE, Buring JE, Sesso HD. Meat intake and the risk of hypertension in middle-aged and older women. Journal of hypertension. 2008;26(2):215-22.
135. Stamler J, Caggiula A, Grandits GA, Kjelsberg M, Cutler JA. Relationship to blood pressure of combinations of dietary macronutrients. Findings of the Multiple Risk Factor Intervention Trial (MRFIT). Circulation. 1996;94(10):2417-23.
136. Stamler J, Liu K, Ruth KJ, Pryer J, Greenland P. Eight-year blood pressure change in middle-aged men: relationship to multiple nutrients. Hypertension. 2002;39(5):1000-6.
137. Cicero AF, Rosticci M, Baronio C, Morbini M, Parini A, Grandi E, et al. Serum LDL cholesterol levels and new onset of arterial hypertension: an 8-year follow-up. Eur J Clin Invest. 2014;44(10):926-32.
138. Karppanen H, Mervaala E. Sodium intake and hypertension. Prog Cardiovasc Dis. 2006;49(2):59-75.
139. Soedamah-Muthu SS, Verberne LD, Ding EL, Engberink MF, Geleijnse JM. Dairy consumption and incidence of hypertension: a dose-response meta-analysis of prospective cohort studies. Hypertension. 2012;60(5):1131-7.
140. Heraclides A, Mishra GD, Hardy RJ, Geleijnse JM, Black S, Prynne CJ, et al. Dairy intake, blood pressure and incident hypertension in a general British population: the 1946 birth cohort. Eur J Nutr. 2012;51(5):583-91.
141. Wang L, Manson JE, Buring JE, Lee IM, Sesso HD. Dietary intake of dairy products, calcium, and vitamin D and the risk of hypertension in middle-aged and older women. Hypertension. 2008;51(4):1073-9.
142. Kratz M, Baars T, Guyenet S. The relationship between high-fat dairy consumption and obesity, cardiovascular, and metabolic disease. Eur J Nutr. 2013;52(1):1-24.
143. Holmberg S, Thelin A. High dairy fat intake related to less central obesity: a male cohort study with 12 years' follow-up. Scand J Prim Health Care. 2013;31(2):89-94.
144. Briasoulis A, Agarwal V, Messerli FH. Alcohol consumption and the risk of hypertension in men and women: a systematic review and meta-analysis. J Clin Hypertens (Greenwich). 2012;14(11):792-8.
145. Okubo Y, Suwazono Y, Kobayashi E, Nogawa K. Alcohol consumption and blood pressure change: 5-year follow-up study of the association in normotensive workers. Journal of human hypertension. 2001;15(6):367-72.
146. Gillman MW, Cook NR, Evans DA, Rosner B, Hennekens CH. Relationship of alcohol intake with blood pressure in young adults. Hypertension. 1995;25(5):1106-10.
147. Sesso HD, Cook NR, Buring JE, Manson JE, Gaziano JM. Alcohol consumption and the risk of hypertension in women and men. Hypertension. 2008;51(4):1080-7.
148. Husain K, Ansari RA, Ferder L. Alcohol-induced hypertension: Mechanism and prevention. World J Cardiol. 2014;6(5):245-52.
149. Marchi KC, Muniz JJ, Tirapelli CR. Hypertension and chronic ethanol consumption: What do we know after a century of study? World J Cardiol. 2014;6(5):283-94.
150. Paplovic LB, Popovic MB, Bijelovic SV, Velicki RS, Torovic LD. Salt Content in Ready-to-Eat Food and Bottled Spring and Mineral Water Retailed in Novi Sad. Srp Arh Celok Lek. 2015;143(5-6):362-8.

151. Powell LM, Nguyen BT, Dietz WH. Energy and nutrient intake from pizza in the United States. Pediatrics. 2015;135(2):322-30.
152. Haidari F, Shirbeigi E, Cheraghpour M, Mohammadshahi M. Association of dietary patterns with body mass index, waist circumference, and blood pressure in an adult population in Ahvaz, Iran. Saudi Med J. 2014;35(9):967-74.
153. Giugliano D, Nappo F, Coppola L. Pizza and vegetables don't stick to the endothelium. Circulation. 2001;104(7):E34-5.
154. Nunez-Cordoba JM, Valencia-Serrano F, Toledo E, Alonso A, Martinez-Gonzalez MA. The Mediterranean diet and incidence of hypertension: the Seguimiento Universidad de Navarra (SUN) Study. American journal of epidemiology. 2009;169(3):339-46.
155. Sleiman D, Al-Badri MR, Azar ST. Effect of mediterranean diet in diabetes control and cardiovascular risk modification: a systematic review. Front Public Health. 2015;3:69.
156. DiNicolantonio JJ, Lucan SC. The wrong white crystals: not salt but sugar as aetiological in hypertension and cardiometabolic disease. Open Heart. 2014;1(1):e000167.
157. Appel LJ, Moore TJ, Obarzanek E, Vollmer WM, Svetkey LP, Sacks FM, et al. A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. The New England journal of medicine. 1997;336(16):1117-24.
158. John JH, Ziebland S, Yudkin P, Roe LS, Neil HA. Effects of fruit and vegetable consumption on plasma antioxidant concentrations and blood pressure: a randomised controlled trial. Lancet. 2002;359(9322):1969-74.
159. Psaltopoulou T, Naska A, Orfanos P, Trichopoulos D, Mountokalakis T, Trichopoulou A. Olive oil, the Mediterranean diet, and arterial blood pressure: the Greek European Prospective Investigation into Cancer and Nutrition (EPIC) study. Am J Clin Nutr. 2004;80(4):1012-8.
160. Nunez-Cordoba JM, Alonso A, Beunza JJ, Palma S, Gomez-Gracia E, Martinez-Gonzalez MA. Role of vegetables and fruits in Mediterranean diets to prevent hypertension. Eur J Clin Nutr. 2009;63(5):605-12.
161. He FJ, Nowson CA, Lucas M, MacGregor GA. Increased consumption of fruit and vegetables is related to a reduced risk of coronary heart disease: meta-analysis of cohort studies. Journal of human hypertension. 2007;21(9):717-28.
162. Eichholzer M, Luthy J, Gutzwiller F, Stahelin HB. The role of folate, antioxidant vitamins and other constituents in fruit and vegetables in the prevention of cardiovascular disease: the epidemiological evidence. Int J Vitam Nutr Res. 2001;71(1):5-17.
163. Roberts WG, Gordon MH. Determination of the total antioxidant activity of fruits and vegetables by a liposome assay. J Agric Food Chem. 2003;51(5):1486-93.
164. Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(11):2954-7.
165. Hu FB. Metabolic profiling of diabetes: from black-box epidemiology to systems epidemiology. Clin Chem. 2011;57(9):1224-6.

Appendix

Table S1: Summary of serum metabolites

Abbreviation	Biochemical name	non-cases of incident T2D Mean	non-cases of incident T2D SD	cases of incident T2D Mean	cases of incident T2D SD
C0	DL-Carnitine	33.711	8.074	38.519	9.873
C10	Decanoyl-L-carnitine	0.239	0.102	0.250	0.098
C10:2	Decadienyl-L-carnitine	0.047	0.016	0.051	0.017
C14:1	Tetradecenoyl-L-carnitine	0.190	0.046	0.197	0.050
C14:2	Tetradecadienyl-L-carnitine	0.025	0.011	0.027	0.012
C16	Hexadecanoyl-L-carnitine	0.111	0.028	0.124	0.033
C16:2	Hexadecadienyl-L-carnitine	0.009	0.006	0.010	0.006
C18	Octadecanoyl-L-carnitine	0.048	0.014	0.049	0.016
C18:1	Octadecenoyl-L-carnitine	0.137	0.039	0.150	0.045
C18:2	Octadecadienyl-L-carnitine	0.063	0.020	0.065	0.022
C2	Acetyl-L-carnitine	6.873	2.845	7.502	3.124
C3	Propionyl-L-carnitine	0.338	0.122	0.408	0.155
C3-DC-M / C5-OH	Methylmalonyl-L-carnitine / Hydroxyvaleryl-L-carnitine	0.031	0.010	0.033	0.009
C5-DC / C6-OH	Glutaryl-L-carnitine / Hydroxyhexanoyl-L-carnitine	0.024	0.008	0.025	0.009
C7-DC	Pimelyl-L-carnitine	0.033	0.013	0.033	0.013
C8:1	Octenoyl-L-carnitine	0.102	0.044	0.124	0.057
C9	Nonayl-L-carnitine	0.051	0.022	0.052	0.024
Arg	Arginine	104.433	23.311	110.351	24.337
Gln	Glutamine	570.772	90.877	569.307	94.314
Gly	Glycine	265.857	87.544	235.776	71.944
His	Histidine	93.092	16.895	89.709	18.281
Met	Methionine	28.446	7.954	28.769	7.435
Orn	Ornithine	94.607	25.226	101.742	26.420
Phe	Phenylalanine	55.788	11.970	59.961	12.418
Pro	Proline	200.016	66.852	212.520	65.246
Ser	Serine	118.372	29.247	110.806	27.274
Thr	Threonine	98.977	31.641	91.642	26.476
Trp	Tryptophan	79.370	12.171	80.662	13.482
Tyr	Tyrosine	79.110	23.312	91.282	25.966
Val	Valine	280.432	72.691	319.233	76.187
xLeu	Leucine/Isoleucin	198.331	59.037	222.373	66.637
PC aa C28:1	Phosphatidylcholine diacyl C28:1	3.747	0.930	3.859	0.923
PC aa C30:0	Phosphatidylcholine diacyl C30:0	5.684	1.842	5.981	1.903
PC aa C32:0	Phosphatidylcholine diacyl C32:0	14.820	3.363	15.557	3.751
PC aa C32:1	Phosphatidylcholine diacyl C32:1	16.739	8.576	21.011	11.137
PC aa C32:2	Phosphatidylcholine diacyl C32:2	5.677	2.032	5.789	1.729
PC aa C32:3	Phosphatidylcholine diacyl C32:3	0.639	0.161	0.654	0.161

Table S1 continued

PC aa C34:1	Phosphatidylcholine diacyl C34:1	236.635	59.864	254.110	71.507
PC aa C34:2	Phosphatidylcholine diacyl C34:2	441.689	96.162	452.264	101.838
PC aa C34:3	Phosphatidylcholine diacyl C34:3	19.242	5.899	19.938	5.936
PC aa C34:4	Phosphatidylcholine diacyl C34:4	2.447	0.919	2.592	0.856
PC aa C36:0	Phosphatidylcholine diacyl C36:0	2.448	0.807	2.383	0.797
PC aa C36:1	Phosphatidylcholine diacyl C36:1	56.764	14.595	62.632	18.574
PC aa C36:2	Phosphatidylcholine diacyl C36:2	280.140	59.199	291.468	64.897
PC aa C36:3	Phosphatidylcholine diacyl C36:3	156.443	37.845	167.857	39.660
PC aa C36:4	Phosphatidylcholine diacyl C36:4	223.791	59.482	233.359	59.929
PC aa C36:5	Phosphatidylcholine diacyl C36:5	32.876	15.962	36.812	17.108
PC aa C36:6	Phosphatidylcholine diacyl C36:6	1.449	0.550	1.500	0.553
PC aa C38:0	Phosphatidylcholine diacyl C38:0	3.261	0.911	3.156	0.872
PC aa C38:1	Phosphatidylcholine diacyl C38:1	0.679	0.425	0.605	0.365
PC aa C38:3	Phosphatidylcholine diacyl C38:3	55.488	14.810	68.060	18.454
PC aa C38:4	Phosphatidylcholine diacyl C38:4	115.223	29.351	128.137	33.731
PC aa C38:5	Phosphatidylcholine diacyl C38:5	58.604	16.020	62.377	18.174
PC aa C38:6	Phosphatidylcholine diacyl C38:6	105.909	31.185	109.292	33.869
PC aa C40:2	Phosphatidylcholine diacyl C40:2	0.384	0.146	0.359	0.119
PC aa C40:3	Phosphatidylcholine diacyl C40:3	0.596	0.208	0.569	0.189
PC aa C40:4	Phosphatidylcholine diacyl C40:4	3.930	1.052	4.394	1.349
PC aa C40:5	Phosphatidylcholine diacyl C40:5	10.991	3.300	12.597	4.097
PC aa C40:6	Phosphatidylcholine diacyl C40:6	32.516	10.588	37.645	12.605
PC aa C42:0	Phosphatidylcholine diacyl C42:0	0.631	0.185	0.543	0.153
PC aa C42:1	Phosphatidylcholine diacyl C42:1	0.315	0.085	0.272	0.070
PC aa C42:2	Phosphatidylcholine diacyl C42:2	0.210	0.066	0.191	0.064
PC aa C42:4	Phosphatidylcholine diacyl C42:4	0.203	0.047	0.193	0.045
PC aa C42:5	Phosphatidylcholine diacyl C42:5	0.437	0.142	0.428	0.136
PC aa C42:6	Phosphatidylcholine diacyl C42:6	0.731	0.169	0.718	0.167
PC ae C30:0	Phosphatidylcholine acyl alkyl C30:0	0.441	0.132	0.405	0.125
PC ae C30:1	Phosphatidylcholine acyl alkyl C30:1	0.419	0.202	0.379	0.197
PC ae C30:2	Phosphatidylcholine acyl alkyl C30:2	0.148	0.049	0.151	0.051
PC ae C32:1	Phosphatidylcholine acyl alkyl C32:1	2.968	0.673	2.748	0.575
PC ae C32:2	Phosphatidylcholine acyl alkyl C32:2	0.764	0.186	0.713	0.165
PC ae C34:0	Phosphatidylcholine acyl alkyl C34:0	1.923	0.545	1.890	0.560
PC ae C34:1	Phosphatidylcholine acyl alkyl C34:1	10.716	2.510	10.329	2.500
PC ae C34:2	Phosphatidylcholine acyl alkyl C34:2	14.191	3.698	12.504	3.116
PC ae C34:3	Phosphatidylcholine acyl alkyl C34:3	8.894	2.502	7.377	2.048
PC ae C36:0	Phosphatidylcholine acyl alkyl C36:0	0.794	0.253	0.794	0.256
PC ae C36:1	Phosphatidylcholine acyl alkyl C36:1	9.865	2.386	9.603	2.392
PC ae C36:2	Phosphatidylcholine acyl alkyl C36:2	18.306	4.558	16.282	4.231
PC ae C36:3	Phosphatidylcholine acyl alkyl C36:3	10.050	2.444	9.180	2.142
PC ae C36:4	Phosphatidylcholine acyl alkyl C36:4	19.140	4.965	18.802	4.655
PC ae C36:5	Phosphatidylcholine acyl alkyl C36:5	11.460	3.126	11.212	2.950
PC ae C38:0	Phosphatidylcholine acyl alkyl C38:0	2.381	0.735	2.360	0.710
PC ae C38:1	Phosphatidylcholine acyl alkyl C38:1	1.266	0.347	1.245	0.347
PC ae C38:2	Phosphatidylcholine acyl alkyl C38:2	2.277	0.569	2.118	0.530

Table S1 continued

PC ae C38:3	Phosphatidylcholine acyl alkyl C38:3	5.046	1.170	5.003	1.136
PC ae C38:4	Phosphatidylcholine acyl alkyl C38:4	14.806	3.310	13.789	3.054
PC ae C38:5	Phosphatidylcholine acyl alkyl C38:5	18.606	4.259	17.847	3.985
PC ae C38:6	Phosphatidylcholine acyl alkyl C38:6	8.455	2.211	8.343	2.047
PC ae C40:1	Phosphatidylcholine acyl alkyl C40:1	1.483	0.388	1.390	0.365
PC ae C40:2	Phosphatidylcholine acyl alkyl C40:2	2.328	0.564	2.263	0.562
PC ae C40:3	Phosphatidylcholine acyl alkyl C40:3	1.226	0.251	1.132	0.237
PC ae C40:4	Phosphatidylcholine acyl alkyl C40:4	2.444	0.551	2.178	0.479
PC ae C40:5	Phosphatidylcholine acyl alkyl C40:5	4.082	0.864	3.687	0.823
PC ae C40:6	Phosphatidylcholine acyl alkyl C40:6	5.730	1.455	5.234	1.444
PC ae C42:1	Phosphatidylcholine acyl alkyl C42:1	0.352	0.094	0.341	0.093
PC ae C42:2	Phosphatidylcholine acyl alkyl C42:2	0.706	0.173	0.681	0.180
PC ae C42:3	Phosphatidylcholine acyl alkyl C42:3	0.887	0.208	0.785	0.191
PC ae C42:4	Phosphatidylcholine acyl alkyl C42:4	0.997	0.243	0.831	0.206
PC ae C42:5	Phosphatidylcholine acyl alkyl C42:5	2.453	0.556	2.105	0.462
PC ae C44:3	Phosphatidylcholine acyl alkyl C44:3	0.110	0.033	0.102	0.030
PC ae C44:4	Phosphatidylcholine acyl alkyl C44:4	0.402	0.105	0.346	0.095
PC ae C44:5	Phosphatidylcholine acyl alkyl C44:5	1.831	0.489	1.562	0.442
PC ae C44:6	Phosphatidylcholine acyl alkyl C44:6	1.222	0.330	1.026	0.279
lysoPC a C14:0	lysoPhosphatidylcholine acyl C14:0	5.654	0.968	5.787	0.951
lysoPC a C16:0	lysoPhosphatidylcholine acyl C16:0	107.796	24.013	109.995	24.058
lysoPC a C16:1	lysoPhosphatidylcholine acyl C16:1	3.297	1.094	3.523	1.105
lysoPC a C17:0	lysoPhosphatidylcholine acyl C17:0	1.993	0.598	1.769	0.531
lysoPC a C18:0	lysoPhosphatidylcholine acyl C18:0	28.911	7.945	29.620	7.444
lysoPC a C18:1	lysoPhosphatidylcholine acyl C18:1	18.525	5.250	16.863	4.578
lysoPC a C18:2	lysoPhosphatidylcholine acyl C18:2	33.382	12.828	27.178	10.262
lysoPC a C20:3	lysoPhosphatidylcholine acyl C20:3	2.338	0.720	2.529	0.734
lysoPC a C20:4	lysoPhosphatidylcholine acyl C20:4	5.954	1.716	5.824	1.741
lysoPC a C28:1	lysoPhosphatidylcholine acyl C28:1	0.653	0.242	0.628	0.234
SM OH C14:1	Hydroxysphingomyelin C14:1	7.841	2.152	7.594	2.139
SM OH C16:1	Hydroxysphingomyelin C16:1	4.065	1.128	3.994	1.212
SM OH C22:1	Hydroxysphingomyelin C22:1	15.141	3.973	15.575	4.062
SM OH C22:2	Hydroxysphingomyelin C22:2	13.156	3.437	12.629	3.445
SM OH C24:1	Hydroxysphingomyelin C24:1	1.608	0.483	1.597	0.489
SM C16:0	Sphingomyelin C16:0	116.038	24.077	113.403	24.695
SM C16:1	Sphingomyelin C16:1	18.242	4.012	18.932	4.287
SM C18:0	Sphingomyelin C18:0	25.632	6.072	27.440	7.103
SM C18:1	Sphingomyelin C18:1	12.641	3.190	13.468	3.746
SM C20:2	Sphingomyelin C20:2	0.771	0.416	0.762	0.458
SM C24:0	Sphingomyelin C24:0	25.545	6.488	26.361	6.555
SM C24:1	Sphingomyelin C24:1	49.692	13.841	47.312	14.088
SM C26:0	Sphingomyelin C26:0	0.233	0.078	0.227	0.078
SM C26:1	Sphingomyelin C26:1	0.540	0.170	0.537	0.172
H1	Hexose	4606.580	902.802	5495.670	1624.820

Table S2: Composition of food groups

Food or food group	Food items included in group
Whole-grain bread	Whole-grain bread, dark and whole-grain rolls
Other bread	Rye bread, wheat bread, mixed bread, pale rolls, crisp bread, croissants
Muesli	Whole-grain, breakfast cereal, muesli
Corn flakes	Corn flakes, crisps
Pasta, rice	Cooked pasta, cooked rice
Vegetarian dishes	Vegetarian dishes, vegetarian spreads
Chips, salt sticks	Chips, flips, salt sticks, crackers
Pizza	Pizza, onion tart, quiche
Cake, cookies	Fruit cake, pound cake, sponge cake, cream cake, flan, brioches, pastries, sweet particles, biscuits, cookies, pancakes
Confect	Chocolate, candy bars, pralines, sugar in coffee and tea, ice cream
Sweet bread spreads	Jam, honey, chocolate spread, peanut butter
Eggs	Boiled eggs, fried eggs, omelettes
Fresh fruit	Apple, pear, peach, nectarine, cherry, plum, plums, grapes, strawberries, currants, raspberries, blackberries, bananas, kiwi, mango, fresh pineapple, orange, Grape fruit, mandarin
Canned fruit	Fruit compote, Canned fruit
Raw vegetables	Cucumber, radish, cabbage, carrots, seeds, sprouts, peppers, chilli pepper, tomato, raw onion, lettuce, endive, chinese cabbage, mixed salad
cabbage	Cauliflower, red cabbage, white cabbage, kohlrabi, broccoli, other cabbage
Cooked vegetables	Tomatoes, tomato sauce, sweet peppers, zucchini, eggplant (aubergine), spinach, carrots, asparagus, pea-carrot vegetable mix, leeks, and celery, broccoli, cauliflower, red and white cabbage, and kohlrabi, green peas, green beans, and pea/bean/lentil stew
Garlic	Raw or fried/cooked garlic (YES/NO answer)
Mushrooms	Fresh mushrooms, mushroom dishes
legumes	Green peas, green beans, lentil soup, pea soup, bean stew
Cooked potatoes	Salted potatoes, jacket potatoes, mashed potatoes, potato salad, dumplings
Fried potatoes	French fries, potato fritters, fried potatoes
Nuts	Nuts
Low-fat dairy products	Milk, dairy drinks, yoghurt, fruit yoghurt, sour milk, kefir, quark, herb quark (fat<=1.5%)
high-fat dairy products	Milk, dairy drinks, yoghurt, fruit yoghurt, quark, herb quark (fat>3.5% or no matter answer or fat>=20%), whipped cream
Low-fat cheese	Cream cheese, Gouda, Emmental, Tilsiter, Camembert, Brie, Gorgonzola (reduced-fat or skim stage)
high-fat cheese	Cream cheese, Gouda, Emmental, Tilsiter, Camembert, Brie, Gorgonzola, cheese (normal/creamfat, double fat / no matter answer)
Water	Tap water, mineral water
Coffee	Coffee with caffeine (black, with milk, with condensed milk, milk with sweetener)
Decaffeinated coffee	Decaffeinated coffee (black, with milk, with condensed milk, with sweetener)
Tea	Black tea, green tea, fruit and herbal teas (pure, with milk, with condensed milk, with sweetener, lemon juice)

Table S2 continued

Fruit juice	citrus (ingredient salad dressing), apple-, orange-, grapefruit-, grape-, cherry-, or pineapple juice, multivitamin drinks
Low-energy soft-drinks	Cola, soda (calorie-reduced)
high-energy soft-drinks	Cola, soda (normal), non-alcoholic beer, malt beer
Beer	Beer
Wine	Wine, fruit wine, champagne
Spirits	Spirits
Other alcoholic beverages	Desert wine, liquor, aperitif
Butter	Butter used as a bread spread and for food preparation
Margarine	Margarine used as a bread spread and for food preparation
Vegetable oils	Vegetable fat used for food preparation (frying, dressing, etc.), e.g. Sunflower oil, olive oil, safflower oil
Other fat	Animal fat for cooking (meat/fish, and vegetables), grease for the preparation of meat/fish, and vegetables, no matter answer
Sauce	Ketchup, brown and white sauce, salad dressing, sauce for vegetables
Dessert	Pudding, Fruit curd, ice cream sundaes, sweet soufflé
Fish	Fish, canned fish, smoked fish, fish sticks
Poultry	Fried, grilled, or roasted chicken or turkey
Meat	Pork cutlets, chops, steak, fillet, roast pork, pork goulash, Sliced, kassler, pork ribs, cooking of pork, pork knuckle, knuckle, pork belly, boulette, hamburgers, meatloaf, meat sauce, hash, liver, calf's liver, lamb, rabbit, beef steak, fillet, loin, roast beef, boiled meat, beef rolls, beef stew, beef
Processed meat	Liver sausage, sausage spread, soft sausage, salami, hard sausage, smoked meat, ham, ham sausage, lyoner, hunting sausage, blood sausage, black pudding, brawn sausage, bratwurst, frankfurter, bockwurst, knackwurst, meat sausage
Soup	Vegetable/potato stew, vegetable soup, meat/fish soup, broth, thickened soup

Table S3: Summary of food groups

Variable	Name of food group	all control Mean	Std Dev	all case Mean	Std Dev	men control Mean (SD)	men case Mean (SD)	women control Mean (SD)	women case Mean (SD)
group1	whole grain bread [g/d]	48.82	56.20	45.84	55.21	45.26 (61.48)	41.38 (58.61)	50.36 (53.69)	47.95 (53.45)
group2	other bread [g/d]	125.60	79.78	133.02	80.11	173.41 (90.81)	180.42 (90.93)	104.90 (64.29)	110.70 (63.23)
group3	grain flakes, grains, muesli [g/d]	6.54	16.45	4.94	14.31	5.76 (17.39)	3.74 (11.51)	6.87 (16.01)	5.51 (15.42)
group4	cornflakes, crisps [g/d]	2.02	5.78	1.60	5.92	1.66(5.35)	1.19 (4.45)	2.17 (5.95)	1.80 (6.49)
group5	pasta, rice [g/d]	17.55	15.80	15.80	13.80	18.41 (16.41)	16.56 (15.39)	17.18 (15.52)	15.44 (12.98)
group6	vegetarian dishes [g/d]	1.80	6.63	0.87	3.04	1.46 (5.66)	0.63 (2.60)	1.95 (7.01)	0.99 (3.22)
group8	pizza [g/d]	8.37	10.75	6.66	8.82	9.27 (12.49)	6.93 (8.76)	7.97 (9.88)	6.54 (8.85)
group9	cake, cookies [g/d]	65.21	67.12	63.28	60.86	76.71 (75.74)	70.21 (67.98)	60.23 (62.37)	60.02 (56.95)
group11	sweet bread spreads [g/d]	11.92	12.81	12.41	12.70	13.82 (14.20)	13.36 (11.84)	11.10 (12.08)	11.96 (13.06)
group14	canned fruit [g/d]	16.21	22.43	18.50	24.19	18.43 (24.49)	19.07 (23.73)	15.25 (21.41)	18.24 (24.41)
group15	raw vegetables [g/d]	57.85	45.90	58.53	49.29	47.79 (39.10)	47.43 (36.36)	62.20 (47.90)	63.76 (53.55)
group16	cabbage [g/d]	13.63	13.34	13.11	12.45	13.38 (14.09)	11.94 (11.62)	13.74 (13.01)	13.66 (12.80)
group17	cooked vegetables [g/d]	28.45	17.97	28.99	17.74	26.65 (17.63)	26.01 (16.61)	29.23 (18.06)	30.39 (18.09)
group19	mushrooms [g/d]	1.95	2.25	2.04	2.49	1.91 (2.39)	1.84 (2.23)	1.97 (2.18)	2.14 (2.60)
group20	legumes [g/d]	21.24	18.30	22.97	19.92	27.91 (22.90)	28.77 (22.90)	18.35 (15.00)	20.23 (17.72)
group21	potatoes [g/d]	76.67	47.31	80.94	46.09	91.38 (52.45)	87.33 (48.99)	70.30 (43.39)	77.92 (44.36)
group24	low-fat dairy products [g/d]	91.67	181.71	101.48	181.04	72.02 (177.26)	78.57 (174.56)	100.17 (182.96)	112.27 (183.13)
group25	high-fat dairy products [g/d]	112.86	170.44	99.15	153.85	117.92 (187.32)	93.31 (146.92)	110.67 (162.56)	101.91 (157.01)
group26	low-fat cheese [g/d]	5.21	13.00	6.36	15.56	4.11 (12.58)	5.05 (13.40)	5.68 (13.15)	6.97 (16.46)
group27	high-fat cheese [g/d]	29.90	26.41	29.14	25.87	34.08 (30.21)	31.86 (29.97)	28.09 (24.36)	27.85 (23.60)
group28	water [g/d]	403.84	422.15	422.92	425.56	321.01 (372.49)	330.48 (361.80)	439.70 (437.10)	466.46 (446.11)
group31	tea [g/d]	274.86	367.34	253.54	351.86	229.79 (324.97)	206.53 (314.32)	294.38 (382.62)	275.68 (366.32)
group32	fruit juice [g/d]	201.82	226.21	197.49	224.40	197.17 (234.72)	173.59 (207.48)	203.83 (222.41)	208.75 (231.21)
group34	high-energy soft drinks [g/d]	40.83	130.50	44.88	128.93	72.67 (181.74)	76.51 (172.52)	27.04 (97.35)	29.99 (98.83)
group36	wine [g/d]	55.83	88.81	53.08	94.21	52.96 (88.84)	48.45 (100.19)	57.07 (88.78)	55.25 (91.24)
group37	spirits [g/d]	1.86	7.34	2.51	7.60	4.02 (12.23)	5.33 (11.72)	0.92 (3.08)	1.18 (3.86)
group40	margarine	15.05	15.17	15.62	15.68	17.87 (17.62)	19.03 (19.42)	13.82 (13.79)	14.01 (13.28)
group41	other vegetable fat [g/d]	3.53	3.40	3.38	3.44	3.14 (3.15)	3.02 (3.11)	3.71 (3.49)	3.55 (3.57)
group43	sauce [g/d]	12.19	11.39	12.22	11.45	13.73 (12.77)	13.78 (12.60)	11.52 (10.67)	11.48 (10.79)
group45	fish [g/d]	22.79	26.18	23.49	25.31	27.70 (32.02)	27.75 (32.36)	20.66 (22.87)	21.48 (20.92)
group46	poultry [g/d]	12.13	12.32	12.39	10.68	14.48 (14.13)	14.11 (12.15)	11.11 (11.30)	11.58 (9.81)
group47	meat [g/d]	39.12	28.95	41.34	27.16	53.30 (36.89)	51.70 (31.42)	32.99 (22.06)	36.46 (23.38)
group48	processed meat [g/d]	56.72	45.68	60.76	44.39	80.64 (59.87)	80.69 (55.64)	46.36 (32.92)	51.37 (34.18)
group49	soup [g/d]	38.21	36.38	39.92	37.84	42.18 (40.35)	45.72 (44.61)	36.49 (34.38)	37.19 (33.88)

Danksagung

Die vorliegende Arbeit wurde in der Abteilung Epidemiologie am Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke erstellt. Auf diesem Wege möchte ich allen Personen danken, die zum Gelingen meiner Dissertation beigetragen haben.

Besonderer Dank gilt Prof. Dr. Heiner Boeing und Dr. Dagmar Drogan, die mir die Möglichkeit boten diese Arbeit im Rahmen der EPIC-Potsdam Studie anfertigen zu können. Zudem möchte ich beiden danken für die interessanten, wissenschaftlichen Gespräche, die Motivation und Unterstützung! Außerdem möchte ich mich bei Prof. Dr. Rainhard Busse für die Möglichkeit der Promotion an der TU-Berlin bedanken.

Ein großes Dankeschön an Janine Wirt, Anna Flögel und Sven Knüppel für die kritischen Anmerkungen, die dazu beigetragen haben diese Arbeit zu verbessern.

Des Weiteren geht ein großes Dankeschön an Ellen, die Datenmanagerin der Abteilung, an Sven Knüppel und Wolfgang Bernigau für die fundierte Beantwortung meiner statistischen Fragen und an all meinen lieben Bürokolleginnen für die stets freundschaftliche Arbeitsatmosphäre.

Ein besonderer Dank gilt meiner Familie und meinen Freunden, die stets für mich da waren, mir bestmögliche Unterstützung geboten haben und damit maßgeblich zum Gelingen dieser Arbeit beitrugen.