Deutsches Institut für Ernährungsforschung Potsdam-Rehbrücke
Abteilung Molekulare Epidemiologie

Linking whole-grain bread, coffee, and red meat to the risk of type 2 diabetes: Using metabolomics networks to infer potential biological mechanisms

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Epidemiologie"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Clemens Wittenbecher

1. Gutachter: Prof. Dr. Matthias B. Schulze
2. Gutachter: Prof. Dr. Daniel Witte
3. Gutachter: Prof. Dr. Edith Feskens

Potsdam, den 26. April 2017

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-404592
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-404592

Structure

Tables

Figures

Definitions

Abbreviations

ADA	American Diabetes Association
BMI	Body Mass Index
CI	Confidence Intervall
DAG	Directed Acyclic Graph
DALY	Disability-Adjusted Life Years
EPIC	European Prospective Investigation into Cancer and Nutrition
FPR	False Positive Rate
Gln	Glutamine
Gly	Glycine
HbA1c	glycated hemoglobin
HOMA-IR	Homeostasis Model Assessment Insulin Resistance
HR	Hazard Ratio
ICD	International Classification of Disease
Ile	Isoleucine
Leu	Leucine
OGTT	Oral Glucose Tolerance Test
Phe	Phenylalanine
QUICKI	Quantitative Insulin Sensitivity Check Index
TDR	True Discovery Rate
TPR	True Positive Rate
Trp	Tryptophan
Tyr	Tyrosine
WHO	World Health Organization

Abstract

Background: Consumption of whole-grain, coffee, and red meat were consistently related to the risk of developing type 2 diabetes in prospective cohort studies, but potentially underlying biological mechanisms are not well understood. Metabolomics profiles were shown to be sensitive to these dietary exposures, and at the same time to be informative with respect to the risk of type 2 diabetes. Moreover, graphical network-models were demonstrated to reflect the biological processes underlying high-dimensional metabolomics profiles.

Aim: The aim of this study was to infer hypotheses on the biological mechanisms that link consumption of whole-grain bread, coffee, and red meat, respectively, to the risk of developing type 2 diabetes. More specifically, it was aimed to consider network models of amino acid and lipid profiles as potential mediators of these risk-relations.

Study population: Analyses were conducted in the prospective EPIC-Potsdam cohort (n = 27,548), applying a nested case-cohort design (n = 2731, including 692 incident diabetes cases). Habitual diet was assessed with validated semiquantitative food-frequency questionnaires. Concentrations of 126 metabolites (acylcarnitines, phosphatidylcholines, sphingomyelins, amino acids) were determined in baseline-serum samples. Incident type 2 diabetes cases were assed and validated in an active follow-up procedure. The median follow-up time was 6.6 years.

Analytical design: The methodological approach was conceptually based on counterfactual causal inference theory. Observations on the network-encoded conditional independence structure restricted the space of possible causal explanations of observed metabolomics-data patterns. Given basic directionality assumptions (diet affects metabolism; metabolism affects future diabetes incidence), adjustment for a subset of direct neighbours was sufficient to consistently estimate network-independent direct effects. Further model-specification, however, was limited due to missing directionality information on the links between metabolites. Therefore, a multi-model approach was applied to infer the bounds of possible direct effects. All metabolite-exposure links and metabolite-outcome links, respectively, were classified into one of three

categories: direct effect, ambiguous (some models indicated an effect others not), and no-effect.

Cross-sectional and longitudinal relations were evaluated in multivariable-adjusted linear regression and Cox proportional hazard regression models, respectively. Models were comprehensively adjusted for age, sex, body mass index, prevalence of hypertension, dietary and lifestyle factors, and medication.

Results: Consumption of whole-grain bread was related to lower levels of several lipid metabolites with saturated and monounsaturated fatty acids. Coffee was related to lower aromatic and branched-chain amino acids, and had potential effects on the fatty acid profile within lipid classes. Red meat was linked to lower glycine levels and was related to higher circulating concentrations of branched-chain amino acids. In addition, potential marked effects of red meat consumption on the fatty acid composition within the investigated lipid classes were identified.

Moreover, potential beneficial and adverse direct effects of metabolites on type 2 diabetes risk were detected. Aromatic amino acids and lipid metabolites with even-chain saturated (C14-C18) and with specific polyunsaturated fatty acids had adverse effects on type 2 diabetes risk. Glycine, glutamine, and lipid metabolites with monounsaturated fatty acids and with other species of polyunsaturated fatty acids were classified as having direct beneficial effects on type 2 diabetes risk.

Potential mediators of the diet-diabetes links were identified by graphically overlaying this information in network models. Mediation analyses revealed that effects on lipid metabolites could potentially explain about one fourth of the whole-grain bread effect on type 2 diabetes risk; and that effects of coffee and red meat consumption on amino acid and lipid profiles could potentially explain about two thirds of the altered type 2 diabetes risk linked to these dietary exposures.

Conclusion: An algorithm was developed that is capable to integrate single external variables (continuous exposures, survival time) and high-dimensional metabolomics-data in a joint graphical model. Application to the EPIC-Potsdam cohort study revealed that the observed conditional independence patterns were consistent with the a priori mediation hypothesis: Early effects on lipid and amino acid metabolism had the potential to explain large parts of the link between three of the most widely discussed diabetes-related dietary exposures and the risk of developing type 2 diabetes.

Zusammenfassung

Hintergrund: Evidenz aus prospektiven Kohortenstudien belegt, dass der gewohnheitsmäßige Verzehr von Vollkorn, Kaffee und rotem Fleisch mit dem Risiko an Typ 2 Diabetes zu erkranken assoziiert ist. Dieser Risikobeziehung eventuell zugrunde liegende Mechanismen sind allerdings noch weitgehend unklar. Des Weiteren wurde gezeigt, dass Metabolitenprofile im Blut durch die oben genannten Ernährungs-expositionen beeinflusst werden und außerdem in Zusammenhang mit dem Typ 2 Diabetesrisiko stehen. Zusätzlich wurde beschrieben, dass grafische Netzwerkmodelle von Metabolitenprofilen die zugrunde liegenden Stoffwechselprozesse gut abbilden.

Zielstellung: Das Ziel dieser Arbeit war es, Hypothesen bezüglich biologischer Mechanismen zu generieren, die die Assoziationen des Vollkornverzehrs, des Kaffeekonsums und des Fleischverzehrs mit dem Typ 2 Diabetesrisiko erklären könnten. Im speziellen sollten Aminosäure- und Lipidprofile als mögliche Mediatoren des Risikozusammenhangs untersucht werden.

Studienpopulation: Analysen wurden auf Grundlage von Daten aus der prospektiven EPIC-Potsdam Kohortenstudie (n=27,548) durchgeführt, wobei ein Fall-Kohorten-Design verwendet wurde (n=2317, darunter 692 inzidente Typ 2 Diabetesfälle). Ernährungsgewohnheiten wurden mit einem validierten, semiquantitativen Verzehrshäufigkeitsfragebogen erfasst. Die Konzentrationen von 126 Metaboliten (Aminosäuren, Acylcarnitine, Sphingomyeline und Phosphatidylcholine) wurden zur Basiserhebung genommen Blutproben gemessen. Inzidente Typ 2 Diabetesfälle wurden im Rahmen einer aktiven Folgeerhebung detektiert und verifiziert. Die mediane Dauer des berücksichtigten prospektiven Erhebungszeitraums lag für diese Studie bei 6,6 Jahren.

Aufbau der Analysen: Die theoretische Grundlage für den methodischen Ansatz dieser Arbeit bildete die kontrafaktische Theorie der Kausalinferenz. Die in Netzwerken kodierte konditionale Unabhängigkeitsstruktur wurde genutzt, um den Raum möglicher Modelle zu begrenzen, die die beobachteten Zusammenhänge zwischen den Metaboliten erklären könnten. Unter Annahme weniger

grundlegender Effektrichtungen (von der Ernährung auf die Netzwerke gerichtete Effekte; von den Netzwerken auf das Diabetesrisiko gerichtete Effekte) genügt die Adjustierung für eine Teilmenge der direkten Nachbarn im Netzwerk, um netzwerkunabhängige direkte Effekte konsistent zu schätzen. Eine weitere Spezifizierung der Modelle war allerdings aufgrund fehlender Richtungsinformationen zu den Metaboliten-abhängigkeiten nicht möglich. Deshalb wurde ein Multi-Modellierungsansatz gewählt, um die Grenzen möglicher Effekte zu schlussfolgern. Alle möglichen Ernährungs-Metaboliten-Beziehungen und Metaboliten-Typ 2 Diabetesrisiko-Beziehungen wurden dadurch in eine der folgenden drei Kategorien klassifiziert: Direkter Effekt, Unklar, Kein Effekt.

Querschnittsbeziehungen wurden in multivariabel adjustierten linearen Regressionsmodellen untersucht. Longitudinale Zusammenhänge wurden mit Cox-Regressionsmodellen geschätzt. Alle Modelle wurden für Alter, Geschlecht, Body-Mass-Index, prävalente Hypertonie, Ernährungs- und Lebensstilfaktoren und die Einnahme von Medikamenten adjustiert.

Ergebnisse: Der Verzehr von Vollkornbrot stand im Zusammenhang mit niedrigeren Konzentrationen gesättigter und einfach ungesättigter Fettsäuren. Kaffee stand in Beziehung zu niedrigeren Konzentrationen verzweigtkettiger und aromatischer Aminosäuren und hatte potentielle Effekte auf das Fettsäureprofil in den Lipidmetaboliten. Rotes Fleisch zeigte einen Zusammenhang mit niedrigeren Glyzinspiegeln und mit höheren Konzentrationen verzweigtkettiger Aminosäuren. Außerdem stand das Fettsäureprofil in den verschieden Gruppen von Lipidmetaboliten in Zusammenhang mit dem Fleischverzehr.

Des Weiteren wurden potentielle Effekte der Metabolite auf das Typ 2 Diabetesrisiko gefunden. Aromatische Aminosäuren und Lipidmetabolite mit geradzahligen, gesättigten (C14-C16) und mit spezifischen mehrfach ungesättigten Fettsäureseitenketten standen mit einem erhöhten Typ 2 Diabetesrisiko in Beziehung. Glyzin, Glutamin und Lipidmetabolite mit einfach ungesättigten und anderen mehrfach ungesättigten Fettsäureseitenketten zeigten einen günstigen Zusammenhang mit dem Diabetesrisiko.

Mögliche Mediatoren der Beziehung der Ernährungsexpositionen wurden identifiziert, indem diese Informationen in gemeinsamen grafischen Modellen integriert wurden. Mediationsanalysen zeigten, dass

die möglichen Effekte von Vollkornverzehr auf die Lipidmetabolite ungefähr ein Viertel des günstigen Einflusses von Vollkornverzehr auf das Diabetesrisikos erklären könnten. Die möglichen Effekte von Kaffeekonsum und von Fleischverzehr auf Aminosäuren und Lipidmetabolite könnten jeweils ungefähr zwei Drittel der Zusammenhänge mit dem Diabetesrisiko erklären.

Schlussfolgerung: Grundlage für die Ergebnisse dieser Arbeit war die Entwicklung eines Algorithmus, der externe Faktoren (kontinuierlich Expositionsvariablen, Ereigniszeit-Daten) und hochdimensionale Metabolitenprofile in einem gemeinsamen grafischen Modell integriert. Die Anwendung dieses Algorithmus auf Daten aus der EPIC-Potsdam Kohortenstudie hat gezeigt, dass die beobachteten konditionalen Unabhängigkeitsstrukturen mit der *a priori* Mediationshypothese konsistent waren. Der frühe Einfluss auf den Aminosäure- und Lipidstoffwechsel könnte die beobachteten Zusammenhänge zwischen drei wichtigen Ernährungsfaktoren und dem Risiko an Typ 2 Diabetes zu erkranken zu großen Teilen erklären.

1 Introduction

1.1 Overview of the chapter

The following section of this chapter outlines public health issues with diabetes mellitus type 2 to underscore the importance of successful prevention approaches (1.2). In terms of prevention, observational evidence suggests that specific foods are involved in the etiology of diabetes mellitus type 2. Trials on potentially underlying biological mechanisms, however, remain inconclusive so far. According evidence will be subject of the central part of this chapter (1.3). The last section will make a case for systems epidemiology based complex models. Moreover, metabolomics will be introduced as tool that is sensitive to habitual diet and informative with regard to type 2 diabetes risk (1.4).

1.2 Diabetes mellitus type 2

1.2.1 Definition and diagnostic criteria

Type 2 diabetes is acquired relative lack of insulin which is generally based on both, impaired beta cell function and peripheral insulin resistance [4]. In combination, these two pathomechanisms lead to a slowly developing dysregulation of blood glucose levels. Accordingly, current diagnostic criteria for type 2 diabetes of the World Health Organization (WHO) are fasting plasma glucose ≥7.0 mmol/L (126 mg/dL) or plasma glucose ≥11.1mmol/L (200mg/dL) two hours after oral tolerance test (OGTT) with 75g glucose or glycated hemoglobin (HbA1c) ≥6.5% [5]. The American Diabetes Association (ADA) accepts random plasma glucose ≥11.1 mmol/L (200mg/dL) in combination with classic symptoms of hyperglycemia or hyperglycemic crisis as diagnostic criterion [6]. The European Association for the Study of Diabetes (EASD) refers to the WHO-criteria. The current code for type 2 diabetes mellitus in the international classification of disease (ICD) system is ICD-10-GM-2017 E11 [7].

1.2.2 Prevalence

In 2015, an estimated 415 million adults (aged 20-79 years) worldwide lived with diabetes [8]. In Europe, over 90% of these cases were considered as type 2 diabetes [9]. This corresponds to an age-adjusted global diabetes prevalence of 8.8% (95%CI 7.2% - 11.3%), and approximately 8.3% prevalence of type 2 diabetes [8]. Estimations are that among all type 2 diabetes cases in 2015, approximately 193 million persons were not aware of the disease, i.e. they were not medically diagnosed. This is particular worrisome because early diagnosis of type 2 is considered an important step in efficient management of the disease and controlling the risk of complications [10]. Another 318 million adults were estimated to being affected by impaired glucose tolerance in 2015, putting them at high risk of developing overt type 2 diabetes in the near future. If these trends are not drastically mitigated over 640 million people are projected to suffer from diabetes by 2040 [8].

1.2.3 Expenditures

Estimated healthcare expenditures amounted to between 673 billion and 1,197 billion spent on treatment of diabetes and its direct complications in 2015, corresponding to 12% of worldwide total healthcare expenditures [8]. Treatments of diabetic complications were the major driver of this heavy burden on global healthcare systems. There is an ongoing trend of rising healthcare costs of type 2 diabetes. This trend is expected to continue particularly because of population growth and increasing prevalence rates in low- and middle-income countries [8]. Progress and development should hopefully allow better access to healthcare for large population groups. However, large disparities in healthcare spending per diabetes case across global regions prevail. More than 75% of total diabetes cases occurred in low- and middle income countries, yet less than 20% of total healthcare expenditures due to diabetes were spent on these cases [8].

Besides the direct costs of diabetes treatment, indirect costs were attributed to diabetes-related mortality and disability. It was estimated that diabetes caused around 5.0 million deaths in 2015 [8]. In Africa, South and South-East Asia and parts of South America the majority of these deaths occurred in people under the age of sixty [8]. Disability-adjusted life years (DALYs) constitute a composite measure of disease burden capturing both, premature death and prevalence and severity of

disease-related disabilities. For 2010 it was estimated that 47 million (95% CI 41 – 55 millions) DALYs were attributable to diabetes [11].

The financial burden with diabetes also affects individuals. Particularly in countries with less well-developed public healthcare systems the socioeconomic situation of type 2 diabetes patients and their families is often severely compromised [12]. This is due to treatment costs but importantly also due loss of income because of disability or death.

These figures largely rely on extrapolations and have to be interpreted with caution. The data is most reliable for Western Europe and the Northern America. Estimates from other global regions often rely on few studies in particular contexts and the validity of such estimates is at least matter of debate. The general message, however, is clear. Type 2 diabetes is a major public health concern. Due to worldwide population growth and increased lifespan, type 2 diabetes prevalence will continue to rise. This implies large benefits of effective preventive actions on modifiable risk factors on the population level.

1.2.4 Prevention

The sharp increase of global diabetes prevalence over the last decades is partly attributable to a prolonged lifespan [13]. In addition, a changed lifestyle plays a role. Industrialization, urbanization and computerization go along with elevated exposure to particular stressors and pollutants. Furthermore, infrastructural changes favor a sedentary lifestyle. Taken together, these factors are considered environmental driver that contributed to increasing diabetes prevalence over the last decades [14]. Moreover, food production and supply chains have undergone drastic changes on the global scale. This has led to and is currently leading to nutrition transition in most countries of the world. Changes in the common diet are generally characterized by easier accessibility of energy-dense foods and higher intake levels of processed foods [14]. Thereby dietary compositions have changed with general global trends towards higher intake of starch, free sugars, saturated and trans-fatty acids, foods from animal origin, salt and preservatives and lower contents of fiber and phytochemicals, a dietary pattern which has been associated with increased type 2 diabetes incidence in prospective cohort studies [13]. Undoubtedly, diabetes prevention is of top public health priority. Yet many of the outlined global changes that effect diabetes prevalence cannot or should not be reversed. Hence diabetes prevention on the

macro-level needs to be complemented by precise interventions on smaller scales that compensate for potentially adverse environmental influences [15].

Primary prevention of type 2 diabetes on the individual level is possible and cost effective [16]. The Finnish Diabetes Prevention Study demonstrated a sustained reduction of type 2 diabetes incidence by lifestyle intervention in a high-risk group of overweight participants with impaired glucose tolerance [17]. In this trial, 522 men and women were randomly assigned to an intervention group and a control group. The intervention aimed to achieve weight loss of >5%, to limit the intake of saturated fat (<10%), to enhance intake of fiber (>15g per 1000 kcal), and to promote moderate physical activity. The mean duration of the intervention was 3 years. The diabetes incidence was reduced by 58% in the intervention group [17] and this risk reduction was largely sustained after 3 years of follow-up [18]. A similar reduction of diabetes incidence was achieved by a lifestyle intervention focusing on weight loss and physical activity in the American Diabetes Prevention Program [19].

In summary, due to the undebated public health relevance, successful type 2 diabetes prevention programs are of top priority. Evidence-based allocation of resources to particular preventive strategies, however, requires sound knowledge on the factors that drive type 2 diabetes risk. On the population level, quantification of the effects of major risk drivers can help to prioritize interventions on these factors [20]. On the individual level, precise and effective prevention must rely on understanding the causal role that major risk factors play in disease development [21]. To put it simply, manipulation of a risk factor will only affect disease occurrence if the factor plays a causal role in the disease development. Furthermore, effectiveness-estimates for preventive interventions will at most be as accurate as estimates on the causal role of the manipulated factor in the target population.

1.3 Diet and type 2 diabetes etiology

1.3.1 Dietary composition and type 2 diabetes risk

The main topic of this thesis is the impact of dietary composition on diabetes development. The understanding of the impact of specific dietary components on type 2 diabetes development is currently limited

by the fact that consistently observed long-term risk relations are not well explained by the knowledge on short-term metabolic effects of these foods. This will be illustrated by contrasting the evidence from prospective epidemiological studies with evidence from dietary intervention trials.

The Spanish PREDIMED study demonstrated that a complex intervention with a strong focus on dietary composition reduced the diabetes incidence by over 50% in patients with prevalent cardiovascular diseases without leading to substantial weight loss [22]. The PREDIMED dietary intervention promoted high consumption of olive oil, fruits, vegetables, legumes, fish, and sofrito (homemade tomato sauce), reduced intake of total meat (particularly fresh and processed red meat), avoidance of butter, cream, fast food, sweets, pastries, and sugar-sweetened beverages and moderate consumption of red wine [23]. Diabetes prevention by modification of the dietary composition is in line with meta-analyses of prospective cohort studies. Consistent associations of dietary habits with type 2 diabetes incidence have been observed on the nutrient-, the food-, and the dietary pattern level [13,24,25].

In this work, three food groups receive particular consideration as risk factors for type 2 diabetes: habitual consumption of whole-grain, coffee and red meat. These food groups were selected as exposures of interest in the present study because of two main reasons. Firstly, all three were consistently related to diabetes risk in meta-analyses of prospective cohort studies. The aggregated evidence suggests reduced type 2 diabetes risk in relation to high consumption of whole-grain and coffee. High habitual consumption of red meat, in contrary, was associated with elevated risk of type 2 diabetes [13]. Secondly, these three food groups were also linked to type 2 diabetes risk in the study population that provided data for this thesis [26]. For other dietary exposures such as consumption of soft drinks, white rice, dairy products and green leafy vegetables there is evidence for a link to type 2 diabetes risk in other populations [13]. However, they were on average not consumed in relevant amounts at baseline or not clearly linked to diabetes risk in the present study cohort.

1.3.2 Whole-grain and fiber

The proposed protective effect of whole-grain consumption on type 2 diabetes development largely relies on observational studies. The association of dietary fiber intake with type 2 diabetes risk was evaluated

in a large meta-analysis of sixteen prospective cohorts with a total of 572,665 participants including 36,578 incident type 2 diabetes cases [27]. This meta-analysis included data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort study which was the primary data-source for the present study. In a dose-response analysis, additional 10g total fiber intake per day was associated with a 9% lower diabetes risk (95%CI 0.87 – 0.96). If analyses were restricted to cereal fiber (including fiber from whole-grain bread) the relative diabetes risk was reduced by 25%per 10g additional daily intake (95%CI 0.65-0.86) [27]. The focus on fiber in these studies implicates the hypothesis that the fiber content is primarily responsible for the health-benefits by whole-grain consumption. It was argued, however, that the favorable metabolic impact of whole-grain cereals might likely go beyond the isolated effect of fiber. Probably other bioactive secondary compounds in the endosperm of grains play a synergistic role [28]. However, meta-analyzes of prospective cohort studies primarily from the U.S. showed a beneficial association of whole-grain intake with reduced type 2 diabetes risk [29]. The extensively validated German Diabetes Risk Score [30,31] that was derived from the EPIC-Potsdam cohort yielded a hazard ratio of 0.92 (95%CI 0.85 – 0.99) associated with 50g (≈ one slice) higher intake of whole-grain bread per day [30]. Thus, evidence from observational studies suggests a diabetes-protective effect of high habitual whole-grain consumption over years.

Dietary intervention studies tested the effect of daily consumption of fiber or whole-grain products in course of weeks to months on parameters of glucose metabolism. Taken together these studies did not indicate a direct effect of whole-grain intake on glucose disposal and insulin sensitivity [32-36]. In particular the two largest randomized trials did not observe an effect of whole-grain intake on markers of glucose homeostasis. In a parallel-arms intervention trial in the U.S. 266 participants were randomized to one of three groups. In two arms whole-grain foods were supplied in different amounts. The third arm served as control group. No effect of whole-grain content of the diet on insulin-sensitivity was detected by the Quantitative Insulin Sensitivity Check Index (QUICKI) method [33]. In a study of comparable size (n=206) in Great Britain participants were similarly randomized to one of three study arms [36]. The two intervention groups received whole-grain products in similar amounts over 12 weeks. The diets between the two intervention groups, however, differed in the

oat content of the whole-grain products. The third control group was advised to restrict their fiber intake. No significant differences in fasting blood glucose concentrations and Homeostasis Model Assessment Insulin Resistance (HOMA-IR) were observed. Still significantly lower blood pressure was detected after the intervention [36]. Relatively short-term effect of whole-grain intake on glucose metabolism was reported in context of a weight loss intervention [37]. A group of 31 obese participants received hypocaloric diets to induce weight loss. In one of two groups the intervention diet was supplemented with powdered, double-fermented whole-grain. Improvement in HOMA-IR was significantly better in the whole-grain intervention arm after adjusting for weight loss [37]. These results, however, are probably less relevant for the general population because of the caloric restriction and the unusual form of whole-grain administration.

Interestingly, the notion of null findings on a potential whole-grain effect on glucose metabolism does not apply to interventions in patients with prevalent type 2 diabetes. A meta-analysis of randomized controlled trials from 2013 concluded that evidence from 13 trials indicated beneficial effects of increased dietary fiber intake over at least 8 weeks on HbA1c levels and fasting blood concentration in diabetic patients. Fiber-rich dietary interventions reduced on average absolute HbA1c values by 0.55%, and lowered blood glucose concentrations by 10mg/dL [38]. Recent trials are in line with these findings. A recent Chinese intervention trial further corroborated beneficial effects of high dietary fiber intake on blood sugar homeostasis in type 2 diabetes patients [39]. Participants were randomized to one of four groups, either control or different levels of fiber intake (30g to 100g). Beneficial metabolic effects of the whole-grain content of the intervention diets followed a dose-response relation and comprised sustained improvements in insulin sensitivity, HbA1c, and blood lipids [39]. It should be noted that high whole-grain intake was part of several complex lifestyle interventions that successfully reduced diabetes incidence [18,40] but it is difficult to trace back the effect to the single components of the intervention.

To summarize, observational studies suggest a protective effect of dietary fiber intake on type 2 diabetes. Moreover, interventions with fiber-rich diets showed the potential to ameliorate disturbed blood glucose homeostasis in type 2 diabetes patients. Whole-grain interventions over several weeks in participants without prevalent type 2

diabetes, however, did not result in improved insulin sensitivity or lower blood sugar concentrations.

1.3.3 Coffee

A potential role for high coffee consumption in lowering type 2 diabetes risk was likewise raised by observations in prospective cohort studies. A recent meta-analysis of prospective cohort studies on total, caffeinated and decaffeinated coffee identified 28 studies with a total of 1,109,272 participants and median follow-up time of 11 years including 45,335 incident type 2 diabetes cases [41]. Included studies were from the U.S. (13 studies), Europe (11 studies) and Asia (4 studies) and the overall quality of the evidence was high (median Newcastle-Ottawa Scale-sore: 7 of 9 possible points). The pooled analyses of the evidence indicated a strong inverse association between total coffee consumption and the risk of developing type 2 diabetes. A non-linear dose-response curve indicated for example a 25% reduction in the relative risk of developing type 2 diabetes in persons with consumption of 4 cups coffee per day compared to non-consumers. The dose-response relation between coffee consumption and reduced diabetes risk seemed not to rely on caffeine content and was relatively independent of the covariables included as potential confounders [41].

Observed long-term beneficial associations between high coffee consumption and low diabetes incidence do not easily match with results from coffee-interventions in the range of several weeks in non-diabetic participants. Result from different randomized trials are somewhat inconsistent with reports of moderate improvements in glucose disposal after several weeks of frequent coffee consumption [42] but null results on glucose load after OGTT in other studies [43,44]. Further coffee interventions with intermediate duration (up to 8 weeks) did not detect beneficial effects on glucose homeostasis but reported beneficial effects on markers of liver metabolism [45] and subacute inflammation [46]. Somewhat puzzling with respect to a proposed protective role against diabetes is the established deteriorating effect of acute caffeine challenge on insulin sensitivity [47] which seems to be robust against habituation.

In summary, evidence from observational studies suggests a beneficial effect of high habitual coffee consumption on the risk of developing type 2 diabetes. Intervention trials show that this potential diabetes-protective role of coffee cannot be attributed to the acute effect of coffee intake on insulin sensitivity. Also coffee intake of intermediate

duration seems not to have direct effects on markers of glucose homeostasis in non-diabetic participants.

1.3.4 Processed and unprocessed red meat

A potential adverse effect of unprocessed and processed red meat on type 2 diabetes risk was again inferred from observations in prospective cohort studies. Across these studies processed (red) meat consumption was consistently associated with elevated type 2 diabetes risk. In a pooled analysis of three large prospective cohort studies from the U.S. with a total of 204,157 participants including 13,759 incident diabetes cases 32% higher risk of type 2 diabetes per additional daily serving of processed red meat (HR 1.32, 95%CI 1.25 – 1.40) was estimated [48]. Portion-sizes for processed meat used in this study ranged from 28g (bacon) to 45g (e.g. hot dogs or hamburgers). The EPIC-Interact study provided the largest European prospective analyses on meat intake and diabetes risk with 11,559 incident diabetes cases and 14,529 non-cases. In this study, 50g increments in daily processed meat intake were associated with 12% higher risk of type 2 diabetes (HR 1.12, 95%CI 1.05 - 1.13) [49]. Habitual intake of unprocessed red meat showed an equally directed but somewhat weaker association with diabetes risk. Additionally, the two types of meat exposure were collapsed into a single food group (for convenience referred to as red meat for the remainder of this work). Risk estimates for the pooled analysis in the U.S. cohorts were 1.12 (95%CI 1.08 - 1.16) and 1.14 (95%CI 1.10 - 1.18) per daily serving (85g) of unprocessed red meat and total red meat [48], respectively, and 1.08 (95%CI 1.03 – 1.13) and 1.09 (95%CI 1.05 – 1.13) per 50g increments of unprocessed red meat and total red meat, respectively, in the EPIC-Interact study [49]. These findings were in line with meta-analyses that relied on a larger range of studies including data from Asia [48,50-52]. Additional analyses prospective cohort studies including substitution modeling [48], analysis of change [53], and mediation analysis [54] further supported the association of high meat intake and development of type 2 diabetes.

This observational evidence is not clearly linked to results from randomized trials that investigated metabolic effects of red meat dietary interventions. Randomized intervention studies in the range of several weeks that used lean red meat as major protein source compared to other animal protein sources reported null findings on an effect of animal protein source on glucose homeostasis [55-57]. In contrary, a

randomized cross-over trial that compared strictly controlled animal protein-based diet with plant protein-based diet each over four weeks in fifteen postmenopausal non-diabetic but at risk women found improved glucose homeostasis in the non-meat group [58]. In addition a randomized crossover-trial in 25 young, iron-deficient women compared diets each over 8 weeks that only differed in red meat vs. oily fish content [59]. Improved insulin sensitivity was found in response to the oily fish dietary intervention [59].

Randomized trials in patients with type 2 diabetes indicated that diets favoring plant protein over animal protein had beneficial effects on glucose control [60]. Which role red meat in particular played for these findings, however, remains speculative. A meta-analysis of randomized trials in diabetic patients with dietary interventions over 4 - 8 weeks found moderate improvements in HbA1c (-0.15%; 7 studies, 149 participants), fasting glucose (-0.53 mmol/L; 8 studies, 197 participants), and fasting insulin (-10.09 pmol/L; 5 studies, 118 participants) in plant-protein groups compared with animal-protein diet control groups. Animal protein diets generally included high amounts of red meat. The authors, however, claimed the need of larger well conducted trials because of heterogeneity of results and suggestive evidence for publication bias [60]. A long-term randomized trial compared soy protein-based with animal protein-based dietary interventions. In this trial 41 type 2 diabetes patients with nephropathy were randomized to one of two parallel study arms. Fasting plasma glucose was significantly improved the soy-protein group after four years (mean change -18 ± 3mg/dL in soy protein vs. +11 ± 2mg/dL in the control group) along with improvements in markers of cardiovascular health and renal function [61].

To summarize, habitual red meat consumption was consistently related to an elevated type 2 diabetes risk in prospective cohort studies. Results from randomized trials that tested red meat-rich intervention diets are not consistent with some showing improvements in markers of glucose homeostasis and others not. Interpretation with respect to the observational evidence is further complicated by the fact that intervention studies mostly focused on red meat as protein source. Several trials actively aimed to avoid effects of altered lipid composition on glucose homeostasis by choosing lean meat cuts [62,63]. Also, the control diet in trials including red meat-rich intervention diets was fairly

heterogeneous, ranging from manufactured soy-protein, over white meat and lean fish, to oily fish.

Taken together, evidence from prospective cohorts clearly points towards a relation of the habitual consumption of whole-grain bread, coffee, and red meat with type 2 diabetes risk. Intervention studies, however, do not indicate a generalizable short-term effect of these dietary components on markers of glucose homeostasis. Interpretation of the observational findings should be concerned with potential sources of bias such as residual confounding. The interpretation of results from dietary intervention trials with respect to population based observations, however, might be compromised by the design of the trials. The intervention studies, for example, often target high-risk study groups; complexity of dietary composition and compensatory dietary behavior complicate definition, administration and monitoring of intervention and control treatments; the choice of surrogate markers for hard endpoints is often debatable; and the duration of the trials seems often not to clearly rely on biological reasons. Foods are complex exposures and type 2 diabetes is a slowly developing disease preceded by several stages of disturbed metabolic conditions. Models that can capture the obvious complexity of the relation between long-term dietary habits and type 2 diabetes risk to a certain degree might help to better integrate observational and interventional evidence on the topic. Beyond that complex models of dietary effects on type diabetes risk could also help to identify gaps in the available evidence and thereby inform aims and design of future studies.

1.4 Systems epidemiology

1.4.1 Complex systems

This work is about the use of complex observations on molecular phenotypes under real-life conditions to generate biological hypotheses on biological mechanisms that could link diet to type 2 diabetes incidence. Conceptually this approach was referred to as systems epidemiology [64,65]. It was argued that metabolomics can be a key tool to uncover the biology underlying observed diet-disease links [65-67]. The combination of sensitivity to dietary exposures and relevance for

type 2 diabetes incidence qualifies metabolomics as a tool to investigate the mechanisms that might link these two entities.

An early attempt to formalize a general perspective on systems theory with a focus on biology was the *General Systems Theory* by Ludwig van Bertalanffy (1909-1972), which was developed between World Wars I and II [68]. Today contributions to the understanding of systems on a general level come mainly from the field of *complex systems sciences* [69]. While important distinctions were made [70] the basic understanding that systems behavior arises from specification of the interactions between its parts remains. Both approaches share that abstract models, at best mathematically formalized, are at the core of understanding and possibly modifying complex systems behavior [71].

Regarding living organisms, systems approaches emphasize that characteristic organizational principles differentiate between living systems and the non-living world [72]. As for any system, the organization of living systems cannot be deduced from its parts (*antireductionism*). The structure of living systems is primarily determined by self-maintaining interactions of their components. Only to a relatively low degree it is determined by the environment, a principle that was labeled *autopoiesis* [73]. The molecular composition of an average biological cell, e.g., is completely renewed about 10^4 times over its life-cycles but the inherent organization remains stable [74]. Thus living systems constantly incorporate information and material from the environment but as soon as they are incorporated these parts are integrated in the organizational structure of the organism. Amongst other implications, this shifts the attention from states and molecules to relations and dynamics.

Studying complex systems needs appropriate tools. Mathematical models play a key role in this regard [71]. Emphasis on complexity and wholeness of systems should not obscure, however, that translating observations into models necessarily involves reductionism. Limitations on our ability to picture complexity are imposed by limited capacities to measure, process and interpret information. Network models are suitable tools to provide guidance in focusing on the relevant information [75,76]. In systems biology, networks are used to break systems into modular subsystems [77] and to identify a manageable number of key components of a system that can be used to approximate the systems internal state as a whole [78,79].

Applications of network approaches in epidemiology have received considerable attention (e.g. [80,81]) and various examples of metabolomics networks derived from blood screenings in observational human cohorts are available [82-84]. It should be noted that the use of certain tools is not what is perceived as a *systems* approach in this work. The working definition of a *systems perspective* on epidemiological observations used in this work is one that integrates information from various levels according to a model that reflects complex interrelations between humans' internal states and the environmental challenges they are exposed to. Moreover, observations on the key variables of the model should enable predictions or explanations on behavior of the system as a whole. Behavior in this sense includes the transition from the healthy into the diseased state in longitudinal studies in the bio-medical field.

1.4.2 Metabolomics

The human metabolism is a subsystem of the human organism. Regulated interactions between its components (i.e. metabolites) maintain an elaborate equilibrium on a systemic level. On a molecular level systemic feedback e.g. via endocrine or neuronal processes regulates enzyme and transport activities. Moreover metabolic states interact with behavioral states. On the one hand metabolic activities provide the fuel for physical and cognitive activities and deliver building blocks for morphological maintenance and adaptation. On the other hand behavior is organized to compensate consumption of metabolic substrates and ameliorate disturbances of metabolic equilibrium. Thus metabolic activities are sensitive to environmental challenges. Particularly diet plays a major role because it supplies substrates for major metabolic processes. A long-term unbalanced diet has the potential to disturb the metabolic balance. Type 2 diabetes is classified as a metabolic disease. The physiopathology of type 2 diabetes can be summarized as an inability to maintain the metabolic equilibrium that controls physiological systemic glucose levels in the healthy state. Thus the constant metabolic impact of the habitual diet can be viewed as an intermediary step in the effect it exerts on diabetes development.

Metabolomics approaches can be regarded as snapshots of the metabolic state [85]. These snapshots can be narrow or wide (e.g. targeted vs. untargeted approaches), the picture can have different resolutions (e.g. qualitative vs. quantitative approaches), and it can be

taken on different locations (e.g. blood vs. other tissue samples), with a different focus (e.g. water-soluble vs. lipophilic compounds) and in different frequencies (e.g. single time-points vs. time series). Technically, metabolomics rely on coupling liquid or gas chromatography with mass spectrometry [86]. Definitions of metabolomics often refer to the approach as to assess all small molecules (small molecules <1500 Da) present in a biological sample (cell, tissue or organ) at a specific time-point [85]. This, however, is a too optimistic definition for metabolomics applications to human studies that typically assess hundred(s) of metabolites. Even though the precise number of metabolites in human blood is still unknown, estimates based on untargeted screenings surpass the ten thousand [86]. Progress of analytical chemistry and combination of different analytical approaches constantly expands the number of metabolites that are detectable [86]. Due to the highly dynamic nature and the compartmentalization of the metabolome, however, any measurement will still be limited to a very limited reflection of the complexity of the underlying metabolic processes. Metabolomics applications in human cohort studies, for example, almost exclusively rely on blood or urine samples from single time-points. Therefore metabolomics applications in human cohorts might be appropriately defined as a subfield of analytical chemistry that aims to simultaneously measure relatively broad spectra of small molecules (metabolites) in biological samples [28].

Dietary composition and metabolomics

Metabolomics applications to nutrition sciences can be subdivided as follows. First, metabolomics approaches were used as a novel dietary assessment tool to surveille the compliance to nutritional interventions and for discovery of biomarkers of intake of single foods or adherence to dietary patterns in observational studies. Though promising results were achieved on specific foods [87,88], food group [88-90], and dietary pattern [91-94] level, such an approach is generally viewed as particularly useful when complemented by traditional dietary assessment tools [95]. In other human nutrition studies, metabolomics was applied to elucidate the impact of diet on metabolic processes [96-98]. The two motivations differ important in some regards. Biomarkers of dietary intake need to be specific for the exposure and should reflect the time under exposure. At best they should allow estimating quantitatively the amount of exposure which is hindered by the interindividual variation of metabolic processes

[99]. For etiological research sensitivity of metabolite levels to dietary challenges is the major concerns and this is where the two motivations converge.

The general notion on sensitivity of the metabolome to dietary challenges extends to evidence for an association of metabolic profiles with the particular foods considered as exposure in this study. Whole-grain bread consumption was associated with alterations in the serum concentrations of phosphatidylcholines and acylcarnitines in previous analyses in EPIC-Potsdam [83,100]. This observation was supported by a cross-over dietary challenge trial in 15 healthy participants. In this trial whole-grain had an acute effect on the circulating concentrations of several amino acids, phosphatidylcholines, and lysophosphatidylcholines [101]. Moreover, a cross-over trial in thirty three postmenopausal women from Finland demonstrated that an intervention with rye bread (≥20% of the total energy) over eight weeks lowered branched-chain amino acid levels compared to refined grain bread over eight weeks [102].

Most studies investigating metabolomics markers of coffee consumption focused on coffee-derived secondary compounds and their metabolites [87,103-105]. Still, previous analyzes in the EPIC-Potsdam study, however, detected associations between coffee consumption and serum concentrations of sphingomyelins, phosphatidylcholines and amino acids [83,100,106]. Other observational cohort studies also reported associations of habitual coffee consumption with acylcarnitines [107,108], sphingomyelins [108], and lysophosphatidylcholines [109].

Several observational and interventional studies indicate an effect of red meat consumption on lipid and amino acid metabolism. Previous findings in EPIC-Potsdam comprise red meat-related alterations in the sphingomyelin- and glycerophospholipid-compartments as well as in the amino acid profiles [54,100]. These findings were in agreement with analyzes in the EPIC-Oxford cohort that compared metabolomic profiles between vegan and vegetarians vs. omnivores [110,111]. Compared to non-meat consumers, acylcarnitines, glycerophospholipids and sphingolipids were markedly elevated [110] and amino acid profiles were altered in meat eaters [111]. Several intervention studies showed alterations of lipid [90] and amino acid profiles [112-114].

Metabolomics and type 2 diabetes risk

Beyond sensitivity to dietary habits, the metabolomics data used in this study are informative with respect to type 2 diabetes risk. Previous

studies in EPIC-Potsdam showed that accurate prediction of type 2 diabetes incidence was possible based on circulating concentrations of amino acids, glycerophospholipids, sphingomyelins and acylcarnitines and these findings were replicated in an independent cohort [115,116]. A recent meta-analysis of six to nine (number depending on the amino acid in question) prospective cohort studies from Europe, the U.S. and Asia evaluated the longitudinal association of metabolomics-assessed amino acids with type 2 diabetes incidence. The study comprised data from up to 8000 participants including 1940 incident cases of type 2 diabetes [117]. Blood concentrations of branched-chain amino acids (valine, leucine, and isoleucine) and aromatic amino acids were associated with elevated type 2 diabetes risk. The circulating concentrations of glycine and glutamine in contrary were associated with reduced risk of type 2 diabetes. Lipid metabolites were evaluated with respect to type 2 diabetes risk in at least 18 cohorts with numerous significant findings [117]. Due to large variety of lipid metabolites and less harmonized measurements, however, across study comparisons are more complicated. Significant findings on acylcarnitines and type 2 diabetes risk were reported from several cohorts [118,119]. Besides aforementioned replicated associations with type 2 diabetes incidence in EPIC-Potsdam the relevance of glycerophospholipids for type 2 diabetes development was indirectly illustrated by analyzes of the fatty acid profile in this lipid compartment [120,121]. Profiles of saturated and polyunsaturated fatty acids glycerophospholipids were strongly associated with type 2 diabetes risk in large scale studies in EPIC-Interact [120,121]. Circulating concentrations of specific sphingomyelins have been associated with a prediabetic insulin resistant state [122] and were strongly affected several months after bariatric surgery in morbidly obese patients in remission of type 2 diabetes [123].

Taken together the available evidence indicates that circulating concentrations of amino acids, acylcarnitines, sphingolipids and glycerophospholipids are sensitive to dietary exposures, in particular to consumption of whole-grain, coffee, and unprocessed and processed red meat. This conclusion relies on evidence from observational cohorts as well as intervention studies in humans. In addition, metabolomics applications in prospective cohort studies detected associations of circulating concentrations of amino acids, acylcarnitines, sphingolipids and glycerophospholipids with type 2 diabetes risk which implicates

amino acid and lipid metabolism in early biological processes that predispose for development of type 2 diabetes. Therefore, lipid- and amino acid-focused metabolomics seems to be applicable as sensors for both, dietary exposures and early diabetes-relevant metabolic alterations.

Modeling metabolomics data

From a modeling perspective metabolomics data pose several challenges. The high intercorrelation between metabolites is biologically informative. To analyze and to communicate the information content of the correlation structure, however, is not trivial. Besides classical data reduction approaches such as principal component analysis, systems biology provides useful graphical tools [124]. Metabolomics data-driven network models have been demonstrated to correspond well with knowledge-based charts of human metabolism [84]. Beyond that network-models have elucidated unknown enzymatic links and have helped to chemically identify unknown metabolites [125].

Conceptual ambiguities remain as to analyzing external variables (such as diet or disease risk) in relation to metabolomics data. The strong intercorrelation between metabolites data corresponds to mechanistical links in metabolic pathways. In other words, most metabolites are sensitive to the levels of several other metabolites. From a methodological perspective this implies severe concerns with the role of confounding. The fact that single metabolite levels commonly integrate information from several pathways needs to be considered. In analyses that aim to interpret metabolites as markers for biological meaningful activities the influence of other pathways should be controlled. Therefore, etiological studies on the metabolites and external variables might better take into account the correlation structure of metabolomics data. Over the past decades epidemiological methods were refined to formally approach the issue of confounding and other sources of bias in complex systems of interdependent factors.

1.5 Causal Inference

1.5.1 Counterfactual thinking

There is an intriguing overlap between network-models of complex systems and the graphical approaches used in counterfactual-based causal inference from observational data. This overlap forms the

methodological basis of this work. Therefore a brief introduction to the counterfactual concept of causality and to the relevance causal inference theory for the analysis of observational data will occupy the remainder of this chapter.

Summarizing or commenting the philosophical debate on concepts of causality is beyond the scope of this work. As a general note, counterfactual thinking is one of several partly competing approaches to define causality. Philosophical debates on the nature of causality date back to ancient Greece and a universal definition of causation is still not agreed upon [126]. The counterfactual theory of causation is one of several relevant theories of causation, and prominent contributions to its formalization were made by David Lewis [127,128]. Basically a cause is defined as a factor that would change the occurrence of the effect in some respect if everything else in the universe remained the same but the cause (contrary to the facts) had been different [127]. A formal analysis of causal question according to this concept is not as trivial as it sounds. Causal questions need to be stated in the form of *would outcome E have been different (say E*) if at some point in the past the state of the potential cause C was manipulated (say to the state C*).* Judea Pearl and others (see [129] for comprehensive citations), however, formalized a well-defined mathematical framework to analyze causal claims in observational settings according counterfactual-based causal models.

1.5.2 Observing causal relations?

Methodological literature on modern epidemiology has pinpointed that for biomedical research in observational settings the focus must lay on estimating effects rather than statistical hypothesis testing [130,131]. Apart from evaluating randomness as potential explanation for observed data patterns, epidemiological modeling has to consider structural sources of biases [132]. In particular, issues of effect directionality including concepts of confounding [133], mediation [21,134], interaction [135] and collider bias [129] need to be addressed in the model building process. This implies the epidemiological model being generally dependent on prior knowledge and assumptions on the nature of the modeled relation. The model must also consider potential sensitivity of the relation of interest to other influential factors and the degree to which information on these factors is available.

In case prior knowledge suffices to unambiguously specify a causal model and sufficient information on the relevant variables is

available, estimates from that model can be given a causal interpretation [129]. The theory to specify assumptions and to identify causal models in complex systems of interrelated variables has been developed in the causal inference literature over the last decades. Apart from well-defined exposure and outcome, analysis of causal effects requires consideration of the direction of effects between the modeled variables. In this regard directed acyclic graphs (DAGs) are a key-tool. DAGs are graphical models that specify directed links between variables and can be viewed as visualization of structured systems of functional dependencies. Therefore translation of DAGs into algebraic models is straightforward and the corresponding framework is well-defined [129]. Examples of applications of causal modeling in epidemiological studies to date were mostly carried out in the restricted potential outcome framework [136-138] which defines exposures as (theoretically) feasible interventions [139]. There is no notion on human feasibility in counterfactual concepts per se. Therefore the original contributions on counterfactual-based structural models by Pearl are in principle not restricted to scenarios where a human intervention is theoretically possible [140].

Counterfactual-based restricted potential outcome models are arguably the best developed formal framework to derive quantitative estimands on causal effects in settings where they apply [139]. They cannot substitute, however, alternative approaches to evaluate causality of observed relationships because of their rather narrow conceptualization of well-defined exposures and outcomes [141]. To pose one example, social movements have changed health inequities across social classes and must therefore be considered important determinants (or causes) of population health [142]. The population-health effects of such movements, however, can only be understood with respect to the interaction with the social environment in the specific historical context. As illustrated by this example, important epidemiological problems might resist formal analysis in restricted potential outcome models due to their wide-spread and context-specific effects and cyclic nature.

Another major obstacle for counterfactual-based causal inference in epidemiological studies is that prior knowledge is seldom (if ever) complete in the sense that it suffices to define a singular causal model. Counterfactual-based causal quantities inferred from observations are always derived, however, with reference to a specific causal model. It is important to keep in mind that causal quantities are only as valid as the

underlying model is. Therefore, counterfactual models depend on deep subject-matter knowledge. In this respect, inference to the best explanation [143] can inform a working model, which in turn can be tested for consistency with observational data. Accordingly, Hills famous viewpoints [144] are still of use, for example by evaluating consistency of association over different source populations or coherence of results from different study types. With the concept of *triangulation,* the latter has recently received a more formal definition in epidemiology. Triangulation puts a focus on making use of the different sources of bias in different study designs [141]. Ideally, the systematic combination of designs mutually excludes all relevant sources of bias and leaves a causal relation as only explanation of associations coherently observed across studies of complementary designs [141].

Still another approach to handle incomplete knowledge on the data-generating mechanisms is to specify the family of causal models that are possible given the prior knowledge and the observations. In other words the space of possible causal relations underlying the observations between two variables is restricted although no single model is defined. Inference is then based on exploring this space in a multi-model procedure rather than on one single presumably correct estimate [145]. This approach was taken in the present work.

1.6 Summary

Increasing prevalence of type 2 diabetes is a major public health concern on the global scale. Public health actions to prevent type 2 diabetes are justified by the tremendous costs of the disease and they are encouraged by the notion that effective diabetes prevention is possible. It is generally accepted that the habitual diet is an important determinant of type 2 diabetes risk. If it comes to the level of specific foods, however, the biological mechanisms that might generate the observed relations with type 2 diabetes risk remain largely speculative. This thesis propagates to integrate metabolomics-based network-models in the analysis of observed links between habitual consumption of specific foods and future type 2 diabetes risk. The aim is to generate hypothesis on the underlying biological mechanisms that are consistent with the observed data patterns.

2 Hypothesis and study question

In the following rational (2.1) and hypothesis (2.2) of the study are summarized and operationalized into primary and secondary aims (2.3).

2.1 Rational

Effective type 2 diabetes prevention strategies are urgently needed. Based on observational evidence and results of complex lifestyle interventions habitual diet is considered a prime target for preventive interventions. The biological role that specific foods paly in predisposing for type 2 diabetes development is not well understood. Elucidation of the biological mechanisms that link dietary habits to type 2 diabetes development will enable evidence-based nutrition interventions to prevent and treat type 2 diabetes.

2.2 Hypothesis

This work relies on the assumption that habitual consumption of specific foods (whole-grain bread, coffee, and unprocessed and processed red meat) causally affects the risk of developing type 2 diabetes. It is further hypothesized that this effect on type 2 diabetes risk is (partially) mediated by prolonged metabolic alterations due to chronic exposure to the respective foods. The study hypothesis is summarized in Figure 1.

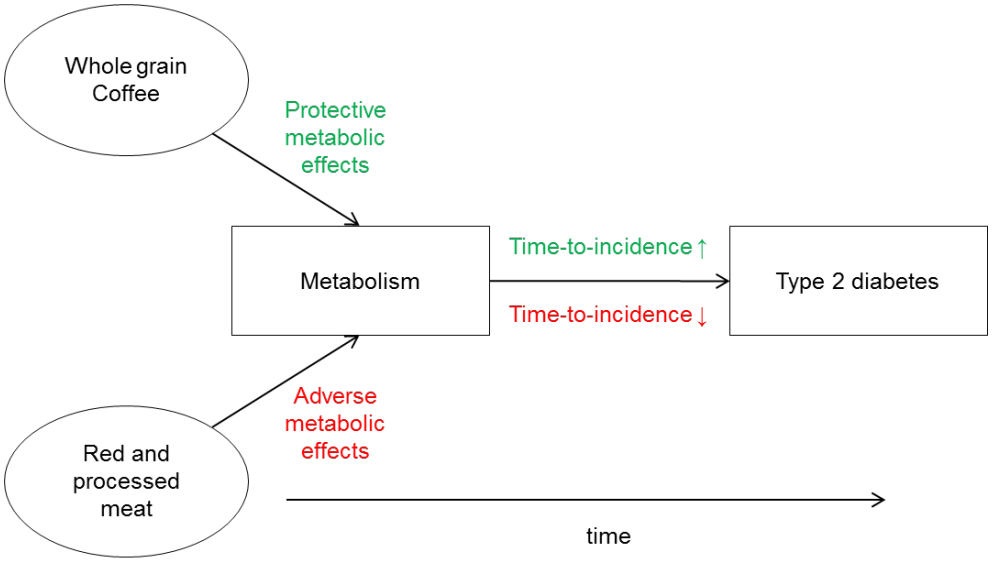


Figure 1: Study hypothesis

2.3 Aims

The primary aim of this study was the identification of potential metabolic links between exposure to dietary risk factors and latter type 2 diabetes incidence based on metabolomics networks. This global aim included following steps:

1. To evaluate the common variance among metabolically closely related metabolites in relation to both, exposure to dietary risk factors and type 2 diabetes incidence.
2. To identify effects of habitual consumption of diabetes-related foods on the metabolomics network.
3. To identify effects of the metabolomics network on type 2 diabetes risk.
4. To evaluate paths in the metabolomics network as potential mediators of an effect of specific foods on type 2 diabetes risk.

A secondary study aim was to develop an algorithm to link external variables (food consumption, time-to-diabetes incidence) to data-driven network models of metabolism. This was a precondition to achieve the primary aim but the developed method might be of use beyond its application to this work. This included evaluation of key assumptions and tools in simulation studies before building the developed method upon them.

3 Data sources and methods

3.1 Overview of the chapter

This chapter will give a brief introduction into causal inference terminology (3.2), describe simulation studies and the EPIC-Potsdam cohort study as data sources (3.2), and lay out the statistical design including methodological developments of this work (3.4).

3.2 Excurse: Causal inference theory

Counterfactual-based causal inference theory is the conceptual foundation of this work. The analytical design relied on the well-defined relation between graphical causal models and conditional independence structures of multivariate distributions. In a supplemental chapter the formal definitions of *causal model* and *causal effect* and the criteria for *effect identifiability* from observational data are provided, from which the methodological approach taken in the present study was logically deduced. In the following, basic concepts and terminology of graphical causal models will be outlined in an illustrative manner.

To begin with, consider the simplest causal claim: X causes Y. Expressing this causal claim in a causal graph is straightforward:

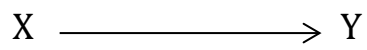


Figure 2: Simple causal graph
with two variables and one effect

We say that X has a direct effect on Y, thus they are connected by an arrow emanating from X and pointing into Y. The variable X is called a parent of Y. The variable Y is called a child of X. It should be noted that this simple causal relationship cannot be expressed in standard statistical notation. X = Y implies Y = X. Statistical relations describe coincidences whereas causal models define mechanistical dependencies.

To put an example: Rain causes wet streets (rain → wet streets). According to this causal claim we would expect to always observe wet streets in relation to rainfall. Observation of rain without wet streets would require another causal explanation of how the mechanism was

blocked. (Maybe the street runs under a bridge.) An alternative explanation would be a misspecified causal model. This would be hardly considered an explanation in the example. But in most research contexts the subject-matter knowledge is not as sound. Commonly, we have to deal with a certain degree of uncertainty about the causal model.

Now consider two additional factors, a parental variable P which has a causal effect on both, X and Y, and a descending variable D which is causally affected by both, X and Y. This is expressed in the following graph:

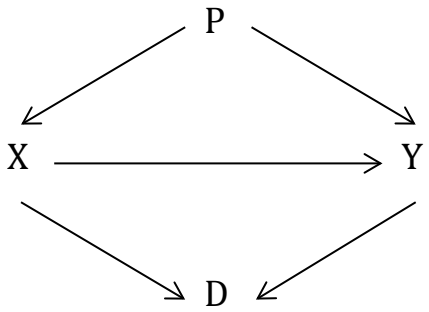


Figure 3: Causal graph with a main effect, a confounder, and a collider

With regard to the causal effect of X on Y, the variable P (which is parent of both, X and Y) is a confounder. It constructs a relationship between X and Y that does not correspond to a causal path between the two. In the contrary, P is the causal factor. If the aim was to observe the causal effect between X any Y we would need to control for the levels of P. In an observational study such control can be exhibited for example by stratification according to P or most commonly by adjusting of a regression X on Y for P. The descendent variable D, however, is a collider with respect to the causal effect of X on Y. Controlling for a common consequence, i.e. a collider, introduces bias. Adjusting for a collider should be avoided.

Last consider a mediating variable M, which is affected by X and in turn has an effect on Y:

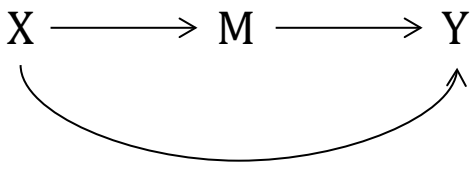


Figure 4: Causal graph including a direct effect and a mediated effect

We say that M is a mediator of the causal effect of X on Y. The total effect of X on Y is thus constituted of two components: an indirect or mediated proportion which can be explained by the effect of X on M; and a proportion which is independent of the effect of X on M. The former is called *indirect effect*, the latter *direct effect*, and the sum of the two is the *total effect* of X on Y.

Let us now put together these variable types in a common causal diagram:

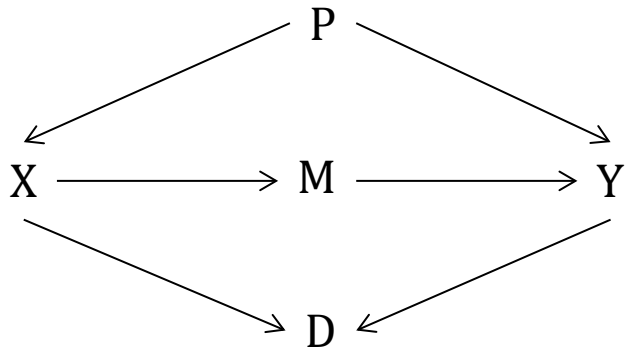


Figure 5: Directed acyclic graph with a mediator, a collider, and a confounder

This causal diagram is called a *directed acyclic graph* or shortly *DAG*. Directed because the arrows specify the direction of effects from the variable from which they emanate towards the variable in which they point into. Acyclic because no loops are contained, i.e. claims of the type A causes B causes C causes A are not included. It should be noted that the absence of arrows specifies the absence of direct effects.

If we delete the directionality information from the causal diagram we obtain the skeleton of the underlying DAG:

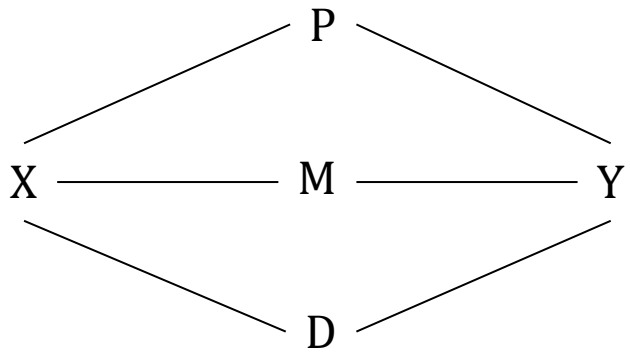


Figure 6: Skeleton of a DAG

The *skeleton of a DAG* specifies the family of causal models that imply the same conditional dependencies on observation of the included variables.

Therefore, the underlying causal model, i.e. the data-generating mechanisms, cannot be inferred from observational data. Still the family of causal models, i.e. the skeleton of the underlying DAG, can be inferred from observations. The skeleton above includes the information for example the information that P and M are conditionally independent, i.e. that they are not directly mechanistically linked. In other words, observations on conditional dependencies cannot define the data-generating model but can still be used to reduce the space of possible causal explanations. (Every explanation that involves a dependency between P and M for example would not be consistent with the observed data if the data-generating DAG belonged to the family specified by the skeleton above.) It might be possible to direct parts of the links, e.g. by chronologically structured data-assessment or by subject-matter knowledge, but others not. Partially directed acyclic graphs further reduce the space of possible causal explanations for observed data patterns.

DAGs are not limited in terms of number of variables to include. DAGs are also not specific on the type of functional relationship that underlies the links. Therefore, DAGs can be translated in structural equation models of any size and with any kind of parametric or non-parametric link-functions.

3.3 Data sources

3.3.1 Simulation studies

Completely specified causal structures

Causal structures with four variables were manually defined as DAGs including indirect and direct effects along with collider variables. According to these, data-generating models joint Gaussian distributions of random variables were simulated on 1000 observations using the *dag*R-package [146] and under the assumption of exclusively linear relations. Details on the setting of parameters were reported along with the results (4.2.1).

The well-defined causal structures included four variables: an exposure X_i; an outcome X_j; a parental variable P that was affected by the exposure X_i and affected the outcome X_j; and a descendent D that was affected by both, the exposure X_i and the outcome X_j. Based on

these four variables, two types of data-generating causal systems were defined, one entailing a direct effect of the exposure on the outcome and one without a direct effect between the two. Effect sizes were set to absolute values of |0.15| for all effects emanating from the exposure X_i, and to |0.30| for other effects involving the outcome X_j and all relevant combinations of signs of the model effects were simulated.

Random causal structures

Simulating data according to a random directed acyclic graph

Simulation algorithms implemented in the *pcalg*-package were used to generate random DAGs and joint distributions of random variables according to these graphical models [147]. For a given sparseness parameter (or connection probability) between zero (i.e., no connected nodes) and one (i.e., all nodes are interlinked) this approach draws a topologically ordered DAG with randomly directed edges. This graphical model is translated into a structural equation system. Model-effects correspond to standardized coefficients of the data-generating linear structural equation model [148].

Varying the parameter-settings to generate the directed acyclic graph

Settings of parameters in the simulation procedure that were suspected to have an impact on sensitivity and specificity of the PC-algorithm were systematically varied to cover the ranges observed in EPIC-Potsdam metabolomics networks. These parameters included the characteristics of the simulated DAGs: number of variables (*network size*); connection probability between variables (*network density*); and strength of the modeled effects that generated these connections (*effect strength*). Furthermore impact of the number of observations (*sample size*) on performance of the PC-algorithm was evaluated in simulations.

Simulations to evaluate the dependency of the PC-algorithm on *network sizes* and *network densities* were based on the following scenarios. Data-generating models with 11, 26, and 81 variables (*network sizes*) were considered. The *network density* was simulated in the range of an average number of direct neighbors of network variables (local density criterion) between one and seven. For each simulation-run the average number of neighbors per node was fixed and 100 data-generating models were randomly generated with constant parameter settings. This procedure was repeated by increasing the average number of neighbors per node

from 1 (sparsest scenario) to 7 (densest scenario) in steps of 0.5. Effect strengths were randomly generated (range 0.15 to 0.80) and Gaussian data were simulated on 2000 observations.

To evaluate performance of the PC-algorithm according to *effect strength* all standardized regression coefficients in data-generating structural model equations were fixed to a specific value for each set of 100 simulations. This value was stepwise increased from 0.01 to 0.91 in steps of 0.05. This framework was applied to three different *network densities* (2, 4, and 7 neighbors per node on average) and three different *network sizes* (11, 26, and 81 nodes) resulting in nine (3x3) scenarios. Per simulation Gaussian data were generated on 2000 observations.

Varying the sample-size

In addition to varying the parameter-settings of the data-generating model also the number of simulated observations per data-generating model was varied. In this set of simulations networks with 11, 26, and 81 nodes, respectively, were used, the average number of neighbors per node was set to four and model-effects were randomly generated in the range from 0.2 to 0.7. For each setting 100 data-generating models were created randomly. The simulated *sample size* per data-generating model was gradually increased from 25 observations to one million observations (steps: 25; 50; 100; 250; 500; 1000; 2000; 4000; 8000; 12,000; 25,000; 50,000; 100,000; 500,000; 1,000,000).

Evaluating the performance of the PC-algorithm

Performance of the PC-algorithm was evaluated according to sensitivity and specificity of the algorithm to detect links between variables that correspond to data-generating mechanisms. Three indicators were calculated. *True positive rate* is equivalent to *sensitivity* and is defined as ratio of true positives (number of detected links that corresponded to data-generating mechanisms) relative to real positives (data-generating mechanisms in the underlying DAG).

$$True\ positive\ rate = \frac{true\ positives}{real\ positives} = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

False positive rate is defined as ratio of false positives (number of links that did not correspond to data-generating mechanisms) relative to all real negatives (pairs of variables not connected by a data-generating mechanism). False positive rate is thus an inverse indicator of specificity.

$$False\ positive\ rate = \frac{false\ positives}{real\ negatives} = 1 - specificity$$

The *true discovery rate* or *positive predictive value* indicates the fraction of all findings that correspond to true links.

$$True\ discovery\ rate = \frac{true\ positives}{true\ positives + false\ positives}$$

The *true discovery rate* therefore assessed the likelihood that a detected link corresponded to a data-generating mechanism.

3.3.2 The EPIC-Potsdam cohort study

Overview

This subsection will explain design and data assessment of the EPIC-Potsdam study and is organized chronologically. After some general information on the multicenter EPIC-study, recruitment of EPIC-Potsdam participants and protocols of the baseline assessment will be outlined, including description of the assessment of habitual diet and lifestyle routines, physical examination and blood sampling protocol. Thereafter follow-up procedures of EPIC-Potsdam participants will be explicated including detection and verification strategies for incident type 2 diabetes. Then the construction of the diabetes case-cohort for molecular phenotyping will be specified. Finally chemical measurements of biological molecules in the blood will be described.

The European investigation into cancer and nutrition study

The EPIC-Potsdam study is part of the multicenter EPIC-study [149]. The design of the EPIC-study was dedicated to investigate the impact of habitual diet on the risk of developing chronic diseases [149,150]. In one of the largest epidemiological cohort studies worldwide 521,448 study participants were recruited in a cooperative effort of 23 research institutions located in 10 European countries [151].

The core protocol in all EPIC centers comprised baseline assessment of information on lifestyle and diet, and anthropometric measurements [150]. Furthermore biological samples were taken at baseline of the majority of study participants. All centers followed up

participants with a focus on the detection of incident chronic diseases including cardiometabolic events and cancer.

Recruitment

The city of Potsdam in Brandenburg, federal state of Germany, had been selected to host one of the two German EPIC centers. The target population of the EPIC-Potsdam study was defined as the general population of middle to older age, i.e. in an age range of 35 to 64 for women and from 40 to 64 for men at the time of recruitment, living in the city of Potsdam and the surrounding municipalities [152]. Therefore, the study region comprised a larger city with above 100,000 inhabitants as well as small towns and rural areas. Random samples of inhabitants meeting the age criteria were periodically provided by the registration offices of the study region and by the end of the recruitment period 100% of eligible individuals in the study region were contacted [152].

The study was approved by the ethics committee of the Medical Society of the State of Brandenburg. Prerequisite for study participation was an *a priori* signed informed consent [152]. The consent covers biomedical research in the public interest and could and can be withdrawn by the study participant at any time for the future without further explanations. In addition, individuals eligible for participation had to complete the basic interview as well as questionnaires on diet and lifestyle to be included in the study [149]. Of the invited individuals 22.7% participated in the study [152].

In Potsdam the recruitment phase took place between August 1994 and September 1998 [152]. In total 27,548 participants from the general population were included in the study, comprising 16,644 women and 10,904 men. The large majority of participants were in the targeted age range from 35 to 64 years. Compared to population survey data, EPIC-Potsdam participants tended to have more favorable socio-economic and health indicators [152].

Baseline assessment

Recruitment and assessment tools

Potential study participants who responded to the invitation letter and agreed to a personal appointment received a food frequency questionnaire [153] and a questionnaire on non-dietary lifestyle aspects [149] per mail, approximately ten days prior to the visit at the

examination center. The questionnaires were completed at home and brought to the examination center. At the examination center a physical examination and a personal interview were carried out by trained personnel. Baseline assessment tools of the EPIC Potsdam study are summarized in Table 1. All baseline assessment tools were monitored according to predefined procedures. These procedures covered qualitative and quantitative aspects of data quality and took into account potential technical and human sources of bias [154].

Table 1: Baseline assessment tools of the EPIC-Potsdam study [adapted from [154]]

Data assessment area	*Applied assessment tool*
Habitual diet	Self-administered food frequency questionnaire
	Computer-guided face-to-face correction interview for the food frequency questionnaire
	24-hour recall (EPIC-Soft) (subsample)
Lifestyle and medication	Self-administered questionnaire, scanner-readable
	Computer-guided face-to-face interview
Anthropometry and blood pressure	Standardized measurement procedures by trained medical personnel
Molecular markers	Blood sampling and specimen storage in liquid nitrogen / deep freezers for biochemical analyses

The habitual diet

Dietary assessment in the EPIC-Potsdam study comprised a self-administered food frequency questionnaire, a computer-guided interview to correct missing or implausible information in the questionnaire and for a subsample of study participants repeated 24-hours dietary recalls.

The food frequency questionnaire was designed to assess individual habitual intake estimates for food groups, single food items and nutrients. Based on German national survey data on the detailed food consumption of a population-based sample over one week, a food list of 158 foods and mixed dishes was compiled that contained all foods that notably contributed to food group and nutrient intake at the population level [155].

For each food item in the list participants were asked to indicate whether they consumed it or not during the last year [155]. In case they reported to having consumed the food item, participants were further

asked to estimate intake frequency and usual portion size. Participants could select their typical portion size from a range of predefined portion sizes. For food items for which the habitual portion size was not easily expressible in usual household measures, colored pictures were prepared to aid the participants' estimate on the usually consumed portion size. For the intake frequency participants were to choose among a frequency (1-6 times) and a time frame (per day / week / months / year). The available frequency and portion size categories were designed to cover the range of reported frequencies and portion sizes in German national survey data. Furthermore, for some foods information on seasonal variation of the intake was considered [155].

The self-administered food frequency questionnaire was completed at home. After scanning, missing and implausible information was reviewed and corrected in a face-to-face interview with the participant. This correction procedure led to complete and logically consistent dietary data on all participants [154].

Participants tended to overestimate intake frequencies of food items from food groups including many slightly different items relative to intake frequencies of food items from groups with few items only [156]. Therefore, summary questions on the overall intake of global food groups were used to calibrate reported intake frequencies of items within the food groups [157]. Technically, reports of intake frequencies of single food items were linearly adjusted for reported intake frequency of the corresponding food group. Weighting factors were calculated by dividing reported intake frequency of the global food group by the sum of reported intake frequencies of all single items within that group. Intake frequencies of single food items were then calibrated by multiplication with the group-specific weighting factor [157].

On the single food level, this study included information from the calibrated food frequency questionnaire on the exposure to diabetes-related dietary items, including the habitual intake of unprocessed and processed red meats, whole-grain bread, and coffee consumption [26,158]. Adjustment for potential confounding effects by other foods also relied on food frequency questionnaire information and models were adjusted for total energy intake calculated over reported intake levels of all food items with the use of the German Federal Food Code [159].

No single method is considered gold standard to assess habitual diet in large human studies. Therefore, the quality of the intake estimated

by the food frequency questionnaire used in the EPIC-Potsdam study was evaluated with regard to consistency and relative validity. Reproducibility was checked by repeated administration of the same questionnaire to the same participants six months apart. Relative validity was appraised against repeated 24 hours dietary recalls [153,155,157,160]. In the 24 hours recalls participants were asked to report in a computer-assisted face-to-face interview all food and drinks consumed within the last 24 hours [161]. A subsample of participants underwent a series of interviews spread across days of the week and seasons [155].

Measures of reproducibility and relative validity of the EPIC-Potsdam food frequency questionnaire are summarized in Table 2. On the food group level, the reproducibility as indicated by the intraindividual correlation of reported intake levels from the repeated food frequency questionnaires was moderate to good, ranging from 0.49 for bread to 0.89 for alcoholic beverages. Correlation between intake levels estimated with frequency questionnaire and intake levels estimated by repeated 24 hours recalls indicated moderate to good validity, ranging from 0.65 for meat to 0.86 for alcoholic beverages [155]. On an aggregated level, total energy and dietary fiber intake estimates from the frequency questionnaire and the dietary recalls were moderately correlated, whereas a high correlation was assessed for total protein intake [157,162]. In comparison to protein intake objectively measured by urinary nitrogen excretion, dietary protein intake was underreported in the frequency questionnaire by about 23 % [162]. Total energy intake was underreported by approximately 22 % when compared to energy expenditure objectively measured with the doubly labeled water method [162]. The amount of misreporting seemed not to depend, however, on the intake levels [162].

Table 2: Reproducibility and validity of intake levels estimated with the EPIC-Potsdam food frequency questionnaire [adapted from [155] and [162]]

Dietary exposure	*Reproducibility*[*]	*Relative validity*[#]
Bread	0.49[§]	0.77
Cereals	0.73	0.70
Processed meat	0.73	0.70
Meat	0.77	0.65
Coffee, tea	0.71	0.70
Total energy	0.68	0.65
Dietary fiber	0.64	0.50

*The same food frequency questionnaire was repeated after six months

#Relative to intake estimates from a series of 24 hours dietary recalls

§Spearman rank order and Pearson (energy, protein, and fiber intake) correlation coefficients; all such values

Lifestyle habits

The lifestyle questionnaire was composed of questions on family circumstances and on socioeconomic indicators, i.e. educational background and working situation. The personal interview was designed to assess details on health-related non-dietary behaviors and on the health status and the medical history of the participants. It included questions on smoking status, physical activity, health status, and medication. Prevalence and history of chronic diseases was also assessed.

Phenotypical traits

At the study center, participants underwent a physical examination by trained medical personnel. The physical examination included various anthropometric measurements [163], of which data on weight and height were used in this study. Weights of participants in light underwear and with emptied bladder was measured with an electronic digital scale (Soehnle, type 7720/23, Murrhardt, Germany) accurate to 100 g. Heights were measured with a flexible anthropometer to the nearest 0.1 cm. Body mass index was calculated as body weight in kilograms divided by squared height in meters. Measurement error of the anthropometric measures was neglectable in comparison to the between person variance. Reliability coefficients above 0.99 for intra-interviewer as well as

between-interviewer variability indicated superior reproducibility of all included anthropometric parameters [163].

Blood pressure was measured with an automated oscillometric device (BOSO Oscillomat, Jungingen, Germany). In a standardized procedure participants rested in an upright sitting position and a series of three consecutive measures was performed with time intervals of two minutes between the single measurements and the device located on the right upper arm [164,165]. Thus, three measures of systolic and diastolic blood pressure, respectively, were obtained per participant. An average over the second and third measurement used because this was shown to being the most stable and consistent parameter of blood pressure [166]. Participants with systolic blood pressure ≥140 mm Hg or diastolic systolic blood ≥90 mm Hg, or both, with self-reported hypertension diagnosis, or with self-reported use of antihypertensive medication at baseline were classified as cases of prevalent hypertension.

Blood sampling

Blood samples were collected of approximately 95.7% of the participants [149]. In total, thirty milliliters of venous blood was taken of each participant who agreed. Thereof, twenty milliliters were stored with citrate (plasma) and ten milliliters were stored without adding any anticoagulant (serum). The blood samples were fractioned into serum, plasma, buffy coat, and erythrocytes. Fractions were aliquoted into straws of 0.5 mL each and stored in tanks of liquid nitrogen (approximately –196°C) and in deep freezers (approximately –80°C), respectively, until further analyses. Sampling and handling of blood was realized according to a highly standardized protocol by qualified and specifically trained medical personnel [149].

Longitudinal data collection

The follow-up procedure

Approximately every two years, information on the study participants was collected in course of an active follow-up procedure [167]. Participants were mailed a questionnaire with amongst others questions on current medication and on the incidence of 24 chronic diseases within the follow-up time. Furthermore, for incident diseases age at diagnosis as well as place of diagnosis and the treating physician were assessed. This study included follow-up information until the end of August 2005.

Several measures were taken to maximize response rates in the EPIC-Potsdam study and to validate the self-reported information [167]. Participants were reminded by telephone calls and additional letters if they did not respond within two weeks to the initial follow-up letter, and reminder activities were continued for up to one year if necessary. If neither the participant nor close relatives were available, vital status and eventually new contact data were derived from local registration offices. The intense follow-up strategy resulted in information on the vital status of close to 100% of the study population, and the final follow-up rate in the first follow-up round was 96%, comprising responders to the questionnaire and identified deaths [167]. In the follow-up rounds two and three response rates were 95% and 91%, respectively. The fourth follow-up round was ongoing at the censoring date (31 August 2005) with a preliminary response rate of 90% [168].

Incident diabetes mellitus type 2

Participants were classified as non-verified (or potential) incident case of type 2 diabetes if the follow-up questionnaire gave any indication of new onset type 2 diabetes. For participants without documented prevalent type 2 diabetes, self-reports on a diagnosis of diabetes mellitus type 2 within the follow-up period, taking antidiabetic drugs, or being dietary treated because of diabetes mellitus type 2 were considered as evidence of potential incident type 2 diabetes. Systematic information sources for incident cases were self-reports of a type 2 diabetes diagnosis, type 2 diabetes-relevant medication, and dietary treatment due to type 2 diabetes during follow-up. Furthermore, additional information from death certificates or from random sources was obtained, such as the tumor centers, physicians, or clinics that provided assessments from other diagnoses. Although self-reports of type 2 diabetes were generally reliable, by including other sources of information, the completeness of case ascertainment was even improved. Once a participant was identified as a potential case, disease status was further verified by sending a standard inquiry form to the treating physician. Only physician-verified cases with a diagnosis of type 2 diabetes (International Classification of Diseases, 10th revision code: E11) and a diagnosis date after the baseline examination were considered confirmed incident cases of type 2 diabetes [169].

The diabetes case-cohort

A case-cohort was constructed within the EPIC-Potsdam cohort. The case-cohort design was established as resource efficient way of molecular phenotyping in large prospective cohort studies [170]. In short, a statistically representative subsample of the full cohort, the so called subcohort, was randomly drawn for molecular studies. In addition to the cases randomly included in the subcohort, all other incident cases of the targeted disease, in this case type 2 diabetes, were included in molecular studies. Applying the appropriate statistical analyses this study design offers valid effect estimates and conserved statistical efficiency compared to the full cohort but largely reduces the costs and the use of resources [170,171]. Another advantage is that the subcohort can be in principle used as control group for several outcomes.

The case-cohort was based on all participants who provided blood at the baseline-examination in the EPIC-Potsdam cohort (n=26,444). Follow-up information was considered until 31st of August 2005 (censoring date for longitudinal analyses), corresponding to a mean follow-up of 7 years. In this timespan, a total of 820 type 2 diabetes cases with complete biological material for molecular phenotyping had been identified. According to appropriate sample size calculations the subcohort representative for the EPIC-Potsdam cohort comprised 2,500 randomly selected participants, among them 74 randomly selected participants with incident type 2 diabetes during follow-up (internal cases). Participants were excluded due to prevalent type 2 diabetes at baseline (n=122); prevalent cancer, myocardial infarction or stroke (n=238); missing information or inconsistent information on baseline covariables (n=4); missing follow-up information (n=59); incomplete biological material (n=14); and missing or implausible metabolomics data (n=78). The analytical type 2 diabetes-cases cohort used for the present study therefore comprised 2731 participants. Descriptive statistics and cross-sectional analyses relied on 2092 members of the subcohort (representative sample of healthy participants in the full cohort). Longitudinal analyses further included 692 incident type 2 diabetes cases of which 53 were also members of the random subcohort (internal cases).

Biomarker measurements

Serum concentrations of 163 metabolites were determined in plasma samples from baseline in the case-cohort applying a kit-based targeted

metabolomics approach. The AbsoluteIDQ™ p 150 Kit (Biocrates Life Sciences AG, Innsbruck, Austria) used isotope-labeled internal standards to quantify the targeted metabolites. High throughput flow injection analysis tandem mass spectrometry (FIA-MS/MS) technique was applied to measure metabolite specific signals. This kit based approach simultaneously targeted the quantification of hexoses (sum of six-carbon monosaccharides), 14 amino acids, 92 glycerophospholipids, 15 sphingomyelins, and 41 acylcarnitines. The group of phospholipids comprised lysophosphatidylcholines, diacyl phosphatidylcholines, and acyl-alkyl phosphatidylcholines (Figure 7). Metabolites were only included in the analyses after surpassing preset reliability criteria in a pilot study. Metabolites were excluded because of mean concentrations below the limit of detection (n = 30) or because of high analytical variance (n = 6).

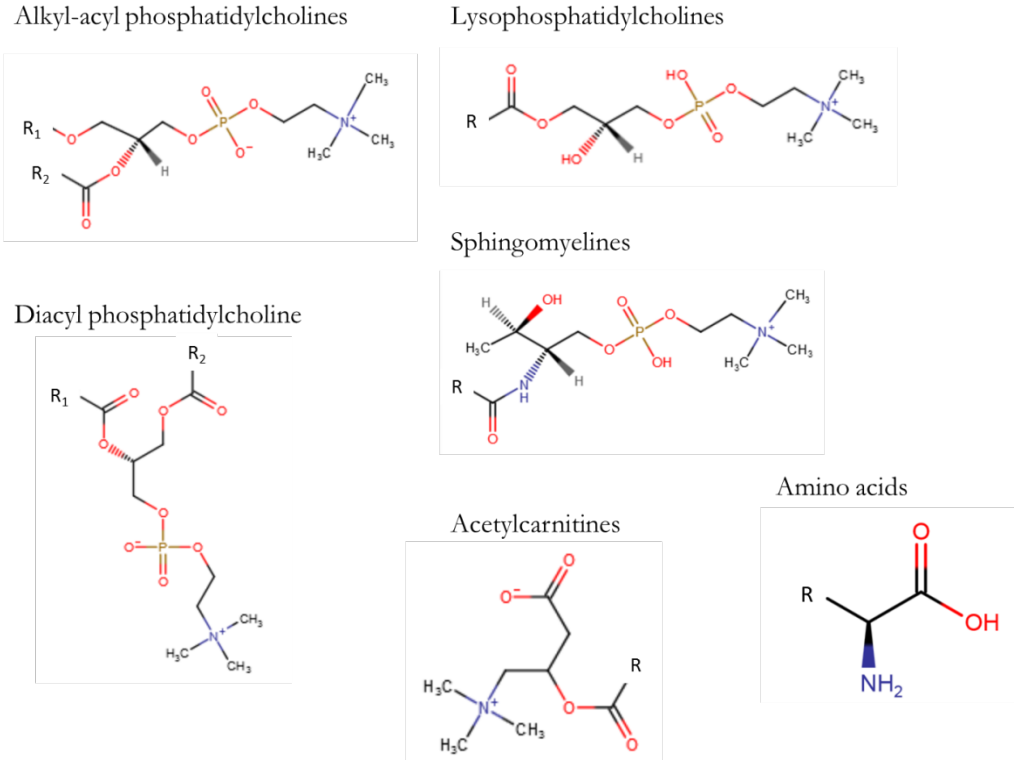


Figure 7: Molecular formulas of the targeted metabolite groups; R: fatty acid residual. Source: Human Metabolome Database (www.hmdb.ca/metabolites)

The targeted metabolomics approach was conducted at the Genome Analysis Center of the Helmholtz Zentrum München. Samples were prepared as indicated in the manufacturer's protocol (Biocrates user's manual UM-P150). The procedure was described in detail previously [172]. Briefly, 10 µL serum was inserted into a filter on a 96-well sandwich plate. The plate already contained stable isotope-labeled internal standards. Amino acids were derivated with 5%

phenylisothiocyanate reagent. Extraction of metabolites (including isotope-labeled internal standards) was realized by adding 5 mmol/L ammonium acetate in methanol. After centrifugation through a membranous filter and dilution with mass spectrometry running solvent, final extracts were analyzed by flow injection analysis mode tandem mass spectroscopy (FIA-MS/MS). Metabolites were quantified in mmol/L by relating their signals to those of the isotope-labeled internal standards. For validation procedures of the method, and analytical specifications the reader is referred to Biocrates manual AS-P150 and previous publications [172]. For the lipid metabolites, fatty acid residues were abbreviated Cx/y, where x represented the cumulative number of carbon atoms in fatty acid residues and y the cumulative number of double bonds. Whenever a single fatty acid was contained in the lipid metabolite the standard short annotation for fatty acids (Cx:y) was used to indicate chain length (x) and number of desaturations (y).
Selection of the targeted set of metabolites by the manufacturer was based on the robustness of their measurements. Hence, uncertainty of the measurements was below 10% for most of the metabolites and accuracy was relatively high with all included metabolites ranging between 80% and 115% of their theoretical values. Within-plate and between-plate coefficients of variation based on the median analytical variance were 7.3% and 11.3%, respectively, for the analyzed samples [173]. Run-order effects were accounted for by randomization of the sample-sequence during measurements, regardless of the case status.

In a previous study reliability of the measurements was investigated by conducting repeated measurements within a small subsample of the study population. Most metabolites were found to have moderate (>0.40) to high (>0.70) intraclass correlation coefficients (ICCs) over a 4-months period [173], and metabolites with an ICC < 0.40 were not considered in the analyses. Consequently, single measurements of the included metabolites were assumed to be applicable for epidemiological risk assessment.

3.4 Statistics & algorithms

This section will describe the statistical procedure including transformation and standardization of the data (3.4.1), factor analysis (3.4.2), estimating causal networks (3.4.3), and regression models used for cross-sectional and longitudinal analysis (3.4.4). Furthermore, the

methods applied for effect decomposition in terms of mediation analysis will be explicated (3.4.6).

3.4.1 Data processing

Dietary information on solid foods was energy standardized. Technically, daily intake individual of each food in grams was divided by daily energy intake. Models were further adjusted for daily energy intake according to the nutrient density method [174]. Beverages were not energy standardized. Assessment of coffee in categories of cups (150 mL) per day or per week resulted in a multimodal distribution. Based on this coffee intake-distribution a categorical variable with seven levels (0-6) was generated: <0.5 cups/day; ≥0.5 and <1.8 cups/day; ≥1.8 and <2.8 cups/day; ≥2.8 and <3.8 cups/day; ≥3.8 and <4.8 cups/day; ≥4.8 and <7.8 cups/day; ≥7.8 cups/day. Dietary exposure-variables were standardized to two standard deviations as unit of variance. Thus effect-estimates for food-exposures indicated change in outcome-level observed in relation to two standard deviations higher exposure-level.

Serum concentrations of metabolites were log-transformed and standardized on all phenotypical characteristics that were regarded as potential confounders. Standardization was done by regressing age, sex, BMI [kg/m²], and prevalence of hypertension (yes/no) on each metabolite and using the residual variance in further models. It should be noted that this procedure had the same effect on significance of estimates compared to adjusting for phenotypical traits in multivariable models. Still the remaining variance in the metabolite-residuals was explained to a larger extent by external factors (e.g. diet). Metabolites were then z-standardized (mean of zero and standard deviation of one).

Metabolomics data were parted into subgroups according to biochemically closely related metabolites: amino acids, acylcarnitines, sphingomyelins, lysophosphatidylcholines, diacyl phosphatidylcholines, and alkyl-acyl phosphatidylcholines. Apart from biological plausibility subgrouping was further supported by the fact that covariance was particular high within these groups and they tended to cluster in previous network-analyses in EPIC-Potsdam [83]. Furthermore, a comparison across four independent cohorts revealed that the links between metabolites were very stable within these metabolite groups but that different cohorts were not well comparable with respect to between metabolite group links (Stefan Dietrich, 2017, unpublished).

3.4.2 Factor analysis

Factor analysis was used to estimate the level of hidden variables based on measured variables that were assumed to being sensitive to the factor levels [175,176]. The hypothesis was that the largest dimension of common variance among groups of biochemically closely related metabolites was indicative for the all-over synthesis and turnover level of that group. It was therefore *a priori* aimed for a one factor solution per metabolite group. Still scree plots and the Eigenvalues were additionally evaluated according to consistency with the one factor solution [175].

As appropriate for the case-cohort design, the study sample to derive the factor loadings was restricted to the random subcohort (which is representative for the full cohort). Subsequently, standardized individual factor scores were imputed in external cases based on the standardized scoring coefficients. Factor analysis was applied restricted to metabolite groups, i.e. separately among amino acids, acylcarnitines, sphingomyelins, lysophosphatidylcholines, diacyl phosphatidylcholines, and alkyl-acyl phosphatidylcholines, respectively. Analyses were based on the metabolite-residuals standardized for the participants' age, sex, BMI, and prevalence of hypertension.

3.4.3 Estimating skeletons of acyclic directed graphs: the PC-algorithm

An implementation of the PC-algorithm [177] was used to estimate the skeleton of the data-generating DAG within metabolite groups. The skeleton of a DAG is the undirected graph that is common to a family of causal models characterized inducing the same conditional independence structure on the observed variables (i.e., the equivalence class of causal models). Application of the PC-algorithm implied the assumption of observed joint distributions being faithful to underlying DAGs but this is generally the case for multivariate normal distributions [148]. Table 3 describes the population version of the PC-algorithm.

Table 3: PC-algorithm [177], Table adapted from [148]

The PC-algorithm part I (to estimate the skeleton of the underlying DAG)

INPUT: Vertex Set *V*, Conditional Independence Information

Form the complete undirected graph $\tilde{C}$ on the vertex set V.
$l = -1;\ C = \tilde{C}$
repeat
  $l = l + 1$
  repeat
    Select a (new) ordered pair of nodes *i*, *j* that are adjacent in C such that $|adj(C,i) \setminus \{j\}| \geq l$
    repeat
      Choose (new) $\boldsymbol{k} \subseteq adj(C,i) \setminus \{j\}\ with\ |\boldsymbol{k}| \geq l$
      if *i* and *j* are conditionally independent given **k then**
        Delete edge *i*; *j*
        Denote this new graph by *C*
      end if
    until edge *i*; *j* is deleted or all $\boldsymbol{k} \subseteq adj(C,i) \setminus \{j\}\ with\ |\boldsymbol{k}| = l$ have been chosen
  until all ordered pairs of adjacent variables *i* and *j* such that $|adj(C,i) \setminus \{j\}| \geq l$ and $\boldsymbol{k} \subseteq adj(C,i) \setminus \{j\}\ with\ |\boldsymbol{k}| \geq l$ have been tested for conditional independence
until for each ordered pair of adjacent nodes *i*, *j*: $|adj(C,i) \setminus \{j\}| < l$

OUTPUT: Estimated skeleton *C*, separation sets *S*

Order-dependency which was still an issue for the original version was resolved for the applied version of the PC-algorithm [178]. In addition applicability of the used implementation of the algorithm to big observational datasets was demonstrated [145,179]. The application of the algorithm to random samples of the population involved decision-making. Therefore, conditional independence needed to be estimated [148] and significance testing was applied to the inner if-condition. As the algorithm was applied to Gaussian data this decision relied on Fisher's z-transform [148]. Outcome of the algorithm was a graphical model G (network) in which pairs of variables were connected only if they were dependent conditional on any subset of other network-variables. This network was passed to the NetCoupler-algorithm (3.4.5).

3.4.4 Multi-model procedures

Multiple models to infer the confidence range of possible effects

Inference on the relation between external variables and metabolites was based on a multi-model procedure. The estimated skeleton of the DAG did not allow inferring a single correct adjustment set to estimate

network-independent direct effects. Still at least one subset of direct neighbors of any metabolite in the estimated skeleton was theoretically sufficient to block all network-mediated effects and was therefore also sufficient to estimate network-independent direct relations of that metabolite with external variables. Hence the applied multi-model strategy ensured to certainly including the correct estimate in a multitude of possible estimates. Inference was then based on summaries over these estimates.

Cross-sectional analyses

Multiple multivariable-adjusted linear regression models were used to estimate ranges of possible direct effects of dietary exposures E on metabolites based on the metabolite network. Each metabolite (Met_o) was defined as outcome, and a fixed set of potential confounders $\mathrm{C} = \{C_1, \ldots, C_q\}$ was included as covariables in all models. The adjacency set $\mathrm{adj}(\mathrm{Met}_o) = \{V_1, \ldots, V_i\}$ was defined as the set of direct neighbors of Met_o in the metabolomics network. In the flexible model part, the dependency of the metabolite on exposure levels $\mathrm{Met}_o \sim E$ was adjusted for all possible subsets of the adjacency set $Met_o|\mathrm{adj}(\mathrm{Met}_o)_n \sim E$. Given i direct neighbors, 2^i models were calculated, corresponding to a distinct model for each element of the power set of $\mathrm{adj}(\mathrm{Met}_o)$.

$$\mathrm{Met}_o\,[E, \mathrm{adj}(\mathrm{Met}_o)_n, C] = \beta_0 + \beta_E * E + \sum_{k=k_{min}}^{p} \beta_k * \mathrm{V}_k + \sum_{l=1}^{q} \beta_l * C_l + \varepsilon$$

With $0 \leq p \leq i$ and $k_{min} = 0$ if $\mathrm{adj}(\mathrm{Met}_o)_n = \{\}$, and $k_{min} = 1$ otherwise. Accordingly for each exposure-metabolite pair 2^i potential effect estimates were generated with i being the number of direct neighbors of the metabolite in the metabolomics network.

The fixed set of potential confounders comprised age [years], sex, BMI [kg/m²], hypertension (yes/no) (respected by standardization, see 3.4.1), sports [h/week], biking [h/week], fasting status (3 stages: fasted, not-eaten-but-drunken, non-fasted), smoking (4 stages: never smoker, former smoker, current smoker <20 Units/day, current heavy smoker >20 Units/day), education (4 stages: no certificate, skilled worker, professional school, college of higher education/university), moderate alcohol consumption (10–40 g/day), antihypertensive medication (yes/no), lipid-lowering medication (yes/no), total energy intake [MJ/d], habitual consumption of: other diabetes-related foods (two of the

following whole-grain bread [g/MJ], total meat [g/MJ], coffee [cups/day]); foods (consumption of other bread, cornflakes, pasta & rice, vegetarian dishes, cakes & cookies, confectionary, eggs, raw vegetables, cooked vegetables, garlic, cabbage, mushrooms, fried potatoes, low fat dairy, high fat dairy, low fat cheese, high fat cheese, butter, margarine, other fat, sauce, fish, soup, all in [g/MJ]) and beverages (tea, sugar-sweetened beverages, wine, all in [g/day]) that were significantly correlated with one (or more) of the diabetes-related foods.

Prospective analyses

Longitudinal analyses were conducted applying Cox proportional hazards regression with Prentice weighting as appropriate for case-cohort design [170] with age as underlying time-scale. The multi-model procedure was organized analogous to the cross sectional analyses. The estimates from the multiple models calculated per metabolite corresponded to the range of possible effects of each metabolite on type 2 diabetes risk. Let Met_e be the metabolite considered as potentially type 2 diabetes-relevant exposure and let $\mathrm{adj}(\mathrm{Met}_e) = \{V_1, \dots, V_i\}$ be the adjacency set of this metabolite in the metabolomics network and $\mathrm{C} = \{C_1, \dots, C_q\}$ a set of fixed exogenous confounders. Given i direct neighbors, 2^i models were calculated, corresponding to a distinct model for each element $\mathrm{adj}(\mathrm{Met}_e)_n$ of the power set of $\mathrm{adj}(\mathrm{Met}_e)$.

$$\lambda(t|\mathrm{Met}_e, adj(Met_e)_n, C) = \lambda_0(t) * exp^{\left(\gamma_{\mathrm{Met}_e} * \mathrm{Met}_e + \sum_{k=k_{min}}^{p} \beta_k * V_k + \sum_{l=1}^{q} \beta_l * C_l + \varepsilon\right)}$$

With $0 \leq p \leq i$ and $k_{min} = 0$ if $\mathrm{adj}(\mathrm{Met}_e)_n = \{\}$, and $k_{min} = 1$ otherwise. The fixed set of covariables was the same as in cross-sectional analyses, with the exception of age which was included as strata-variable, corresponding to allowing for random baseline risks (intercepts) but assuming fixed effects within age strata of one year.

The software implementation of the multi-model procedure was based on the generalized linear models function in R. For this work Gaussian link functions (diet → metabolite relations) and Cox proportional hazards functions (metabolite → time-to-type 2 diabetes) were used. The glmulti-package in R basically provided a wrapper for generalized linear models that parted independent variables into three groups: first the main effect of interest (fixed, i.e. part of all generated formulae); second a group of flexible independent variables (flexible means that these

variables were included in some but not all generated formulae); and third a group of fixed confounders. An implemented enumerator in the package derived all possible non-redundant combinations of the flexible independent variables, generated according non-redundant formulae and passed them to R's glm environment. The output was a list with detailed information on and results from all calculated models. This information was used by the NetCoupler-algorithm to classify possible links into direct effects, ambiguous links, and non-direct effects. For this work applications were limited to linear models and Cox models without interactions but the tool can handle other types of link functions and interactions as well.

3.4.5 NetCoupler

An algorithm to identify effects between external variables and causal networks

Secondary aim of this thesis was to develop a method to link dietary data and information on disease risk to metabolomics networks in observational studies. To this end the NetCoupler algorithm was created. Links in DAG-skeleton-like networks mark direct effects. Key issue to couple external variables to an existing causal network structure thus was differentiation between direct and indirect (network-mediated) effects. The algorithm relied on the fact that under Markov assumptions the set of direct neighbors of a metabolite in the metabolomics network necessarily included the set of Markovian parents of that metabolite. Furthermore, it was shown that conditioning on Markovian parents renders a variable (metabolite) independent of all other variables in the network, which rules out any explanation but a direct effect for an association [129], under the assumption of sufficient information on the data-generating mechanisms. Missing directionality information on links between metabolites, however, implied that only limited partial information on the data-generating causal model was available. Unambiguous identification of the Markovian parents among the direct neighbors of metabolites was thus not possible. This was a challenge because adjusting for descendants implied the possibility to bias effect estimates. With these considerations at hand, a multi-model approach was pursued adjusting the relation of interest for all non-redundant combinations of direct neighbors of the involved metabolite in the metabolite network. Inference on network-independent direct effects

between metabolites and external variables was based on summarizing results from this multi-model procedure.

Different nature of the relation between diet and metabolite levels and between metabolite levels and disease incidence was assumed. Diet was postulated to potentially affect metabolite concentrations (diet → metabolite). Metabolite levels in turn were assumed to potentially affect time-to-type 2 diabetes incidence (metabolite → type 2 diabetes risk). This had implications for the design of the algorithm. Hence, two versions were developed, NetCoupler.IN to link external variables that were assumed to potentially affect network variables, and NetCoupler.OUT to link external variables that were potentially affected by network variables.

NetCoupler.IN

The NetCoupler.IN-algorithm retrieved the range of possible direct effects of potentially influential *exogenous* factors on network variables. In this study, the direct effects of specific foods on metabolite concentrations were modeled given the metabolic network structure.

Table 4: NetCoupler.IN retrieves direct effects of exogenous variables on network variables

NetCoupler.IN

Input: DAG-skeleton G (Graphical model)
observations on: network variables M, exogenous exposure X; confounders C

start $DE = \{\}$
repeat
  add all new direct effect to DE
    start $AMB = M$
    repeat
      s***elect*** *a variable* $M_i \in AMB$
      select *all nodes adjacent to* M_i *in* G, $adj(G,\ M_i)$
        repeat
          select a subset $S \in adj(G,\ M_i)$
          estimate $\widehat{M}_i \sim X \mid S, DE, C$
          add effect estimate PE *for* X *on* M_i to CS_i
        until no further non-redundant S can be selected from $adj(G,\ M_i)$
      if (lower bound of $CS_i > 0$ or upper bound $CS_i < 0$) and $sign(pe_1) = sign(pe_2)$ for every pair of estimates in CS_i: classify M_i as affected by X
      else if (lower bound of $CS_i > 0$ or upper bound $CS_i < 0$) and ($0 \in CS_i$ or $sign(pe_1) \neq sign(pe_2)$ for any pair of estimates in CS_i): classify M_i as ambiguous with respect to X
      else classify M_i as non-affected by X
      end if
    until all $M_i \in AMB$ have been selected
until no further M_i is classified as DE

Output: *Confidence set CS for* effects of X on M based on G

classification of every $M_i \in M$ as affected ($X \rightarrow M_i$), non-affected ($X \quad M_i$) or ambiguous (X---M_i)

|=conditional on; $adj(G,\ M_i)$=set of nodes adjacent to M_i in G ≡direct neighbors of M_i in G;

Table 4 shows a version of NetCoupler.IN that assumes full information on the source population. Application to limited samples of the source implied hypothesis-testing and following decision rules were applied:

1. Consider exposure-metabolite pairs only if significantly associated at false discovery rate-controlled p-value <0.1 based on a model *not adjusted* for adjacent metabolites ($S = \{\}$, marginal model)
2. Consider effect estimates as 0 if p-value >0.05

Thus with respect to a given exposure a metabolite was classified as non-affected if the false discovery rate-adjusted p-value was non-significant in the marginal model. Metabolites were classified exposure-affected with

significant false discovery rate-controlled p-value in the marginal model and all estimates in the confidence set significant and consistent. Other metabolites were classified ambiguous with respect to the exposure.

The outer loop of the algorithm included identified directly affected metabolites into the fixed model part. Then the multi-model procedure was repeated on the still ambiguous metabolites to check whether further unambiguous classification was possible based on that additional information. In the applied version of the algorithm this was limited to metabolites in the same connected components. Connected components were defined as group of two or more directly linked metabolites that were all associated with the exposure in the marginal model. The rational was that indirect effects could only be mediated by metabolites that were themselves (directly or indirectly) affected by the exposure. These metabolites were expected to be marginally associated with the exposure. In theory, one could construct scenarios in which incidental cancellation of several direct and indirect effects concealed exposure-dependency in the marginal model [129]. Still high abundance of incidental cancellation was considered unlikely based on observed correlation structures and the chance to unambiguously resolve such complicated scenarios in the applied modeling approach was low. Therefore the pragmatic decision was taken to adjust still ambiguous metabolites only for directly affected metabolites identified within the same connected component.

NetCoupler.OUT

The OUT-version of the NetCoupler-algorithm retrieved direct effects of network variables on later occurring events. In this study, the direct effects of metabolite concentrations at baseline on time-to-diabetes incidence were modeled given the metabolic network structure.

Table 5 reveals analogy of the OUT-version of the algorithm to the IN-version. Reversing assumptions on effect directionality ($M_i \leftarrow$ X but $M_i \rightarrow t(E)$), however, had implications for decision rules. Effect directionality from the network towards the outcome implied other network-variables as potential confounders of the effect of M_i on $t(E)$. Some network-variable M_c might have introduced spurious association by affecting metabolite M_i and time-to-event E ($M_i \leftarrow M_c \rightarrow t(E)$). In the IN-version (that assumed effect-directionality from the exposure towards the network) network adjustments were used to sort out indirect effects mediated by another network variable say M_M ($X \rightarrow M_M \rightarrow M_i$).

Table 5: NetCoupler.OUT generates estimates on direct effects network-variables on time-to-event data

NetCoupler.OUT

Input: DAG-skeleton G (Graphical model)
observations on: network variables M, time-to-event $t(E)$; confounders C.

start with $DE = \{\}$
repeat
  add all new direct effect to DE
    start with $AMB = M$
    repeat
      s*elect* *a variable* $M_i \in M$
      select *all nodes adjacent to* M_i *in* G $adj(G,\ M_i)$
        repeat
          select a subset $S \in adj(G,\ M_i)$
          estimate $t(E) \sim M_i | S, C$
          add effect estimate PE for X on M_i to CS_i
        until no further non-redundant S can be selected from $adj(G,\ M_i)$
      if (lower bound of $CS_i > 0$ or upper bound $CS_i < 0$) and $sign(pe_1) = sign(pe_2)$ for every pair of estimates in CS_i: classify Met_i as affecting the risk of E
      else if (lower bound of $CS_i > 0$ or upper bound $CS_i < 0$) and ($0 \in CS_i$ or $sign(pe_1) \neq sign(pe_2)$ for any pair of estimates in CS_i): classify M_i as ambiguous with regard to risk of E
      else classify M_i as non-affecting risk of E
      end if
    until all M_i have been selected
until no further M_i is classified as DE

Output: confidence set CS for effects of M on $t(E)$ based on G
classification of every $M_i \in M$ as effector ($M_i \rightarrow t(E)$), non-effector ($M_i \quad t(E)$) or ambiguous (M_i---$t(E)$).

|=conditional on; $adj(G,\ M_i)$=set of nodes adjacent to M_i in G ≡direct neighbors of M_i in G;

It should be noted that an indirect effect unlike confounding is not biased. Decision rules were adjusted to avoid bias:

1. Consider metabolite-outcome pairs only if significantly associated at false discovery rate-controlled p-value <0.1 based on a model *adjusted for all adjacent metabolites* ($S = \text{adj}(M_i)$)
2. Consider effect estimates as 0 if p-value >0.05

Thus metabolites were classified as affecting type 2 diabetes risk if the false discovery rate-adjusted p-value was significant in the model adjusted for the full adjacency set and all estimates in the confidence set were significant and consistent. Metabolites were considered to not directly affect type 2 diabetes risk if they were not significantly associated

in the model adjusted for the full adjacency set. All other metabolites were classified ambiguous with respect to type 2 diabetes risk.

NetCoupler: stepwise-application example

Figure 8 and Figure 9display NetCoupler in operation in a stepwise fashion. This example was based on data on sphingomyelins, red meat consumption, and time-to-diabetes incidence. Layout of graphs here aimed to explain the logic of the algorithm, and graphs based on the same data will be presented in a content-oriented layout in the next chapter (4.3.3 - 4.3.5).

First the PC-algorithm was used to estimate the metabolite network based on metabolomics data (skeleton of the underlying DAG). Second preliminary links (dashed lines, colored border) of red meat-exposure with metabolites were identified based on significance of the marginal associations. Third information from the confidence set of possible model-effects was used to classify one of the links as direct effect (arrow pointing towards the network). Fourth the direct effect was added to the fixed model part and the multi-model procedure was repeated. It should be noted that all marginally associated metabolites belonged to the same connected component. This explained the marginal association of two metabolites (unambiguously non-affected, dashed lines were deleted). No new direct effect was detected and the multi-model procedure for the exposure was stopped. Preliminary links and direct effects of metabolites on type 2 diabetes risk were identified in steps five and six in an analogous multi-model procedure. In step seven the information was displayed in a joint graphical exposure-metabolites-disease model. For some external variable-metabolite pairs multi-model information remained inconclusive (dashed lines), for some there was no indication of a direct effect (no link), and some were classified as direct effects (arrows). Direction of the arrows was based on *a priori* assumptions.

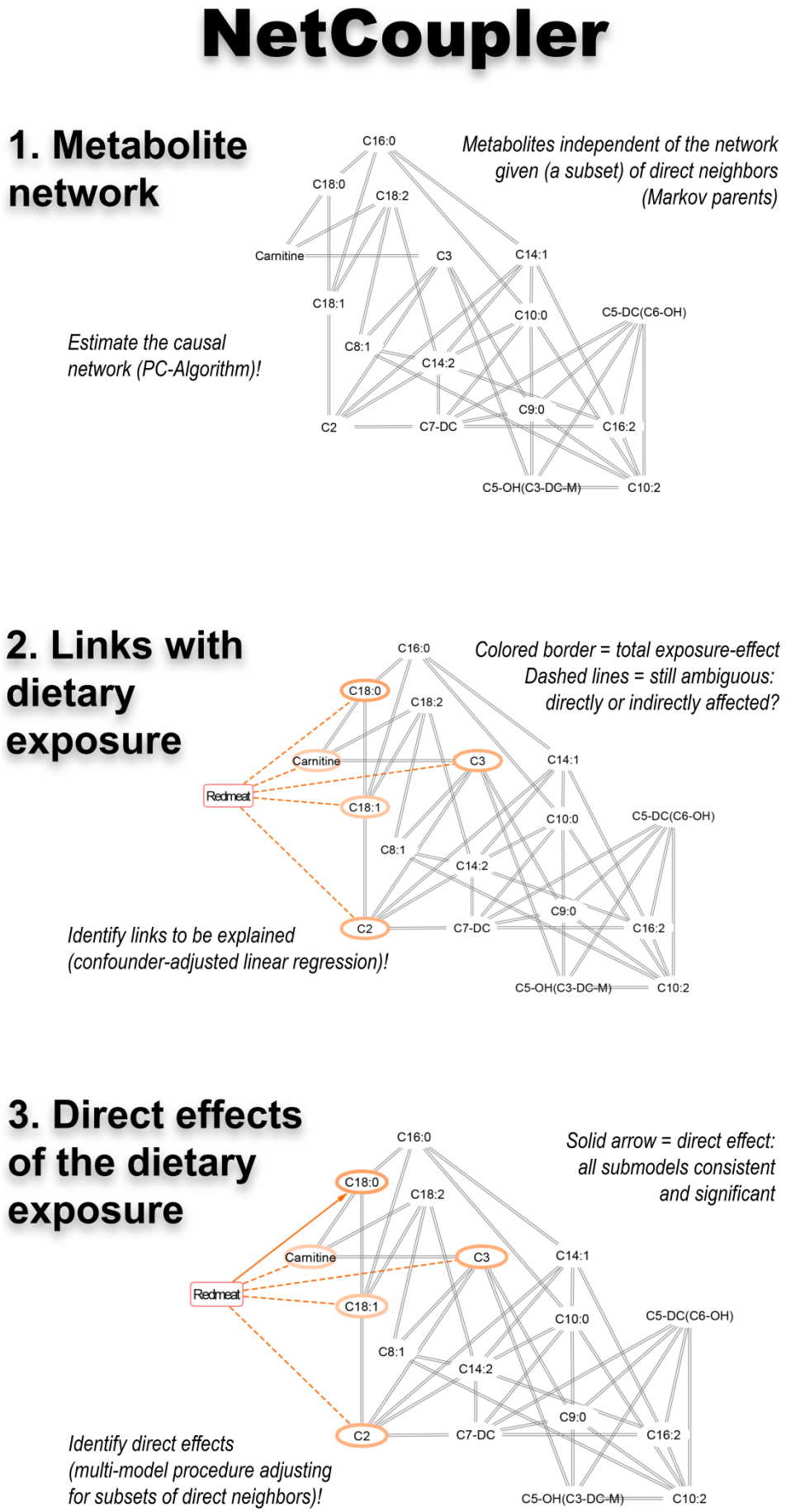


Figure 8: NetCoupler schematic application example (part I).

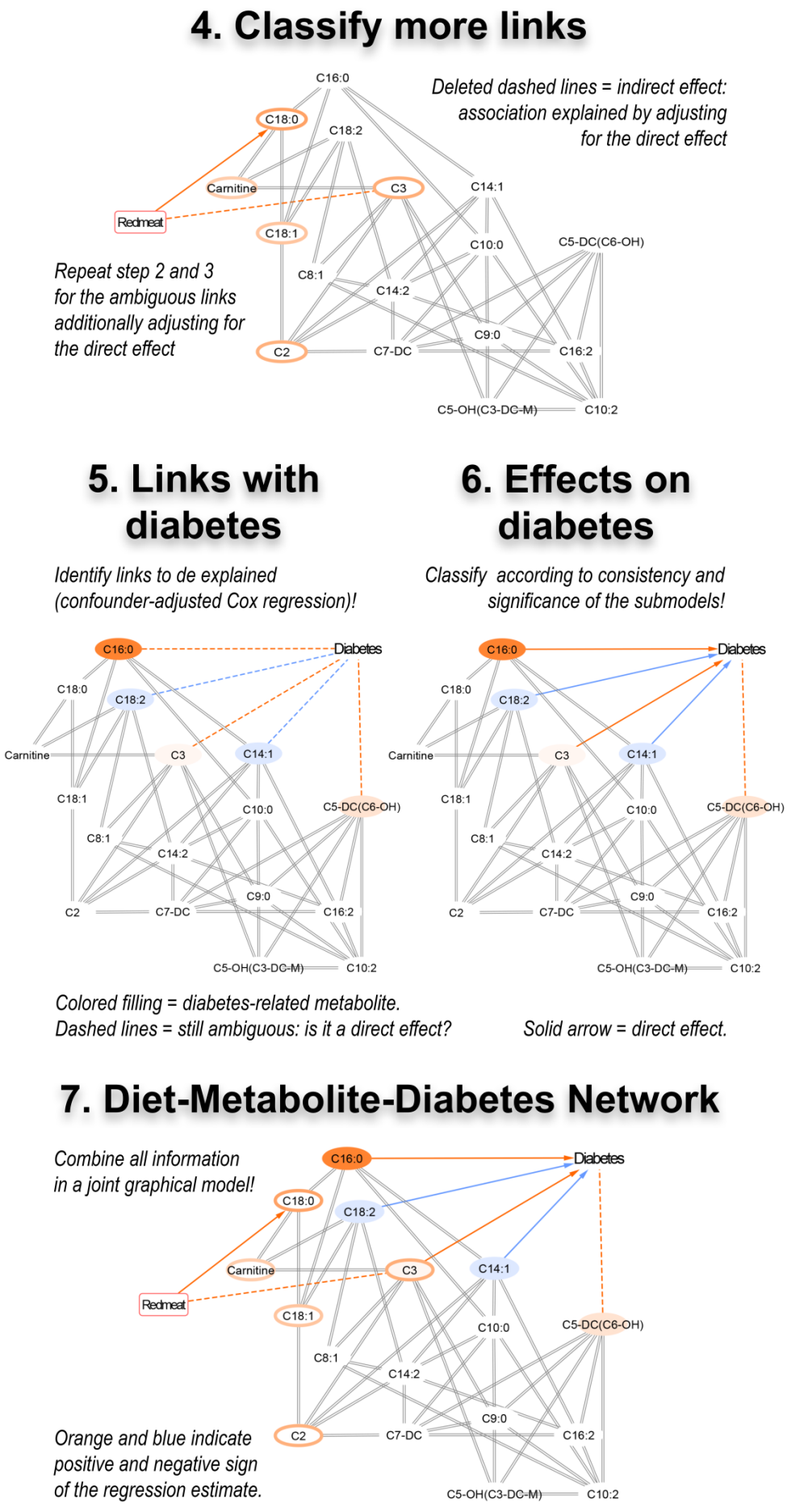


Figure 9: NetCoupler schematic application example (part II).

3.4.6 Mediation analysis

Potential mediating paths were selected manually according to following selection criteria: Metabolites were considered as potential mediators if they were (i) within an exposure-connected component (definition of connected components see 3.4.5) and (ii) classified as potential direct effectors of type 2 diabetes risk based on the multi-model procedure. Furthermore, (iii) effect directions needed to be consistent with a mediation hypothesis. Formally the product of the regression coefficient from the exposure-metabolite model and the regression coefficients from the metabolite-diabetes model was required to have the same sign as the regression coefficient from the exposure-diabetes model. Metabolites that fulfilled the three criteria were regarded as potential mediators of the exposure-diabetes relation.

The proportion potentially mediated by the selected mediators was estimated. Therefore, the network independent variation of the metabolite was estimated by adjusting the metabolite for all direct neighbors in the metabolite-network that were not on the shortest path from the exposure to the directly diabetes-linked mediator. The resulting residuals were used for decomposition of the total exposure-effect into a part that was potentially explainable by the selected mediator (i.e., an indirect effect or mediated proportion) and a part that was independent of the selected mediator (i.e., a direct effect or non-mediated proportion). Technically, two fully confounder-adjusted Cox-models were calculated, one with and the other without adjusting for the network-independent variation of the mediator. The proportion potentially mediated was estimated as difference between non-adjusted and adjusted exposure-estimates relative to the non-adjusted exposure-estimate. This is a valid approach to estimate mediated proportions based on proportional hazards models in scenarios where the outcome is rare [180]. Measures of central tendency and variation of the proportion mediated were obtained as median and 2.5^{th} and 97.5^{th} percentile from a bootstrapping-procedure with a sampling rate of eighty percent and 1000 repetitions. A logical upper bound of 100% was set for the proportion mediated.

3.4.7 Software applications

Descriptive analyses and factor analyses were performed with SAS software (Version 9.4, Enterprise Guide 6.1, SAS Institute Inc., Cary, NC, USA).

Other statistical analyses were performed in the R environment (R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria). The corrplot-package was used to generate (partial) correlation plots (https://CRAN.R-project.org/package=corrplot).The dagR-package was used to simulate data according to predefined causal graphs (https://CRAN.R-project.org/package=dagR).The pcalg-package was used to simulate and to estimate causal networks ([147], https://CRAN.R-project.org/package=pcalg).The glmulti-package was used to generate multi-model estimates (https://CRAN.R-project.org/package=glmulti).The igraph-package was used to identify connected components within networks (http://igraph.org/).The RCytoscape-package was used to export network-files to Cytoscape (https://www.bioconductor.org/packages/release/bioc/html/RCytoscape.html). The dplyr-package was used to manipulate data-frames (https://CRAN.R-project.org/package=dplyr). The ReporteRs-package was used to generate formatted tables (https://CRAN.R-project.org/package=ReporteRs). All packages were used in a version updated on January 16th 2017.

Cytoscape version 3.4 was used to visualize and analyze the output-networks ([181], http://www.cytoscape.org).

Adobe Photoshop CC 2014 (www.adobe.com) was used for the final layout of figures.

4 Results

4.1 Overview of the chapter

This chapter comprises a section with results on quantitative impact of collider bias and on the performance of the PC-algorithm based on simulated data (4.2) and a section with results on the mediating role of metabolomics network substructures that connect dietary exposures to type 2 diabetes incidence that was generated in the prospective EPIC-Potsdam cohort study (4.3). The structure reflects the chronology of the workflow. Simulation studies were used to *a priori* evaluate assumptions and test applicability of available tools to develop the NetCoupler-algorithm. Then the algorithm was developed and applied to human cohort data.

4.2 Illustrating concepts & testing tools: results from the simulation studies

4.2.1 Analyzes of bias in completely specified causal structures

In System 1 (S1), X_i had a direct effect on X_j (Figure 10). Furthermore, X_i affected variable P which was parent of X_j, and variable D which was descendent of X_j. Effect sizes were set to absolute values of $|0.15|$ for all effects emanating from X_i (α, δ, γ), and to $|0.30|$ for other effects involving X_j (β, ζ). All relevant combinations of signs of the effects in the system were considered. The effect of X_i on X_j was target parameter of the estimation procedure.

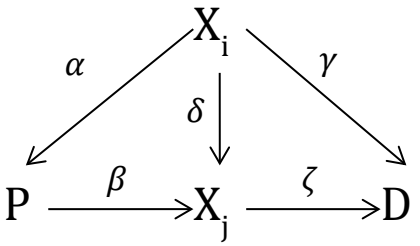


Figure 10: Data-generating model S1

According to causal inference theory the total effect (τ) of X_i on X_j was correctly estimated by a model not adjusted for covariables ($X_j = \tau * X_i + \varepsilon$). No confounding mechanisms were active because none of the arrows aimed at X_i (*backdoor criterion*). The total effect in this case was composed of two portions, a direct effect (δ) and an indirect effect ($\iota = \alpha * \beta$). To estimate the direct effect of X_i on X_j, the model

needed to be adjusted for the levels of P (to subtract the indirect portion of the effect mediated by P), but not for the levels of D. Thus the model $X_j = \delta * X_i + \iota * P + \varepsilon$ was expected to generate a valid estimate on the direct effect ($\hat{\delta}$).

In the first set of simulations of S1, the direct effect δ was set to +0.15 and the components of the indirect effect had equally positive signs ($\alpha = +0.15,\ \beta = +0.30$). The direct effect was well approximated in a regression model adjusted for P ($\hat{\delta} = 0.149$, M12 in Table 6). It should be noted that the estimate of X_i from the unadjusted model closely resembled the expectations on the total effect ($\hat{\tau} = 0.194$, M11), which was composed of the direct effect ($\hat{\delta}$), and the indirect portion mediated by P ($\hat{\iota} = \hat{\tau} - \hat{\delta} = 0.045$). Consequently, the estimated indirect effect also reflected *a priori* theoretical expectation of $\alpha * \beta = 0.3 * 0.15 = 0.045$. Whereas M11 and M12 provided valid estimates for different dimensions of the effect of X_i on X_j, adjusting for D introduced bias into the effect estimate of X_i on X_j (M13-M16). The fact that D was affected by both X_i and X_j rendered it a collider in relating these two variables (*d-separation*). The estimate for the effect of X_i on X_j was biased towards the null if the signs of the effects of X_i and X_j on D were the same (sgn(γ)=sgn(ζ), M13 and M14). On the contrary, effect inflation was observed in collider-adjusted models if the signs of the effects of X_i and X_j on D were different (sgn(γ) ≠sgn(ζ), M15 and M16). Notably, the presence of a collider in the model also compromised the efficacy of controlling for a parental variable (e.g., the strongest bias towards the null was observed in a model adjusted for the parental variable P and the descendent D, rather than in a model adjusted for D only; compare M13 and M14). It should also be noted that albeit the effect estimates were biased in models adjusted for collider-variable D, all models M11-M16 consistently indicated a significant positive effect of X_i on X_j.

In the second set of simulations of S1, the direct effect δ was again set to +0.15 but the components of the indirect effect had different signs now ($\alpha = -0.15, \beta = +0.30$) corresponding to a negative indirect effect. Again, non-adjusted model M21 and parental variable-adjusted model M22 delivered valid estimates for the total and the direct effect, respectively (Table 6). Generally, as described above adjusting for the collider D introduced bias. Still, there were some particularities to be pointed out here. Firstly, adjusting for an indirect

effect and a collider could lead to additive effect attenuation. The effect of X_i on X_j in model M23, e.g., was still significant (p=0.04). Slight changes of the data-generating effect sizes or of the significance cutoff (to account for multiplicity, e.g.) would have obscured, however, presence of a direct effect X_i on X_j. Secondly, various indirect mechanisms have the potential to incidentally cancel out. In M25, e.g., positive collider bias compensated for the negative indirect effect. Therefore, an incorrectly specified model delivered an estimate that quantitatively closely resembled the true underlying direct effect.

Table 6: Estimates for the effect of X_i on X_j from differently adjusted regression models based on simulated data according to System 1

ID	Regression model (adjustments)	Estimate	*p-Value*	Bias
Positive indirect effect: sgn(α)=sgn(β)				
M11	$X_j = X_i$	0.194*	3.7E-16	+0.045#
M12	$X_j = X_i + P$	0.149	1.0E-10	reference
	Equal signs of the effects of X_i and X_j on D: sgn(γ)=sgn(ζ)			
M13	$X_j = X_i + D$	0.122	9.5E-08	-0.027
M14	$X_j = X_i + P + D$	0.088	7.2E-05	-0.061
	Unequal signs of the effects of X_i and X_j on D: sgn(γ)≠sgn(ζ)			
M15	$X_j = X_i + D$	0.221	<2e-16	+0.072
M16	$X_j = X_i + P + D$	0.177	<2e-16	+0.028
Negative indirect effect: sgn(α)≠sgn(β)				
M21	$X_j = X_i$	0.107	5.7E-06	-0.042
M22	$X_j = X_i + P$	0.149	1.0E-10	reference
	Equal signs of the effects of X_i and X_j on D: sgn(γ)=sgn(ζ)			
M23	$X_j = X_i + D$	0.046	0.04	-0.103
M24	$X_j = X_i + P + D$	0.088	7.2E-05	-0.061
	Unequal signs of the effects of X_i and X_j on D: sgn(γ)≠sgn(ζ)			
M25	$X_j = X_i + D$	0.143	4.3E-10	-0.006
M26	$X_j = X_i + P + D$	0.177	2.9E-15	0.028

Data were generated by structural equation models according to the DAG depicted in Figure 10. Effect sizes were set to absolute values of $|0.15|$ for all effects emanating from X_i (α, δ, γ), and to $|0.30|$ for other effects involving X_j (β, ζ). Signs were varied, i.e. positive and negative regression coefficients were varied as indicates in the subheadings.*Regression coefficient of X_i; #Bias if interpreted as direct effect of X_i on X_j was calculated as difference to the reference model.

In contrast to S1, X_i had no direct effect on X_j in System 2 (S2, Figure 11). Analogously to causal structure S1, however, X_i affected a variable P, which was parent of X_j, and a variable D, which was descendent of X_j . As laid out above, the total effect was expected to be validly estimated by the non-adjusted model, whereas the model adjusted for the mediating variable P was expected to correctly indicate absence of a direct effect of X_i on X_j.

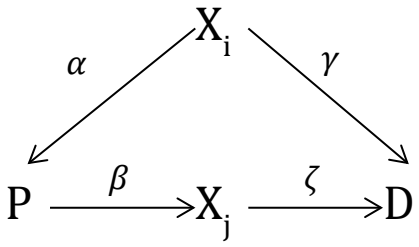


Figure 11: Data-generating model S2

In the first set of simulations of S2, the components of the indirect effect had equally positive signs ($\alpha = +0.15, \beta = +0.30$). Absence of a direct effect was inferable from the regression model adjusted for P ($\hat{\delta} = -0.001$, M12 in Table 7), whereas the non-adjusted model provided an accurate estimate on the total effect ($\hat{\tau} = 0.044$, M11), which was due to an indirect mechanism in this case. Again, the estimated indirect effect corresponded well with theoretical expectations ($\alpha * \beta = 0.3 * 0.15 = 0.045$). Albeit accurately estimated, this solely indirect effect would not have been considered significant at a 95% confidence level. Similar to S1, including the collider variable D into the model again introduced bias (M13-M16). It should be noted that effect estimates of X_i on X_j, were unstable with some suggesting positive and others suggesting negative effects. Moreover, given that the sought quantity was the direct effect of X_i on X_j, bias of the estimate was not always ameliorated by including the parental variable P along with the collider D (compare M13 and M14).

Analogous to S1, the second set of simulations of S2 evaluated collider bias in the presence of a negative indirect effect ($\alpha = -0.15, \beta = +0.30$). As expected, the models M21 and M22 provided accurate estimates of the total and the direct effect, respectively. Unsurprisingly, in absence of a direct effect, the sign of the indirect effect did not make a difference with regard to the absolute bias introduced by adjusting for collider-variable D. The pattern of unstable associations was comparable to the first set of S2 simulations. Results from either set included positive and negative estimates. Furthermore, some models indicated a significant effect of X_i on X_j due to collider-bias in either set of S2 simulations.

Table 7: Estimates for the effect of Xi on Xj from differently adjusted regression models based on simulated data according to System 2

ID	Regression model (adjustments)	Estimate	*p-Value*	Bias
Equal signs of the components of the indirect effect: sgn(α)=sgn(β)				
M11	$X_j = X_i$	0.044*	0.064	0.045#
M12	$X_j = X_i + P$	-0.001	0.963	reference
Equal signs of the effects of X_i and X_j on D: sgn(γ)=sgn(ζ)				
M13	$X_j = X_i + D$	-0.014	0.544	-0.015
M14	$X_j = X_i + P + D$	-0.049	0.027	-0.050
Unequal signs of the effects of X_i and X_j on D: sgn(γ)≠sgn(ζ)				
M15	$X_j = X_i + D$	0.083	<0.001	0.082
M16	$X_j = X_i + P + D$	0.038	0.086	0.037
Unequal signs of the components of the indirect effect: sgn(α)≠sgn(β)				
M21	$X_j = X_i$	-0.046*	0.050	-0.048**
M22	$X_j = X_i + P$	0.002	0.940	reference
Equal signs of the effects of X_i and X_j on D: sgn(γ)=sgn(ζ)				
M23	$X_j = X_i + D$	-0.095	<0.001	0.097
M24	$X_j = X_i + P + D$	-0.049	0.027	0.050
Unequal signs of the effects of X_i and X_j on D: sgn(γ)≠sgn(ζ)				
M25	$X_j = X_i + D$	0.000	0.986	0.001
M26	$X_j = X_i + P + D$	0.043	0.054	-0.042

Data were generated by structural equation models according to the DAG depicted in Figure 11. Effect sizes were set to absolute values of $|0.15|$ for all effects emanating from X_i (α, δ, γ), and to $|0.30|$ for other effects involving X_j (β, ζ). Signs of effects were varied as indicated the subheadings.*Regression coefficient of X_i; #Bias if interpreted as direct effect of X_i on X_j was calculated as difference to the reference model.

To summarize, simulation of data according to manually defined DAGs with four variables was used to evaluate the possible ramifications of collider bias. In most cases, adjusting for a collider produced quantitatively biased effect estimates. Still, based on significance testing, presence of a direct effect was correctly indicated by most models (S1), whereas some models always correctly indicated the absence of a direct effect when this was the underlying truth (S2). In some unusual cases an indirect effect and collider bias incidentally cancelled each other out leading to accurate estimates on the direct effect. Simulations also illustrated the potential of collider bias to obscure true effects or to suggest a direct effect where none was present. These concerns would particularly apply if inference on direct effects was based on a single

model without having valid knowledge of the underlying data-generating mechanisms. If any non-significant submodel would have been considered as evidence against a direct effect, to summarize over the estimates from all possible models, however, would have consistently indicated a direct effect in S1, and revealed its absence in S2.

4.2.2 Discovering causal structures in larger random networks

Overview

In this section, the causal inference algorithm later applied to discover metabolomics networks (PC-algorithm) was tested on simulated data. Settings of parameters in the simulation procedure, that might have influenced sensitivity and specificity of the PC-algorithm, were systematically varied to cover the ranges observed in EPIC-Potsdam metabolomics networks. Conceptually, network-simulations were divided into two parts: firstly, variation of the parameters that determined the random generation of the underlying DAG; secondly, variation of the sample size. The first part corresponded to applying the evaluated algorithm to varying biological settings, whereas the second part accounted for modifications of the study design.

Dependency of the PC-algorithm on the underlying DAG

Network density and network size

Figure 12 shows dependencies of the performance of the PC-algorithm on *network densities* in the different *network sizes*. Overall, the true positive rate as indicator of sensitivity of the PC-algorithm was good to excellent (>0.75) for networks with up to four (11 nodes network) or five (larger network) neighbors per node on average, with a tendency to be more accurate and precise in the larger networks. In the densest simulated scenarios, with on average seven neighbors per node, the true positive rate was moderate. The false positive rate as indicator of specificity was close to zero and thus generally neglectable over the whole range of modeled connection probabilities.

In accordance with very low false positive rates and moderate to excellent true positive rates, the PC-algorithm generally showed excellent true discovery rates in networks of moderate to high densities. Only in large and sparse networks (81 nodes, 1-2 neighbors per node) a relevant fraction of false positives among all detected links was assessed.

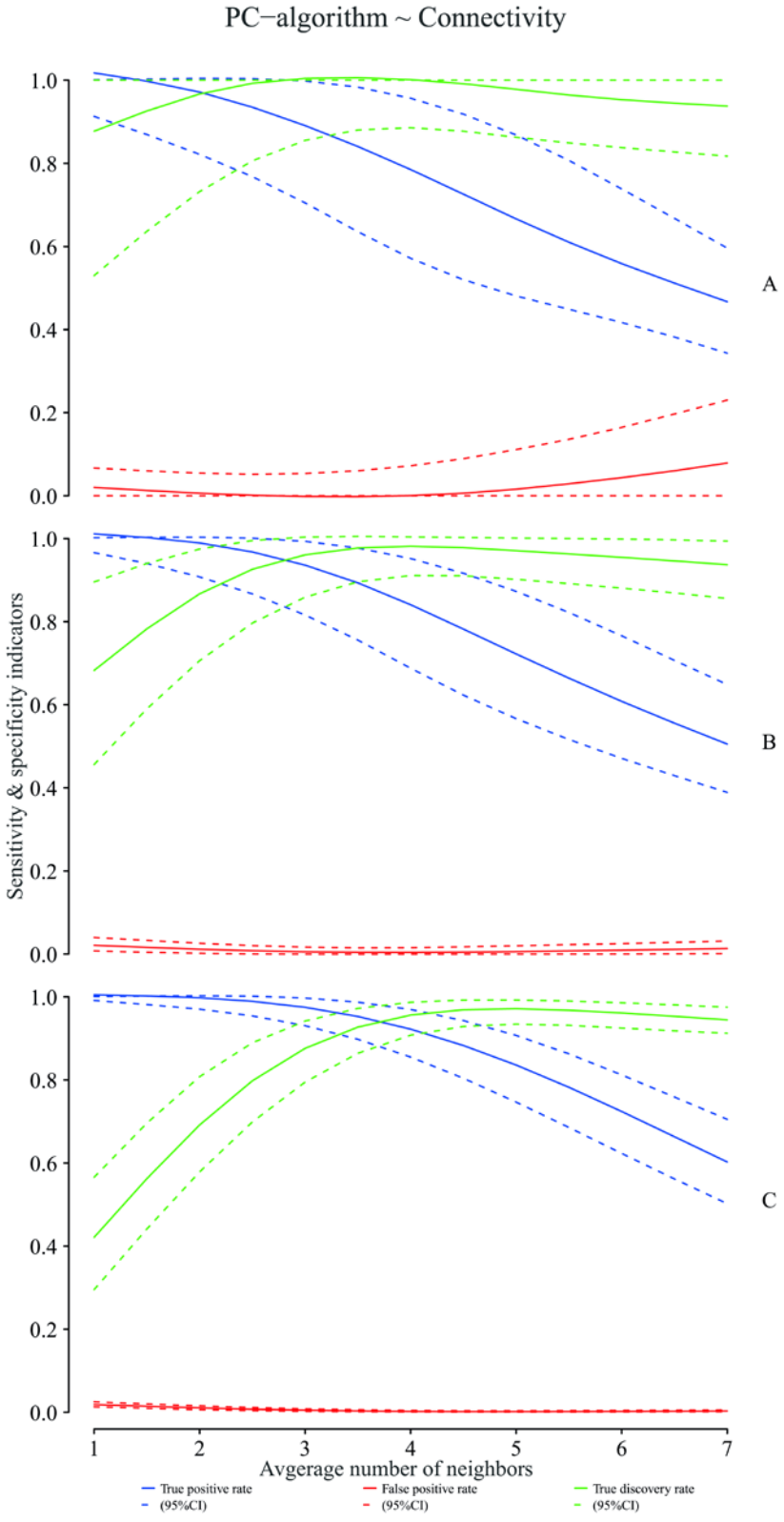


Figure 12: Performance of the PC-algorithm according to network size and density. For each simulation-run the average number of neighbors per node was fixed and 100 DAGs were randomly generated. Rates (95 % CI) = Median (2.5^{th}, 97.5^{th} percentile). The average number of neighbors per node was increased from 1 to 7 by 0.5, and spline functions were fitted to the data points. A: 11 nodes; B: 26 nodes; C: 81 nodes.

Effect strength

Figure 13 shows that in sparse networks with two neighbors on average (row I), sensitivity of the PC-algorithm to detect effects over 0.1 was excellent across all three simulated network sizes (column A = 11 nodes, column B = 26 nodes, column C = 81 nodes). True positive rates were consistently one ore close to one. In addition, in sparse settings, specificity for effects above 0.1 was very high, as indicated by false positive rates close to zero in simulated sparse scenarios over all network sizes. The fraction of true discoveries among all detected links was accordingly high. Only in large sparse networks with weak effects (C-I in Figure 13) a relevant fraction of false discoveries among all links were identified, which was indicated by moderate true discovery rates for effects below 0.5.

For networks with four neighbors on average (row II in Figure 13) good to excellent sensitivity was assessed over a wide range of effect strengths in the three evaluated network-sizes. Only in small networks (11 nodes) with strong effects (>0.5), the true positive rate dropped below 0.75. In general, there was a tendency to higher sensitivity for weak effects in smaller networks. Strong effects were more accurately detected in larger networks. False positive rates were very low with estimates of zero or very close to zero and true discovery rates were accordingly high.

In the densest networks, the PC-algorithm was considerably less sensitive for detect strong effects (row III in Figure 13), particularly in small networks (column A). The false positive rates and true discovery rates, however, were excellent. This suggested difficulty of the PC-algorithm to differentiate between direct and indirect effects in settings where each network-variable was strongly affected by a large fraction of the other network variables. The ability of the PC-algorithm to detect effects below 0.1 was only moderate. False positive rates were not affected by small effect sizes and were consistently zero or close to zero.

To summarize, the PC-algorithm showed excellent sensitivity and specificity for discovering the skeleton of the data-generating DAG for models with an average number of direct neighbors per node below four (smallest simulated network) or five (larger networks). In dense models (>5 neighbors per node on average), the sensitivity of the PC-algorithm was moderate. In large and very sparse networks (≤2 neighbors per node on average), despite high true positive and low false positive rates, a

relevant fraction of all discoveries was false. Good to excellent sensitivity of the PC-algorithm was observed for the detection of moderately weak effects (≥0.1), regardless of network density and also moderate to strong effects in moderately dense networks of moderate density (≤4 neighbors per node on average). Sensitivity of the PC-algorithm was moderate to detect medium to strong effects in very dense networks (7 neighbors per node on average), particularly, if this setting was modeled in small networks (11 nodes).

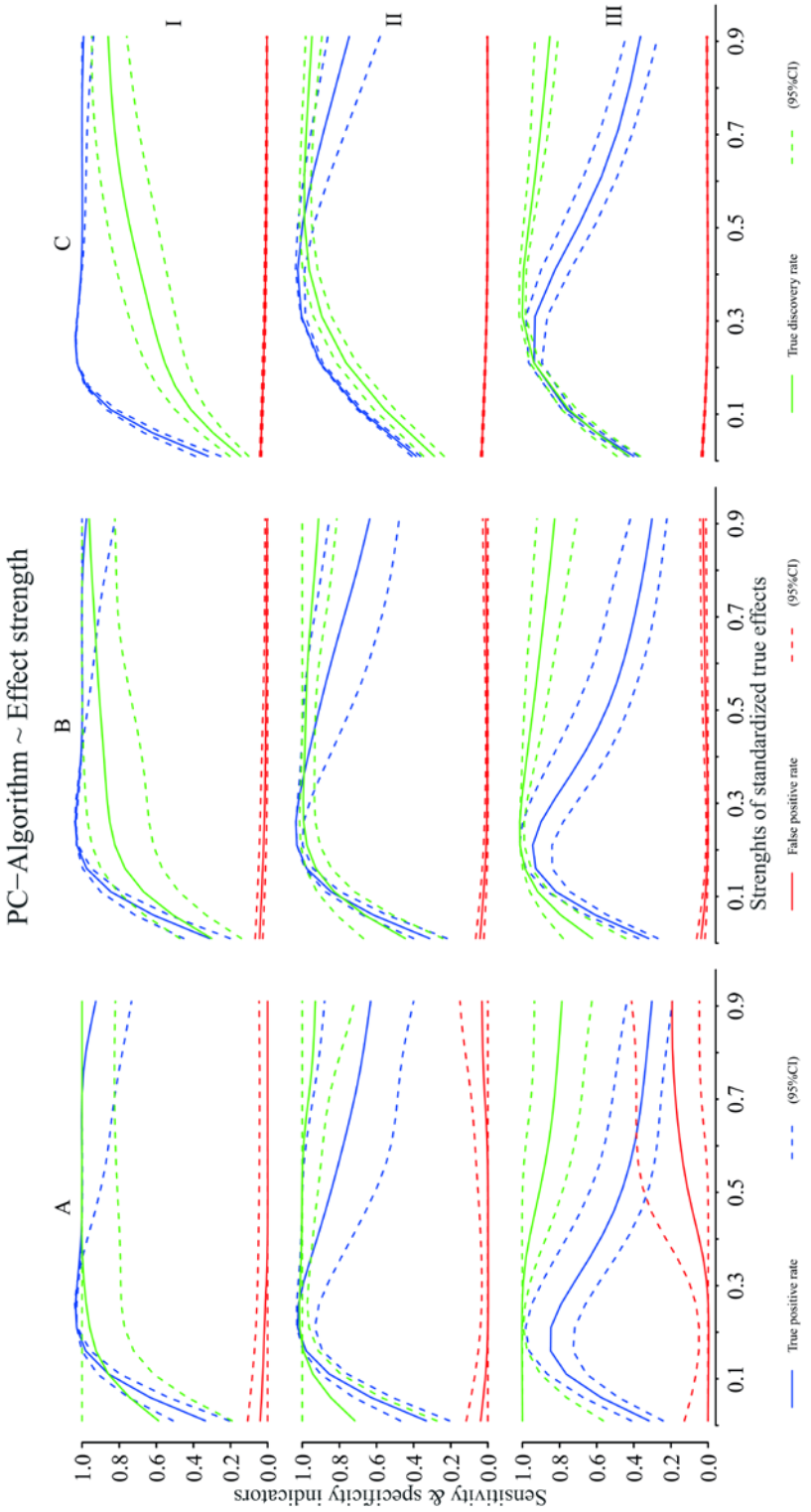


Figure 13: Performance of the PC-algorithm according to effect strengths and network size. For each simulation-run, model effects were fixed at the same level and 100 DAGs were randomly generated. Rates (95 % CI) are median (2.5th, 97.5th percentile) of these models. Effect strength was increased from 0.01 to 0.91 by 0.05 and spline functions were fitted to the resulting data points. Rows: *Network densities*– I = 2, II = 4, and III = 7 neighbors per node on average. Columns: *Network sizes*- A: 11 nodes; B: 26 nodes; C: 81 nodes.

Dependency of the PC-algorithm on the sample size

Figure 14 shows performance of the PC-algorithm according to the number of observations (*sample size*). True positive rates of close to one or one indicated excellent sensitivity of the PC-algorithm, whenever estimation procedures were based on 2000 or more observations. Furthermore, false positive rates were close to zero or zero regardless of sample size, and accordingly true discovery rates were consistently excellent. With considerably less than 2000 observations, however, sensitivity of the PC-algorithm markedly dropped. In a sample of 1000 observations, e.g., the true positive rates were only moderate in all simulated network sizes.

It should be noted that the impact of different settings of the significance threshold (α-level) was already evaluated in simulations [147].

Taken together, these simulations indicated applicability of the PC-algorithm to infer the equivalence class of data-generating structures (skeleton of DAGs) based on joint Gaussian distribution generated by one DAG of this class. Prerequisites for an excellent performance were network size between 11 and 81 variables, between 3 and 5 neighbors per node on average, moderate to strong model effects (0.1 to 0.7), and a sample size of at least 2000 observations. The sensitivity and specificity of the PC-algorithm to reveal the skeleton of the data-generating DAG was excellent in the range of parameters, which was expected for applications in the EPIC-Potsdam cohort study. Therefore, the PC-algorithm was considered a valid tool to estimate causal metabolomics networks in the current study.

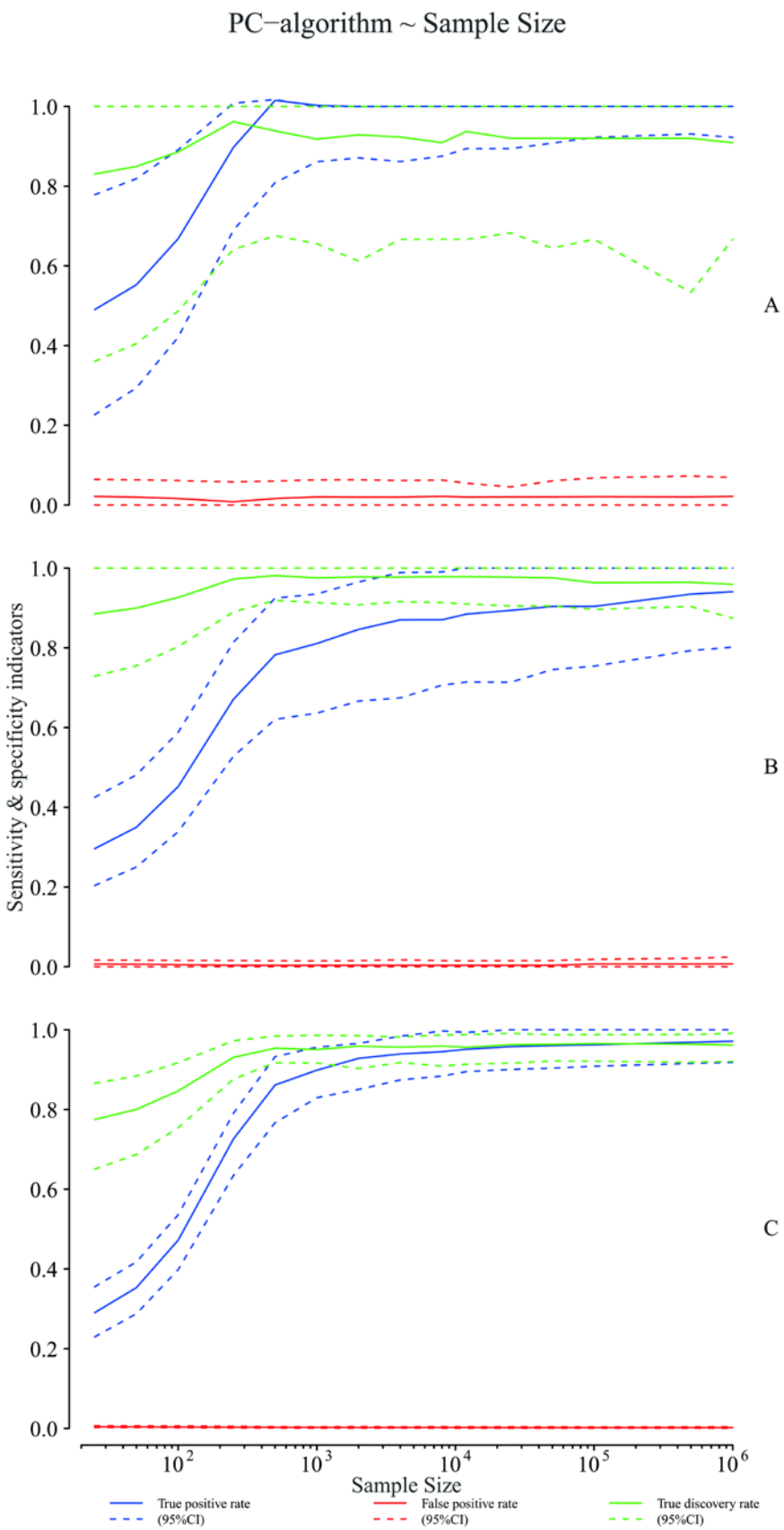


Figure 14: Performance of the PC-algorithm by *sample size* (number of simulated observations). Average neighbors per node = 4; effect strength randomly generated in the range from 0.2 to 0.7; *Network sizes:* A = 11, B = 26, and C = 81 nodes.

4.3 Metabolic links between habitual diet and type 2 diabetes risk: results from the EPIC-Potsdam cohort

4.3.1 Distributions, confounding structure, and covariance

Whole-grain bread

The median habitual consumption of whole-grain bread was 3.2g per MJ (IQR 0.9-8.4). The distribution of intake levels was right-tailed with over 10% non-consumers and the upper five percent with 19g whole-grain bread intake per MJ or above. Table 8 displays the distribution of potential confounders over five categories according to whole-grain bread intake (exposure). Categories were built by splitting the representative sample of the EPIC-Potsdam cohort (subcohort, n=2092 after exclusions) into subgroups at quintiles of the exposure-distribution. The resulting five groups were labeled Q1–Q5 (lowest to highest whole-grain bread intake).

The percentage of women was markedly higher in categories with higher whole-grain bread intake, and so was the percentage of participants with academic education. In categories of higher whole-grain bread intake, average daily energy intake was markedly lower and average red meat intake, percentage of smokers, and percentage of participants with hypertension was also lower. BMI was lowest in the highest category but comparable over the others; average coffee intake was lowest in the two highest categories; lipid-lowering medication was least frequent in categories with moderately high whole-grain bread consumption (Q3 and Q4).

Table 8: Confounding structure over categories according to whole-grain bread consumption

	Categories according to WGB consumption					
	Q1	Q2	Q3	Q4	Q5	ALL
Participants (n)	418	419	418	419	418	2092
WGB [g/MJ]	0.1 (0.3)*	1.2 (0.7)	3.2 (1.4)	7 (2.7)	14.8 (7.1)	3.2 (7.5)
Women	43%#	57%	63%	70%	77%	62%
Age	51.5 (15.4)	48 (14.5)	48 (15.3)	48.4 (15.9)	50.5 (16)	49 (15.6)
BMI [kg/m^2]	25.4 (5.7)	25.7 (5)	25.3 (4.8)	25.7 (5.3)	24.8 (5.2)	25.4 (5.2)
Sports [h/week]	4.5 (6.5)	4 (6)	5 (6)	5 (6)	4.5 (6.5)	4.5 (6)
Coffee [cups/day]	3 (2)	3 (2)	3 (2)	2 (2)	2 (3)	2 (2.5)
Red meat [g/MJ]	12.6 (7)	12.2 (5.9)	10.9 (5.8)	10.2 (7.1)	9.7 (6.4)	11.1 (6.7)
Total energy [MJ/day]	9 (3.9)	8.4 (4)	8.2 (3.3)	8.1 (3.1)	7.5 (2.6)	8.1 (3.4)
Alcohol: 10-40 g/day	33%	38%	33%	37%	32%	35%
Fasted	32%	29%	27%	29%	25%	28%
Education						
None or in training	5%	2%	2%	3%	3%	3%
Vocational training	41%	32%	36%	29%	30%	34%
Technical school	20%	25%	22%	25%	29%	24%
University	34%	41%	40%	42%	38%	39%
Smoking status						
Never smoker	37%	45%	46%	56%	53%	47%
Former smoker	36%	32%	33%	28%	32%	32%
Smoker <20 U/day	16%	17%	17%	12%	11%	15%
Smoker ≥20 U/day	11%	6%	4%	3%	4%	6%
Hypertension [Yes]	49%	48%	45%	46%	44%	46%
Lipid-lowering medication [Yes]	4%	5%	3%	2%	5%	4%
Antihypertensive medication [Yes]	15%	17%	16%	19%	18%	17%

*Median (IQR), all such values; #Column percentages, all such values.

Coffee

The median coffee intake was 2 cups per day (IQR: 1½-4). Over 25% of the participants had an average coffee consumption of approximately two cups per day, coffee intake of below half a cup per day, and one, three and four cups per day, were found in 10% to 20% of the participants, respectively. Higher coffee intake levels were less frequently observed. Five categories were built according to consumption of coffee in cups per day. Table 9 displays averages and frequencies of potential confounders within these groups. Fewer women and lower BMI were observed in the category of very high coffee consumption. Average daily energy intake and percentage of participants with moderate alcohol consumption and percentage of smokers were higher in participants who drank more coffee. Frequency of academic education was lower in participants who drank more coffee. Medication (lipid-lowering and antihypertensive) was least frequent among participants in the highest category of coffee consumption.

Table 9: Confounding structure over categories according to coffee consumption

	Categories according to coffee consumption[§]					
	≤1.5 Cup	2 Cup	3 Cups[ǀ]	4 Cups	≥5 Cups	All
Participants (n)	523	553	375	417	224	2092
Women	61%[#]	67.6%	60.8%	66.9%	42.9%	62%
Age	48.9 (16.2)*	48 (15.2)	47.7 (14.7)	52.8 (15.4)	48.7 (13.7)	49 (15.6)
BMI [kg/m^2]	25.1 (5.7)	25.3 (4.9)	25.3 (4.5)	25. 7 (5.8)	26.1 (5.1)	25.4 (5.2)
Sports [h/week]	4.5 (7)	4.5 (5.5)	4.5 (6)	5.5 (6.5)	4 (5.3)	4.5 (6)
Whole-grain bread [g/MJ]	3.6 (8.9)	3.6 (7.7)	2.6 (6.1)	3.3 (7.1)	1.9 (5.7)	3.2 (7.5)
Red meat [g/MJ]	10.1 (6.4)	11.1 (6.7)	11.7 (6.8)	11.4 (6.1)	11.7 (7.7)	11.1 (6.7)
Total energy [MJ/day]	7.9 (3.3)	7.9 (3)	8.3 (3.4)	8.1 (3.3)	9.5 (4.5)	8.1 (3.4)
Alcohol: 10-40 g/day	30%	33 %	35%	40%	37%	35%
Fasted	28%	29%	26%	32%	25%	28%
Education						
None or in training	4%	3%	3%	2%	2%	3%
Vocational training	29%	35%	33%	35%	38%	34%
Technical school	25%	22%	23%	27%	24%	24%
University	42%	39%	41%	36%	36%	39%
Smoking status						
Never smoker	53%	50%	47%	50%	28%	47%
Former smoker	31%	36%	30%	30%	33%	32 %
Smoker <20 U/day	11%	12%	18%	14%	24%	15%
Smoker ≥20 U/day	4%	2%	6%	6%	16%	6%
Hypertension [Yes]	46%	48%	46%	44%	47%	46%
Lipid-lowering medication [Yes]	3%	5%	5%	4%	2%	4%
Antihypertensive medication [Yes]	19%	17%	15%	19%	11%	17%

[§]Coffee consumption in cups per day; [ǀ]category contains participants that reported consumption of 3.5 cups/day (≈50%); [#]Column percentages, all such values; *Median (IQR), all such values.

Red Meat

The median red meat intake was 11.1 g/MJ (IQR 8.0-14.7). Energy-standardized habitual red meat consumption was approximately normally distributed. Categories according to habitual red meat consumption were built as described for whole-grain bread above. Fewer women, lower average sportive activity and consumption of whole-grain bread and lower percentage of academic education were observed in higher categories of red meat consumption. Average daily energy intake and frequencies of moderate alcohol consumption and smoking were higher in categories of higher red meat intake. Coffee consumption was highest in the two highest categories of red meat consumption.

Table 10: Confounding structure over categories according to total meat consumption

	Categories according to red meat consumption					
	Q1	Q2	Q3	Q4	Q5	ALL
Participants (n)	418	419	418	419	418	2092
Red meat [g/MJ]	5.6 (2.4)	8.7 (1.3)	11.1 (1.2)	13.8 (1.7)	18.8 (4.2)	11.1 (6.7)
Age	48.4 (16.7)[#]	50.3 (15.3)	51.4 (16.6)	48.5 (16)	47.8 (12.5)	49 (15.6)
Women	73%[§]	64%	63%	53%	57%	62%
BMI [kg/m^2]	25.0 (5.3)	24.9 (5.2)	25.2 (4.7)	25.5 (5.4)	26.2 (5.4)	25.4 (5.2)
Sports [h/week]	5.0 (5.5)	5.0 (7.5)	4.5 (6.5)	4.5 (6.0)	4.0 (6.0)	4.5 (6.0)
Whole-grain bread [g/MJ]	5.6 (11.1)	3.8 (7.5)	3 (7)	1.9 (5.5)	2.2 (5.8)	3.2 (7.5)
Coffee [cups/day]	2 (2)	2 (3)	2 (2)	3 (2)	3 (2)	2 (2.5)
Total energy [MJ/day]	7.7 (3.1)	8.2 (3.2)	7.9 (3.1)	8.4 (3.5)	8.5 (3.9)	8.1 (3.4)
Fasted	30%	28%	28%	28%	27%	28%
Alcohol: 10-40 g/day	26%	35%	34%	39%	40%	35%
Education						
None or in training	3%	3%	4%	3%	3%	3%
Vocational training	30%	33%	35%	33%	37%	34%
Technical school	25%	22%	25%	24%	24%	24%
University	42%	42%	36%	41%	36%	39%
Smoking status						
Never smoker	52%	51%	48%	45%	41%	47%
Former smoker	33%	31%	32%	32%	33%	32%
Smoker <20 U/day	11%	13%	15%	16%	18%	15%
Smoker ≥20 U/day	3%	5%	5%	7%	8%	6%
Hypertension [Yes]	42%	43%	47%	48%	52%	46%
Lipid-lowering medication [Yes]	4%	3%	4%	4%	5%	4%
Antihypertensive medication [Yes]	18%	16%	17%	15%	19%	17%

[#]Median (IQR), all such values; [§]Column percentages, all such values.

Distribution and covariance of metabolites

Distributions of metabolite serum concentrations in the EPIC-Potsdam subcohort have been described before [116]. Figure 15 plots the correlation (lower-left triangle) and the partial correlation (upper right triangle) structure among lysophosphatidylcholines, which were overall strongly intercorrelated. Strongest partial correlations, however, were observed between pairs of metabolites for which the chain-length of the fatty acid residue differed by two carbon atoms (e.g. C16:0 and C18:0, or C16:1 and C18:1) or between pairs of metabolites for which bound fatty acids differed by one desaturation (e.g. C18:1 and C18:2, or C20:3 and C20:4). Comparable observations were made in the other metabolite groups and the corresponding correlation-partial correlation plots are shown in the Annex (Supplementary Figure 1 and Supplementary Figure 2, 8.1).

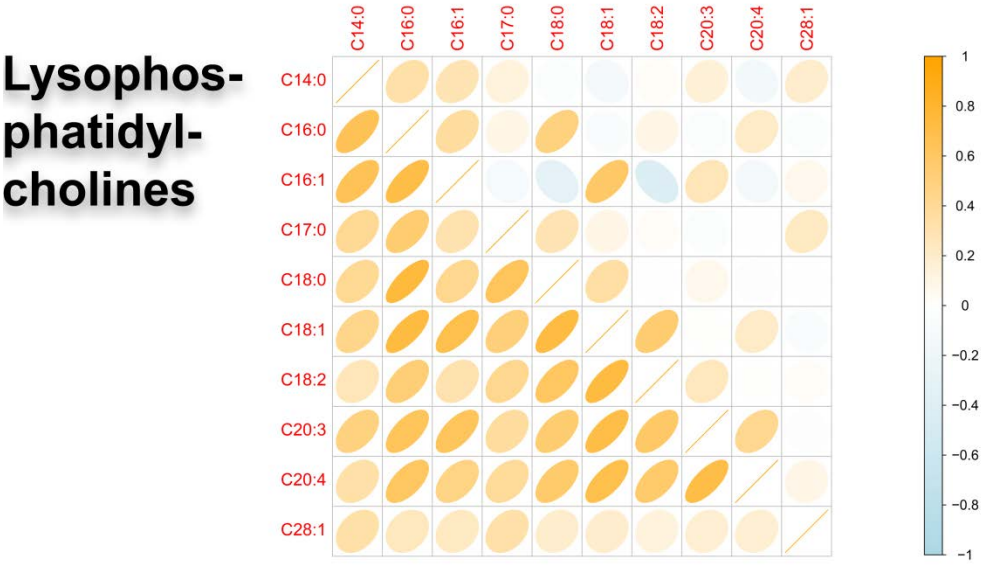


Figure 15: Correlation (below the diagonal) and partial correlation (adjusted for all other metabolites within the group, above the diagonal) among lysophosphatidyl-cholines; It should be noted that analogous plots for other metabolite-groups are shown in the Annex (8.1).

Type 2 diabetes

After exclusions 53 cases (29 men and 24 women) of type 2 diabetes occurred within the subcohort during a median follow-up time of 6.6 years (IQR 6.1-8.7). From the total study population 692 incident type 2 diabetes cases were considered after exclusions (406 men and 286 women), with 639 of those external to the subcohort (377 men and 262 women).

4.3.2 Common variation among metabolites of the same group

Factor analysis within metabolite classes

The first factor derived within the group of 14 amino acids had an Eigenvalue of 23.8 and explained 84% of the variance within the metabolite group. All amino acids except glycine (0.3) loaded above 0.5 on factor 1. The highest loadings were contributed by aromatic and branched-chain amino acids (tyrosine, phenylalanine, tryptophan and valine, leucine/isoleucine) and methionine (all loadings >0.75).

The first factor derived within the group of 17 acylcarnitines had an Eigenvalue of 18.7 and explained 71% of the variance within the metabolite group. The majority of acylcarnitines loaded above 0.5 on factor 1, and only 2 out of 17 acylcarnitines, i.e. carnitine (0.15) and propionylcarnitine (C3) (0.09) had factor loadings below 0.4.

The first factor derived within the group of 14 sphingomyelins had an Eigenvalue of 101 and explained 80% of the variance within the metabolite group. Loadings on factor 1 were above 0.7 for all sphingomyelins, except for sphingomyelin C 20:2 with a loading of 0.22.

The first factor derived within the group of 10 lysophosphatidylcholines had an Eigenvalue of 19.5 and explained 87% of the variance within the metabolite group. All lysophosphatidylcholines loaded above 0.5 on factor 1, with the exception of lysophosphatidylcholines C 28:1 which displayed a factor loading of 0.26.

The first factor derived within the group of 34 diacyl phosphatidylcholines had an Eigenvalue of 281 and explained 58% of the variance within the metabolite group. The majority of diacyl phosphatidylcholines loaded above 0.5 on factor 1, and only 3 out of 34 diacyl phosphatidylcholines had factor loadings below 0.4 (diacyl phosphatidylcholines C 38:1, C 42:0, and C 42:1 with factor loadings of 0.15, 0.26 and 0.35, respectively).

The first factor derived within the group of 37 alkyl-acyl phosphatidylcholines had an Eigenvalue of 198 and explained 61% of the variance within the metabolite group. All alkyl-acyl phosphatidylcholines except alkyl-acyl phosphatidylcholine C 30:0 (0.3) loaded above 0.5 on factor 1. Scree plots used to evaluate the appropriateness of the one factor solution are shown in the Annex (Supplementary Figure 3, 8.2).

Association of dietary exposures with group factors

Habitual consumption of whole-grain bread was associated with significantly lower scores in the diacyl phosphatidylcholine factor 1 (Table 11). Habitual coffee consumption was associated with higher scores in the amino acid factor 1 and diacyl phosphatidylcholine factor 1 but with higher scores in the alkyl-acyl phosphatidylcholine factor 1. Higher red meat consumption was associated with higher scores in the alkyl-acyl phosphatidylcholine factor 1.

Table 11: Association of dietary exposures with metabolite group factors

	Whole-grain bread		Coffee		Red meat	
Factor 1	*Estimate*	*fdr P*	*Estimate*	*fdr P*	*Estimate*	*fdr P*
Amino acids	0.028	0.826	-0.067	0.004	0.007	0.776
Acylcarnitines	0.010	0.826	-0.039	0.115	0.037	0.164
Sphingomyelins	-0.002	0.940	0.024	0.206	-0.033	0.164
Lyso-PCs	-0.012	0.826	0.030	0.178	0.006	0.776
Diacyl PCs	-0.050	0.030	-0.063	<.001	-0.032	0.144
Alkyl-acyl PCs	0.008	0.826	0.052	0.001	0.042	0.034

Variance-standardized betas (*Estimate*) and false discovery corrected p-values (*fdr P*) for an association of dietary exposures with the first common factor within each metabolite class; Estimates were derived from a linear regression model comprehensively adjusted for significantly correlated other metabolite factors and for age, sex, BMI, lifestyle, diet, fasting status at blood draw occasion, and prevalence of hypertension and medication. PC: phosphatidylcholine.

Association of group factors with type 2 diabetes incidence

An elevated risk of type 2 diabetes was observed in relation to higher scores of the amino acid factor 1 with a hazard ratio of 1.24 (95% CI 1.10, 1.41) per standard deviation. An elevated diabetes risk was also found in relation to higher scores in the diacyl phosphatidylcholine factor 1 with a hazard ratio of 1.44 (95% CI 1.23, 1.69) per standard deviation (Table 12). Reduced risk of type 2 diabetes was associated with higher scores in the alkyl-acyl phosphatidylcholine factor 1 (hazard ratio per standard deviation 0.65, 95%CI 0.54, 0.78).

Table 12: Association of metabolite group factors with the risk of developing type 2 diabetes

	Type 2 diabetes incidence	
Factor 1	*HR (95% CI)*	*fdr P*
Amino acids	1.24 (1.10, 1.41)	0.001
Acylcarnitines	1.01 (0.90, 1.15)	0.826
Sphingomyelins	1.05 (0.93, 1.19)	0.531
Lysophosphatidylcholines	0.87 (0.77, 0.99)	0.059
Diacyl phosphatidylcholines	1.44 (1.23, 1.69)	<.001
Alkyl-acyl phosphatidylcholines	0.65 (0.54, 0.78)	<.001

Hazard Ratios (*HR*) per standard deviation in the factor score with 95% Confidence Intervals (*CI*) and corresponding false discovery corrected p-values (*fdr P*) indicate the relative risk of developing type 2 diabetes in relation to the first common factor within each metabolite group; Hazard ratios were derived from a Cox proportional hazard regression model comprehensively adjusted for significantly correlated other metabolite factors and for age, sex, BMI, lifestyle, diet, fasting status at blood draw, prevalence of hypertension and medication.

4.3.3 Linking diet and diabetes incidence to metabolite networks

Amino acids

The plasma amino acid network was sensitive to habitual consumption of coffee and of red meat, however, not sensitive to whole-grain bread consumption (Figure 17). One out of eight associations between dietary exposures and amino acids was classified as direct effect. Five amino acids were linked to type 2 diabetes risk. Four of these amino acids were unambiguously classified as direct effects based on the multi-model procedure.

Acylcarnitines

The plasma acylcarnitine network was affected by red meat and by whole-grain bread consumption, but no evidence for an effect of coffee consumption on acylcarnitine levels was found (Figure 16). Out of four remaining diet-acylcarnitine links one was classified as direct effect, and the others remained ambiguous. Five acylcarnitines were related to type 2 diabetes risk. Four of these links were classified as direct effects.

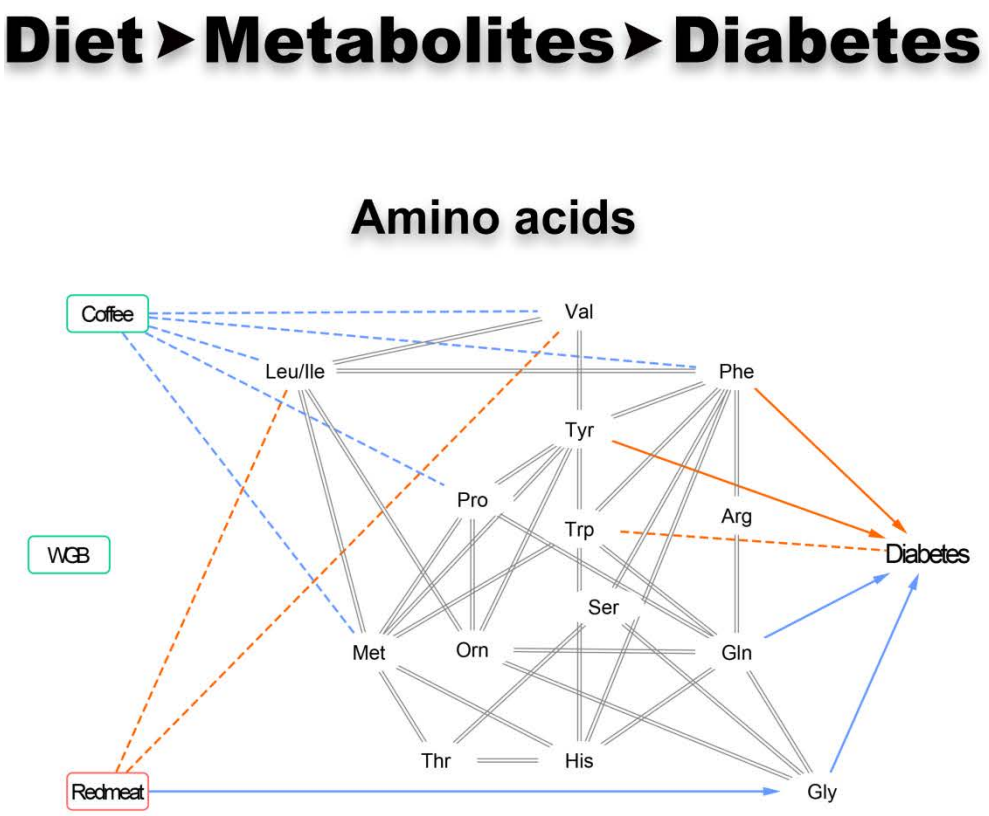


Figure 17: Joint network: diet, diabetes risk, and amino acids

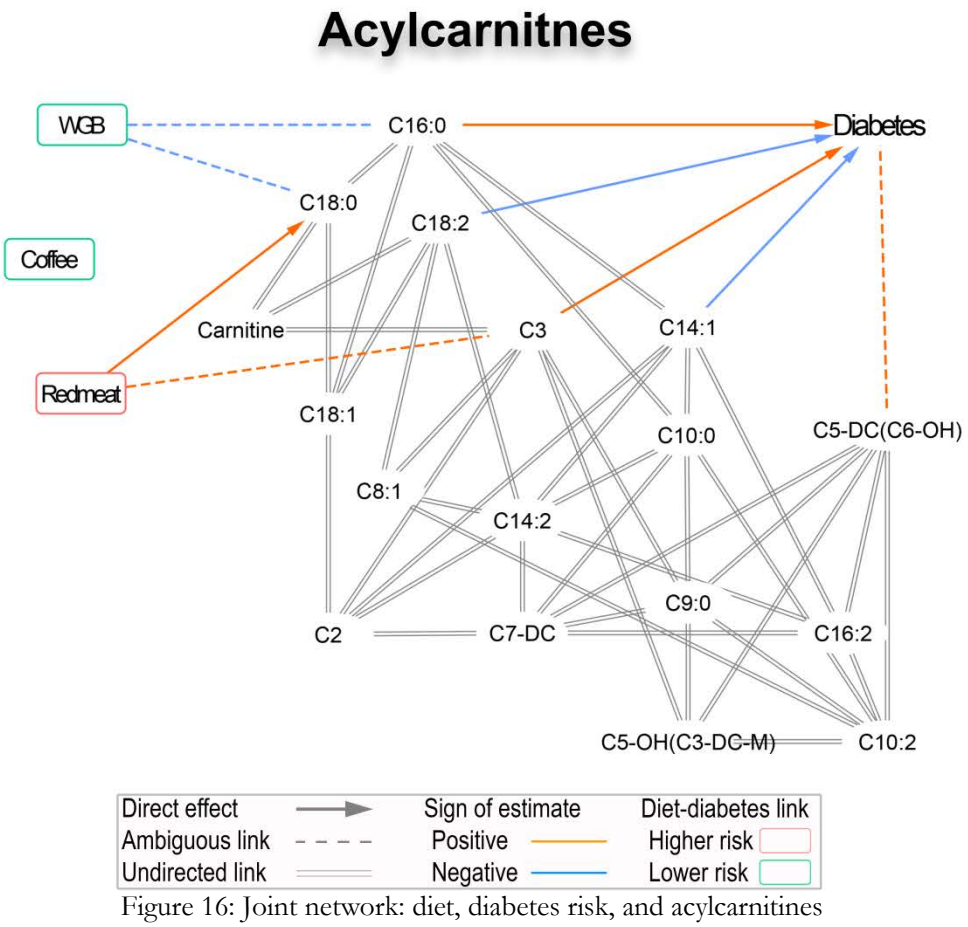


Figure 16: Joint network: diet, diabetes risk, and acylcarnitines

Sphingomyelins

The plasma sphingomyelin network was sensitive to habitual consumption of coffee and red meat, but not to whole-grain bread consumption (Figure 19). Five out of seven links between dietary exposures and plasma sphingomyelin concentrations were unambiguously classified as direct effects. Seven sphingomyelins were network-independently related to type 2 diabetes risk and were therefore classified as direct effects.

Lysophosphatidylcholines

The lysophosphatidylcholine-network was sensitive to all three dietary exposures, i.e. to consumption of whole-grain bread, consumption of coffee, and consumption of red meats (Figure 18). Two out of nine links were unambiguous and thus classified as direct effects. Six lysophosphatidylcholines were linked to type 2 diabetes risk. All six were classified as direct effects based on the multi-model-estimates.

Diacyl phosphatidylcholines

The diacyl phosphatidylcholine-network was also linked to consumption of whole-grain bread, of coffee, and of red meat (Figure 20). Seventeen links between diacyl phosphatidylcholine and one of the dietary exposures were detected. Three of these links were classified as direct effects. Ten diacyl phosphatidylcholines were linked to type 2 diabetes risk. Five of these links were classified as direct effects based on unambiguous multi-model information.

Alkyl-acyl phosphatidylcholines

The alkyl-acyl phosphatidylcholine-network was similarly sensitive to all three dietary exposures (Figure 21). Twelve links of the investigated foods and alkyl-acyl phosphatidylcholine serum concentrations were present in the joint network of which two were classified as direct effects. Ten alkyl-acyl phosphatidylcholines were linked to type 2 diabetes risk. Nine of these links were classified as direct effect according to consistency of the multi-model estimates.

Diet ➤ Metabolites ➤ Diabetes

Sphingomyelins

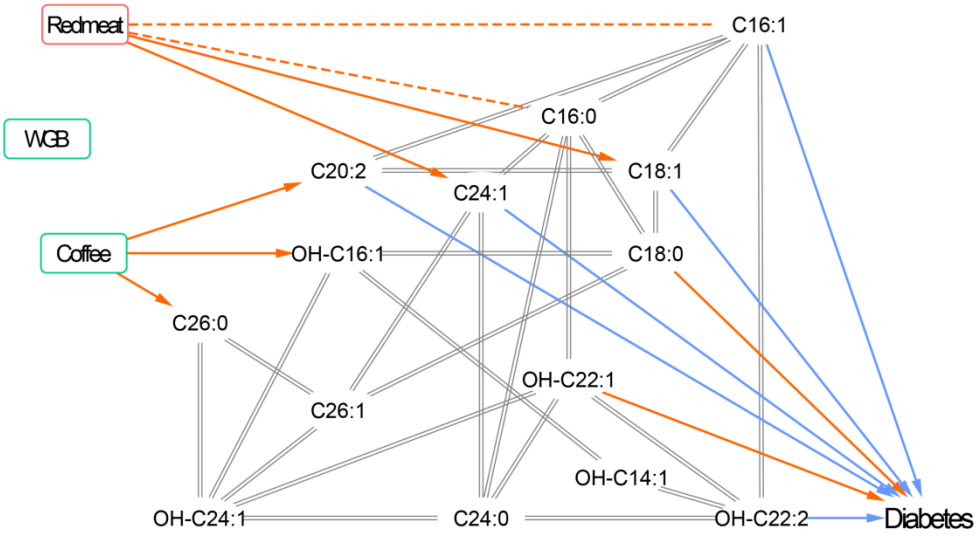
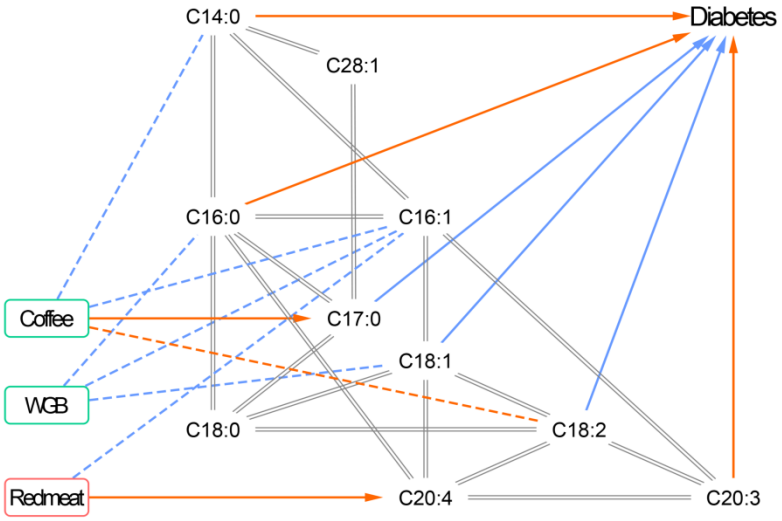


Figure 19: Joint network: diet, diabetes risk, and sphingomyelins

Lysophosphatidycholines



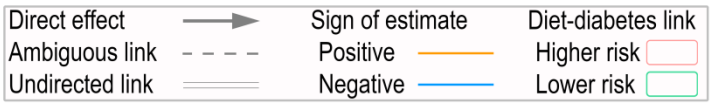


Figure 18: Joint network: diet, diabetes risk, and lysophosphatidylcholines

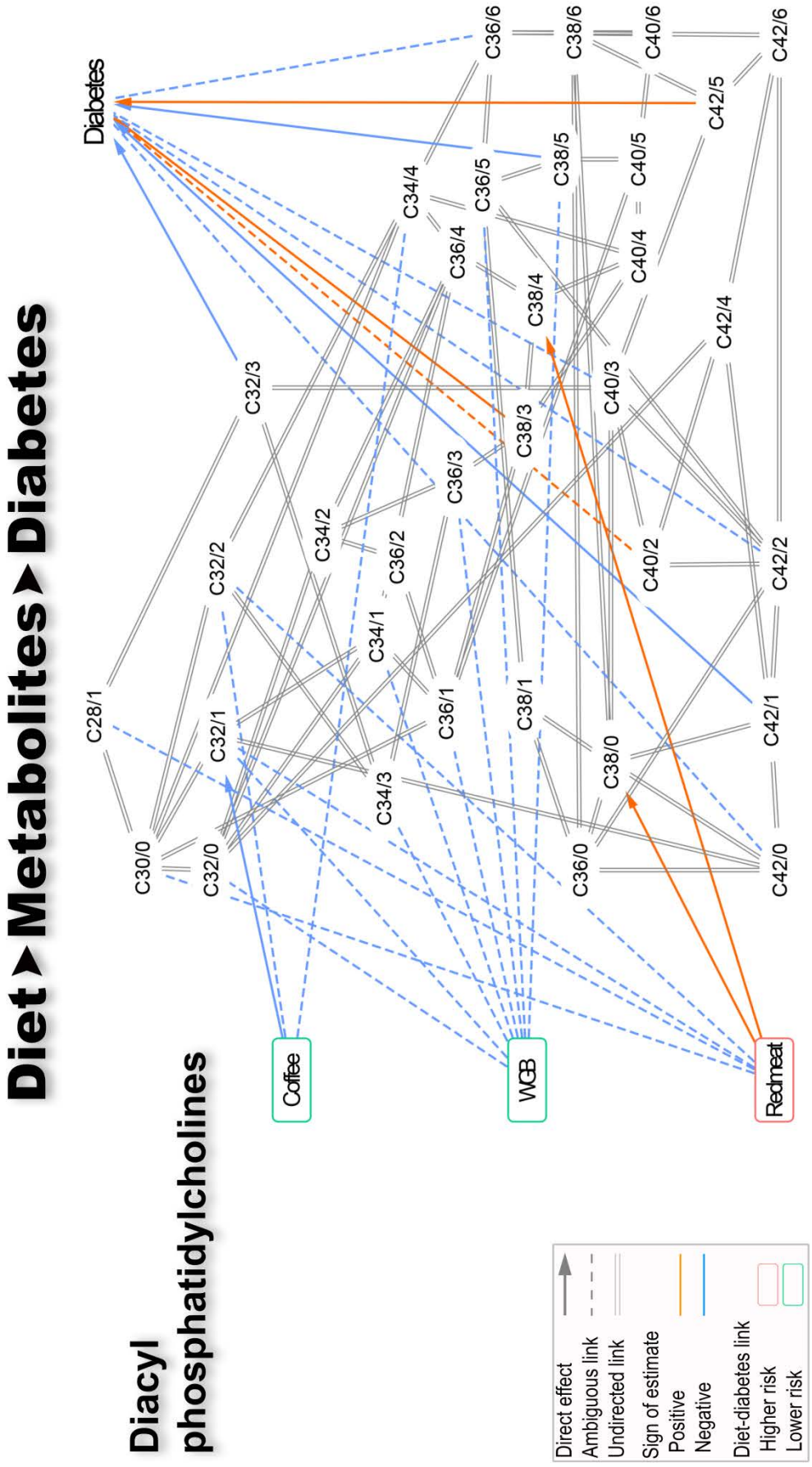


Figure 20: Joint network: diet, diabetes risk, and diacyl phosphatidylcholines

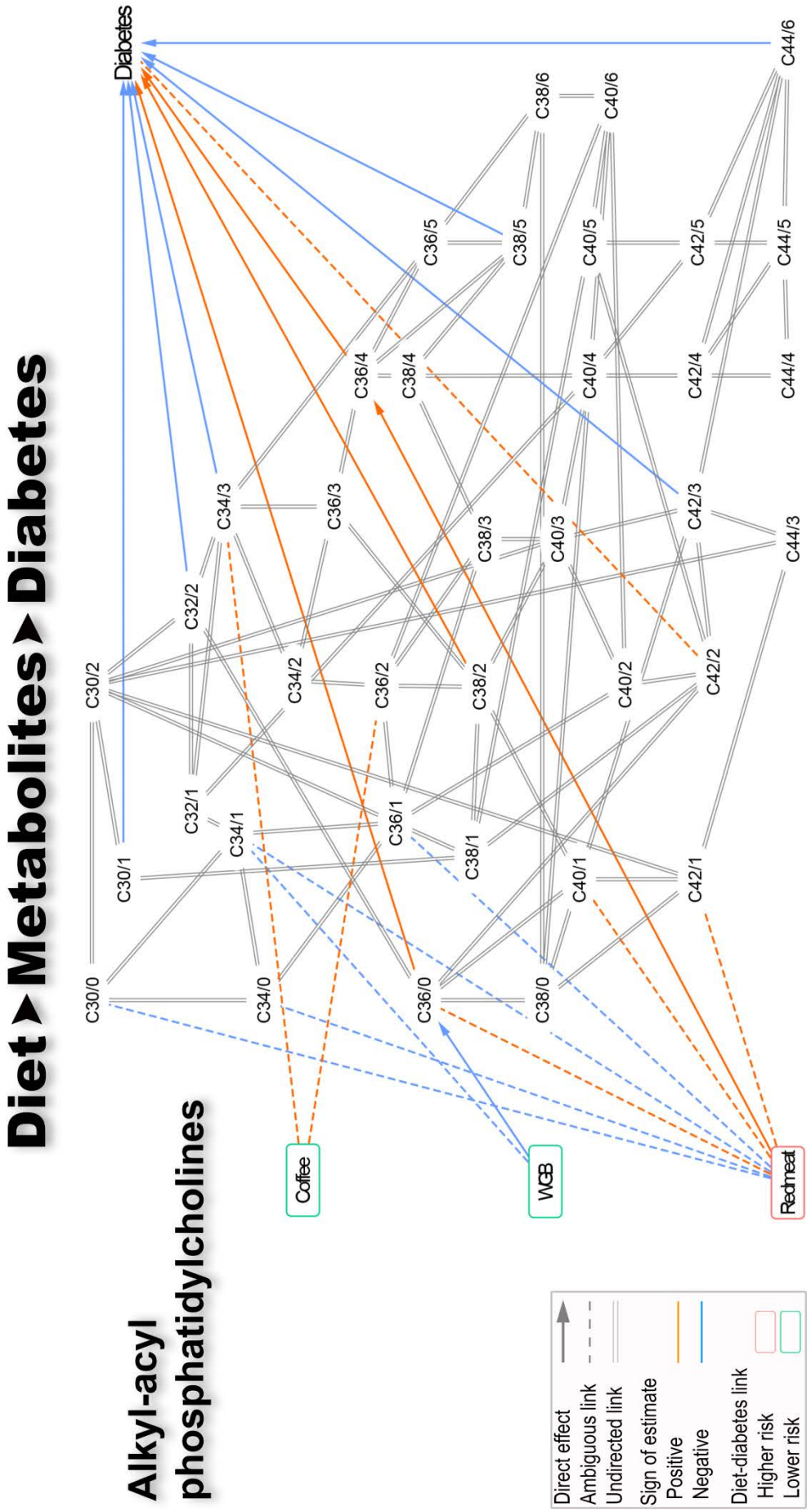


Figure 21: Joint network: diet, diabetes risk, and alkyl-acyl phosphatidylcholines

4.3.4 Direct effects of dietary exposures on metabolites

Digital only supplemental material

Summary information on the multi-model procedure for all diet-metabolite links and comprehensive information on all submodels is provided as *digital only supplemental material* on the accompanying CD.

Whole-grain bread

Habitual whole-grain bread consumption was related to lower levels of various lipid metabolites from different lipid classes (Table 13). Lower serum concentrations of saturated long-chain fatty acid-containing metabolites were observed among acylcarnitines [i.e. palmitoylcarnitine (C16:0) and stearoylcarnitine (C18:0)]; for lysophosphatidylcholine C16:0; for diacyl phosphatidylcholine C32/0; and for alkyl-acyl phosphatidylcholine C36/0. Furthermore, the diacyl phosphatidylcholines C32/1, C34/1 and C36/1, and alkyl-acyl phosphatidylcholine C34/1 contained one saturated fatty acid along with one monounsaturated fatty acid and were also lower concentrated in participants with higher whole-grain bread consumption. Lower serum concentrations in relation to higher whole-grain bread consumption were also observed for the monounsaturated fatty acid-containing lysophosphatidylcholines C16:1 and C18:1. In addition the lower concentrations of diacyl phosphatidylcholines C34/3, C36/3, C36/5, and C38/5 implicated lower abundance of specific polyunsaturated fatty acids in relation to higher consumption of whole-grain bread in this lipid compartment.

Figure 22 shows whole-grain bread connected components extracted from the joint networks. Substructures illustrate that the effects of whole-grain bread on saturated and monounsaturated were interlinked. However, most of the links between whole-grain bread consumption and lipid metabolites could not be unambiguously classified based on multi-model information according to predefined criteria. Consequently, the algorithm did not resolve the exact entry point(s) into the network, i.e. did not differentiate between direct and indirect effects within the identified whole-grain-connected components. An exception was the direct effect of whole-grain bread consumption to lower levels of alkyl-acyl phosphatidylcholine C36/0, which was not explainable by any other alteration in the alkyl-acyl phosphatidylcholine network and thus classified as direct effect. Subnetworks further

indicated that the suggested effects of whole-grain bread consumption on polyunsaturated fatty acids in diacyl phosphatidylcholines was not directly linked with the effects of whole-grain on saturated and monounsaturated fatty acid-containing metabolites within that lipid compartment.

Table 13: Multi-model inference on possible effects of whole-grain bread consumption on metabolite-levels

	Whole-grain bread consumption			
	Estimate range (betas)	*fdr P*	*Upper P*	*Lower P*
Acylcarnitine				
C16:0	-0.13 (-0.15, -0.02)	9.7e-02	4.7e-01	6.2e-04
C18:0	-0.14 (-0.14, -0.05)	9.7e-02	1.9e-01	7.4e-03
Lysophosphatidylcholine				
C16:0	-0.13 (-0.16, -0.04)	7.4e-02	3.2e-01	2.0e-04
C16:1	-0.12 (-0.12, -0.02)	7.4e-02	5.0e-01	2.0e-02
C18:1	-0.12 (-0.14, -0.04)	7.4e-02	3.5e-01	1.3e-04
Diacyl phosphatidylcholine				
C36/1	-0.26 (-0.26, -0.06)	3.1e-05	5.2e-02	1.1e-09
C34/1	-0.24 (-0.24, -0.04)	4.1e-05	1.7e-01	2.4e-06
C32/1	-0.18 (-0.18, 0.03)	5.4e-03	7.2e-01	4.8e-04
C36/5	-0.15 (-0.15, -0.04)	2.7e-02	2.3e-01	1.5e-03
C32/0	-0.15 (-0.15, 0.05)	3.9e-02	9.1e-01	8.2e-04
C38/5	-0.14 (-0.14, -0.01)	4.2e-02	6.3e-01	7.4e-03
C36/3	-0.14 (-0.14, -0.01)	5.0e-02	5.5e-01	2.7e-03
C34/3	-0.13 (-0.13, -0.01)	7.1e-02	7.1e-01	1.9e-03
Alkyl-acyl phosphatidylcholine				
C36/0	-0.19 (-0.19, -0.11)	1.0e-02	4.8e-03	1.1e-04
C34/1	-0.16 (-0.16, -0.04)	4.7e-02	7.3e-02	2.5e-03

Summary of standardized estimates from multiple linear regression models (consumption of whole-grain bread as exposure and concentration of the single metabolite as outcome), with single models corresponding to adjustment for a specific subset of direct neighbors of the respective metabolite in the subgroup-specific metabolite-network; Metabolite concentrations were standardized on age, sex, BMI and prevalence of hypertension, and all models were comprehensively adjusted for lifestyle, diet, fasting status at blood draw occasion, and medication.

Estimate range (betas) indicates the effect of 2 standard deviation higher intake of energy-standardized whole-grain bread on the variance standardized metabolite concentrations (lowest estimate, highest estimate from the multi-model procedure); *fdr P*= false discovery rate corrected p-value (based on the model adjusted for external confounders but not for other network-variables); *upper/lower P* = highest and lowest p-values from the multi-model procedure.

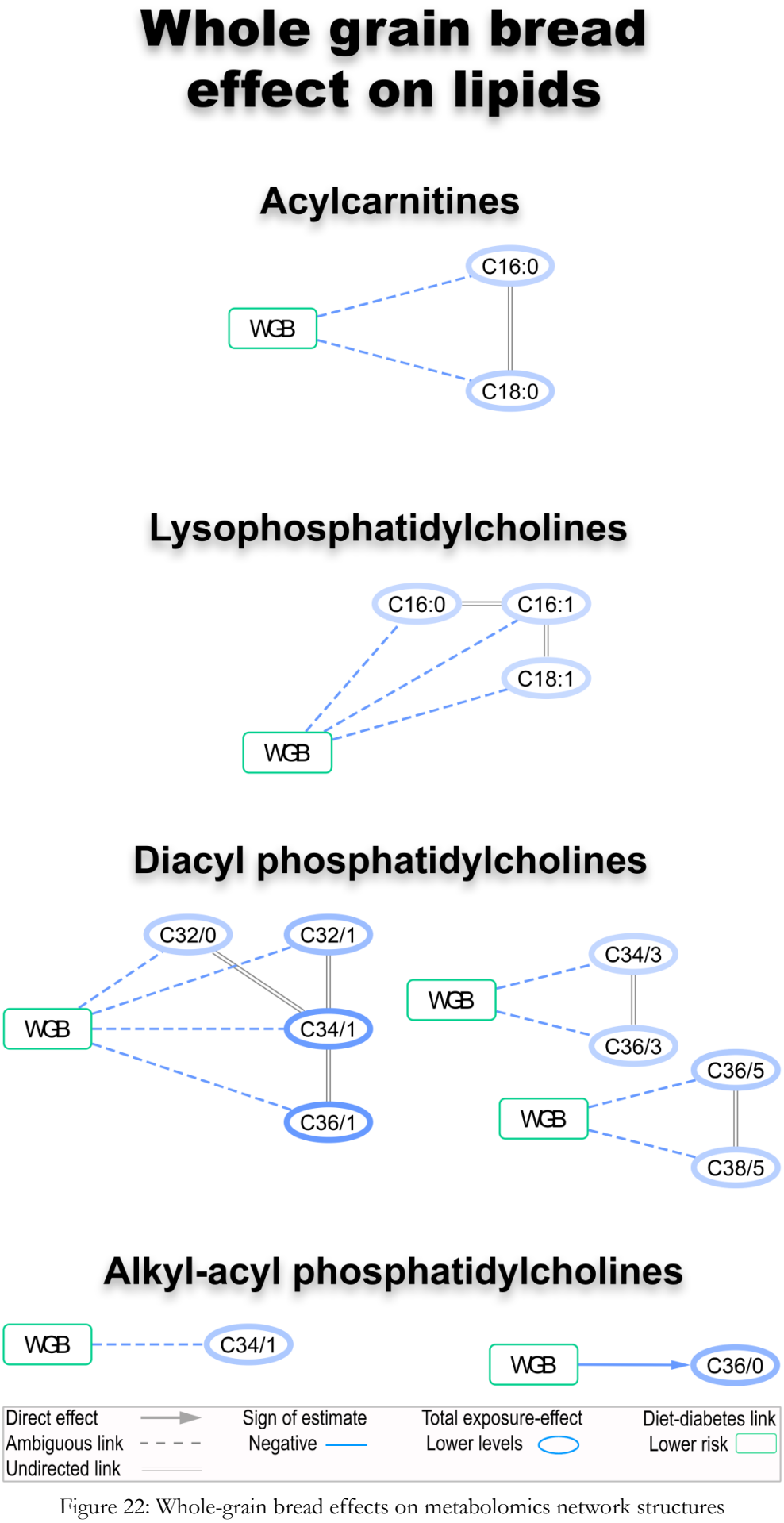


Figure 22: Whole-grain bread effects on metabolomics network structures

Coffee

Habitual coffee consumption was linked to lower serum concentrations of several amino acids, and to alterations of the concentration of metabolites in various lipid compartments (Table 14). Among the amino acids, higher coffee consumption was related to lower concentrations of the branched-chain amino acids valine and leucine/isoleucine, the aromatic amino acid phenylalanine, and of methionine and proline. The extracted subnetwork showed that all coffee-related amino acids were interlinked, i.e. they belonged to a single coffee-connected component (Figure 23). None of the coffee-related amino acids was unambiguously classified as direct effect based on the multi-model estimates.

The effects of habitual coffee consumption on lipid composition differed between lipid compartments. Frequent coffee intake was related to higher levels of eight sphingomyelins, three of which were classified as directly affected based on multi-model information. The effect of coffee intake on higher sphingomyelin C26:0 concentrations was not directly linked to other coffee related alterations in the sphingomyelin network. Other coffee-affected sphingomyelins, however, were interlinked (i.e., they formed one coffee-connected component). The direct effects on hydroxy-sphingomyelin C16:1 and sphingomyelin C20:2 fully explained the coffee-association of the other sphingomyelins within the connected component rendering them indirect effects according to the preset criteria (Table 14). Marginal associations of hydroxy-sphingomyelins OH-C22:2 and OH-C22:1, and sphingomyelins C16:0, C18:0 and C18:1 were explained by considering direct effects within the connected component, i.e. adjusting for hydroxy-sphingomyelin C16:1 and sphingomyelin C20:2.

Within the phosphatidylcholine compartments, coffee intake was related to lower serum concentrations of lysophosphatidylcholines C14:0 and C16:1, and to lower concentrations of diacyl phosphatidylcholines C32/1, C32/2 and C34/3. Diacyl phosphatidylcholine C32/1 was classified as direct effect but was a singleton. The other two were part of the same connected component but remained ambiguous. Lysophosphatidylcholines C17:0 and C18:2, and alkyl-acyl phosphatidylcholine C34/3 and C36/2 were higher in relation to higher coffee consumption. These potential coffee effects were not interlinked and margaric acid (C17:0) enrichment in lysophosphatidylcholines was, therefore, classified as direct effect.

Table 14: Multi-model inference on possible effects of coffee-consumption on metabolite-levels

	Coffee consumption			
	Estimate range (betas)	*fdr P*	*Upper P*	*Lower P*
Amino Acid				
Met	-0.12 (-0.12, -0.00)	4.8e-02	9.9e-01	3.9e-03
Pro	-0.12 (-0.13, -0.06)	4.8e-02	1.6e-01	1.4e-03
Val	-0.11 (-0.11, 0.00)	4.8e-02	9.0e-01	1.4e-02
Leu/Ile	-0.13 (-0.15, -0.02)	4.8e-02	3.2e-01	8.3e-05
Phe	-0.10 (-0.12, -0.01)	6.0e-02	6.7e-01	2.0e-03
Sphingomyelin				
OH-C16:1	0.15 (0.06, 0.15)	7.2e-03	3.6e-03	7.8e-06
C20:2	0.12 (0.09, 0.12)	3.1e-02	4.6e-02	6.6e-03
C26:0	0.11 (0.07, 0.11)	4.9e-02	3.1e-02	6.7e-03
OH-C22:2	-0.04 (-0.04, 0.00)	4.8e-01	9.8e-01	9.5e-02
OH-C22:1	-0.03 (-0.03, 0.02)	4.8e-01	8.0e-01	1.7e-01
C16:0	-0.03 (-0.03, 0.02)	4.8e-01	9.3e-01	2.2e-01
C18:0	-0.03 (-0.03, -0.01)	4.8e-01	7.1e-01	1.2e-01
C18:1	-0.00 (-0.00, 0.04)	9.0e-01	9.0e-01	1.2e-02
Lysophosphatidylcholine				
C17:0	0.14 (0.09, 0.15)	1.4e-02	4.8e-03	1.7e-05
C14:0	-0.13 (-0.13, -0.05)	2.3e-02	1.6e-01	3.0e-04
C16:1	-0.12 (-0.14, -0.04)	2.3e-02	2.6e-01	1.6e-05
C18:2	0.09 (0.05, 0.09)	9.8e-02	1.8e-01	6.8e-03
Diacyl phosphatidylcholine				
C32/1	-0.16 (-0.16, -0.07)	1.2e-02	4.1e-03	6.8e-05
C32/2	-0.13 (-0.13, -0.01)	5.9e-02	5.6e-01	5.2e-03
C34/4	-0.13 (-0.13, -0.01)	5.9e-02	5.6e-01	4.2e-03
Alkyl-acyl phosphatidylcholine				
C34/3	0.13 (0.04, 0.13)	8.3e-02	1.2e-01	2.0e-03
C36/2	0.13 (0.04, 0.13)	8.3e-02	7.2e-02	2.8e-03

Summary of standardized estimates from multiple linear regression models (coffee consumption as exposure and metabolite concentration as outcome), with single models corresponding to adjustment for a specific subset of direct neighbors of the metabolite in the subgroup-specific metabolite-network; Metabolite concentrations: standardized on age, sex, BMI and prevalence of hypertension; All models adjusted for lifestyle, diet, fasting status at blood draw occasion, and medication.

Estimate range (betas) indicates the effect of 3 cups (2 standard deviation) higher coffee intake on the variance standardized metabolite concentrations, adjusted for thus far identified directly affected metabolites within the connected component (lowest estimate, highest estimate from the multi-model procedure); *fdr P*= false discovery rate corrected p-value (based on the model adjusted for external confounders but not for other network-variables); *upper/lower P* = highest and lowest p-values from the multi-model procedure.

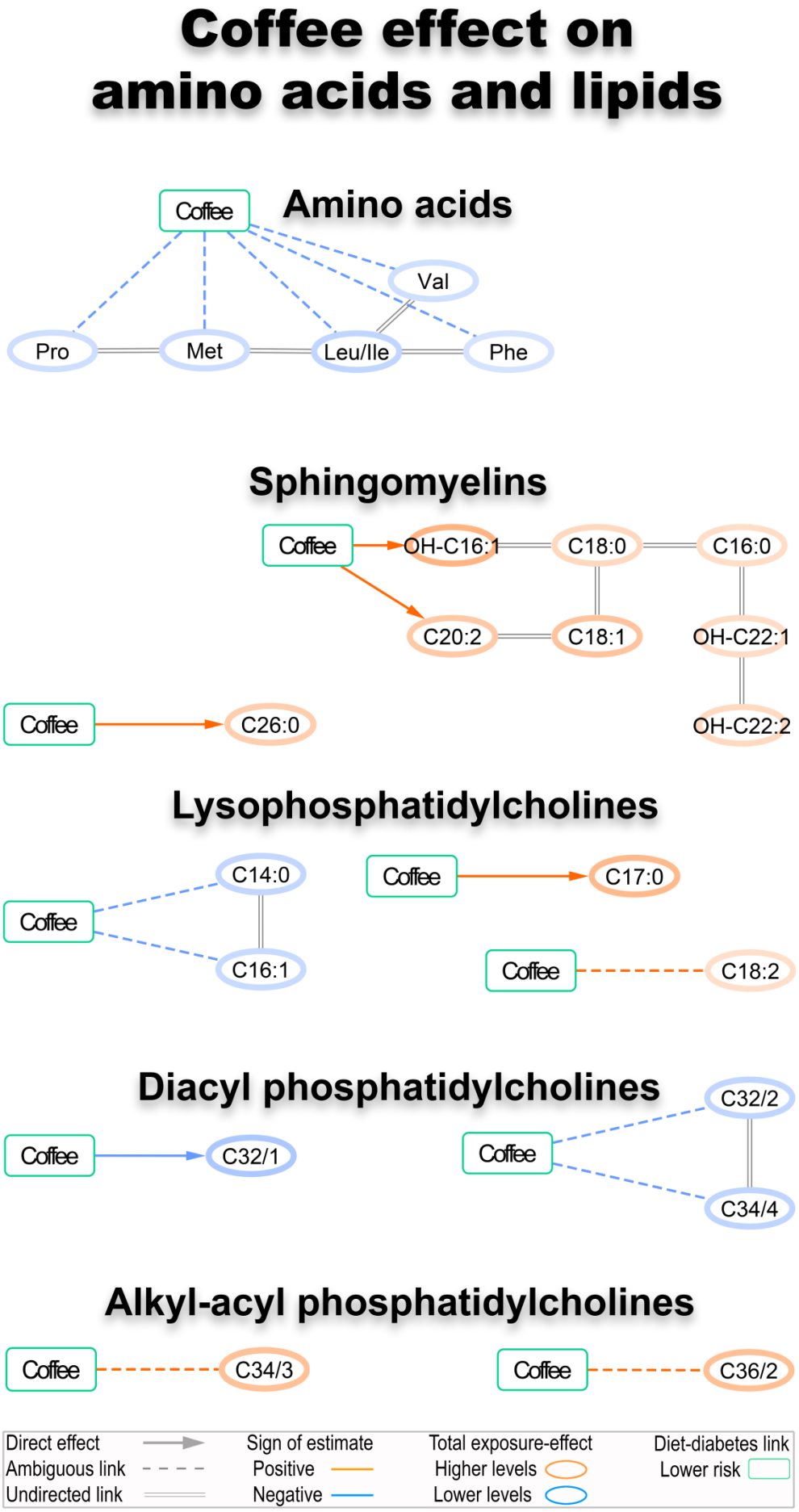


Figure 23: Coffee effects on metabolomics network structures

Red meat

Habitual red meat consumption was linked to alterations in the concentration of particular serum amino acids, and to pronounced alterations in the fatty acid composition of all evaluated lipid compartments (Table 15).

Higher red meat intake was related to higher concentrations of the connected branched-chain amino acids valine and leucine/isoleucine (Figure 24). These links, however, were not unambiguously classified as direct or indirect effects. In contrast, the relation of higher red meat consumption with lower glycine serum concentrations was a singleton and consistent and significant across all submodels. Glycine was thus classified as directly affected by red meat consumption.

Among acylcarnitines, higher red meat consumption was related to higher serum concentrations of a group of five interrelated acylcarnitines, i.e. stearoylcarnitine (C18:0), octadecenoylcarnitine (C18:1), propionylcarnitine (C3:0), acetylcarnitine (C2:0), and carnitine. Among these metabolites, stearoylcarnitine was classified as directly affected by red meat intake based on unambiguous multi-model estimates. Adjusting for stearoylcarnitine explained the association of red meat consumption with all other acylcarnitines within the connected component (therefore classified as indirect effects), except for the association of red meat intake with propionylcarnitine. Here, effect estimates remained positive after adjusting for stearoylcarnitine levels but the highest p-value from the multi-model procedure surpassed the preset significance threshold. The red meat-propionylcarnitine link was therefore classified as ambiguous (direct or indirect effect).

Within the sphingomyelin compartment, red meat consumption was related to elevated levels of six connected metabolites. All contained saturated and monounsaturated fatty acids of long and very long chain-length. Adjusting for sphingomyelins C18:1 and 24:1 (classified as direct effects based on multi-model estimates) explained red meat-associations of sphingomyelins C18:0 and C26:1 (accordingly classified as indirect effects), but not red meat-associations of sphingomyelins C16:0 and C16:1 (which remained thus marked as ambiguous).

In contrast to sphingomyelins and acylcarnitines, red meat consumption was also related to lower concentrations of specific metabolites within the phosphatidylcholine compartments. Primarily metabolites were inversely affected which contained saturated and monounsaturated fatty acids with a chain-length of sixteen carbon atoms

or less (Figure 25). In particular, lower red meat-associated serum concentrations were observed of lysophosphatidylcholine C16:1, diacyl phosphatidylcholines C30/0, C28/1, C32/1, and C32/2 (which all belonged to one red meat-connected component), and alkyl-acyl phosphatidylcholines C30/0, C34/0, C34/1, and C36/1 (also belonging to one red meat-connected component). None of these inversely affected metabolites was unambiguously classified as direct effect based on multi-model estimates.

Still, a larger number of phosphatidylcholines, however, was elevated in relation to higher red meat consumption. Two groups of phosphatidylcholines positively associated with red meat could be distinguished. On the one hand, elevated concentrations of primarily saturated and monounsaturated (very) long-chain (≥C18) fatty acid containing metabolites were detected in the diacyl phosphatidylcholine compartment (C38/0, C38/1 and C36/0, forming a connected component) and in the alkyl-acyl phosphatidylcholine compartment (C36/0, C40/1, C42/1, also forming a connected component). Among diacyl phosphatidylcholines, the red meat link with C38/0 was classified as direct effect, and adjusting for this direct effect explained the red meat association with C38/1 and C36/0 (indirect effects). None of the three above-mentioned alkyl-acyl phosphatidylcholines was unambiguously classified as direct effect based on multi-model estimates.

On the other hand, red meat was positively associated with several polyunsaturated fatty acid-containing phosphatidylcholines. Among lysophosphatidylcholines, red meat directly affected the serum concentrations of C20:4. Furthermore, red meat consumption was related to higher concentrations of the two connected diacyl phosphatidylcholines C38/4 and C36/4. Based on the multi-model estimates C38/4 was classified as directly affected, and adjusting for this direct effect rendered the red meat-association with C36/4 non-significant (indirect effect). Red meat consumption was also positively associated with twelve polyunsaturated fatty acid-containing alkyl-acyl phosphatidylcholines that together formed the largest red meat-connected component (C34/2, C34/3, C36/3, C36/4, C36/5, C38/4, C38/5, C38/6, C40/4, C40/5, C40/6, and C42/5). Among these, only alkyl-acyl phosphatidylcholine C36/4 was classified as directly affected based on multi-model estimates. Adjusting for this direct effect rendered all other links of red meat with metabolites of that connected component non-significant (or even reversed the effect to a slight

significant inverse relation in some cases – which was however not stable over multi-model estimates and thus not further considered for interpretation). Therefore the effect of red meat on all other alkyl-acyl phosphatidylcholines of that connected component were classified as indirect, mediated by the direct effect of red meat consumption on alkyl-acyl phosphatidylcholine C36/4.

Table 15: Multi-model inference on possible effects of red meat-consumption on metabolite-levels

	Red meat consumption			
	Estimate range (betas)	*fdr P*	*Upper P*	*Lower P*
Amino Acid				
Gly	-0.16 (-0.20, -0.15)	4.1e-02	3.0e-03	1.6e-05
Val	0.12 (0.02, 0.14)	7.4e-02	5.2e-01	3.0e-04
Leu/Ile	0.12 (0.01, 0.13)	7.4e-02	5.6e-01	2.0e-04
Acylcarnitine				
C18:0	0.17 (0.08, 0.17)	1.4e-02	4.4e-02	8.3e-04
C3	0.16 (0.07, 0.16)	1.9e-02	8.6e-02	1.2e-03
C2	0.09 (0.08, 0.11)	1.1e-01	5.8e-02	8.0e-03
Carnitine	0.07 (-0.00, 0.07)	1.9e-01	9.6e-01	1.4e-01
C18:1	0.03 (-0.02, 0.06)	5.1e-01	8.1e-01	5.7e-02
Sphingomyelin				
C24:1	0.25 (0.11, 0.25)	7.5e-06	3.7e-05	3.3e-08
C18:1	0.21 (0.05, 0.21)	3.0e-04	1.7e-02	3.1e-05
C16:0	0.16 (-0.04, 0.16)	5.8e-03	9.8e-01	1.5e-03
C16:1	0.14 (-0.03, 0.14)	1.7e-02	9.9e-01	6.7e-03
C18:0	-0.04 (-0.04, 0.00)	1.2e-01	9.9e-01	8.8e-02
C26:1	-0.05 (-0.05, 0.03)	1.3e-01	7.7e-01	1.3e-01
Lysophosphatidylcholine				
C20:4	0.18 (0.15, 0.21)	2.5e-03	2.5e-04	1.1e-09
C16:1	-0.12 (-0.16, -0.05)	9.1e-02	2.0e-01	9.1e-06
Diacyl phosphatidylcholine				
C38/0	0.32 (0.09, 0.33)	4.7e-09	1.8e-03	9.8e-22
C38/4	0.21 (0.08, 0.21)	4.5e-04	3.0e-03	9.8e-09
C30/0	-0.20 (-0.20, -0.03)	7.7e-04	3.2e-01	5.0e-11
C32/1	-0.17 (-0.17, -0.03)	3.8e-03	4.1e-01	1.3e-07
C32/2	-0.18 (-0.18, -0.03)	3.8e-03	4.2e-01	3.4e-11
C28/1	-0.15 (-0.15, -0.04)	1.2e-02	3.3e-01	3.0e-03
C36/4	-0.02 (-0.02, 0.05)	4.0e-01	7.1e-01	1.1e-02
C38/1	0.05 (0.05, 0.07)	5.5e-01	2.8e-01	1.4e-01
C36/0	0.01 (-0.02, 0.05)	8.0e-01	9.3e-01	9.3e-02

	Red meat consumption			
	Estimate range (betas)	*fdr P*	*Upper P*	*Lower P*
Alkyl-acyl phosphatidylcholine				
C36/4	0.43 (0.05, 0.43)	1.2e-15	1.8e-02	3.1e-17
C34/0	-0.13 (-0.13, -0.01)	2.2e-02	5.6e-01	5.3e-03
C30/0	-0.13 (-0.14, -0.03)	2.3e-02	3.7e-01	6.7e-04
C42/1	0.13 (0.04, 0.13)	2.4e-02	2.4e-01	2.0e-03
C36/0	0.12 (0.04, 0.14)	3.7e-02	2.7e-01	5.3e-04
C36/1	-0.12 (-0.13, -0.00)	3.7e-02	8.7e-01	1.5e-05
C40/1	0.12 (0.01, 0.13)	3.7e-02	8.3e-01	5.9e-05
C34/1	-0.11 (-0.14, 0.01)	4.4e-02	9.4e-01	1.8e-05
C40/5	-0.13 (-0.13, -0.01)	1.4e-02	5.6e-01	2.5e-03
C34/2	-0.12 (-0.12, 0.00)	1.4e-02	9.5e-01	2.0e-03
C34/3	-0.12 (-0.13, 0.02)	1.8e-02	9.2e-01	6.2e-04
C38/4	-0.07 (-0.09, 0.01)	3.0e-02	7.5e-01	8.9e-04
C40/4	-0.10 (-0.10, 0.05)	3.8e-02	9.2e-01	1.7e-02
C42/5	-0.09 (-0.09, 0.06)	6.6e-02	7.0e-01	3.1e-03
C36/3	-0.07 (-0.07, 0.04)	6.6e-02	9.2e-01	1.5e-02
C40/6	-0.06 (-0.08, 0.08)	2.6e-01	9.8e-01	4.1e-03
C38/5	0.02 (0.02, 0.05)	3.0e-01	3.7e-01	1.1e-02
C36/5	0.02 (0.00, 0.06)	4.9e-01	8.4e-01	3.6e-02
C38/6	0.02 (-0.01, 0.07)	6.4e-01	9.6e-01	4.2e-04

Summary of standardized estimates from multiple linear regression models (consumption of red meat as exposure and concentration of the single metabolite as outcome), with single models corresponding to adjustment for a specific subset of direct neighbors of the respective metabolite in the subgroup-specific metabolite-network; Metabolite concentrations were standardized on age, sex, BMI and prevalence of hypertension, and all models were comprehensively adjusted for lifestyle, diet, fasting status at blood draw occasion, and medication.

Estimate range (betas) indicates the effect of 2 standard deviation higher red meat intake on the variance standardized metabolite concentrations, adjusted for thus far identified directly affected metabolites within the connected component (lowest estimate, highest estimate from the multi-model procedure); *fdr P*= false discovery rate corrected p-value (based on the model adjusted for external confounders but not for other network-variables); *upper/lower P* = highest and lowest p-values from the multi-model procedure.

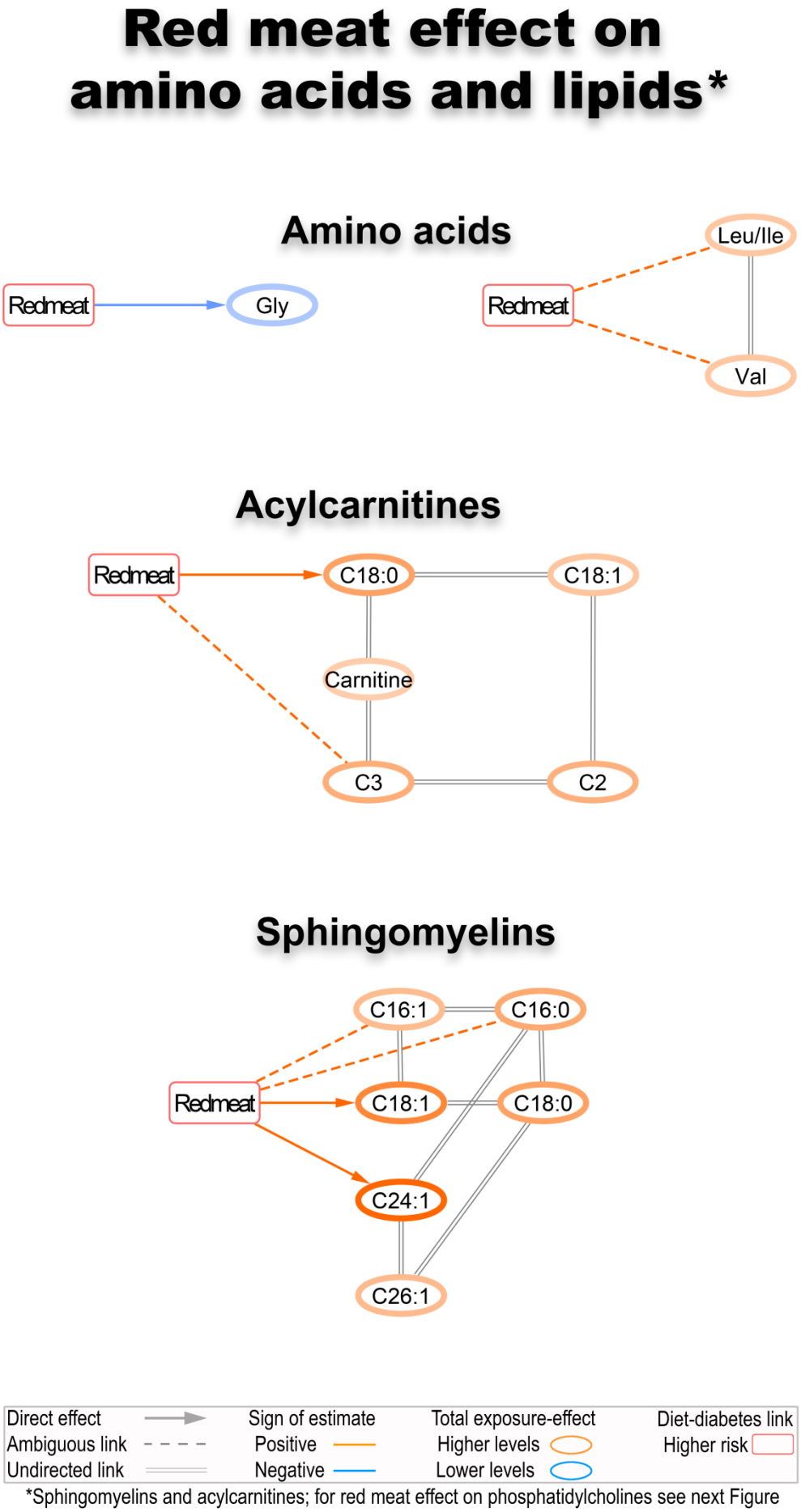


Figure 24: Red meat effects on amino acids, acylcarnitines, and sphingomyelins

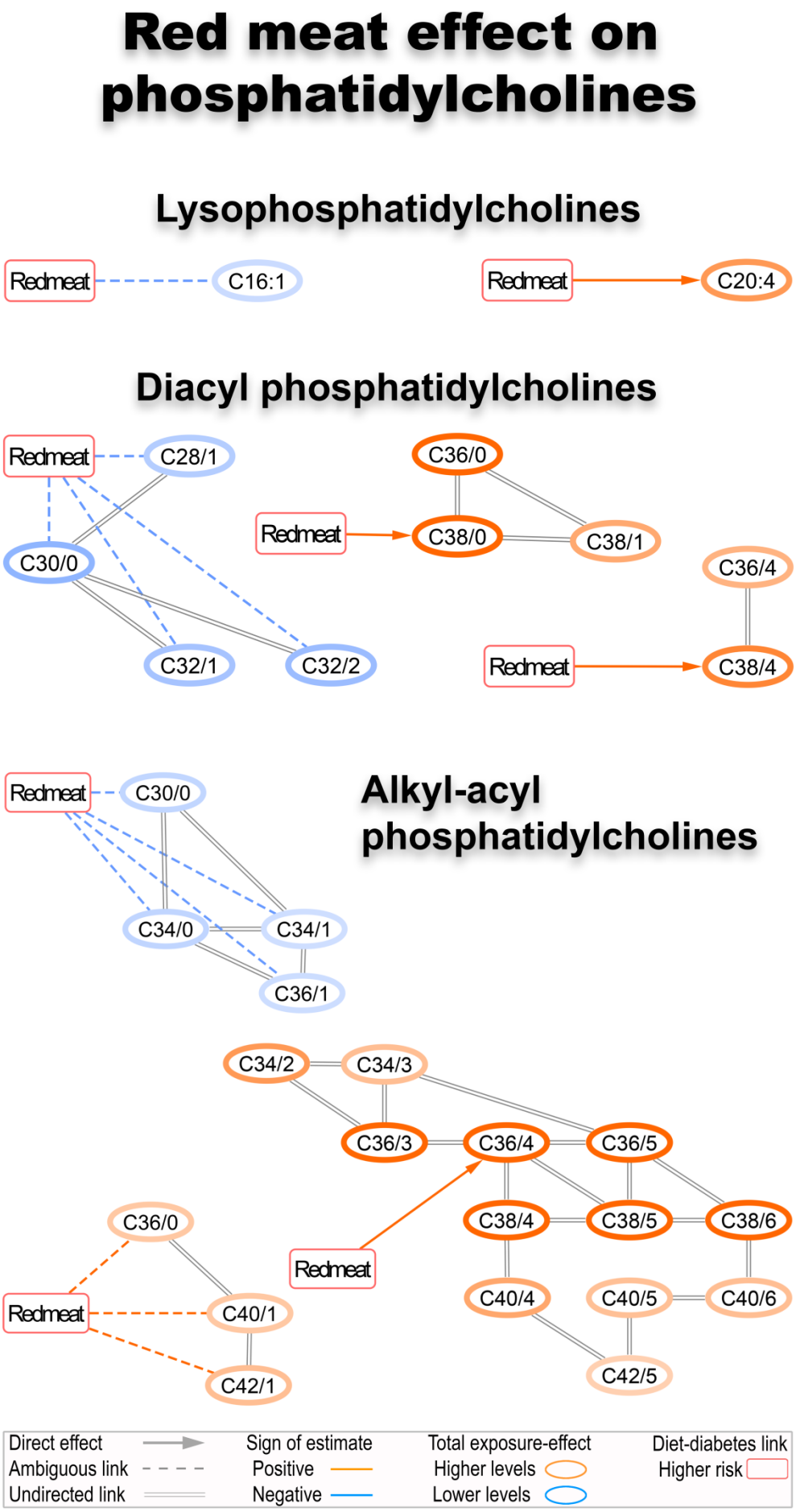


Figure 25: Red meat effects on phosphatidylcholines

4.3.5 Metabolites that directly affected type 2 diabetes risk

Digital only supplemental material

Summary information on the multi-model procedure for all metabolite-diabetes links and comprehensive information on all submodels is provided as digital only supplemental material on the accompanying CD.

Amino acids

Glycine and glutamine were directly related to a reduced risk of type 2 diabetes (Table 16). For glutamine, this relation was revealed only after adjusting for other directly diabetes-related amino acids. Phenylalanine and tyrosine were directly related to an elevated type 2 diabetes risk. For tryptophan, a direct link to higher diabetes risk was suggested, but not all multi-model estimates reached significance and the link was therefore classified as ambiguous. It should also be noted that the two interconnected but differentially diabetes related groups of amino acids, with risk-elevating aromatic amino acids-on the one side and risk-reducing amino acids glycine and glutamine on the other (Figure 26)

Acylcarnitines

Two unsaturated fatty acid-containing acylcarnitines, i.e. tetradecenoyl-carnitine (C14:1) and lineoylcarnitine (C18:2) were related to a reduced risk of type 2 diabetes (Table 16). Palmitoylcarnitine (C16:0) and propionylcarnitine (C3) were directly related to an elevated diabetes risk. Furthermore, a direct link of glutarylcarnitine (C5DC-C6OH) with an elevated diabetes risk was suggested but not all estimates reached the significance threshold. Hence this link was classified as ambiguous. It also should be noted that the direct link between the two directly but differentially diabetes related acylcarnitines C16:0 and C 14:1 (Figure 26).

Sphingomyelins

Five mono- and polyunsaturated fatty acid-containing sphingomyelins were directly related to a reduced risk of developing diabetes (Table 16). In particular, enrichment of sphingomyelins containing C16:1, C18:1, C24:1, C20:2 and of hydroxysphingomyelin C22:2 had a beneficial effect on type 2 diabetes risk. For sphingomyelin C18:1 and hydroxysphingomyelin C22:2, the direct effect was only revealed after adjusting for other directly diabetes-affecting sphingomyelins. An elevated type 2 diabetes risk was observed in participants with higher

serum concentrations of sphingomyelin with stearic acid (C18:0), and of hydroxysphingomyelin C22:1. Both links were classified as direct effect only after adjusting for other directly diabetes-linked sphingomyelins.

It should be noted that two pairs of directly linked sphingomyelins were particularly strongly but oppositely related to type 2 diabetes risk: sphingomyelin C18:0 and C18:1; and hydroxysphingomyelin C22:1 and C22:2 (Figure 26). Metabolites within either pair were highly correlated (the partial correlation coefficient adjusted for all other sphingomyelins was 0.74 between sphingomyelins C18:0 and C18:1, and 0.65 between hydroxysphingomyelins C22:1 and C22:2). Accordingly, full strength of the risk relation was revealed only after mutual adjustment. In the mutually adjusted models, however, on the one hand, each pair of risk estimates for the none/lower unsaturated metabolite were the highest among all metabolites (hazard ratios per standard deviation of 2.55 and 2.98 for sphingomyelin C18:0 and hydroxysphingomyelin C22:1, respectively). On the other hand, the higher unsaturated metabolites showed to be among the strongest relations to reduced risk of type 2 diabetes (hazard ratios per standard deviation of 0.62 and 0.44 for sphingomyelin C18:1 and hydroxysphingomyelin C22:2, respectively).

Phosphatidylcholines

Among phosphatidylcholines, metabolites that were related to an elevated risk of type 2 diabetes can be summarized into two groups: one contained saturated fatty acids with 14 to 18 carbon atoms; and the other contained specific polyunsaturated fatty acids (Table 16, Figure 27). The first group comprised lysophosphatidylcholines that contained myristic acid (C14:0) and palmitic acid (C16:0), respectively, and alkyl-acyl phosphatidylcholine C38/0. The group of high-risk polyunsaturated fatty acid-containing metabolites was constituted by lysophosphatidylcholine C20:3, diacyl phosphatidylcholines C38/3 and C42/5, and alkyl-acyl phosphatidylcholines C36/4 and C38/2. Other phosphatidylcholines with odd chain, monounsaturated, and polyunsaturated fatty acid partly with very long chains were directly related to lower risk of type 2 diabetes. Lysophosphatidylcholines containing margaric acid (C17:0), C18:1 and C18:2; diacyl phosphatidylcholines C32/3, C38/5, and C42/1; and alkyl-acyl phosphatidylcholines C30/1, C32/2, C34/3, C38/5, C42/3, and C44/6 were directly related to a reduced type 2 diabetes risk.

Elevated type 2 diabetes risk was also suggested in relation to higher serum concentrations of diacyl phosphatidylcholine C40/2 and of alkyl-acyl phosphatidylcholine C42/2, and reduced diabetes risk was suggested in relation to higher serum concentrations of diacyl phosphatidylcholines C36/6, C42/0, C40/3 and C42/2. Still, for these metabolites multi-model information did not allow unambiguous classification.

It should be noted that also among phosphatidylcholines there were pairs of linked metabolites that were both directly related to diabetes risk, but one had a beneficial and the other an adverse effect (Figure 27). Highlighting two of these pairs, lysophosphatidylcholines C18:2 and C20:3 were partially correlated (r=0.25) but had an oppositely directed relation to type 2 diabetes risk (hazard ratios per standard deviation of 0.75 for the former and of 1.33 for the latter in mutual adjusted models). Furthermore, alkyl-acyl phosphatidylcholines C36/4 and C38/5, which were also partially correlated (r=0.32), had a particularly strong direct relation with type 2 diabetes risk, with hazard ratios of 2.24 per standard deviation of C 36/4 and of 0.55 per standard deviation of C 38/5.

Table 16: Multi-model inference on possible effects of metabolites on diabetes risk

	RR of type 2 diabetes			
	Estimate range (HRs)	*fdr P*	*Upper P*	*Lower P*
Amino Acid				
Gly	0.74 (0.70, 0.81)	4.6e-05	4.4e-04	1.3e-07
Tyr	1.51 (1.24, 1.74)	4.6e-05	9.2e-03	1.1e-10
Phe	1.35 (1.18, 1.66)	1.7e-03	3.6e-02	7.1e-14
Gln	0.78 (0.72, 0.83)	5.9e-03	1.4e-02	6.8e-06
Trp	1.22 (1.13, 1.22)	3.4e-02	1.6e-01	3.4e-02
Acylcarnitine				
C16:0	1.54 (1.13, 1.58)	5.8e-05	4.4e-02	8.0e-07
C14:1	0.77 (0.72, 0.89)	4.9e-03	4.7e-02	1.1e-05
C18:2	0.77 (0.77, 0.87)	4.9e-03	1.1e-02	3.6e-04
C3	1.18 (1.13, 1.23)	7.2e-02	3.3e-02	2.6e-03
C5-DC(C6-OH)	1.16 (1.13, 1.23)	7.2e-02	1.3e-01	1.9e-03
Sphingomyelin				
C18:0	2.55 (1.43, 2.63)	3.6e-07	1.1e-04	9.1e-10
OH-C22:1	2.98 (1.34, 3.09)	5.2e-07	6.1e-03	2.4e-11
OH-C22:2	0.44 (0.44, 0.56)	7.1e-05	1.2e-04	9.2e-06
C18:1	0.62 (0.62, 0.62)	2.1e-03	2.1e-03	2.1e-03
C24:1	0.71 (0.63, 0.81)	6.6e-03	1.6e-02	1.2e-07
C16:1	0.70 (0.67, 0.84)	9.3e-03	2.2e-02	6.5e-05
C20:2	0.87 (0.82, 0.87)	4.8e-02	3.0e-02	7.2e-04
Lysophosphatidylcholine				
C14:0	1.33 (1.21, 1.33)	1.1e-03	7.2e-04	7.3e-05
C18:2	0.75 (0.66, 0.84)	1.6e-03	3.0e-02	8.7e-08
C17:0	0.78 (0.77, 0.89)	2.9e-03	4.4e-02	2.0e-04
C20:3	1.33 (1.28, 1.33)	5.0e-03	4.2e-03	2.3e-03
C18:1	0.70 (0.69, 0.84)	1.2e-02	4.4e-02	2.2e-03
C16:0	1.35 (1.33, 1.50)	1.9e-02	2.3e-02	2.1e-04
Diacyl phosphatidylcholine				
C38/5	0.47 (0.47, 0.70)	6.9e-05	2.0e-04	6.9e-06
C42/5	1.48 (1.11, 1.48)	1.8e-03	4.6e-02	1.2e-05
C38/3	1.54 (1.33, 1.54)	4.8e-03	1.5e-03	2.5e-07
C42/1	0.73 (0.61, 0.77)	1.9e-02	1.5e-02	3.4e-13
C36/6	0.63 (0.61, 0.97)	3.0e-02	7.7e-01	5.1e-03
C40/3	0.74 (0.74, 0.91)	3.0e-02	2.5e-01	5.1e-03
C42/2	0.80 (0.79, 0.92)	3.5e-02	3.3e-01	1.4e-02
C40/2	1.31 (1.03, 1.34)	3.5e-02	6.8e-01	8.3e-03
C32/3	0.82 (0.75, 0.85)	3.6e-02	2.5e-02	7.6e-05
C42/0	0.82 (0.80, 0.87)	6.8e-02	1.7e-01	3.7e-02

	RR of type 2 diabetes			
	Estimate range (HRs)	*fdr P*	*Upper P*	*Lower P*
Alkyl-acyl phosphatidylcholine				
C34/3	0.49 (0.48, 0.68)	6.0e-11	1.8e-08	5.9e-13
C36/4	2.24 (1.31, 2.28)	8.9e-05	7.8e-03	1.9e-07
C32/2	0.59 (0.59, 0.80)	9.2e-05	2.6e-02	3.7e-05
C42/3	0.66 (0.63, 0.77)	1.6e-03	1.4e-02	6.3e-05
C36/0	1.35 (1.22, 1.40)	2.0e-03	5.7e-03	3.2e-05
C38/5	0.55 (0.55, 0.60)	4.1e-03	2.1e-03	1.4e-03
C44/6	0.69 (0.65, 0.73)	8.1e-03	3.9e-03	1.1e-11
C30/1	0.83 (0.83, 0.86)	1.2e-02	1.0e-02	3.5e-03
C38/2	1.28 (1.19, 1.37)	4.1e-02	4.3e-02	3.7e-05
C42/2	1.24 (1.20, 1.29)	4.5e-02	8.2e-02	1.3e-02

Hazard Ratios (*HR*) (lowest estimate, highest estimate from the multi-model procedure) per standard deviation in serum metabolite concentrations; *fdr P*= false discovery rate corrected p-value (based on the model adjusted for external confounders and for adjacent network-variables); *upper/lower P* = highest and lowest *p-values* from the multi-model procedure. Hazard ratios were derived from a Cox proportional hazard regression model comprehensively adjusted for age, sex, BMI, lifestyle, diet, fasting status at blood draw occasion, prevalence of hypertension and medication.

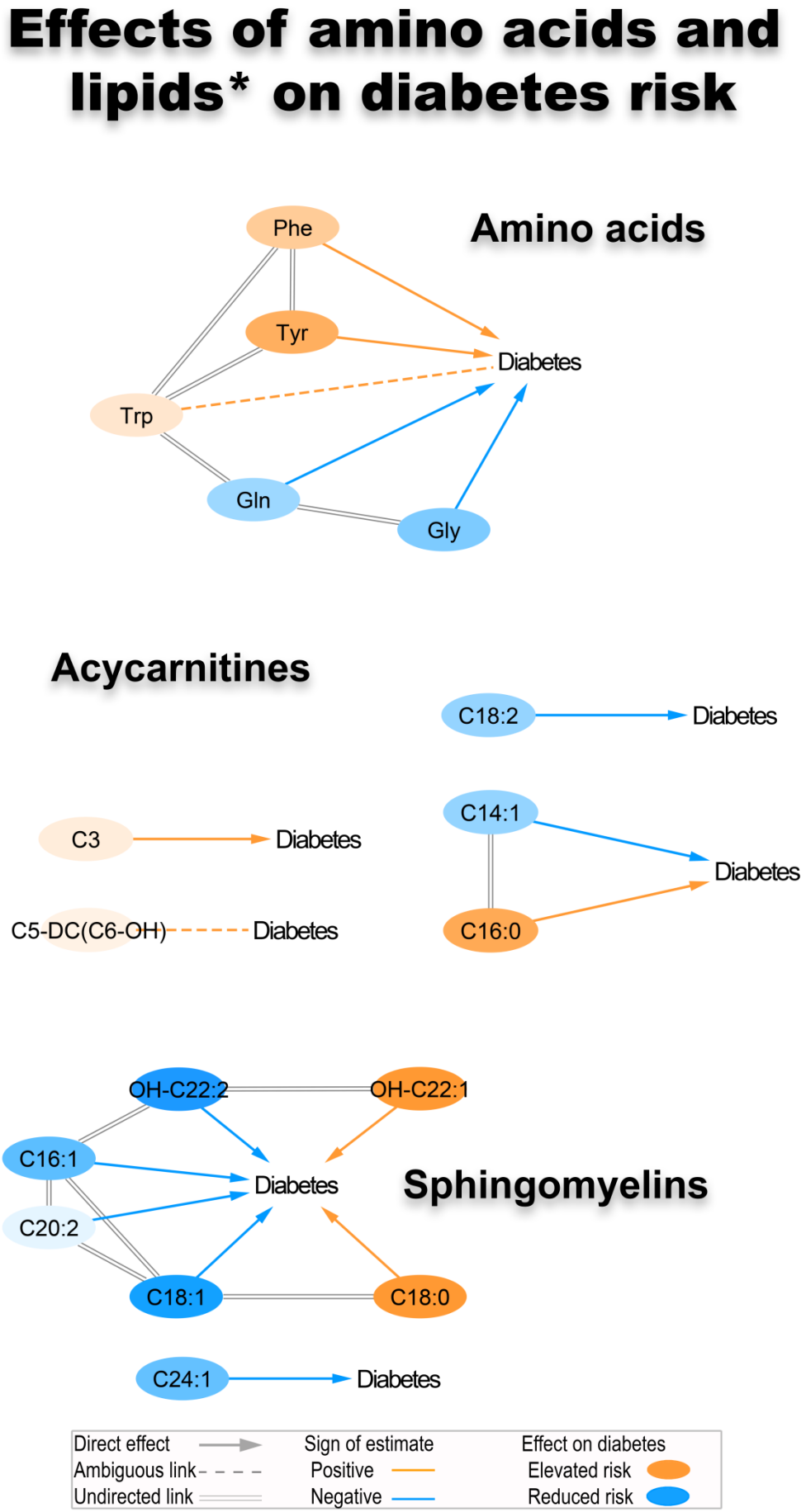


Figure 26: Direct effects on diabetes risk: amino acids, acylcarnitines, sphingomyelins

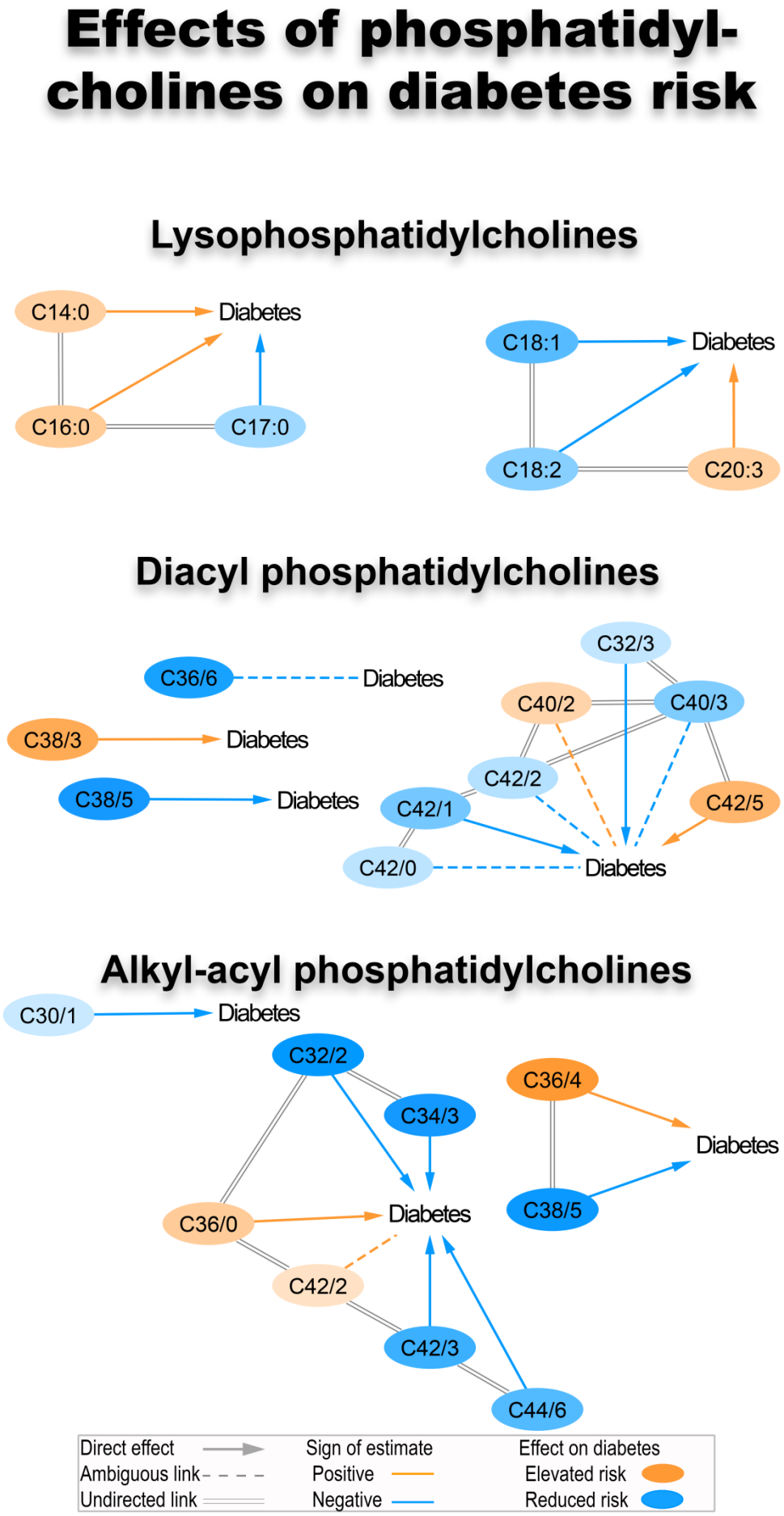


Figure 27: Direct effects on type 2 diabetes risk: phosphatidylcholines

4.3.6 Potential metabolic links between diet and diabetes incidence

Whole-grain bread effects on diabetes risk: potential metabolic mechanisms

All three identified potential metabolic links between whole-grain bread consumption and diabetes risk consistently pointed at lower levels of saturated fatty acids. In particular, serum concentrations of palmitate in the acylcarnitine and the lysophosphatidylcholine compartment, as well as levels of alkyl-acyl phosphatidylcholine C36/0, were lower with higher whole-grain bread consumption, and were at the same time directly related to an elevated diabetes risk.

Only a slight attenuation of the whole-grain bread-effect on diabetes risk was observed after adjusting for lysophosphatidylcholine C16:0. The explainable proportion of the whole-grain effect on type 2 diabetes risk was larger after adjusting for palmitoylcarnitine (C16:0) with 14% effect attenuation and after adjusting for alkyl-acyl phosphatidylcholine C36/0 with 16% effect attenuation. In the mutual mediator-adjusted model 28% of the whole-grain bread effect on diabetes risk was potentially explained by the network-independent variation in selected mediators.

Table 17: Quantitative mediation analysis for whole-grain bread effect on type 2 diabetes risk

Potential Mediator	*Mediator-adjusted HR (95% CI)*	*Proportion explainable (95% bootstrap CI)*
Confounder-adjusted	0.85 (0.67, 1.08)	reference
Acylcarnitine C16:0	0.87 (0.69, 1.10)	14% (4% to 62%)
Lyso-PC16:0	0.86 (0.68, 1.08)	3% (-6% to 15%)
Alkyl-acyl PC C36/0	0.88 (0.69, 1.11)	16% (0% to 62%)
All	0.89 (0.71, 1.13)	28% (11% to 96%)

Potential mediating paths were selected manually based on the joint diet-metabolomics-type 2 diabetes networks. Network independent variation of the metabolite was estimated by adjusting the metabolite for all direct neighbors in the metabolite-network that were not on the shortest path from the exposure to the directly diabetes-linked mediator. The resulting residuals were used to estimate the explainable proportion. Two fully confounder-adjusted Cox models were calculated, with and without adjusting for the network-independent variation of the mediator. The proportion explainable was estimated as difference between non-adjusted and adjusted exposure-estimates relative to the non-adjusted exposure-estimate. The proportion explainable (95% CI) was estimated as median (2.5^{th}, 97.5^{th} percentile) of a bootstrapping-procedure with 1000 replicates and a sampling rate of eighty percent. CI=Confidence Interval.

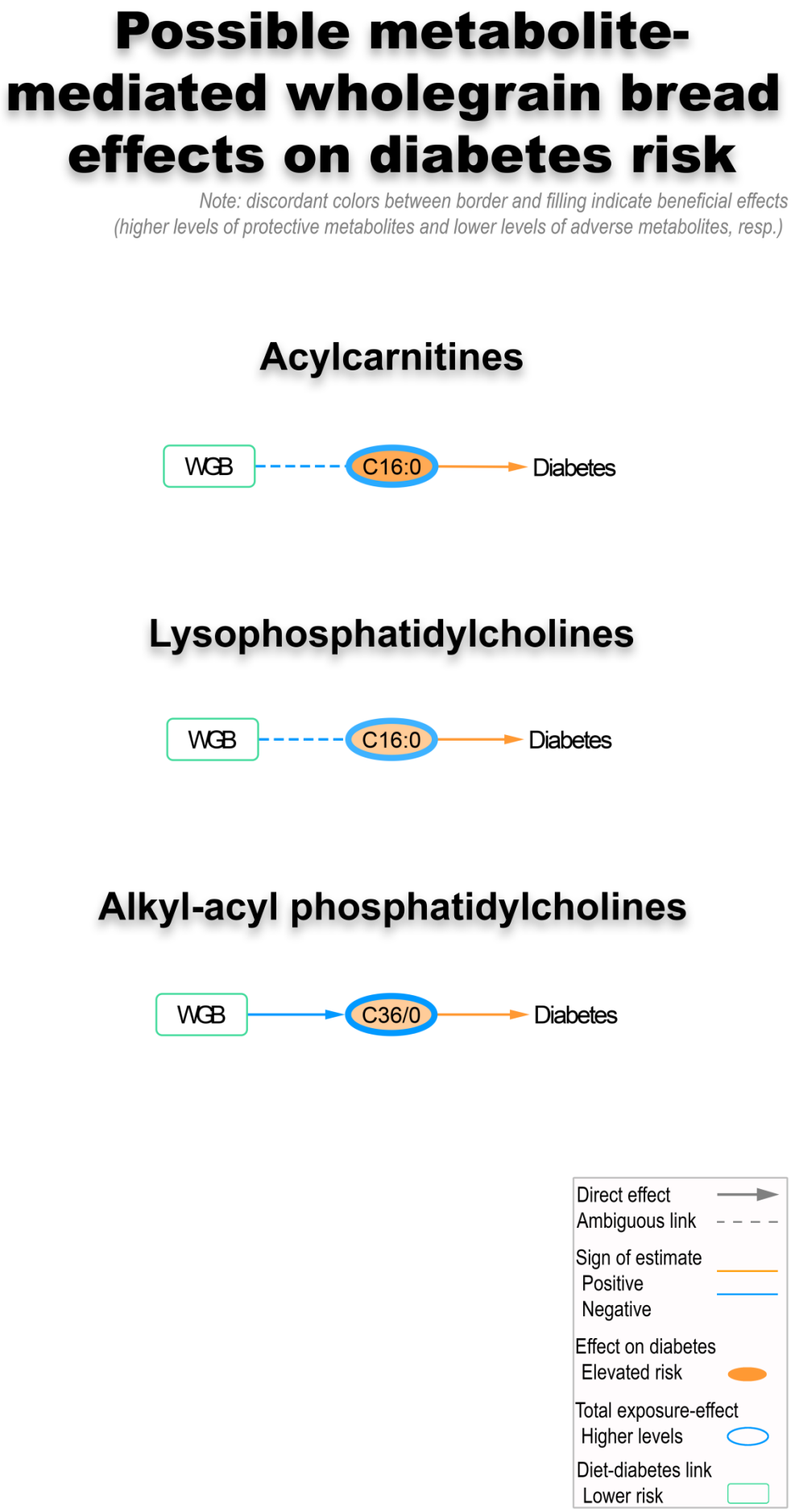


Figure 28: Potential mediators of the whole-grain bread effect on type 2 diabetes risk

Coffee effects on diabetes risk: potential metabolic mechanisms

Six network-independent variations in metabolites were identified as potential metabolic mediators of the assumed coffee-effect on diabetes risk. These were lower concentrations of aromatic amino acids; higher serum concentrations of specific polyunsaturated fatty acid-containing metabolites; alterations in saturated fatty acid containing metabolites.

Lower coffee-related serum concentrations of amino acids involved phenylalanine, which was estimated to have a direct adverse effect on type 2 diabetes risk. This link potentially explained 14% of the coffee-related diabetes risk. Coffee-related higher serum concentration of the beneficial polyunsaturated fatty acid-containing sphingomyelin C20:2, lysophosphatidylcholine C18:2 and alkyl-acyl phosphatidylcholine C34:3 potentially explained 7%, 18% and 33% of the coffee-related type 2 diabetes risk. Among saturated fatty-acid containing metabolites, lysophosphatidylcholine with margaric acid (C17:0) was higher concentrated in plasma in relation to higher coffee consumption and was classified as having a direct beneficial effect on type 2 diabetes risk. Lysophosphatidylcholine with myristic acid (C14:0), in contrast, was lower in relation to higher coffee consumption but directly linked to elevated diabetes risk. Adjusting the coffee-diabetes relation for lysophosphatidylcholine C17:0, however, hardly changed the estimate, whereas adjusting the relation for lysophosphatidylcholine C14:0 potentially explained 11% of the risk reduction by coffee intake

In sum, network-independent variation in metabolites selected as potential mediators had the potential to explain about two thirds of the coffee-diabetes risk relation. The protective coffee effect on type 2 diabetes risk was attenuated by 66% after mutual mediator adjustment.

Table 18: Quantitative mediation analysis for coffee effect on type 2 diabetes risk

Potential Mediator	*Mediator-adjusted HR (95% CI)*	*Proportion explainable (95% bootstrap CI)*
Confounder-adjusted	0.76 (0.62, 0.94)	reference
Phenylalanine	0.79 (0.64, 0.98)	14% (4% to 32%)
Sphingomyelin C20:2	0.78 (0.63, 0.96)	7% (3% to 16%)
Lyso-PC C14:0	0.79 (0.64, 0.97)	11% (5% to 22%)
Lyso-PC C17:0	0.76 (0.62, 0.95)	1% (-7% to 13%)
Lyso-PC C18:2	0.80 (0.65, 0.99)	18% (10% to 33%)
Alkyl-acyl C34:3	0.84 (0.67, 1.04)	33% (17% to 64%)
All	0.94 (0.75, 1.17)	66% (38% to 100%)

Potential mediating paths were selected manually based on the joint diet-metabolomics-type 2 diabetes networks. Network independent variation of the metabolite was estimated by adjusting the metabolite for all direct neighbors in the metabolite-network that were not on the shortest path from the exposure to the directly diabetes-linked mediator. The resulting residuals were used to estimate the explainable proportion. Two fully confounder-adjusted Cox-models were calculated, with and without adjusting for the network-independent variation of the mediator. The proportion explainable was estimated as difference between non-adjusted and adjusted exposure-estimates relative to the non-adjusted exposure-estimate. The proportion explainable (95% CI) was estimated as median (2.5^{th}, 97.5^{th} percentile) of a bootstrapping-procedure with 1000 replicates and a sampling rate of eighty percent. CI=Confidence Interval.

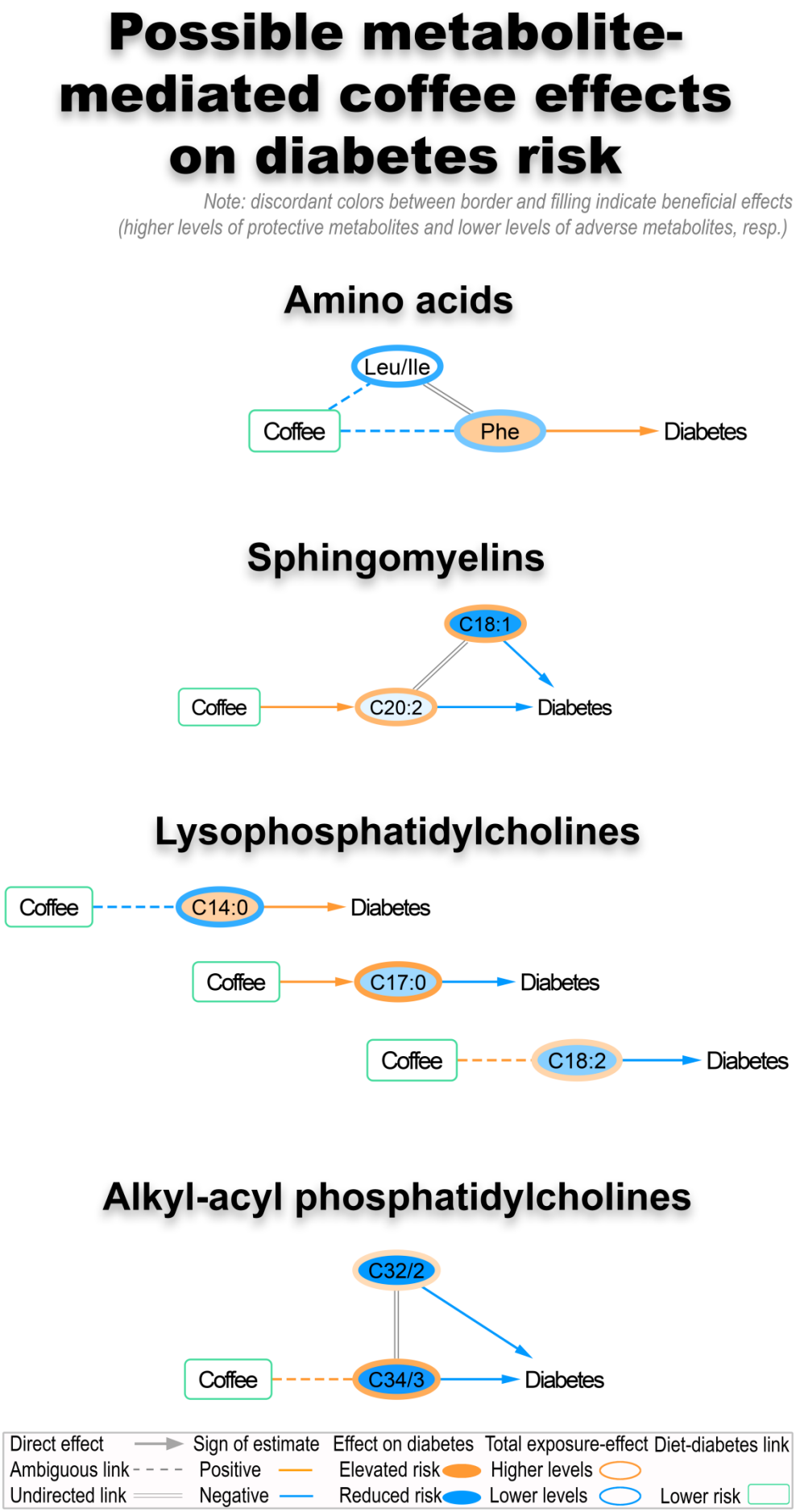


Figure 29: Potential mediators of the coffee effect on type 2 diabetes risk

Red meat effects on diabetes risk: potential metabolic mechanisms

Network-independent variations in concentrations of five metabolites were identified as potential metabolic link between habitual red meat consumption and elevated risk of type 2 diabetes. The potential links comprised elevated levels of saturated and polyunsaturated fatty acids and lower levels of glycine.

Higher consumption of red meat was directly linked to lower circulating glycine concentrations, whereas in turn glycine was related to a reduced diabetes risk. The red meat-related alterations in glycine levels potentially explained 35% of the red meat-related diabetes risk. Among lipid metabolites propionylcarnitine (C3) was preselected as potential metabolic link according to predefined criteria. Mediation analysis did not support, however, a relevant role of that metabolite in metabolically linking red meat consumption to diabetes risk. The long-chain fatty acid-containing sphingomyelin C18:0 and alkyl-acyl phosphatidylcholine C36/0 were higher concentrated in the plasma of participants with higher red meat consumption, and were related to elevated diabetes risk. The proportion of the red meat-related diabetes risk explainable by these two metabolites was 11% and 27%, respectively. Furthermore, red meat consumption was network-independently related to higher serum concentrations of alkyl-acyl phosphatidylcholine C36/4, which in turn was directly linked to higher type 2 diabetes risk. Adjusting the red meat-diabetes relation for that metabolite attenuated the effect by 18%.

The major proportion of the red meat-related type 2 diabetes risk was explainable by network-independent variation in glycine and three saturated and polyunsaturated fatty-acid containing metabolites. The effect-estimate of red meat intake on type 2 diabetes risk was attenuated by 70% after adjusting for these selected potential mediators.

Table 19: Quantitative mediation analysis for the red meat effect on type 2 diabetes risk

Potential Mediator	*Mediator-adjusted HR (95% CI)*	*Proportion mediated (95% bootstrap CI)*
Confounder-adjusted	1.25 (1.00, 1.55)	reference
Glycine	1.16 (0.93, 1.44)	35% (22% to 67%)
Propionylcarnitine (C3)	1.25 (1.00, 1.55)	0% (-7% to 9%)
Sphingomyelin C18:0	1.22 (0.98, 1.51)	11% (0% to 24%)
Alkyl-acyl PC C36/0	1.17 (0.94, 1.46)	27% (14% to 64%)
Alkyl-acyl PC C36/4	1.19 (0.95, 1.49)	18% (2% to 49%)
All	1.06 (0.84, 1.32)	70% (43% to 100%)

Potential mediating paths were selected manually based on the joint diet-metabolomics-type 2 diabetes networks. Network independent variation of the metabolite was estimated by adjusting the metabolite for all direct neighbors in the metabolite-network that were not on the shortest path from the exposure the directly diabetes-linked mediator. The resulting residuals were used to estimate the explainable proportion. Two fully confounder-adjusted Cox-models were calculated, one with and the other without adjusting for the network-independent variation of the mediator. The proportion explainable was estimated as difference between non-adjusted and adjusted exposure-estimates relative to the non-adjusted exposure-estimate. The proportion explainable (95% CI) was estimated as median (2.5^{th}, 97.5^{th} percentile) of a bootstrapping-procedure with 1000 replicates and a sampling rate of eighty percent. CI=Confidence Interval.

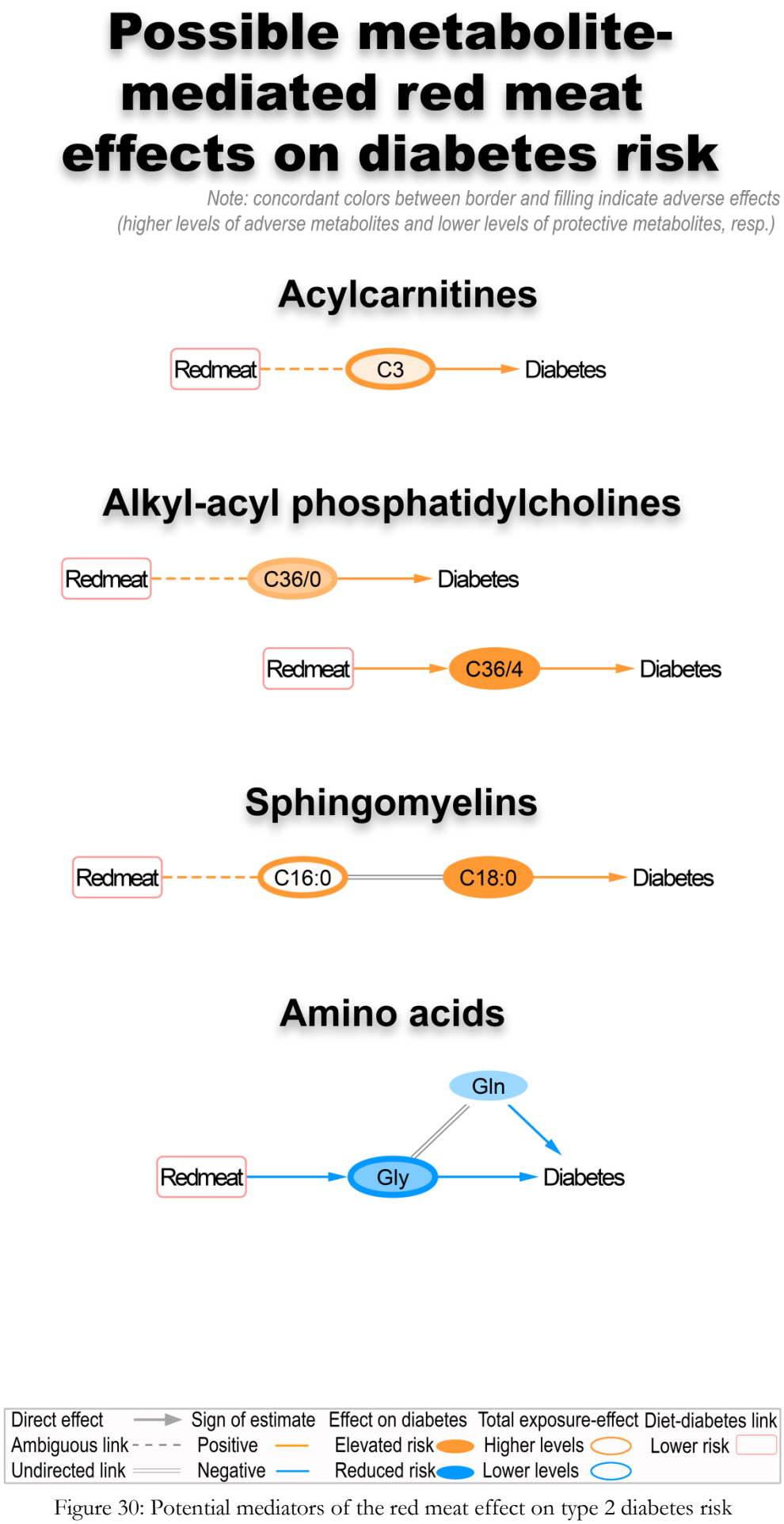


Figure 30: Potential mediators of the red meat effect on type 2 diabetes risk

5 Discussion

5.1 Summary of the results & overview of the chapter

A causal inference framework was developed to detect potential effects of whole-grain bread, coffee, and red meat consumption on metabolomics networks, and to infer potential direct effects of metabolites on type 2 diabetes risk. After successfully testing the validity of the integrated tools, the developed algorithm was applied to data from the prospective EPIC-Potsdam cohort study.

Findings of the present work suggested that consumption of whole-grain bread was related to lower levels of several lipid metabolites with saturated and monounsaturated fatty acids. Coffee was related to lower aromatic and branched-chain amino acids, and had potential effects on the lipid profile and the fatty acid profile within lipid classes. Red meat was linked to lower glycine levels and was related to higher circulating concentrations of branched-chain amino acids. In addition, potential marked effects of red meat consumption on the fatty acid composition within the investigated lipid classes were identified. Moreover, beneficial and adverse effects on type 2 diabetes risk were detected of metabolites within each metabolite group. Aromatic amino acids and lipid metabolites with even-chain saturated fatty acids (C14-C18) and with specific polyunsaturated fatty acid had adverse effects on type 2 diabetes risk. Glycine, glutamine, and lipid metabolites with monounsaturated fatty acid and other polyunsaturated fatty acid were classified as having direct beneficial effects on type 2 diabetes risk.

Potential mediators were identified by graphically overlaying this information in network models. Mediation analysis revealed that effects on lipid metabolites could potentially explain about one fourth of the whole-grain bread effect on type 2 diabetes risk; and that effects of coffee and red meat consumption on amino acid and lipid profiles could potentially explain about two thirds of the altered type 2 diabetes risk linked to these dietary exposures. Thus, analyses of observational data from a large prospective cohort were consistent with the a priori mediation hypothesis: Early effects on lipid and amino acid metabolism showed the potential to explain large parts of the link between three of the most widely discussed diabetes-related dietary exposures and the risk of developing type 2 diabetes.

In the following, the analytical concepts will be discussed, including assumptions and application-related constraints of the analytical design as well as limitations by data quality and potential sources of bias inherent in the study design (5.2). Thereafter, the results of the study will be biologically interpreted (5.3). The chapter will close with a brief discussion of transferability of the findings within research and their relevance for the public and giving an outlook (5.4).

5.2 Analytical concepts

5.2.1 Etiological research in observational settings

The present study aimed to use metabolomics data to identify potential mechanisms that link dietary exposures to type 2 diabetes risk. Methodologically, this task goes beyond classification and prediction. It searches for the explanations of observed data patterns. Many interesting approaches to infer mechanisms from observational data have been developed over the last two decades [139,182]. Still, few examples are available that methodologically approached related research questions [54,106,183], i.e. to use complex biomarker profiles to infer the mechanisms underlying a classical epidemiological exposure-outcome relation in by one comprehensive study design. Network models, however, were not integrated in these studies and, thus, methods were not implemented in a way which was applicable to the present study's question and setting. From a methodological perspective, the large majority of metabolomics applications to large-scale epidemiological studies so far have focused on identifying stable predictors on the single metabolite level [115-119,184-186]. Naturally, interpretations often brought into play biological mechanisms and some observation-based metabolomics findings have already been linked to biological mechanisms. Some of the hypotheses that were generated by metabolomics applications in observational studies were successfully followed up in experimental work [187,188]. Methods, however, are best chosen in accordance to the task. To inferring potential biological mechanisms, a causal inference framework was considered the appropriate choice for the present work.

5.2.2 Factor analysis

One of the major motivations for methodological innovations of the present study was the concern that multiple influences on single metabolite levels might hamper the biological interpretability. Therefore, the analytical design aimed to rule out indirect effects and confounding, and to focus on direct effects that concerned single metabolites. An obvious source of overlapping information was the strong covariance among metabolites of the same metabolite group. On one hand the level of a single metabolite integrates information on metabolic processes that might be rather specific for this particular metabolite. On the other hand metabolites are most commonly also markers of the abundance of the metabolite group as a whole. The approach to control for the local neighborhood of a metabolite in the network should have blocked the influence by the overall group level. First, however, factor analysis was applied to support the hypothesis that the common variance within the metabolite indeed played a role and to investigate this common variance in relation to both, dietary exposures and type 2 diabetes risk.

Factor analysis is a tool to derive information on latent variables that were not directly assessed but that are reflected in the data by impact on groups of measured variables [175,176]. Accordingly, factor analysis was used to evaluate the hypothesis that the high correlation among metabolites of the same biochemical class [84,189] reflected common biological determinants. For this work, metabolites were sorted and looked at in groups characterized by apparent biochemical similarities. Within the groups (amino acids, acylcarnitines, sphingomyelins, diacyl phosphatidylcholines, alkyl-acyl phosphatidylcholines, and lysophosphatidylcholines), the dominant component of common variance was assumed to being induced by shared synthesis, transport and degradation pathways. Therefore, a one factor solution per metabolite group was a priori aimed at. The first factor of common variance was assumed to explain a major proportion of common variance. The plausibility of this assumption was evaluated by examining the scree plots and the Eigenvalues of the maximum factor solution [175]. Indeed major proportions of the variance within metabolite classes were explained by the first factor. In addition, an elbow-like bend in the scree plot after the first factor indicated one factor to be the preferable solution. The variance of well above 50% explained by the first common factor in each of the groups suggested that the knowledge-driven

groupings indeed reflected highly relevant common biological determinants.

5.2.3 Causal inference

Causal modeling

Estimating potential effects rather than associations implied that the estimation procedure was informed by and results were interpreted with respect to a causal model. The subject-matter knowledge that was considered sound enough to inform the causal model was relatively general.

The first group of directionality assumptions specified by the a priori models was that diet affects metabolite concentrations and not the other way around. The claim that metabolite profiles are sensitive to dietary composition was supported by numerous studies [90,91,93,110,190-192]. Still, special cases where metabolite levels in the blood determine dietary behavior cannot be excluded in principle. That such cases might have specifically affected one of the dietary components investigated in this study, however, seems rather unlikely. Broader effects on dietary composition and energy intake should have been blocked because all models were comprehensively adjusted for correlated food intakes and energy intake. Thus, effect directionality from habitual food intake towards the metabolite level is in general the most likely explanation for observed network-independent associations.

TEXTBOX 1: Effect estimates

For regression modeling, dietary exposures were standardized to two standard deviations. This corresponds to marked but frequently observed difference in the exposure level: for example low vs. high, very low vs. average, or average vs. very high consumption of the food. Solid foods were further energy standardized. For participants with an average energy intake of 9 MJ per day, for example, the reported estimates correspond to differences in the average daily intake of approximately 60g for whole-grain bread and of almost 200g for red meat. Coffee consumption was not energy-standardized, reported estimates correspond to a difference of 4 cups per day. Metabolites were standardized to one standard deviation. Estimates on metabolites are therefore comparable on a relative scale.

The second group of directionality assumptions specified by the a priori causal model was that metabolite levels affected type 2 diabetes risk and not the other way round. This relationship was modeled in longitudinal data. Therefore, the second group of directionality assumptions was strongly supported by the prospective design. Chronological structure of the events is considered a reliable criterion for effect directionality.

PC-Algorithm

Assumptions

The PC-algorithm was used to estimate the skeleton of the data-generating DAG based on metabolomics data. This certainly relied on the assumption that the data-generating processes were adequately represented by a DAG. Several studies have applied metabolomics-based network models to observational cohort studies [124]. For exclusively continuous measurements of blood metabolite concentrations, Gaussian graphical models have been used most frequently to estimate conditional independence graphs [84,124]. Gaussian graphical models of metabolomics data have been shown to correspond well with known metabolic reactions [84]. Moreover, Gaussian graphical models have been used to identify unknown metabolites and to detect unknown enzymatic reactions [193]. Recently, Gaussian graphical models that relied on the same targeted metabolomics approach were compared across four cohorts including the EPIC-Potsdam study. The high consistency of network links within the metabolite groups suggested that these links were indeed generated by stable biological mechanisms (*Stefan Dietrich, 2017, unpublished*). In contrast to purely association-based methods to estimate partial correlation networks, causal inference algorithms aim to avoid linking two variables due to conditioning on a collider by more sophisticated adjustment strategies. Otherwise the network-structures are comparable between conditional independence graphs and DAG skeletons [147,148]. Therefore, the assumption that metabolomics data in the blood reflect directed metabolic processes is supported by the literature.

The simulations in this study were performed to evaluate whether the PC-algorithm was suitable to estimate the skeleton of the DAG presumably underlying the metabolomics data. Data were simulated by

structural equations according to random DAGs based on Gaussian link functions. The simulation procedure could be modified, e.g. to more closely resemble metabolic reaction systems. Such simulations, however, were already applied for Gaussian graphical models [84]. These simulations demonstrated that links in conditional independence networks indeed reflected mechanistic links by enzymatic reactions. An exception was that substrates of a common product were often linked in the network albeit not directly connected by a metabolic reaction [84]. This was defined as a collider situation and these links should be omitted by the PC-algorithm [145,148]. Therefore, systematic comparison of the networks generated by the PC-algorithm to those generated by Gaussian graphical modeling would be of interest to generally define overlap and differences, and to derive appropriate application scenarios. Such simulations, however, were not within the scope of this work. Simulation studies on effects of the setting of the alpha level on performance of the PC-algorithm showed that results hardly depended on tuning of this parameter [148].

Moreover, estimating the skeleton of a DAG based on observed data implies the assumption that the data is faithful to the generating DAG. This assumption, however, is generally fulfilled for non-negative multivariate normal distributions and therefore applies to metabolomics data [179].

Applications

Results of simulation studies suggested that the PC-algorithm was suitable to estimate the skeleton of the data-generating DAGs in the settings to which it was applied in the present study. In the range of sizes, densities, and effect strengths that was observed for metabolomics networks in EPIC-Potsdam, the PC-algorithm performed excellently. Furthermore the PC-algorithm reached close to optimal performance at sample sizes of around 2000 observations, which again corresponds well with the present study. It should be noted that this cannot necessarily be generalized to other settings because the performance of the PC-algorithm was clearly dependent on the settings of the network parameters and the size of the study sample. Possible use of the methodology developed in the present study in other settings should always rely on cautious evaluation of the applicability of the tools. In smaller samples, e.g., an algorithm that estimates networks based on few observations with higher robustness might be preferable.

NetCoupler

Assumptions

The NetCoupler-algorithm relied on effect directions specified by the underlying causal model and on the PC-algorithm to estimate the skeleton of DAGs. Therefore, assumptions discussed in the paragraphs above also apply with respect to the NetCoupler-algorithm. In addition, interpretation of the identified links as effects relies on the assumption that confounding was efficiently controlled for by the comprehensive adjustment strategy.

Implications of the different directionality assumptions on the exposure–metabolite relation vs. the metabolite-outcome relation were already explained (3.4.5). To briefly summarize, the NetCoupler.IN-version relies on the assumption that the exposure affects the metabolite levels and not the other way round. Thus, other metabolites are not considered as potential confounders but only as mediators of indirect effects. The NetCoupler.OUT-version, on the contrary, assumes effect directionality from the metabolites towards the outcome. Adjacent metabolites are thus considered as potential confounders. Consequently, ambiguous estimates on exposure-metabolite relations can still be interpreted as total effects, whereas ambiguous estimates on metabolite-outcome relations need to be treated as potentially confounded.

Application

The biological interpretation of the results will be subject of the next section (5.3). Biologically consistent and interpretable findings alone, however, cannot prove that the NetCoupler-algorithm indeed detected biological effects. Improved biological interpretability in comparison to other metabolomics studies can still be graded as indication for meaningful methodological innovations of this work. It should be noted that the aim of the current study was to generate well-defined hypotheses on the metabolic mechanisms that link dietary habits to type 2 diabetes risk, which are consistent with observational data. Options to consolidate these hypotheses by the use of other study designs will be discussed below (5.4.2).

Unambiguous classification relied on the detection of direct effects based on consistent and significant effects across submodels. This was only partly given for diet-metabolite relations. For each food exposure several links remained ambiguous. Two explanations might be

worthwhile to consider. Firstly, dietary effects on metabolites were relatively weak in the first place and adjustment for descendants of the metabolites might have further diluted these effects to result in non-significant estimates. It should be noted that the weak effects, which are further discussed below (5.2.4), might to some extent be artefacts of generally imprecise dietary assessment-tools [194]. More sophisticated search strategies might still be able to detect a few explanatory models that are consistent with the observed patterns of conditional dependencies [179]. Secondly, dietary effects might have included latent factors that affected several metabolites at a time. In this case causal modeling techniques involving inference on latent variables could be considered [178]. As discussed above, the interpretation of ambiguous exposure-metabolite links as total effects was still meaningful.

The classification of metabolite-diabetes risk links as direct and indirect effects was successful and the minority of links remained ambiguous. In the present study, by considering network-adjusted estimates, both, beneficial effects and adverse effects, were revealed within each of the metabolite classes. This agrees with biologically motivated prior expectations, which will be discussed in the next section (5.3). Based on theoretical considerations and supported by the results of the present study, confounding by metabolically-related factors has the potential to conceal biologically interesting links. Metabolomics applications in etiological diabetes research should consider strategies to deal with this issue. The present study offers a computationally feasible and graphically informed approach to generate confidence sets of possible estimates. By graphical specification a priori assumption on effect directionality can be integrated. Links within the graphical model can be translated into equations of any parametric or non-parametric form [195]. Without major changes of the general framework, the approach could be extended to include other link functions, repeated measurements of the exposure, or interaction terms, to give some examples of increased model complexity.

Mediation analysis

Assumptions

Four main assumptions have to be fulfilled to derive quantitatively valid estimates on the natural indirect effect, i.e. the proportion mediated or proportion explained by mediation in the context of this study [196,197].

1. No unmeasured confounding of the exposure-outcome effect.
2. No unmeasured confounding of the exposure-mediator effect.
3. No unmeasured confounding of the mediator-outcome effect.
4. No confounder (measured or unmeasured) of the mediator-outcome effect that is affected by the exposure.

Recent methodological contributions have relaxed some of these assumptions, particularly (4) [198,199]. The major point is that quantitative interpretation of estimates on mediated proportions implies strong assumptions on the absence of confounding. A further limitation is that a different extent of measurement error can also compromise the quantitative estimates on mediated proportions [200]. Most importantly, the concept of mediation again relies on a valid underlying causal model. Because of the hypothesis-generating nature of this work, the term *proportion explainable* is used. In contrary, the term "proportion explained" or "proportion mediated" would only be applicable if a very distinct causal hypothesis strongly supported by subject-matter knowledge was tested in a setting, where all (possibly) relevant information was available. Certainly, the mediation analysis could also be refined. For example, more sophisticated search algorithms could be used to identify additional potential mediators [201,202]. Mediation analysis involving multiple mediators could be extended to include interactions [203-205]. Sophisticated statistics should not distract, however, from the emphasis on a clearly specified causal hypothesis, which mediation analysis underlies the application of mediation analysis to observational data.

5.2.4 Limitations of the data quality & sources of bias

Limited data-assessment

Dietary data was assessed as habitual consumption over the last year by validated food frequency questionnaires. Based on the questionnaires, average intake levels were estimated for 49 food groups [153]. In the present study it was not differentiated between different subtypes of the evaluated food exposures. Possible modifying effects of different total exposure durations (life-time exposure) or differences in the dietary composition in the days before blood sampling were not considered. Fasting status was included as adjustment variable but fasting duration for example was not considered.

Metabolomics were measured on a targeted platform [173]. The targeted metabolite spectrum was shown to be sensitive to the dietary

habits [83,100] and to be informative with respect to type 2 diabetes risk [115,116]. The coverage of the metabolome or a subsystem thereof (e.g. lipid metabolites) by the targeted approach, nevertheless, was very limited. Furthermore, a major limitation of the targeted metabolomics approach of the study was low resolution of information on fatty acid residues. Only the cumulative number of carbon atoms and desaturations over all fatty acids contained in a single metabolite were provided. Neither information on the localization nor information on the stereochemistry of desaturations was available. In metabolites with two fatty acids, the chain-length of the single fatty acids was also not provided but a summary measure of C-atoms from the two fatty acids. Moreover, the metabolomics approach did not differentiate between leucine and isoleucine and provided a summary measure only. In addition, other types of biomarkers were not considered for the present study but could be of interest. Genetic data, for example, could be used to consolidate causality of the observed effect in terms of Mendelian randomization studies [206] or to identify relevant subgroups of participants in terms of interaction analyses [207].

Measurement error

Dietary assessments with validated food frequency questionnaires deliver relatively crude estimates on average intake levels over the year. The dietary assessment tools in EPIC-Potsdam were shown to deliver valid estimates on average intake levels [155,157], but these can be considered as rather imprecise. In complex systems of interrelated variables, scenarios can be constructed were estimates are either inflated or biased towards the null by measurement error [130]. In most cases, however, random noise due to imprecise assessment of the exposure is expected to produce underestimated effects [130]. The metabolomics measurements relied on blood samples from a single time point. Intra-individual biological variation is again considered a source of random noise. It has been estimated that even under the assumption of excellent reliability of a biomarker the imprecision related to single measurements could again lead to relevant underestimation of effect sizes [208,209]. Due to rigorous validation of the case assessment, false positives among the diabetes cases are unlikely. Despite the multi-source assessment strategy, some cases might have remained undetected, which again bears the potential to underestimate real effects. In summary, measurements in large-scale human cohorts are limited in terms of precision. In general,

this imprecision would be expected to produce a tendency to underestimate effects.

Erroneous decisions

In the present study, hypothesis testing was involved in terms of decision making. Test-based decisions are prone to two types of error, the possibility of false discoveries (type I error) and the possibility to overlook real dependencies (type II error). Moreover, each decision in the workflow of this study involved multiple tests.

First, models adjusted for a comprehensive set of confounders were used to select pairs of metabolites and external variables (dietary exposure or type2 diabetes risk) that required explanation. The different metabolite groups reflected different biological hypotheses but within the metabolite groups multiplicity was considered an issue. Within metabolite groups, false discovery rate [210] was controlled at 0.1 within metabolite groups. Thus, a relatively liberal significance-threshold was used at the first screening step. Type II error at the screening step implied that the respective relation was not further consider in the multi-model procedure.

Second, multiple models were calculated and if any of these models was non-significant the relation of the pair was not classified as unambiguous effect. This is to some extent the inverse of the multiplicity problem and the approach taken by the present study is in some respect analogous to Bonferroni correction: regardless of the number of submodels, if any estimate was non-significant the link was not classified as direct effect but remained ambiguous. Thus, a fairly conservative threshold was set at the second classification step. For the metabolite-diabetes relation 34 metabolites were classified as directly affecting type 2 diabetes risk in the present work. This closely resembled the number of metabolites that were considered as predictors of diabetes incidence in a previous regression selection, in which thirty-three metabolites from the same set were significantly associated with type 2 diabetes risk after Bonferroni correction over all tests [116]. Therefore, difference in the findings between the former study and the present work cannot mainly be attributed to threshold-effects. Of course, multiple tests also imply multiple opportunities to commit a type II error.

Two points are important: Firstly, p-values were used to aid the selection procedure for promising explanations as handy summary information on precision and size of the estimate. For single estimates,

the focus on confidence intervals and effect ranges was preferred. Secondly, to partially rely on p-values in the present study was a pragmatic decision. Depending on the study size and the data-type, potential future applications of a similar workflow to other study settings should consider the option to use other selection criteria.

Sources of bias

Cohort studies are generally prone to various sources of bias. Firstly, outcome-related misclassification of the exposure (or vice versa) can lead to biased estimates of the exposure-outcome relation. For example, selective misreporting is related to anthropometric traits, which are in turn related to the risk of type 2 diabetes. Therefore, selective underreporting of dietary intake levels by overweight participants can be a possible source for outcome-related misclassification of the exposure. Bias due to background factors that motivated misreporting should have been improved by comprehensively adjusting statistical models for phenotypical traits and lifestyle factors. Misclassification of the outcome is unlikely due to the rigorous case validation in the EPIC-Potsdam study. Reverse causation will be discussed below, under the heading *misspecification of the causal model*. Confounding generally is a major concern in observational studies. This study has accounted for confounding within the metabolomics network and this is a novelty for metabolomics applications in observational settings. Residual confounding, however, cannot be excluded and has the potential to having biased the results obtained in this study. Unstable confounding mechanisms can be revealed by external validation in similar studies and stable confounding mechanisms can probably be ruled out by combining different study designs [141]. Both approaches will be revisited in the last section of the discussion.

Misspecification of the causal model

Any misspecification of the causal model could have severely compromised the interpretation of the estimates as effects, and the interpretation of the results in general. A priori assumptions were clearly specified and are thus open to discussion. Wherever subject-matter knowledge was not detailed enough, a multi-model approach was taken to restrict the space of possible causal explanations for the observed data patterns. The multi-model procedure was informed by a graphical causal model estimated with the PC-algorithm. Failure of this tool to detect

mechanistic links might have resulted in a misspecified causal model. Furthermore, missing or erroneous information on important variables could have also resulted in a false model. On several occasions, the final causal model remained blurry by classifying effects as ambiguous. Due to the observational nature of the study, any causal claims including the causal models as a whole have a hypothetical character. This limitation is explicitly recognized. However, in nutritional epidemiology with chronic disease endpoints, specification of causal mechanisms consistent with the data is rather uncommon. Specifying a complex causal model that involves mediation hypothesis implies that model assumptions and claims on single links or groups of links are explicitly stated which should ease the falsification or validation in other studies.

5.3 Biological interpretation

5.3.1 Metabolites as pathway sensors

Accumulating evidence indicates that concentrations of metabolites in the blood are predictive markers of type 2 diabetes risk [117,187,211]. Causality of observed links between circulating metabolite concentrations and type 2 diabetes incidence, however, is often questioned. But such an unspecific question is hard to answer in general. Physiopathological processes underlying type 2 diabetes development are not mainly located in the vascular system. Thus, it is rather unlikely that metabolite concentrations in the blood are causal factors in a molecular biological sense. (Some metabolites might exhibit systemic signaling functions with the circulation as major site of action but these can be considered as special cases.) For the following discussion, it might therefore be helpful to put the question more accurately: Are metabolite concentrations in the circulation sensitive markers for metabolic processes and signaling activities (which likely take place in other tissues) that causally affect type 2 diabetes risk?

There is another important consideration attached to the interpretation of circulating metabolites as sensors for pathway activities. The finding that a serum metabolite is sensitive to the activity of a diabetes-related pathway does not necessarily imply specificity of that marker. On the contrary, most metabolite concentrations in the blood likely integrate information on the activity of several pathways in various tissues. Using the example of lipid metabolites, integration of multiple

pathway signals can be well illustrated. Lipid metabolites contain fatty acid residues. The nature of the attached fatty acid clearly depends on the availability of specific fatty acids within cells and tissues. Therefore, the circulating concentration of a particular metabolite includes information on fatty acid metabolism. At the same time, however, the circulating metabolite concentration depends on the availability of the head-group and backbone of that metabolite, e.g. sphingosine in case of sphingolipids or phosphatidylcholine in case of glycerophospholipids. These two processes are not necessarily equally related to type 2 diabetes risk. They might, but maybe in the opposite way - one with a beneficial and the other with an adverse effect on type 2 diabetes risk. This example is an oversimplification. The strong intercorrelation between metabolites implies that any effect on metabolites spreads across the network. Thus, single metabolite concentrations must be assumed to being affected by a variety of influential processes. The example underscores, however, that the aim of identifying the best sensors for pathways that are causally involved in type 2 diabetes development from a metabolomics dataset must rely on a modeling approach that controls for indirect effects by other signals.

Compared to the above discussed connections between metabolites and type 2 diabetes risk, linking dietary determinants to metabolic markers is easier in some regards. Evidently, dietary effects on circulating concentration of metabolites involve a variety of metabolic processes, including facilitated transport over membranes and enzymatically catalyzed reactions. Whenever a metabolite concentration in the blood is sensitive to the exposure level, however, the interpretation as effect is intuitive and plain. For the purpose of mediation analyses a general notion on total effects is, however, not sufficient. Observed diet-related alterations in circulating metabolite concentration need to be considered with respect to possible modulation of diabetes-relevant pathway activities. That is not given by observed effects on circulating metabolite concentrations per se because, as mentioned above, metabolite concentrations in the blood are likely to be sensitive to several metabolic processes in various tissues.

In the following the biological role of circulating metabolites as pathway sensors in relation to both, dietary exposures and type 2 diabetes risk, will be discussed. Therefore, metabolites will be organized in groups according to metabolic processes that might be relevant for type 2 diabetes risk.

5.3.2 Amino acids

Physiology

All tissues metabolize amino acids but the liver is the central site of nitrogen metabolism in the organism [212]. Surplus supply of amino acids with the diet leads to the storage of the carbon chains of amino acids as glucose (gluconeogenesis) or fatty acids (ketogenesis). Exclusively ketogenic amino acids are lysine and leucine. Aromatic amino acids and isoleucine can be used in gluconeogenic and in ketogenic pathways. All other amino acids are exclusively gluconeogenic, which means that their catabolism leads to intermediates of the Krebs cycle. In times of starvation amino acids are used for energy production. Furthermore, amino acids form building blocks for proteins and biologically active molecules, such as hormones [212].

Dietary effects on amino acids

In the present study the first factor of common variance among amino acids was inversely related to coffee consumption. Alterations of amino acid abundance in the blood in response to coffee consumption might be related to a regulatory effect of coffee on hepatic intermediary metabolism, which will be discussed below.

Whole-grain bread consumption was not related to amino acid levels in the blood. Coffee was related to lower circulating concentrations of branched-chain amino acid, phenylalanine, methionine and proline. An inverse relation of coffee intake with phenylalanine in men was reported by a previous study in EPIC-Potsdam [106]. Red meat consumption was related to higher concentrations of circulating branched-chain amino acids. Red meat is an important dietary source of branched-chain amino acids [213]. Furthermore, red meat intake was directly related to lower glycine levels, which is consistent with previous reports from EPIC-Potsdam [54]. Elevated glycine utilization for biosynthesis of heme or creatinine or glutathione in response to high iron intake and related oxidative challenges could be possible explanations of lower glycine levels in relation to high red meat consumption.

Effects of amino acids on type 2 diabetes risk

Metabolomics approaches have reignited the interest in the hypothesis raised by a study from 1969 on a potential etiological role of circulating

branched-chain amino acids in type 2 diabetes pathogenesis [187,214]. The first analysis of metabolomics data in relation to type 2 diabetes incidence by Wang et al. (2011) reported strong associations of branched-chain amino acids with type 2 diabetes risk [215]. A recent meta-analysis that comprised data from up to eight thousand participants including 1,940 incident cases of type 2 diabetes corroborated these findings [117]. Consistently, valine and the sum of leucine and isoleucine were significantly associated with type 2 diabetes risk in EPIC-Potsdam [116]. These associations were rendered non-significant, however, by mutually including metabolites that were associated with type 2 diabetes risk at the single metabolite level a joint regression model [116]. Accordingly, the potential link of branched amino acids with type 2 diabetes risk was not classified as direct effect in the present study. A possible explanation for this unstable association might relate to the measurements. As discussed above (5.2.4), the targeted metabolomics approach in EPIC-Potsdam only provided summary measures of isoleucine and leucine and differentiation between the two was thus not possible. An alternative biological explanation suggests that the link between branched-chain amino acids is of indirect nature and might be explainable by the adverse effect of aromatic amino acids on type 2 diabetes risk. Genetic evidence, however, supported a role for alterations in branched-chain amino acid catabolism in the development of type 2 diabetes [216]. Another Mendelian randomization study suggested that alterations in branched-chain amino acids were most likely early consequences of insulin resistance [217].

The association between aromatic amino acids, tyrosine and phenylalanine in particular, and elevated type 2 diabetes risk was similarly consistent across different prospective cohort studies [117], which was again in line with previous analyzes in EPIC-Potsdam [116]. In the present study phenylalanine and tyrosine were classified as having direct adverse effects on type 2 diabetes risk. Aromatic amino acids are also substrate for ketogenic pathways and might therefore sensitive to the same disturbances of degradation diabetes-related pathways as branched-chain amino acids. According to the results of the current study they might be even better sensors. Phenylalanine and tyrosine are, however, also substrate for the synthesis of a number of very potent signaling molecules including thyroxin and catecholamines [212]. These hormones are key-regulators of systemic energy metabolism and thereby implicated in metabolic homeostasis [218]. In light of the present results and former

observational studies, however, the strong inclination in the field to consider branched-chain amino acid over aromatic amino acids for biological interpretation and follow-up studies seems arbitrary to some extent.

In the present study, circulating concentrations of glycine and glutamine were classified as having direct beneficial effects on type 2 diabetes risk. Both findings are in line with the already cited meta-analysis of prospective cohort studies [117] and are further corroborate by recent studies in Chines cohorts [119,219]. Mendelian randomization studies did not support a direct causal effect of circulating glycine levels on type 2 diabetes risk [220]. However, equipped with the interpretation of metabolites as non-specific pathway sensors, it is evident that Mendelian randomization studies can only be interpreted with regard to the particular pathway by which considered genes affect the metabolite [221]. The consistent association of glycine with type 2 diabetes risk still requires explanation. Glycine is centrally involved in many metabolic key processes, e.g. oxidative stress response, folate metabolism, gluconeogenesis, and of course protein biosynthesis [212]. Therefore, studies on determinants of the circulating glycine concentrations in terms of quantitative contributions by different glycine-producing and -utilizing pathways would be highly desirable. Similar considerations apply to glutamine. Until such evidence is available, speculation on the processes underlying the beneficial direct effect of these two amino acids on type 2 diabetes risk remain difficult.

5.3.3 Acylcarnitines

Physiology

The major biological function of acylcarnitines is commitment of fatty acids to energy-generating oxidation [222]. Acylcarnitines are, however, not all oxidized immediately. Their export from cells is evident by detectability in the blood. Besides exercise, physiological determinants of acylcarnitine concentrations in the circulation include fasting status and dietary composition [223-227].

Dietary effects on acylcarnitines

Among food-exposures, whole-grain bread consumption was related to lower circulating concentrations of long-chain saturated acylcarnitines in the present study, stearoylcarnitine (C18:0) and palmitoylcarnitine

(C16:0) in particular. None of the two was unambiguously classified, however, as directly or indirectly affected due to non-significant but directionally consistent estimates in some network-adjusted submodels. In EPIC-Potsdam inverse associations between whole-grain bread intake and acylcarnitines in general were previously described [83] but not specifically for stearoylcarnitine and palmitoylcarnitine. Whole-grain dietary interventions in humans that investigated acylcarnitine responses were not identified in the literature. Yet an animal study demonstrated a clear effect of fiber supplementation on lipid metabolism including acylcarnitine levels in the muscle [228]. Unfortunately circulating acylcarnitine concentrations were however not reported.

Previous studies observed inverse associations between coffee consumption and long-chain and medium-chain acylcarnitines [107,108]. That was not the case in the present study. A possible explanation of this inconsistency might be the different fasting status between the studies with primarily non-fasted serum samples in the present study.

In the present work a direct effect of red meat consumption on higher serum concentrations of stearoylcarnitine (C18:0) was detected. Stearic acid is one of the most abundant saturated fatty acids in red meat [229]. Therefore, enrichment of stearoylcarnitine in response to high habitual red meat consumption seems plausible. Propionylcarnitine was also associated with red meat consumption and this association was not explained by controlling for the direct effect on stearoylcarnitine. Information on the relation of red meat consumption with higher serum propionylcarnitine concentration was still classified as ambiguous based on consistent but partly non-significant estimates in the submodels. Combined analysis of observational and interventional data in a study on biomarkers of dietary intake identified propionylcarnitine as marker for unprocessed and processed red meat intake [90] , which is in line with other observations [230]. Therefore, non-significance of some estimands might have been a power problem. Another possible explanation is that effect-attenuation in some submodels hints towards a partly indirect effect of red meat intake on propionylcarnitine concentrations in the blood. Propionylcarnitine is a product of branched-chain fatty and amino acid catabolism and this notion coincides with red meat being among the most important dietary sources of branched-chain amino acids [231]. The observation that higher red meat-related levels of

octadecenoylcarnitine (C18:1), acetylcarntine (C2:0) and carnitine (C0) were explainable by the direct red meat-effect on stearoylcarnitine seemed plausible because these metabolites are involved in long-chain fatty acid oxidation [232].

Effects of long-chain acylcarnitines on type 2 diabetes risk

Consistency with other studies

The present study observed direct effects of several acylcarnitines on type 2 diabetes risk. Long-chain saturated fatty acid containing palmitoylcarnitine (C16:0) was classified as having direct adverse effects on type 2 diabetes risk based on multi-model estimates. This metabolite was not among predictive markers for type 2 diabetes risk incidence in a previous regression-based selection procedure on the single metabolite-level in EPIC-Potsdam [116]. Palmitoylcarnitine was selected, however, as predictive diabetes risk-marker in a random survival forest-based selection procedure of predictors for type 2 diabetes risk [115]. Relevance of palmitoylcarnitine for type 2 diabetes risk is also supported by findings in other cohorts. A Chinese study applied targeted metabolomics (52 metabolites) in two case-control samples nested within independent prospective cohorts (1039 and 520 incident type 2 diabetes cases and the same number of controls). This prospective study identified palmitoylcarnitine as one of four metabolites that were consistently associated with type 2 diabetes risk across the two cohorts [119]. Another Chinese study in 2,103 participants aged 50–70 years and including 507 type 2 diabetes cases found elevated type 2 diabetes risk in relation to higher long-chain acylcarnitines [118].

In the present study, unsaturated long-chain and medium-chain acylcarnitines were classified as having a beneficial direct effect on type 2 diabetes incidence. This applies to tetradecenoylcarnitine (C14:1) and octadecadienylcarnitine (C18:2). These results stand in contrast to one of the aforementioned Chinese cohort studies where acylcarnitines C14:1 and C18:2 were markedly higher concentrated in the plasma of participants that later on developed type 2 diabetes [118]. Still the comparison is not straightforward because on the single acylcarnitine level the Chinese study presented unadjusted means in incident type 2 diabetes cases compared to controls only [118]. Moreover, the potential beneficial effect of unsaturated fatty acid containing acylcarnitines on type 2 diabetes risk in the present work was revealed only after

controlling for potential confounding by other acylcarnitines. This also explains why prior analyses in EPIC-Potsdam that did not take into account network information did not detect beneficial associations of acylcarnitines with type 2 diabetes risk [116]. Beneficial effects of acylcarnitines C14:1 and C18:2 on type 2 diabetes risk in the present studies coincide with lower plasma concentrations of these metabolites in patients with prevalent metabolic syndrome and type 2 diabetes compared to healthy controls (n≈40 per group) [233].

TEXTBOX 2: Fatty acid oxidation [2]

Inside cells long-chain fatty acids are determined for energy-generating catabolism by esterification with coenzyme A. For the transport over the mitochondrial membrane they need to be converted into acylcarnitines by *palmitoyltransferases* and shuttled over the membrane by *carnitine acylcarnitine translocases*. Under physiological conditions this is the rate limiting step of fatty acid oxidation. Within mitochondria acylcarnitines are reconverted into acyl coenzyme A. This fuels the mitochondrial energy-generation (*β oxidation*). Resulting acetyl coenzyme A is further utilized as substrate for citrate formation, which feeds the Krebs cycle.

Circulating long-chain acylcarnitine concentrations were thus related to type 2 diabetes risk in several prospective studies. Some heterogeneity of the results might point towards dependencies of the relation between specific acylcarnitines and type 2 diabetes risk on characteristics of the source population. Other possible explanations of inconsistencies are differences in ethnicity, age, health status, fasting status, analytical chemistry and statistical design between the studies. Taken together the available evidence yet suggests that acylcarnitine concentrations in the blood are markers for diabetes-related processes.

Possible mechanisms

Efflux of acylcarnitines from cells is believed to largely depend on the intracellular concentrations [234]. Still, physiological studies comparing acylcarnitine concentrations in arterial and venous blood samples from muscle and liver showed that the release of acylcarnitines to the circulation is complexly regulated. Several tissues contribute differently to the different circulating acylcarnitine species and the contribution

depends on the metabolic status [225,235]. Liver and muscle are considered major sources of circulating acylcarnitines because of the high content of mitochondria. Contribution of distinct tissues to the concentration of specific acylcarnitines in the peripheral blood, however, was not yet studied in detail [222,225]. On a systemic level acylcarnitines take part in energy transport between tissues [225,235]. To some extent acylcarnitines, however, simply leak into the circulation. In metabolically challenged cells acylcarnitine formation and export plays a role to relieve coenzyme A and to thereby secure continuation of intracellular metabolic processes [222,236]. Therefore, high abundance of long-chain acylcarnitines in the circulation was proposed to indicate a mismatch between fatty acid oxidation and the following utilization of resulting tricarboxylic acids in the Krebs cycle [222]. According to this model mitochondrial overload in metabolically active tissues leads to increased systemic levels of acylcarnitines, which could entail adverse metabolic effects.

Accumulating experimental evidence involves signaling activities of long-chain acylcarnitines in diabetes-relevant processes [2]. Physiopathological implications of dysregulated acylcarnitine metabolism are also indicated by observations in patients with genetic defects in fatty acid oxidation [237]. In vitro studies found that higher abundance of acylcarnitines locally affected inflammatory pathways [238,239]. Proinflammatory activities might partly explain observed relations of acylcarnitines with insulin resistance [240,241]. It was also demonstrated in vitro that palmitoylcarnitine blunts insulin-simulated phosphorylation of the serine/threonine-specific protein kinase Akt [242], and modulates protein kinase C activity [243,244] and ion-flux via calcium-depend transmembrane channels [245,246]. Moreover, in vitro models demonstrated that palmitoylcarnitine in high concentrations adversely affects cellular membrane integrity thereby leading to cellular stress responses [247].

Taken together, experimental mechanistical evidence suggests possible direct and indirect modulation of insulin signaling by palmitoylcarnitine. The concentrations of palmitoylcarnitine used in the in vitro model systems were much higher, however, than those found in vivo in the circulation [2]. Yet the actual site of action of palmitoylcarnitine within cells is unclear. It might therefore well be that metabolically produced acylcarnitines have regulatory effects at much lower concentrations compared to externally applied acylcarnitines. This

remains speculative nonetheless based on the available evidence. Mechanistical studies on the effects of unsaturated fatty acid-containing acylcarnitines were not identified in the literature. With regard to the beneficial effects of acylcarnitines C18:2 and C14:1, the suggested regulatory functions of acylcarnitines are of particular interest. Signaling activities would likely be sensitive to the different spatial conformation that distinguishes saturated from unsaturated fatty acids.

To conclude, adverse effects of palmitoylcarnitine on type 2 diabetes risk are in line with findings from other cohorts. Furthermore, in vitro evidence suggests regulatory effects of acylcarnitines on insulin-signaling and other diabetes-related processes. Another line of argument interprets circulating palmitoylcarnitine as a marker of a challenged energy metabolism. The findings on beneficial effects of desaturated fatty acid-containing acylcarnitines (C18:2 and C14:1) in the present study were not reported from other cohorts so far and need thus to be interpreted with caution. The absence of comparable results might be explained by the innovative analytical approach of this study that accounted for potential negative confounding by other acylcarnitines.

Short-chain fatty acid-containing acylcarnitines

Among short-chain fatty acid containing acylcarnitines propionylcarnitine was classified as having a direct adverse effect on type 2 diabetes risk in the present study. A similar risk estimate was reported from the EPIC-Norfolk study 1.15 (95%CI 0.98-1.34) [216]. Moreover, elevated plasma concentrations of propionylcarnitine were reported in relation to impaired glucose disposal [118] and insulin resistance in humans [215]. Higher propionylcarnitine concentrations were also observed in patients with prevalent type 2 diabetes and metabolic syndrome, respectively, compared to healthy controls [233].

Experimental studies on effects of propionylcarnitine on type 2 diabetes risk were not identified. As mentioned earlier in this section, propionylcarnitine is a marker for oxidation of branched-chain carbon chains and should therefore be interpreted apart from long-chain fatty acid utilization. Multi-model information on glutarylcarnitine [C5-DC(C6-OH)] was classified as ambiguous with regard to a direct effect on diabetes risk. Evidence on the metabolic role of this metabolite was not identified.

5.3.4 Sphingomyelins

Physiology

Sphingolipid synthesis in cells commences with condensation of a fatty acid acyl-coenzyme A (particularly palmitoyl-coenzyme A) with an amino acid, i.e. serine, glycine, or alanine. This reaction is catalyzed by the serine-palmitoylcarnitine reductase. Several enzymatic steps lead to the formation of ceramide [248]. Ceramide is precursor for the synthesis of other complex sphingolipids, including sphingomyelin. Sphingomyelins are synthesized by assembly of ceramide with a phosphatidylcholine-moiety [249]. Sphingomyelin is therefore the sphingolipid that structurally most closely resembles glycerophospholipids.

Dietary effects on sphingomyelins

Group levels of sphingomyelins were not sensitive to the investigated food exposures. Furthermore, whole-grain bread had no direct effect on single sphingomyelins. Coffee and red meat consumption, however, directly affected certain sphingomyelins. The direct effect of coffee consumption on unsaturated fatty acid-containing sphingomyelins explained the higher abundance of a large coffee-connected component. In addition, a direct effect on higher levels of C26:0 in sphingomyelins was observed. Higher sphingomyelin levels in relation to coffee consumption were already reported in a previous study [108], in which alterations in sphingomyelins correlated with coffee-related changes in blood lipids. Still, the associations with coffee consumption were stronger for sphingomyelins [108]. These observations might suggest that blood lipids and sphingomyelins are targeted by the same regulatory actions of coffee compounds on lipid metabolism. Red meat consumption also affected an interlinked component of several sphingomyelins, with some overlap to the coffee-connected component. Still, red meat seemed to affect other sphingomyelins, which contained long-chain and very long-chain monounsaturated fatty acid. Higher levels of sphingomyelins in consumers vs. non-consumers of red meat were recently reported from a study applying an untargeted lipidomics approach [190]. Furthermore, enrichment of long- and very long-chain saturated and monounsaturated fatty acids in sphingomyelins matches the fatty acid composition in red meat [250].

Effects of sphingomyelins on type 2 diabetes risk

Ceramide was implicated in insulin resistance [251-255] and inflammatory processes [256-258] by numerous studies. Consistent with the literature sphingomyelin group levels were not related to type 2 diabetes risk. Associations of sphingomyelins with metabolic traits were reported to depend on the fatty acid residues [259]. The investigated food items in the present study were not found to associate with the first component of common variance among sphingomyelins. General tendencies towards beneficial effects of monounsaturated and biunsaturated fatty acids on the risk of type 2 diabetes were in line with observations in phosphatidylcholines and will be discussed in the following section. In this section, the focus is on two particular pairs of sphingomyelins: sphingomyelins C18:0 and C18:1; and hydroxy-sphingomyelins C22:1 and C22:2. For both pairs, the partners were particularly strongly associated with type 2 diabetes risk, but in the opposite directions. The lower unsaturated metabolite (C18:0 and OHC22:1, respectively) had a strong adverse effect, whereas the higher unsaturated partner (C18:1 and OHC22:2, respectively) was highly beneficial. Still, the pairs were strongly correlated so that the strength of the association was only revealed in mutually adjusted models. The mutual dependency also explains why only few of the diabetes risk relations of sphingomyelins, that were detected in the present study, were also observed in previous studies in EPIC-Potsdam that operated at the single metabolite level [115,116]. Interestingly, a recent study on genetic association with metabolite ratios detected an association of diabetes-related variants in the sphingosine-1-phosphate phosphatase 1 gene with the ratio of hydroxy-sphingomyelin C22:1 to C22:2 (*Susanne Jäger, Scientific reports, under review*). A role of sphingosine-1-phosphate signaling in type 2 diabetes development is well established [260-262]. From a functional perspective, such metabolically closely related metabolites with an oppositely directed relation to type 2 diabetes risk could perhaps qualify as sensors for physiopathological relevant metabolic steps, e.g. as markers for enzymatic activities. Such pairs of interest also occurred in other lipid classes (e.g. lysophosphatidylcholines C18:2 and C20:3, and alkyl-acyl phosphatidylcholines C36/4 and C38/5) and are easily identified by visual inspection of the diabetes-linked networks.

5.3.5 Phosphatidylcholines

Physiology

Phosphatidylcholines are among the most abundant phospholipids in human cells and constitute the major lipid fraction in most cellular membranes [263]. The major fraction of phosphatidylcholines is produced by bonding of a CDP-choline moiety to diacylglycerol. Another pathway to produce phosphatidylcholine is conversion of phosphatidylethanolamine by step-wise methylation of the head-group, which particularly in hepatocytes can play a quantitative role [248].

Dietary effects on phosphatidylcholines

Diacyl phosphatidylcholine and alkyl-acyl phosphatidylcholine group levels (as reflected in the first factor of common variance) were sensitive to dietary exposures in the present study. A potential effect of diet on the abundance of whole groups of metabolites could be interpreted in two ways. Either specific foods might modulate substrate availability for synthesis of key components of that metabolite group; or foods might exhibit regulatory functions on synthesis and degradation processes, e.g. by targeting transcription factors or modulating enzymatic activities.

Whole-grain bread consumption was inversely related to the first factor of common variance among diacyl phosphatidylcholines. Observations of an inverse relation of whole-grain consumption with phosphatidylcholines were previously reported from EPIC-Potsdam [74,90]. In the present, study this relation was traced back to diacyl phosphatidylcholines rather than other phosphatidylcholine-groups.

Coffee consumption was inversely related to the first factor of common variation among diacyl phosphatidylcholines but positively related to the first alkyl-acyl phosphatidylcholine-factor. Coffee does not contribute relevant amounts of lipids or other building blocks for lipid metabolites to the diet. Coffee, however, is exceptionally rich in phytochemicals [264]. A regulatory effect of coffee consumption on lipid metabolism in general is well documented in the literature. Coffee is the major dietary source of kahweol and cafestol. The cholesterol- and triglyceride-raising effect of coffee that was demonstrated in intervention trials [265] was mainly attributed to biological actions of these of two diterpenes [266]. Coffee, however, affects lipid metabolism in the liver, the adipose tissue

and in other tissues in a complex way by the action of several compounds [267]. Human intervention studies also revealed beneficial changes in the high density lipoprotein (HDL)- to low density lipoprotein (LDL)-cholesterol ratio in response to coffee consumption [46]. Glycerophospholipid composition was reported to be related to the HDL-to-LDL ratio [268,269]. Human trials further showed that coffee intake had beneficial effects on the composition of LDL-particles including lower contents of apolipoprotein B and AI [46] and higher antioxidant capacity [270-272]. The latter point might be related to observations in the present study where higher coffee intake was related to higher alkyl-acyl phosphatidylcholine levels. These ether-lipids have antioxidant properties [273]. Taken together, the summarized evidence indicates regulatory effects of coffee consumption on lipid metabolism. This might point towards a biological explanation for the observed association of coffee consumption with the common variance in several phosphatidylcholine-groups.

Red meat consumption was related to the first factor of common variance among alkyl-acyl phosphatidylcholines. This is related to previous EPIC-Potsdam analyses where red meat loaded positively on a dietary factor that was associated with higher alkyl-acyl phosphatidylcholine levels [100]. This finding might simply reflect the contribution of red meat to dietary lipid intake. Another possible explanation relates to pro-oxidative processes that were suggested to be induced by red meat-related high heme intake [52]. Upregulation of the antioxidant alky-acyl phosphatidylcholines [273] might hint towards a compensatory response. These interpretations, however, are mostly speculative because other studies have not yet investigated effects of red meat consumption on ether lipid metabolism.

Effects of phosphatidylcholines on type 2 diabetes risk

Within cells the major proportion of incoming fatty acids is bound to glycerol backbones [274,275]. The major route to lipid storage goes from monoacyl glycerol via diacyl glycerol to triacyl glycerol [274]. Triacyl glycerols are metabolically relatively inert but diacyl glycerol levels in cells were involved in insulin resistance [276]. The particular biological role of diacyl glycerols in cells remains, however, matter of debate [277]. One line of reasoning considers diacyl glycerol as precursor for glycerophospholipid synthesis. Glycerophospholipids could therefore be

the functional relevant downstream factors. In this regard, experiments in genetically modified mice are of interest. Muscle-specific knockout of ethanolamine-phosphate cytidylyltransferase in mice lead to an increase in muscular diacyl glycerol content. These animals were nevertheless protected against insulin resistance and had an improved metabolic flexibility [278]. In wildtype animals and humans, high diacyl glycerol levels were associated with disturbed insulin signaling [276]. These findings suggest that high diacyl glycerol levels might be sensors for adverse alterations in downstream glycerophospholipid metabolism rather than being directly involved in insulin signaling.

Consistent with these considerations, alterations in the relative abundance of glycerophospholipids were related to dysregulated energy and glucose metabolism [249]. In the present study, higher diacyl phosphatidylcholine concentrations were associated with higher risk of type 2 diabetes. These findings resemble previous analyses in EPIC-Potsdam where the first component of common variance among all metabolites was dominated by diacyl phosphatidylcholines and was associated with an elevated type 2 diabetes risk [116]. Diacyl phosphatidylcholines were associated with adverse metabolic traits including higher 2-hours glucose after oral challenge [279] and higher BMI [280] in other studies. For other traits, such as non-alcoholic fatty liver disease (NAFLD), the relation with metabolically closely related phosphatidylethanolamines was suggested to be of interest [281,282]. The targeted metabolomics approach of the present study did not cover phosphatidylethanolamines.

Alkyl-acyl phosphatidylcholine concentrations were associated with a reduced type 2 diabetes risk in the present study. Again, these findings are to some extent replication of previous results from EPIC-Potsdam where the second component of common variance among all metabolites was dominated by alkyl-acyl phosphatidylcholine and was associated with a reduced type 2 diabetes risk [116]. Alkyl-acyl phosphatidylcholines were reported to be inversely related with obesity [280], prediabetes, prevalence of type 2 diabetes, and glucose clearance after oral challenge [279]. To sum up, present findings are in line with results from other studies. This suggests that the abundance of glycerophospholipid-subgroups in the circulation is related to the risk of type 2 diabetes.

5.3.6 The role of fatty acid residues

Lipid metabolites and fatty acid composition

Lipid metabolites in the circulation integrate information on fatty acid metabolism. Circulating glycerophospholipids predominantly reflect lipid metabolism in the liver [249]. Fatty acid composition in lipid compartments is sensitive to dietary habits [121,283]. Moreover, several lines of evidence relate specific fatty acids to type 2 diabetes risk [3,120,121,284]. Lipids are structurally highly diverse and metabolomics approaches are still heterogeneous with regard to lipid characterization, which makes across study comparisons of metabolomics findings on lipids difficult [117]. This is different for the highly reliable analyses of total phospholipid-bound plasma fatty acids [121]. To measure total plasma phospholipid fatty acids, phospholipids are separated and acyl side-chains are chemically cleaved from whichever backbone bound to. Then single fatty acids are quantitatively determined by gas chromatography. Results on the role of specific fatty acid residues in phospholipids will therefore be compared to the association of cleaved fatty acids from the phospholipid compartment with the risk of type 2 diabetes. Dietary effects on fatty acid composition will also be discussed in this context.

Saturated and monounsaturated fatty acids

Dietary effects

Whole-grain bread was related to lower concentrations of saturated fatty acids. Particularly palmitate-containing metabolites were lower in several phosphatidylcholine-compartments (i.e., lysophosphatidylcholine C16:0, likely diacyl phosphatidylcholine 32/0, C32/1, C34/1, and alkyl-acyl phosphatidylcholine C34/1) and in acylcarnitines. In addition, stearate-containing metabolites were less abundant in the circulation of participants with high whole-grain bread consumption (acylcarnitine C18:0, lysophosphatidylcholine C18:0, diacyl phosphatidylcholine C36/0). These observations might suggest reduced de novo lipogenesis in relation to high habitual whole-grain bread intake. Evidence from animal model demonstrated modulatory effects of microbiota derived short-chain fatty acids on hepatic de novo lipogenesis by regulating transcription factors in the liver [285-287]. These mechanisms were shown to mediate the effect of fiber intake on hepatic lipid metabolism

in animals [288,289]. Whole-grain rich diet was shown to modulate the microbiota in humans [290].

The most important triggers of de novo lipogenesis still are high sugar and insulin levels in the blood [1]. Whole-grain rich foods are believed to be related to a favorable blood-sugar and insulin response [291]. Insulin-signaling is the major regulator of hepatic lipogenesis. Whole-grain bread consumption might also exhibit direct modulatory effects on insulin secretion. This hypothesis is supported gene diet interactions. The beneficial effect of whole-grain bread consumption seemed to depend on a functional variant of the transcription factor-7-like 2 (TCF7L2) gene and was abolished in carriers of the deleterious variant [292-294]. Genetic polymorphisms in TCF7L2 are believed to trigger early defects in insulin secretion and were consistently associated with type 2 diabetes risk [295]. Reduced de novo lipogenesis seems also a reasonable explanation for the lower concentrations of monounsaturated fatty acid-containing metabolites that were observed in all phosphatidylcholine-groups.

Among saturated fatty acids, coffee consumption was related to higher levels of the odd-chain margaric acid (C17:0), very long-chain cerotic acid (C26:0), and monounsaturated hydroxysphingomyelin C16:1. All three links were classified as direct effects. Sphingomyelins with stearate (C18:0), C18:1 (likely oleate), palmitate (C16:0), and C22:1 were indirectly affected by coffee consumption. On the contrary, myristic acid (C14:0) and C16:1 in lysophosphatidylcholines were inversely linked to coffee

TEXTBOX 3: De novo lipogenesis [1]

De novo lipogenesis characterizes the synthesis of fatty acids from acetyl-coenzyme A. The main products are saturated (C14:0 to 18:0) and monounsaturated fatty acids (C16:1 and C18:1). The biochemical process takes place primarily in the liver. The substrate acetyl-coenzyme A is most commonly derived from carbohydrate catabolism. Thus hepatic de novo lipogenesis is the major route to convert dietary sugars into storage lipids. To this end hepatic de novo lipogenesis is activated by high blood sugar levels and insulin signaling but inhibited by malonyl-coenzyme A and fatty acyl-coenzyme A. The latter prevents futile cycling of carbon chains between fatty acid oxidation and de novo lipogenesis. De novo lipogenesis was implicated in the development of hepatic insulin resistance, non-alcoholic fatty liver-disease and type 2 diabetes.

consumption. The same inverse relations might have also been reflected in the direct coffee effect to lower diacyl phosphatidylcholine C32/1.

Animal models have unraveled that several compounds in coffee have the potential to modulate the activity of key transcription factors of hepatic lipid metabolism. Effects of coffee (and its components caffeine, chlorogenic acid, and polyphenols) to reduce hepatic steatosis in animal models were traced back to altered activities of transcription factors including sterol regulatory element-binding protein 1-c (SREBP1-c) peroxisome-proliferator activated receptors alpha and gamma (PPAR-α and PPAR-γ), cluster of differentiation 36 (CD36), and fatty acid binding protein 4 (FABP4). Accordingly, dependent lipid metabolizing enzymes were also regulated by coffee administration including acetyl-coenzyme A carboxylase-1 (ACC1) and stearoyl-coenzyme A desaturase-1 (SCD1) [296-299]. Taken together, these experiments in animals indicate that fatty acid oxidation and lipid export from the liver is enhanced and de novo lipogenesis is reduced by coffee intake. It might be speculated that the higher content of saturated and monounsaturated fatty acids in sphingomyelins is related to enhanced lipid export from the liver [300] in relation to coffee consumption. Lower phosphatidylcholine levels might rather point towards reduced de novo lipogenesis [1]. Systematic investigation of the metabolomics signature of these traits in the blood in large human samples, however, is still missing.

Phosphatidylcholines with saturated and monounsaturated fatty acids of 14 to 16 carbon atoms chain-length were lower in relation to red meat consumption. This might again be interpreted as reflection of reduced hepatic de novo lipogenesis. Reduced hepatic lipogenesis would be expected to some extent in response to high intake of fat-rich foods [1]. Long-chain saturated fatty acids and monounsaturated fatty acids of 18 and more C-atoms were enriched in relation to red meat consumption across all analyzed lipid groups. In particular stearate-containing metabolites were classified as directly affected by red meat consumption (acylcarnitine C18:0 and diacyl phosphatidylcholine C38/0). This seems also plausible because red meat importantly contributes to stearate content of mixed diets [301].

Effects on type 2 diabetes risk

In the present study, enrichment of several saturated fatty acid containing lipid metabolites was estimated to have direct adverse effects

on type 2 diabetes risk. This applies to myristic acid (C14:0) in lysophosphatidylcholines; to palmitic acid (C16:0) in acylcarnitines and lysophosphatidylcholines; and to stearic acid (C18:0) in sphingomyelins. Furthermore, saturated fatty acids (C16:0, C18:0, and/or C20:0) contained in alkyl-acyl phosphatidylcholine C36/0 had a direct effect on type 2 diabetes risk. The largest prospective study on the association of individual saturated phospholipid-bound fatty acids with type 2 diabetes incidence was conducted in the EPIC-Interact cohort [120]. This study included 12,403 participants with incident type 2 diabetes and a subcohort of 16,154 participants in a case cohort design. Thus, results are generalizable to the EPIC-source cohort with 340,234 European participants. Among saturated fatty acids in the phospholipid compartment, myristic acid (C14:0), palmitic acid (C16:0), and stearic acid (C18:0) were associated with a higher type 2 diabetes risk. The strongest risk relation was observed for palmitic acid (C16:0) with a hazard ratio of 1.26 (95%CI 1.15 - 1.37) per standard deviation.

Heptadecanoic acid (C17:0) lysophosphatidylcholine was associated with reduced type 2 diabetes risk in the present study. This is also supported by the association of odd-chain saturated fatty acids with reduced risk of type 2 diabetes in EPIC-Interact. Among odd-chain fatty acids, the strongest risk reduction was associated with heptadecanoic acid [HR per standard deviation 0.67 (95%CI 0.63 - 0.71)] [120].

Very long-chain fatty acids were very likely contained in diacyl phosphatidylcholine C42/1 and perhaps also in alkyl-acyl phosphatidylcholine C42/3. In the present work, both metabolites were classified as having direct beneficial effects on type 2 diabetes risk. This is in line again with EPIC-Interact analyses, where very long-chain fatty acids were associated with reduced type 2 diabetes risk [120].

In the present study, lipid metabolites from several lipid compartments that contained monounsaturated fatty acids were classified as having direct beneficial effects on type 2 diabetes risk. This applies to C14:1 in acylcarnitines; to C16:1 sphingomyelins; to C18:1 lysophosphatidylcholines and sphingomyelins; to C24:1 in sphingomyelins; and to C42/1 in diacyl phosphatidylcholines and C30/1 alkyl-acyl phosphatidylcholines. Data on the longitudinal association of monounsaturated fatty acids with type 2 diabetes risk are sparser and the picture is more heterogeneous [283,302].

TEXTBOX 4: Fatty acid desaturation and elongation [3]

Fatty acid desaturases are important endogenous determinants of the fatty acid profiles in cells and tissues. The stearoyl coenzyme A desaturase catalyzes the conversion of saturated into monounsaturated fatty acid; Δ5 and Δ6 desaturases constitute the rate limiting factors of conversion of ω-3 and ω-6 polyunsaturated fatty acid into higher unsaturated products. The two enzymes have different substrate specificities. The grade of desaturation determines the spatial structure of fatty acids thereby their physiological function. Desaturase activities have been implicated in type 2 diabetes development.

Polyunsaturated fatty acids

Dietary effects

Among polyunsaturated fatty acids, whole-grain bread consumption was related to lower circulating concentrations of diacyl phosphatidylcholines C34/3, C36/3, C36/5, and C38/5. None of these links were classified as direct whole-grain bread effect. In contrary to the very consistent picture for the whole-grain bread effect on saturated fatty acids and monounsaturated fatty acids across metabolite-groups, these findings stood alone and were therefore not further interpreted.

Coffee was related to higher levels of several metabolites that contained fatty acids with two desaturations. The higher coffee-related levels of lysophosphatidylcholine C18:2 most likely correspond to linoleic acid enrichment, and higher concentrations of alkyl-acyl phosphatidylcholines C34/3 and C36/2, and the direct effect on sphingomyelin C20:2 might point to the same or metabolically closely related fatty acids. The potential regulatory effect of coffee on hepatic de novo lipogenesis was already discussed above in relation to saturated fatty acids. It should be noted that it is the ratio of palmitic to linoleic acid that is used as indirect index for de novo lipogenesis in the liver [303].

Red meat was associated with higher concentrations of lysophosphatidylcholine C20:4, which most likely contained arachidonic acid. The interpretation is again straightforward because red meat is

among the most important dietary sources for arachidonic acid [159,250]. This finding seems to be related to the higher concentrations of diacyl phosphatidylcholine C38/4 and alkyl-acyl phosphatidylcholine C36/4 that were also linked to high red meat intake in the present study. These metabolites likely contained arachidonic acid or metabolically closely related fatty acids.

Effects on type 2 diabetes risk

Among polyunsaturated fatty acids, enrichment of C18:2 (likely linoleic acid) in lysophosphatidylcholines and acylcarnitines was classified as having a direct beneficial effect on type 2 diabetes risk. In line with these observations, significant inverse association of linoleic acid (18:2ω6) with the risk of type 2 diabetes were reported from EPIC-Interact [121]. An attractive hypothesis is that the beneficial effects of diacyl phosphatidylcholine C32/3, and of alkyl-acyl phosphatidylcholines C32/2, C34/3, and C42/3 might also reflect the beneficial associations of linoleic acid and eicosadienoic acid (20:2ω6) with type 2 diabetes risk. Possibly, these metabolites also (partly) contain alpha-linolenic acid (18:3ω3) which was also related to a lower risk of type 2 diabetes in EPIC-Interact [121]. Comparisons regarding polyunsaturated fatty acids remain somewhat speculative, however, because of the limitation of the targeted metabolomics approach of the present study. Localization of the desaturation was not resolved and chain-length of the individual fatty acids was not specified for metabolite with two bound fatty acid residues. Other lipid metabolites with polyunsaturated fatty acid were classified as having a direct adverse effect on type 2 diabetes risk: C20:3 in lysophosphatidylcholines; C38/3 and C42/5 in diacyl phosphatidylcholines; and C36/4 in alkyl-acyl phosphatidylcholines. In EPCI-Interact, adverse associations with type 2 diabetes risk were detected for γ-linolenic acid (18:3ω6), dihomo-γ-linolenic acid (20:3ω6), docosatetraenoic acid (22:4ω6). Therefore effects of fatty acid residues in lysophosphatidylcholines correspond intriguingly well with results from the EPIC-Interact cohort [121]. For the other metabolites with two fatty acids bound comparison comparisons with cleaved fatty acids are more speculative again. It seems still likely that above mentioned metabolites partly contain long-chain fatty acids with three or four desaturations that might correspond to above mentioned omega-6 polyunsaturated fatty acids. Arachidonic acid was not significantly associated with type 2

diabetes risk [121], which is also in line with the present observation that lysophosphatidylcholine C20:4 was not related to type 2 diabetes risk.

With respect to the importance of fatty acid residues for type 2 diabetes risk, results from the present study were well in line with EPIC-Interact results on cleaved fatty acid from plasma phospholipids [121]. It should be noted that the role of fatty acid residues was not apparent from results of a p-value based regression selection on the single metabolite level [116]. The primary aim of this previous study was diabetes prediction and in this respect it was successful. For biological interpretation, however, the most obvious pattern was that all diacyl phosphatidylcholines were related to elevated type 2 diabetes risk whereas all alkyl-acyl phosphatidylcholines were related to reduced risk of type 2 diabetes. There were some suggestions of rather adverse association of saturated fatty acids with shorter chains and rather beneficial association of polyunsaturated fatty acids with longer chains. The picture on the role of fatty acid residues remained blurry and the interpretation remained vague. Consistent patterns across metabolite groups could not be easily inferred from the results [116]. Prediction of type 2 diabetes was slightly improved by applying a random survival forest-based metabolite selection [115]. Learning predictive metabolites with this machine did not, however, did also not facilitate greatly the biological interpretation of the results.

The network adjustment approach was intended to block confounding mechanisms, e.g. influence of the metabolite group level on type 2 diabetes risk or other metabolites that affected type 2 diabetes risk. Consistency of the results on fatty acid residues across metabolite groups and with analyses of cleaved fatty acids suggests that this aim was achieved. A major advantage of the network adjustment approach taken by the present study over analyses of cleaved fatty acid is that the information on the backbone is conserved. The backbone locates fatty acids in the cell and is thus biologically informative.

5.3.7 Fatty acids in lipid compartments

Lipid trafficking

Biosynthesis of glycerophospholipids and cholesterol within the cell primarily takes place within the endoplasmatic reticulum [304]. Cholesterol is immediately trafficked towards the plasma membrane. Ceramide synthesis is also localized in the endoplasmatic reticulum [305].

Glycerophospholipids and ceramide are trafficked to the Golgi apparatus. This is where sphingolipids are synthesized from the ceramide backbone [306]. From the Golgi apparatus lipids are further trafficked to the plasma membrane and other subcellular domains via vesicular and alternative routes. Endocytosis routes organize the revers transport of lipids from the cell surface towards inner organelles via early endosomes, late endosomes and lysosomes [307].

Along these biosynthetic and degradation routes the lipid composition is differentially regulated. The membrane of the endoplasmatic reticulum is dominated by glycerophospholipids [304]. Contents of sterols and sphingolipids increase along the outlined intracellular trafficking routes. The inner layer of the plasma membrane is still relatively rich in glycerophospholipids whereas the outer layer is dominated by sphingolipids and cholesterol [304]. The complex regulation extends from intracellular organization to differences between cell types and tissues [308].

A functional perspective on subcellular fatty acid localization

The highly organized and regulated lipid composition clearly has major functional implications. Lipid composition in membranes modifies the content and activity of membrane proteins and membrane-associated proteins (i.e. receptors, transporters and signaling factors) by different mechanisms [309]. Lipid signaling per se has been implicated in diabetes pathogenesis in many ways. Metabolites of arachidonic acid and related polyunsaturated fatty acids (e.g. thromboxanes, prostaglandins, leukotrienes, and lipoxins) modulate inflammatory responses [310] and were implicated in insulin resistance [311,312] and β-cell decay [313,314].

These lipid mediators, however, are very short-lived and mainly exhibit autocrine and paracrine functions [311]. These reflections support the hypothesis that the subcellular localization of the polyunsaturated fatty acid substrates for inflammatory-mediator production matters. The evidence on lipid trafficking implicates that this information is partly inferable from the backbone of fatty acid-containing metabolites. The short-living nature of inflammatory lipid signals makes it particularly difficult to find markers for these processes in observational human studies. According to these considerations, phospholipid remodeling has been centrally implicated in acute and chronic inflammation by a line of experimental works over the last decade [315]. Possibly the results of this work on specific

polyunsaturated fatty acid-containing metabolites with adverse effects and others with beneficial effects on type 2 diabetes risk might point towards sensors of lipid-mediated inflammation. To validate such sensors would involve considerable effort, including mechanistical studies and tracer studies in humans. Given that lipid-mediated inflammation has been proposed to be sensitive to diet [316-319] and can be pharmacologically targeted [320] it seems warranted to further investigate in this direction.

Lipid signaling was also implicated in other diabetes-relevant pathways, among them direct effects on insulin signaling [312,321], binding of transcription factors [322], and modulation of the endoplasmatic reticulum stress response [323]. These pathways provide alternative explanations for the effects of saturated and unsaturated fatty acids on type 2 diabetes risk that were observed in the present study. Developing stable biological markers applicable to epidemiological studies would open a range of opportunities to study the relevance of lipid signaling in humans under real-life conditions.

5.3.8 Mediation

Whole-grain bread and type 2 diabetes risk

In the current study, high whole-grain bread consumption was linked to lower levels of three saturated fatty acid-containing lipid metabolite, which in turn were related to higher type 2 diabetes risk. Adjusting the whole-grain bread-type 2 diabetes association for the network-independent residual variance in the metabolite levels attenuated the whole-grain bread effect on type 2 diabetes by more than one quarter. This observation is consistent with the hypothesis of an important contribution of the metabolic pathway reflected by these metabolites to mediate the beneficial effect of whole-grain bread on type 2 diabetes risk.

As discussed above (5.3.6), the lower concentrations of saturated fatty acids in relation to high whole-grain bread intake might reflect reduced de novo lipogenesis. Possible explanations for reduced hepatic lipogenesis in response to whole-grain bread consumption could be an effect on the microbiota-derived short-chain fatty acids, and subsequent regulatory actions of short-chain fatty acids; or direct effects on insulin signaling, for example by modulatory effects on insulin secretion.

Coffee and type 2 diabetes risk

Frequent coffee consumption was related to lower phenylalanine levels, and to alterations of lipid metabolites. More specifically, people who drank more coffee had higher levels of several polyunsaturated fatty acid-containing sphingomyelins and phosphatidylcholines, and lower levels of myristic acid (C14:0) in lysophosphatidylcholines. The polyunsaturated fatty acid-containing metabolites in turn beneficially affected type 2 diabetes risk, whereas phenylalanine and lysophosphatidylcholine C14:0 had direct adverse effects on type 2 diabetes risk. Adjusting for the network-independent residual variance in the circulating levels of these metabolites attenuated the coffee effect on type 2 diabetes by two thirds. This observation is consistent with the hypothesis of the selected potential mediators reflecting key pathways that link coffee consumption to type 2 diabetes risk.

A role of circulating phenylalanine in type 2 diabetes development is suggested by observational evidence but, so far, this hypothesis has hardly been investigated in mechanistical studies. A protective role of particular polyunsaturated fatty acid on type 2 diabetes risk is well supported by the literature [121,283]. For example, linoleic acid (18:2ω6) was related to anti-inflammatory effects [318], improved insulin sensitivity [324], and anti-oxidative effects [325,326]. Unfortunately, the targeted metabolomics approach did not provide the location and conformation of unsaturations in fatty acids. In similar European populations, linoleic acid (18:2ω6) was on average at least tenfold higher concentrated compared to other polyunsaturated fatty acid with two or three unsaturations within the phospholipid compartment [121]. Therefore, it seems likely that observed alterations were at least partly attributable to alteration in the content of linoleic acid in the investigated metabolite groups. Accordingly, above discussed regulatory impact of coffee consumption on hepatic lipid metabolism (5.3.5, 5.3.6) might affect polyunsaturated fatty acid metabolism (including enrichment of linoleic acid), which could in turn have beneficial effects on various diabetes-related processes including local insulin sensitization and anti-inflammatory actions. The lower levels of myristic acid in lysophosphatidylcholines in relation to high coffee consumption might point towards reduced hepatic de novo lipogenesis, which would also be in line with a beneficial effect of coffee consumption on hepatic lipid metabolism.

A previous study on potential metabolomics-based biomarkers as potential mediators of the link between coffee consumption and risk of type 2 diabetes in EPIC-Potsdam took another approach [106]: Metabolites were considered as potential biomarkers only if they were selected as independent predictors of type 2 diabetes risk in a p-value-based regression selection before [116]. Whereas results agreed upon a potential mediating role of phenylalanine with the present study, none of the investigated lipid metabolites was selected as potential mediator in the former analyses [106]. This might have two reasons. Firstly, most of the lipid metabolites that were identified as potential mediators in the present study were simply not included in the former work (due to the focus on previously selected significant predictors). Secondly, consideration of network-adjusted residual variance should have helped to focus on the unconfounded direct effects. Results of the current study in line with experimental evidence suggested complex effects of coffee consumption on lipid metabolism. Therefore, control of network-inherent indirect influences might have revealed real effects that were not visible by considering single metabolites apart from the network information.

Red meat and type 2 diabetes risk

Red meat was directly related to lower circulating glycine concentrations. Glycine in turn was classified as having a beneficial effect on type 2 diabetes risk. Furthermore, red meat consumption was directly linked to higher levels of saturated and polyunsaturated fatty acid-containing metabolites, which in turn adversely affected type 2 diabetes risk. Adjusting for the network-independent residual variance in the serum concentration of these metabolites attenuated the red meat effect on type 2 diabetes by 70%. The observed effect attenuation is consistent with the hypothesis of the selected potential mediators being sensitive to pathways that are centrally involved in linking red meat consumption to type 2 diabetes risk.

The role of lipid signaling in diabetes development was discussed above. Among saturated fatty acids, particularly palmitate has been implicated in cellular stress responses [327,328] and inflammation [329-331] in the liver. Polyunsaturated fatty acid metabolites are even more directly involved in inflammatory responses. Ω-6-polyunsaturated fatty acids are precursors of many very potent local inflammatory signals, i.e. eicosanoids [310,323]. Moreover, eicosanoid-mediated inflammation has

been related to β-cell decay [313,314]. Again, it is important to clarify that the exact structure of fatty acids was not provided by the applied analytical chemistry. Data on cleaved fatty acids from phospholipids in large population-based European samples showed that beside linoleic acid arachidonic acid (20:4ω6) and dihomo-γ-linoleic acid (20:3ω6) are by far the highest concentrated polyunsaturated fatty acids with four or less unsaturations [121]. Therefore, one of the two is likely contained in alkyl-acyl phosphatidylcholine C36/4. Only dihomo-γ-linoleic acid (20:3ω6) was related to risk of type 2 diabetes on the level of cleaved fatty acids from phospholipids [121]. Analyses in Dutch and German cohorts were not consistent with a role for C-reactive protein in mediating the red meat-related type 2 diabetes risk [332,333]. Therefore, defining possibly more specific markers for red meat-related subacute inflammatory processes is of considerable interest. According to the discussion above, however, lipid signaling might also affect diabetes pathogenesis by non-inflammatory pathways, e.g. by interfering directly with insulin signaling (5.3.7).

A red meat effect on lipid signaling could be related to the lipid content of red meat, which might provide bioactive lipids or substrate for synthesis thereof. Another potential explanation for red meat effects on lipid signaling involves regulatory factors. For example, oxidative stress related to heme-iron might trigger cellular stress responses that involve systemic lipid signaling [334]. A quite novel line of reasoning involves the non-human sialic acid N-glycolylneuraminic acid, which is primarily taken up by consumption of meat from large mammals [335] and subsequently incorporated in membranes of meat eaters [336]. Membrane-standing N-glycolylneuraminic acid has the potential to trigger inflammatory immune responses [337,338]. Several alternative hypotheses on red meat compounds that could affect cellular and systemic stress signals are available, which were comprehensively reviewed elsewhere [52,213,336,339].

In a previous study in EPIC-Potsdam, the same metabolomics dataset was considered for the identification of mediators of the red meat-associated type 2 diabetes risk on a single metabolite level. This former study identified several lipid metabolites as potential mediators [54]. By the comparison of the results from this former to the present study, some points can be further clarified.

First, there were some differences in the red meat-related metabolites between the studies but the major interpretation was similar:

Red meat was linked to enrichment of long-chain saturated and specific polyunsaturated fatty acids in several lipid groups. By the network-adjusted selection approach of the present study, however, this information was more consistently shown across different lipid groups.

Second, the selected lipid mediators seemed to partly incorporate similar information across the studies although different metabolites were selected. For example, diacyl-phosphatidylcholines C36/4 and C38/4 (selected in as mediators in the former study) might contain closely related or the same fatty acids as alkyl-acyl phosphatidylcholine C36/4 (selected as mediator in the present study). Alkyl-acyl phosphatidylcholine C36/4 was not selected as a potential mediator the single metabolite level although the metabolite was associated with red meat consumption. The reason was that an association with type 2 diabetes risk was not detected in the previous study [54]. Meanwhile, the interpretation of this difference between the studies might be easily comprehensible. Particularly for alkyl-acyl phosphatidylcholines, the beneficial effect of the ether lipids on the group level concealed all adverse effects of specific fatty acid residues within the group if the network information was not considered. Confounding by the group level was blocked by the network adjustment. The potential mediating role of saturated fatty acid enrichment in alkyl-acyl phosphatidylcholines was also not detected on the single metabolite level in the former study [54], but explained a major proportion of the red meat effect on type 2 diabetes risk in the present work.

Third, the proportion of the red meat effect on type 2 diabetes risk explainable by lipid mediators was improved by the network-adjusted approach of the present study. For example, the proportion mediated by network-adjusted residuals of alkyl-acyl phosphatidylcholine C36/4 (18%) was larger compared to the proportions mediated by diacyl phosphatidylcholines C36/4 and C38/4 (≤14%). The notion of a larger explainable proportion also applied to selected lipid mediators in general, which suggests that network-informed metabolite residuals might be more accurate sensors for red meat- and diabetes-related metabolic processes.

Glycine was estimated to explain a major proportion of red meat-related type 2 diabetes risk. Glycine is a central molecule in intermediary metabolism and, by that, involved in a variety of diabetes-related metabolic processes (5.3.2). Identification of glycine as potential mediator of the red meat effect on type 2 diabetes risk replicates findings

from the previous mediation analysis on the single metabolite level [54]. Again, the explainable proportion was slightly increased by using network-adjusted residuals. More strikingly, however, variance of the estimated proportion mediated by glycine was markedly reduced in the present study. In the previous study, evaluating the stability of the proportion of the red meat-related type 2 diabetes risk mediated by glycine in a bootstrapping procedure yielded a 95% confidence interval that covered the whole range from below five percent up to hundred percent. The majority (95%) of bootstrapping estimates on the mediated proportion by using network-adjusted glycine residuals in the present analysis fell into the range of roughly twenty to seventy percent. Increased precision of the estimates on mediated proportions in the present compared to the previous study was observed in general. The marked increase in precision could be interpreted as another hint towards having identified more accurate sensors for biological effects in the current study. In general, biological effects should be stable over subgroups of the study population.

5.4 Outlook

5.4.1 Integrating evidence: systems perspective

The systems perspective on diet as risk factor for type 2 diabetes allowed subdividing the black-box model (*diet* → *diabetes*) into parts (*diet* → *metabolomics network* and *metabolomics network* → *type 2 diabetes*). This allows distinct study designs to investigate the separate model parts. Trials, e.g., could be used to evaluate the diet-metabolite links, whereas the metabolite-diabetes connections could be further investigated with the help of instrumental variables. Therefore, systems epidemiology can inform translational research on dietary risk factors for chronic diseases. Integration of study designs according to observation-based complex models is a promising approach to generate valid and relevant knowledge.

5.4.2 Validation & Translation

The current study aimed to generate hypotheses on biological mechanisms. Biological mechanisms can be expected to be stable across populations. However, generalizability of findings from a single study, even if it is conducted in a large population-based sample such as the

EPIC-Potsdam cohort, can always be questioned. Therefore, results from the current study need to be validated.

Evaluation of the consistency of the results over independent cohort studies is warranted. The mediation model implies that consistency of the diet-metabolomics links could be evaluated separately from the consistency of the metabolomics-type 2 diabetes links. This allows independent replication of the separate model parts according to the available data, possibly relying on several cohorts. Metabolomics networks can also be used to evaluate whether different study populations are biologically comparable. If a cohort had for example a different ethnical background, observing the same metabolomics network structure would still suggest the same data-generating mechanisms. Therefore, such a cohort would qualify for replication. Different network-structures would require explanation.

For validation of the diet-metabolite links (diet → metabolomics network), intervention studies could be performed. Randomized trials that replicate a proposed diet-metabolite link would rule out confounding as explanation for the observations. In this case, specific mechanisms could be evaluated in detail. For example, dose-response curves of metabolites could be compared between different subtypes of the investigated exposures: different whole-grain varieties, different brewing types of coffee, or different types of unprocessed and processed red meat. As another example, metabolic flux experiments could be performed to trace the mechanisms by which dietary exposures induce changes in the metabolic profiles. Even though the described study types have already been broadly applied in nutrition sciences, the explicit design of trials based on a causal model that is consistent with large-scale observational data could add novel aspects.

In order to validate the part of the model dealing with the link between metabolomics networks and type 2 diabetes incidence, the use of instrumental variables is of interest. The directionality assumption implied that diet was treated a as quasi-instrumental variable with respect to metabolite-disease links in the current study. Observed diet-disease links are of course prone to various sources of bias, such as residual confounding, measurement error, and misreporting. The same modeling framework, however, could be applied to identify direct effects of more reliable instruments (e.g. genetics or pharmacological interventions) on metabolomics networks. Furthermore, non-human models, such as animal models or cell cultures, could be considered to investigate

potential biological mechanisms. As discussed above, metabolomics networks could be used to evaluate comparability between the model system and the human situation. Evaluation of a human biomarker in a non-human model system implies the assumption that it reflects the same biological processes in both settings. Confidence in that assumption could be considerably strengthened by demonstrating that the biomarker is integrated in a similar network of local dependencies in both, observational and experimental data.

5.4.3 Public relevance

Based on the available evidence, it is recommended to include whole-grain products in the daily diet, and to reduce red meat consumption to a moderate level are justified (see, e.g., www.dge.de). The results of the current study are in line with these recommendations. Furthermore, it is supported that coffee has a rather beneficial metabolic effect.

Elucidating the biological mechanisms underlying observed diet-disease links is relevant from a public health perspective. Accurate cost-effectiveness calculations for preventive interventions, for example, rely on valid causal effect estimates [340]. Therefore, a quantitative understanding of the biological mechanisms linking diet to disease occurrence is important to inform and prioritize public health actions. The identification of metabolic mediators of the health effects of dietary exposures might also help to identify groups for which changes in the habitual diet are particularly useful, and thereby allow giving more precise dietary recommendations. Moreover, markers of the metabolic mechanisms linking dietary exposures to disease risk could illustrate the success of changes in the habitual diet on the individual level.

6 Conclusions

Secondary aim of the current work was to establish a methodological approach to link data on dietary habits and time-to-event data to metabolomics networks. Conceptually, the analytical design was based on causal inference theory. The developed NetCoupler-algorithm generated joint graphical models that illustrated direct effects of dietary exposures on metabolite networks, and network-independent direct effects of metabolites on the risk of developing type 2 diabetes. Based on observations in the EPIC-Potsdam-cohort study, biologically coherent and consistent information was obtained. Beyond current applications, the modeling framework might be useful to integrate high-dimensional biomarker profiles in etiological epidemiological research in other studies in the future.

The primary aim of the present work was to evaluate complex lipid and amino acid profiles in the circulation as mediators of the effect of whole-grain bread, coffee, and red meat, respectively, on the risk of developing type 2 diabetes. Skeletons of causal networks within metabolite groups were estimated based on the observed conditional dependency patterns. These network models were assumed to integrate information on metabolic processes taking place at the tissue level. Dietary exposures and type 2 diabetes risk were linked to these metabolomics networks in a multi-model procedure. Paths through the network were evaluated as potential mediators of the effect of dietary exposures on type 2 diabetes risk. For each of the dietary exposures, metabolites of interest as potential mediators were identified.

For whole-grain bread, observations were consistent with a partial mediation hypothesis. The whole-grain bread effect on type 2 diabetes risk might be partly mediated by an effect of whole-grain bread consumption on lipid metabolism, most likely in the liver. Lower levels of saturated fatty acids, palmitate in particular, were observed in acylcarnitines and phosphatidylcholines. Adjusting for these potential mediators attenuated the whole-grain bread relation to the risk of type 2 diabetes by about one fourth.

Furthermore, the observed conditional independence structures were consistent with the hypothesis of a major role of lipid and amino acid metabolism in mediating the coffee effect on type 2 diabetes risk. High coffee consumption was related to higher levels of polyunsaturated

fatty acids, among them likely linoleic acid, in sphingomyelins and alkyl-acyl phosphatidylcholines, and lower levels of phenylalanine and lysophosphatidylcholine C14:0. Adjusting for these potential mediators attenuated coffee-related type 2 diabetes risk by about two thirds.

Lastly, observations were also consistent with the hypothesis that red-meat related alterations in lipid and amino acid metabolism mediate the major proportion of the red meat-related type 2 diabetes risk. Red meat-related higher levels of saturated fatty acids, particularly stearate, and of alkyl-acyl phosphatidylcholine C36/4, and of lower levels of glycine had the potential to explain seventy percent of the red meat-related type 2 diabetes risk.

Taken together, these observations suggest that three widely discussed diabetes-related dietary factors might influence systemic lipid and amino acid metabolism long before type 2 diabetes occurs. Possibly, the identified metabolic mediators are sensors for early biological processes that are triggered by the habitual dietary behavior and that protect of, or predispose for, type 2 diabetes development. Thus, the current thesis enhanced the quest for the biological mechanisms that link consumption of whole-grain bread, coffee, and whole-grain bread to type 2 diabetes risk by providing mediation hypotheses that are consistent with the observations in a large prospective cohort study.

7 Literature

References

1. Sanders FW, Griffin JL (2016) De novo lipogenesis in the liver in health and disease: more than just a shunting yard for glucose. Biol Rev Camb Philos Soc 91: 452-468.
2. McCoin CS, Knotts TA, Adams SH (2015) Acylcarnitines—old actors auditioning for new roles in metabolic physiology. Nature reviews Endocrinology 11: 617-625.
3. Kroger J, Schulze MB (2012) Recent insights into the relation of Delta5 desaturase and Delta6 desaturase activity to the development of type 2 diabetes. Curr Opin Lipidol 23: 4-10.
4. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, et al. (2015) Type 2 diabetes mellitus. Nat Rev Dis Primers 1: 15019.
5. United Nations, International Diabetes Federation, „ editor (2006) Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. Geneva, Switzerland.
6. Association AD (2014) Diagnosis and classification of diabetes mellitus. Diabetes Care 37 Suppl 1: S81-90.
7. Classification of Diseases (ICD) (2016) ICD-10.
8. International Diabetes Federation (2015) IDF Diabetes Atlas. Belgium: International Diabetes Federation.
9. Holman N, Young B, Gadsby R (2015) Current prevalence of Type 1 and Type 2 diabetes in adults and children in the UK. Diabet Med 32: 1119-1120.
10. Chatterjee S, Khunti K, Davies MJ Type 2 diabetes. The Lancet.
11. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, et al. (2012) Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380: 2197-2223.
12. Kengne AP, June-Rose McHiza Z, Amoah AG, Mbanya JC (2013) Cardiovascular diseases and diabetes as economic and developmental challenges in Africa. Prog Cardiovasc Dis 56: 302-313.
13. Ley SH, Hamdy O, Mohan V, Hu FB (2014) Prevention and management of type 2 diabetes: dietary components and nutritional strategies. Lancet 383: 1999-2007.
14. Hu FB (2011) Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes Care 34: 1249-1257.
15. Ezzati M, Riboli E (2012) Can noncommunicable diseases be prevented? Lessons from studies of populations and individuals. Science 337: 1482-1487.
16. American Diabetes Association (2017) Prevention or Delay of Type 2 Diabetes. Diabetes Care 40: S44-S47.

17. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, et al. (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. New England Journal of Medicine 344: 1343-1350.
18. Lindstrom J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, et al. (2006) Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study.
19. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, et al. (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 346: 393-403.
20. Pearl J (2010) An introduction to causal inference. Int J Biostat 6: Article 7.
21. Pearl J (2012) The causal mediation formula--a guide to the assessment of pathways and mechanisms. Prev Sci 13: 426-436.
22. Salas-Salvado J, Bullo M, Babio N, Martinez-Gonzalez MA, Ibarrola-Jurado N, et al. (2011) Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. Diabetes Care 34: 14-19.
23. Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, et al. (2013) Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 368: 1279-1290.
24. Schwingshackl L, Hoffmann G, Lampousi AM, Knuppel S, Iqbal K, et al. (2017) Food groups and risk of type 2 diabetes mellitus: a systematic review and meta-analysis of prospective studies. Eur J Epidemiol.
25. Jannasch F, Kroger J, Schulze MB (2017) Dietary Patterns and Type 2 Diabetes: A Systematic Literature Review and Meta-Analysis of Prospective Studies. J Nutr.
26. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, et al. (2007) An Accurate Risk Score Based on Anthropometric, Dietary, and Lifestyle Factors to Predict the Development of Type 2 Diabetes. Diabetes Care 30: 510-515.
27. Consortium TI (2015) Dietary fibre and incidence of type 2 diabetes in eight European countries: the EPIC-InterAct Study and a meta-analysis of prospective studies. Diabetologia 58: 1394-1408.
28. Ross AB (2015) Whole grains beyond fibre: what can metabolomics tell us about mechanisms? Proc Nutr Soc 74: 320-327.
29. de Munter JS, Hu FB, Spiegelman D, Franz M, van Dam RM (2007) Whole grain, bran, and germ intake and risk of type 2 diabetes: a prospective cohort study and systematic review. PLoS Med 4: e261.
30. Muhlenbruch K, Ludwig T, Jeppesen C, Joost HG, Rathmann W, et al. (2014) Update of the German Diabetes Risk Score and

external validation in the German MONICA/KORA study. Diabetes Res Clin Pract 104: 459-466.
31. Paprott R, Muhlenbruch K, Mensink GB, Thiele S, Schulze MB, et al. (2016) Validation of the German Diabetes Risk Score among the general adult population: findings from the German Health Interview and Examination Surveys. BMJ Open Diabetes Res Care 4: e000280.
32. Andersson A, Tengblad S, Karlstrom B, Kamal-Eldin A, Landberg R, et al. (2007) Whole-grain foods do not affect insulin sensitivity or markers of lipid peroxidation and inflammation in healthy, moderately overweight subjects. J Nutr 137: 1401-1407.
33. Brownlee IA, Moore C, Chatfield M, Richardson DP, Ashby P, et al. (2010) Markers of cardiovascular risk are not changed by increased whole-grain intake: the WHOLEheart study, a randomised, controlled dietary intervention. Br J Nutr 104: 125-134.
34. Giacco R, Clemente G, Cipriano D, Luongo D, Viscovo D, et al. (2010) Effects of the regular consumption of wholemeal wheat foods on cardiovascular risk factors in healthy people. Nutr Metab Cardiovasc Dis 20: 186-194.
35. Saltzman E, Das SK, Lichtenstein AH, Dallal GE, Corrales A, et al. (2001) An oat-containing hypocaloric diet reduces systolic blood pressure and improves lipid profile beyond effects of weight loss in men and women. J Nutr 131: 1465-1470.
36. Tighe P, Duthie G, Vaughan N, Brittenden J, Simpson WG, et al. (2010) Effect of increased consumption of whole-grain foods on blood pressure and other cardiovascular risk markers in healthy middle-aged persons: a randomized controlled trial. Am J Clin Nutr 92: 733-740.
37. Rave K, Roggen K, Dellweg S, Heise T, tom Dieck H (2007) Improvement of insulin resistance after diet with a whole-grain based dietary product: results of a randomized, controlled cross-over study in obese subjects with elevated fasting blood glucose. Br J Nutr 98: 929-936.
38. Silva FM, Kramer CK, de Almeida JC, Steemburgo T, Gross JL, et al. (2013) Fiber intake and glycemic control in patients with type 2 diabetes mellitus: a systematic review with meta-analysis of randomized controlled trials. Nutr Rev 71: 790-801.
39. Li X, Cai X, Ma X, Jing L, Gu J, et al. (2016) Short- and Long-Term Effects of Wholegrain Oat Intake on Weight Management and Glucolipid Metabolism in Overweight Type-2 Diabetics: A Randomized Control Trial. Nutrients 8.
40. Lindstrom J, Peltonen M, Eriksson JG, Louheranta A, Fogelholm M, et al. (2006) High-fibre, low-fat diet predicts long-term weight loss and decreased type 2 diabetes risk: the Finnish Diabetes Prevention Study. Diabetologia 49: 912-920.

41. Ding M, Bhupathiraju SN, Chen M, van Dam RM, Hu FB (2014) Caffeinated and decaffeinated coffee consumption and risk of type 2 diabetes: a systematic review and a dose-response meta-analysis. Diabetes Care 37: 569-586.
42. Ohnaka K, Ikeda M, Maki T, Okada T, Shimazoe T, et al. (2012) Effects of 16-week consumption of caffeinated and decaffeinated instant coffee on glucose metabolism in a randomized controlled trial. J Nutr Metab 2012: 207426.
43. MacKenzie T, Comi R, Sluss P, Keisari R, Manwar S, et al. (2007) Metabolic and hormonal effects of caffeine: randomized, double-blind, placebo-controlled crossover trial. Metabolism 56: 1694-1698.
44. van Dam RM, Pasman WJ, Verhoef P (2004) Effects of coffee consumption on fasting blood glucose and insulin concentrations: randomized controlled trials in healthy volunteers. Diabetes Care 27: 2990-2992.
45. Wedick NM, Brennan AM, Sun Q, Hu FB, Mantzoros CS, et al. (2011) Effects of caffeinated and decaffeinated coffee on biological risk factors for type 2 diabetes: a randomized controlled trial. Nutr J 10: 93.
46. Kempf K, Herder C, Erlund I, Kolb H, Martin S, et al. (2010) Effects of coffee consumption on subclinical inflammation and other risk factors for type 2 diabetes: a clinical trial. Am J Clin Nutr 91: 950-957.
47. Shi X, Xue W, Liang S, Zhao J, Zhang X (2016) Acute caffeine ingestion reduces insulin sensitivity in healthy subjects: a systematic review and meta-analysis. Nutr J 15: 103.
48. Pan A, Sun Q, Bernstein AM, Schulze MB, Manson JE, et al. (2011) Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. Am J Clin Nutr 94: 1088-1096.
49. Bendinelli B, Palli D, Masala G, Sharp SJ, Schulze MB, et al. (2013) Association between dietary meat consumption and incident type 2 diabetes: the EPIC-InterAct study. Diabetologia 56: 47-59.
50. Aune D, Ursin G, Veierod MB (2009) Meat consumption and the risk of type 2 diabetes: a systematic review and meta-analysis of cohort studies. Diabetologia 52: 2277-2287.
51. Micha R, Wallace SK, Mozaffarian D (2010) Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. Circulation 121: 2271-2283.
52. Feskens EJ, Sluik D, van Woudenbergh GJ (2013) Meat consumption, diabetes, and its complications. Curr Diab Rep 13: 298-306.
53. Pan A, Sun Q, Bernstein AM, Manson JE, Willett WC, et al. (2013) Changes in red meat consumption and subsequent risk of type 2

diabetes mellitus: three cohorts of US men and women. JAMA Intern Med 173: 1328-1335.
54. Wittenbecher C, Muhlenbruch K, Kroger J, Jacobs S, Kuxhaus O, et al. (2015) Amino acids, lipid metabolites, and ferritin as potential mediators linking red meat consumption to type 2 diabetes. Am J Clin Nutr 101: 1241-1250.
55. Turner KM, Keogh JB, Clifton PM (2015) Red meat, dairy, and insulin sensitivity: a randomized crossover intervention study. Am J Clin Nutr 101: 1173-1179.
56. Sayer RD, Wright AJ, Chen N, Campbell WW (2015) Dietary Approaches to Stop Hypertension diet retains effectiveness to reduce blood pressure when lean pork is substituted for chicken and fish as the predominant source of protein. The American Journal of Clinical Nutrition 102: 302-308.
57. Aadland EK, Graff IE, Lavigne C, Eng Ø, Paquette M, et al. (2016) Lean Seafood Intake Reduces Postprandial C-peptide and Lactate Concentrations in Healthy Adults in a Randomized Controlled Trial with a Crossover Design. The Journal of Nutrition 146: 1027-1034.
58. van Nielen M, Feskens EJ, Rietman A, Siebelink E, Mensink M (2014) Partly replacing meat protein with soy protein alters insulin resistance and blood lipids in postmenopausal women with abdominal obesity. J Nutr 144: 1423-1429.
59. Navas-Carretero S, Perez-Granados AM, Schoppen S, Vaquero MP (2009) An oily fish diet increases insulin sensitivity compared to a red meat diet in young iron-deficient women. Br J Nutr 102: 546-553.
60. Viguiliouk E, Stewart SE, Jayalath VH, Ng AP, Mirrahimi A, et al. (2015) Effect of Replacing Animal Protein with Plant Protein on Glycemic Control in Diabetes: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. Nutrients 7: 9804-9824.
61. Azadbakht L, Atabak S, Esmaillzadeh A (2008) Soy Protein Intake, Cardiorenal Indices, and C-Reactive Protein in Type 2 Diabetes With Nephropathy. A longitudinal randomized clinical trial 31: 648-654.
62. Daly RM, O'Connell SL, Mundell NL, Grimes CA, Dunstan DW, et al. (2014) Protein-enriched diet, with the use of lean red meat, combined with progressive resistance training enhances lean tissue mass and muscle strength and reduces circulating IL-6 concentrations in elderly women: a cluster randomized controlled trial. Am J Clin Nutr 99: 899-910.
63. Hodgson JM, Ward NC, Burke V, Beilin LJ, Puddey IB (2007) Increased lean red meat intake does not elevate markers of oxidative stress and inflammation in humans. J Nutr 137: 363-367.

64. Cornelis MC, Hu FB (2013) Systems Epidemiology: A New Direction in Nutrition and Metabolic Disease Research. Curr Nutr Rep 2.
65. Hu FB (2011) Metabolic profiling of diabetes: from black-box epidemiology to systems epidemiology. Clin Chem 57: 1224-1226.
66. Shah SH, Newgard CB (2015) Integrated metabolomics and genomics: systems approaches to biomarkers and mechanisms of cardiovascular disease. Circ Cardiovasc Genet 8: 410-419.
67. Bictash M, Ebbels TM, Chan Q, Loo RL, Yap IK, et al. (2010) Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. J Clin Epidemiol 63: 970-979.
68. Von Bertalanffy L (1950) The theory of open systems in physics and biology. Science 111: 23-29.
69. Bar-Yam Y (2002) General features of complex systems. Encyclopedia of Life Support Systems (EOLSS), UNESCO, EOLSS Publishers, Oxford, UK.
70. Richardson K (2004) The problematisation of existence: towards a philosophy of complexity. Nonlinear Dynamics Psychol Life Sci 8: 17-40.
71. Wolkenhauer O (2014) Why model? Frontiers in Physiology 5.
72. Zeleny M (1977) Self-organization of living systems: A formal model of autopoiesis. International journal of general system 4: 13-28.
73. Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5: 187-196.
74. Zeleny M (1981) Autopoiesis today. CYBERNETICS 9: 3.
75. Mitchell M (2006) Complex systems: Network thinking. Artificial Intelligence 170: 1194-1212.
76. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.
77. Alon U (2006) An introduction to systems biology: design principles of biological circuits: CRC press.
78. Ellner SP, Guckenheimer J (2011) Dynamic models in biology: Princeton University Press.
79. Liu Y-Y, Slotine J-J, Barabási A-L (2013) Observability of complex systems. Proceedings of the National Academy of Sciences 110: 2460-2465.
80. Christakis NA, Fowler JH (2008) The collective dynamics of smoking in a large social network. New England journal of medicine 358: 2249-2258.
81. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. New England journal of medicine 357: 370-379.
82. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, et al. (2014) An atlas of genetic influences on human blood metabolites. Nat Genet 46: 543-550.

83. Floegel A, Wientzek A, Bachlechner U, Jacobs S, Drogan D, et al. (2014) Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. Int J Obes (Lond) 38: 1388-1396.
84. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 5: 21.
85. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL (2011) Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc Rev 40: 387-426.
86. Zamboni N, Saghatelian A, Patti GJ (2015) Defining the metabolome: size, flux, and regulation. Mol Cell 58: 699-706.
87. Rothwell JA, Fillatre Y, Martin JF, Lyan B, Pujos-Guillot E, et al. (2014) New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. PLoS One 9: e93474.
88. Scalbert A, Brennan L, Manach C, Andres-Lacueva C, Dragsted LO, et al. (2014) The food metabolome: a window over dietary exposure. Am J Clin Nutr 99: 1286-1308.
89. Zheng H, Clausen MR, Dalsgaard TK, Bertram HC (2015) Metabolomics to Explore Impact of Dairy Intake. Nutrients 7: 4875-4896.
90. Cheung W, Keski-Rahkonen P, Assi N, Ferrari P, Freisling H, et al. (2017) A metabolomic study of biomarkers of meat and fish intake. Am J Clin Nutr.
91. Garcia-Perez I, Posma JM, Gibson R, Chambers ES, Hansen TH, et al. (2017) Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. Lancet Diabetes Endocrinol.
92. Wu GD, Compher C, Chen EZ, Smith SA, Shah RD, et al. (2016) Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut 65: 63-72.
93. Vazquez-Fresno R, Llorach R, Urpi-Sarda M, Lupianez-Barbero A, Estruch R, et al. (2015) Metabolomic pattern analysis after mediterranean diet intervention in a nondiabetic population: a 1- and 3-year follow-up in the PREDIMED study. J Proteome Res 14: 531-540.
94. Bondia-Pons I, Martinez JA, de la Iglesia R, Lopez-Legarrea P, Poutanen K, et al. (2015) Effects of short- and long-term Mediterranean-based dietary treatment on plasma LC-QTOF/MS metabolic profiling of subjects with metabolic syndrome features: The Metabolic Syndrome Reduction in Navarra (RESMENA) randomized controlled trial. Mol Nutr Food Res 59: 711-728.
95. Bhupathiraju SN, Hu FB (2017) One (small) step towards precision nutrition by use of metabolomics. Lancet Diabetes Endocrinol.

96. Barron R, Bermingham K, Brennan L, Gibney ER, Gibney MJ, et al. (2016) Twin metabolomics: the key to unlocking complex phenotypes in nutrition research. Nutr Res 36: 291-304.
97. Swann JR, Claus SP (2014) Nutrimetabonomics: nutritional applications of metabolic profiling. Sci Prog 97: 41-47.
98. Guertin KA, Moore SC, Sampson JN, Huang WY, Xiao Q, et al. (2014) Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. Am J Clin Nutr 100: 208-217.
99. Van Dam R, Hunter D (2013) Biochemical indicators of dietary intake. Nutritional epidemiology 3: 150-212.
100. Floegel A, von Ruesten A, Drogan D, Schulze MB, Prehn C, et al. (2013) Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. Eur J Clin Nutr 67: 1100-1108.
101. Nielsen KL, Hartvigsen ML, Hedemann MS, Laerke HN, Hermansen K, et al. (2014) Similar metabolic responses in pigs and humans to breads with different contents and compositions of dietary fibers: a metabolomics study. Am J Clin Nutr 99: 941-949.
102. Moazzami AA, Bondia-Pons I, Hanhineva K, Juntunen K, Antl N, et al. (2012) Metabolomics reveals the metabolic shifts following an intervention with rye bread in postmenopausal women--a randomized control trial. Nutr J 11: 88.
103. Guertin KA, Loftfield E, Boca SM, Sampson JN, Moore SC, et al. (2015) Serum biomarkers of habitual coffee consumption may provide insight into the mechanism underlying the association between coffee consumption and colorectal cancer. Am J Clin Nutr 101: 1000-1011.
104. Redeuil K, Smarrito-Menozzi C, Guy P, Rezzi S, Dionisi F, et al. (2011) Identification of novel circulating coffee metabolites in human plasma by liquid chromatography-mass spectrometry. J Chromatogr A 1218: 4678-4688.
105. Zheng Y, Yu B, Alexander D, Steffen LM, Boerwinkle E (2014) Human metabolome associates with dietary intake habits among African Americans in the atherosclerosis risk in communities study. Am J Epidemiol 179: 1424-1433.
106. Jacobs S, Kroger J, Floegel A, Boeing H, Drogan D, et al. (2014) Evaluation of various biomarkers as potential mediators of the association between coffee consumption and incident type 2 diabetes in the EPIC-Potsdam Study. Am J Clin Nutr 100: 891-900.
107. Menni C, Zhai G, Macgregor A, Prehn C, Romisch-Margl W, et al. (2013) Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. Metabolomics 9: 506-514.

108. Altmaier E, Kastenmuller G, Romisch-Margl W, Thorand B, Weinberger KM, et al. (2009) Variation in the human lipidome associated with coffee consumption as revealed by quantitative targeted metabolomics. Mol Nutr Food Res 53: 1357-1365.
109. Miranda AM, Carioca AA, Steluti J, da Silva ID, Fisberg RM, et al. (2016) The effect of coffee intake on lysophosphatidylcholines: A targeted metabolomic approach. Clin Nutr.
110. Schmidt JA, Rinaldi S, Ferrari P, Carayol M, Achaintre D, et al. (2015) Metabolic profiles of male meat eaters, fish eaters, vegetarians, and vegans from the EPIC-Oxford cohort. Am J Clin Nutr 102: 1518-1526.
111. Schmidt JA, Rinaldi S, Scalbert A, Ferrari P, Achaintre D, et al. (2016) Plasma concentrations and intakes of amino acids in male meat-eaters, fish-eaters, vegetarians and vegans: a cross-sectional analysis in the EPIC-Oxford cohort. Eur J Clin Nutr 70: 306-312.
112. Ross AB, Svelander C, Undeland I, Pinto R, Sandberg AS (2015) Herring and Beef Meals Lead to Differences in Plasma 2-Aminoadipic Acid, beta-Alanine, 4-Hydroxyproline, Cetoleic Acid, and Docosahexaenoic Acid Concentrations in Overweight Men. J Nutr 145: 2456-2463.
113. Burd NA, Gorissen SH, van Vliet S, Snijders T, van Loon LJ (2015) Differences in postprandial protein handling after beef compared with milk ingestion during postexercise recovery: a randomized controlled trial. Am J Clin Nutr 102: 828-836.
114. Neacsu M, Fyfe C, Horgan G, Johnstone AM (2014) Appetite control and biomarkers of satiety with vegetarian (soy) and meat-based high-protein diets for weight loss in obese men: a randomized crossover trial. Am J Clin Nutr 100: 548-558.
115. Dietrich S, Floegel A, Troll M, Kuhn T, Rathmann W, et al. (2016) Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol 45: 1406-1420.
116. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, et al. (2013) Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes 62: 639-648.
117. Guasch-Ferre M, Hruby A, Toledo E, Clish CB, Martinez-Gonzalez MA, et al. (2016) Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. Diabetes Care 39: 833-846.
118. Sun L, Liang L, Gao X, Zhang H, Yao P, et al. (2016) Early Prediction of Developing Type 2 Diabetes by Plasma Acylcarnitines: A Population-Based Study. Diabetes Care 39: 1563-1570.
119. Qiu G, Zheng Y, Wang H, Sun J, Ma H, et al. (2016) Plasma metabolomics identified novel metabolites associated with risk of type 2 diabetes in two prospective cohorts of Chinese adults. Int J Epidemiol 45: 1507-1516.

120. Forouhi NG, Koulman A, Sharp SJ, Imamura F, Kroger J, et al. (2014) Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the EPIC-InterAct case-cohort study. Lancet Diabetes Endocrinol 2: 810-818.
121. Forouhi NG, Imamura F, Sharp SJ, Koulman A, Schulze MB, et al. (2016) Association of Plasma Phospholipid n-3 and n-6 Polyunsaturated Fatty Acids with Type 2 Diabetes: The EPIC-InterAct Case-Cohort Study. PLoS Med 13: e1002094.
122. Tulipani S, Palau-Rodriguez M, Minarro Alonso A, Cardona F, Marco-Ramell A, et al. (2016) Biomarkers of Morbid Obesity and Prediabetes by Metabolomic Profiling of Human Discordant Phenotypes. Clin Chim Acta 463: 53-61.
123. Arora T, Velagapudi V, Pournaras DJ, Welbourn R, le Roux CW, et al. (2015) Roux-en-Y Gastric Bypass Surgery Induces Early Plasma Metabolomic and Lipidomic Alterations in Humans Associated with Diabetes Remission. PLoS One 10: e0126401.
124. Krumsiek J, Bartel J, Theis FJ (2016) Computational approaches for systems metabolomics. Curr Opin Biotechnol 39: 198-206.
125. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, et al. (2012) Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genet 8: e1003005.
126. Beebee H, Hitchcock C, Menzies P (2009) The Oxford handbook of causation: Oxford University Press.
127. Lewis D (1974) Causation. The journal of philosophy 70: 556-567.
128. Lewis D (2000) Causation as influence. The Journal of Philosophy 97: 182-197.
129. Pearl J (2009) Causality: Cambridge university press.
130. Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology: Lippincott Williams & Wilkins.
131. Rothman KJ, Greenland S (2005) Causation and causal inference in epidemiology. Am J Public Health 95 Suppl 1: S144-150.
132. Pearl J (2011) Invited commentary: understanding bias amplification. Am J Epidemiol 174: 1223-1227; discussion pg 1228-1229.
133. VanderWeele TJ, Shpitser I (2013) On the definition of a confounder. Ann Stat 41: 196-220.
134. VanderWeele TJ (2016) Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health 37: 17-32.
135. VanderWeele TJ, Tchetgen Tchetgen EJ (2014) Attributing effects to interactions. Epidemiology 25: 711-722.
136. VanderWeele TJ, Hernan MA (2013) Causal Inference Under Multiple Versions of Treatment. J Causal Inference 1: 1-20.
137. Danaei G, Pan A, Hu FB, Hernan MA (2013) Hypothetical Midlife Interventions in Women and Risk of Type 2 Diabetes. Epidemiology 24: 122-128.

138. Boada LD, Henriquez-Hernandez LA, Luzardo OP (2016) The impact of red and processed meat consumption on cancer and other health outcomes: Epidemiological evidences. Food Chem Toxicol 92: 236-244.
139. VanderWeele TJ (2017) On Causes, Causal Inference, and Potential Outcomes. Int J Epidemiol.
140. Pearl J (2013) Structural counterfactuals: a brief introduction. Cogn Sci 37: 977-985.
141. Lawlor DA, Tilling K, Davey Smith G (2017) Triangulation in aetiological epidemiology. Int J Epidemiol.
142. Krieger N, Davey Smith G (2016) Response: FACEing reality: productive tensions between our epidemiological questions, methods and mission. International Journal of Epidemiology 45: 1852-1865.
143. Krieger N, Davey Smith G (2016) The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. International Journal of Epidemiology 45: 1787-1808.
144. Hill AB (1965) THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? Proc R Soc Med 58: 295-300.
145. Maathuis MH, Colombo D, Kalisch M, Bühlmann P (2010) Predicting causal effects in large-scale systems from observational data. Nature Methods 7: 247-248.
146. Breitling LP (2010) dagR: a suite of R functions for directed acyclic graphs. Epidemiology 21: 586-587.
147. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47: 1-26.
148. Kalisch M, Bühlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research 8: 613-636.
149. Boeing H, Wahrendorf J, Becker N (1999) EPIC-Germany--A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. Ann Nutr Metab 43: 195-204.
150. Riboli E, Kaaks R (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol 26 Suppl 1: S6-14.
151. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, et al. (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr 5: 1113-1124.
152. Boeing H, Korfmann A, Bergmann MM (1999) Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 43: 205-215.

153. Brandstetter BR, Korfmann A, Kroke A, Becker N, Schulze MB, et al. (1999) Dietary habits in the German EPIC cohorts: food group intake estimated with the food frequency questionnaire. European Investigation into Cancer and Nutrition. Ann Nutr Metab 43: 246-257.
154. Kroke A, Bergmann MM, Lotze G, Jeckel A, Klipstein-Grobusch K, et al. (1999) Measures of quality control in the German component of the EPIC study. European Prospective Investigation into Cancer and Nutrition. Ann Nutr Metab 43: 216-224.
155. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J (1997) 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. Int J Epidemiol 26 Suppl 1: S59-70.
156. Haraldsdottir J (1993) Minimizing error in the field: quality control in dietary surveys. Eur J Clin Nutr 47 Suppl 2: S19-24.
157. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J (1997) 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. Int J Epidemiol 26 Suppl 1: S71-81.
158. Schulze MB, Manson JE, Ludwig DS, Colditz GA, Stampfer MJ, et al. (2004) Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. Jama 292: 927-934.
159. Dehne LI, Klemm C, Henseler G, Hermann-Kunz E (1999) The German Food Code and Nutrient Data Base (BLS II.2). European Journal of Epidemiology 15: 355-358.
160. Boeing H, Bohlscheid-Thomas S, Voss S, Schneeweiss S, Wahrendorf J (1997) The relative validity of vitamin intakes derived from a food frequency questionnaire compared to 24-hour recalls and biological measurements: results from the EPIC pilot study in Germany. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol 26 Suppl 1: S82-90.
161. Voss S, Charrondiere UR, Slimani N, Kroke A, Riboli E, et al. (1998) [EPIC-SOFT a European computer program for 24-hour dietary protocols]. Z Ernahrungswiss 37: 227-233.
162. Kroke A, Klipstein-Grobusch K, Voss S, Moseneder J, Thielecke F, et al. (1999) 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 70: 439-447.

163. Klipstein-Grobusch K, Georg T, Boeing H (1997) Interviewer variability in anthropometric measurements and estimates of body composition. International Journal of Epidemiology 26: S174.
164. Kroke A, Liese AD, Keil U, Boeing H (1999) Arterial hypertension and glycemia in non-diabetic subjects: is there an association independent of obesity? Diabetes Metab Res Rev 15: 99-105.
165. Kroke A, Fleischhauer W, Mieke S, Klipstein-Grobusch K, Willich SN, et al. (1998) Blood pressure measurement in epidemiological studies: a comparative analysis of two methods. Data from the EPIC-Potsdam Study. European Prospective Investigation into Cancer and Nutrition. J Hypertens 16: 739-746.
166. Schulze MB, Kroke A, Bergmann MM, Boeing H (2000) Differences of blood pressure estimates between consecutive measurements on one occasion: implications for inter-study comparability of epidemiologic studies. Eur J Epidemiol 16: 891-898.
167. Bergmann MM, Bussas U, Boeing H (1999) Follow-Up Procedures in EPIC-Germany – Data Quality Aspects. Annals of Nutrition and Metabolism 43: 225-234.
168. Schienkiewitz A, Schulze MB, Hoffmann K, Kroke A, Boeing H (2006) Body mass index history and risk of type 2 diabetes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. Am J Clin Nutr 84: 427-433.
169. Spranger J, Kroke A, Möhlig M, Hoffmann K, Bergmann MM, et al. (2003) Inflammatory cytokines and the risk to develop type 2 diabetes results of the prospective population-based European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. Diabetes 52: 812-817.
170. Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73: 1-11.
171. Barlow WE, Ichikawa L, Rosner D, Izumi S (1999) Analysis of case-cohort designs. Journal of clinical epidemiology 52: 1165-1172.
172. Römisch-Margl W, Prehn C, Bogumil R, Röhring C, Suhre K, et al. (2012) Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics 8: 133-142.
173. Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, et al. (2011) Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. PLoS One 6: e21103.
174. Willett WC, Howe GR, Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. Am J Clin Nutr 65: 1220S-1228S; discussion 1229S-1231S.
175. Hatcher L (1994) A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling.

176. Brown JD (2009) Statistics Corner Questions and answers about language testing statistics: Principal components analysis and exploratory factor analysis—Definitions, differences, and choices.
177. Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search: MIT press.
178. Colombo D, Maathuis MH (2014) Order-independent constraint-based causal structure learning. Journal of Machine Learning Research 15: 3741-3782.
179. Nandy P, Maathuis MH, Richardson TS (2014) Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. arXiv preprint arXiv:14072451.
180. VanderWeele TJ (2011) Causal mediation analysis with survival data. Epidemiology 22: 582-585.
181. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2: 2366-2382.
182. Pearl J (2010) Causal Inference. Journal of Machine Learning Research-Proceedings Track 6: 39-58.
183. Guasch-Ferre M, Zheng Y, Ruiz-Canela M, Hruby A, Martinez-Gonzalez MA, et al. (2016) Plasma acylcarnitines and risk of cardiovascular disease: effect of Mediterranean diet interventions. Am J Clin Nutr 103: 1408-1416.
184. Lu Y, Wang Y, Ong CN, Subramaniam T, Choi HW, et al. (2016) Metabolic signatures and risk of type 2 diabetes in a Chinese population: an untargeted metabolomics study using both LC-MS and GC-MS. Diabetologia 59: 2349-2359.
185. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, et al. (2012) Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 8: 615.
186. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, et al. (2011) Metabolite profiles and the risk of developing diabetes. Nat Med 17: 448-453.
187. Newgard CB (2017) Metabolomics and Metabolic Diseases: Where Do We Stand? Cell Metab 25: 43-56.
188. Meikle PJ, Summers SA (2017) Sphingolipids and phospholipids in insulin resistance and related metabolic disorders. Nat Rev Endocrinol 13: 79-91.
189. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2012) Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. J Proteome Res 11: 4120-4131.
190. Carrizo D, Chevallier OP, Woodside JV, Brennan SF, Cantwell MM, et al. (2017) Untargeted metabolomic analysis of human serum samples associated with different levels of red meat consumption: A possible indicator of type 2 diabetes? Food Chem 221: 214-221.
191. Gibbons H, O'Gorman A, Brennan L (2015) Metabolomics as a tool in nutritional research. Curr Opin Lipidol 26: 30-34.

192. O'Gorman A, Morris C, Ryan M, O'Grada CM, Roche HM, et al. (2014) Habitual dietary intake impacts on the lipidomic profile. J Chromatogr B Analyt Technol Biomed Life Sci 966: 140-146.
193. Quell JD, Römisch-Margl W, Colombo M, Krumsiek J, Evans AM, et al. (2017) Automated pathway and reaction prediction facilitates in silico identification of unknown metabolites in human cohort studies. Journal of Chromatography B.
194. Willett W (2012) Nutritional epidemiology: Oxford University Press.
195. Pearl J (2012) The causal foundations of structural equation modeling. DTIC Document.
196. Vansteelandt S, Vanderweele TJ (2012) Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. Biometrics 68: 1019-1027.
197. Pearl J (2012) The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prevention Science 13: 426-436.
198. Tchetgen Tchetgen EJ, Vanderweele TJ (2014) Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Epidemiology 25: 282-291.
199. Vanderweele TJ, Vansteelandt S, Robins JM (2014) Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology 25: 300-306.
200. VanderWeele TJ, Valeri L, Ogburn EL (2012) The role of measurement error and misclassification in mediation analysis. Epidemiology (Cambridge, Mass) 23: 561.
201. Shukla A, Singh TR (2017) Computational Network Approaches and Their Applications for Complex Diseases. Translational Bioinformatics and Its Application: Springer. pp. 337-352.
202. Sommer C (2014) Shortest-path queries in static networks. ACM Computing Surveys (CSUR) 46: 45.
203. Valeri L, Vanderweele TJ (2013) Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychol Methods 18: 137-150.
204. VanderWeele TJ (2013) A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology 24: 224-232.
205. VanderWeele T, Vansteelandt S (2014) Mediation Analysis with Multiple Mediators. Epidemiological Methods 2.
206. Franks PW, Atabaki-Pasdar N (2017) Causal inference in obesity research. J Intern Med 281: 222-232.
207. Franks PW, McCarthy MI (2016) Exposing the exposures responsible for type 2 diabetes and obesity. Science 354: 69-73.
208. Wittenbecher C, di Giuseppe R, Biemann R, Menzel J, Arregui M, et al. (2015) Reproducibility of Retinol Binding Protein 4 and Omentin-1 Measurements over a Four Months Period: A

Reliability Study in a Cohort of 207 Apparently Healthy Participants. PLoS One 10: e0138480.
209. Pierce BL, VanderWeele TJ (2012) The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. Int J Epidemiol 41: 1383-1393.
210. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological): 289-300.
211. Pallares-Mendez R, Aguilar-Salinas CA, Cruz-Bautista I, Del Bosque-Plata L (2016) Metabolomics in diabetes, a review. Ann Med 48: 89-102.
212. Löffler G, Petrides P, Heinrich P (2007) Biochemie und Pathobiochemie Springer. Heidelberg.
213. Wolk A (2017) Potential health hazards of eating red meat. J Intern Med 281: 106-122.
214. Felig P, Marliss E, Cahill GFJ (1969) Plasma Amino Acid Levels and Insulin Secretion in Obesity. New England Journal of Medicine 281: 811-816.
215. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, et al. (2009) A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. Cell Metabolism 9: 311-326.
216. Lotta LA, Scott RA, Sharp SJ, Burgess S, Luan Ja, et al. (2016) Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. PLOS Medicine 13: e1002179.
217. Mahendran Y, Jonsson A, Have CT, Allin KH, Witte DR, et al. (2017) Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels. Diabetologia.
218. Mancini A, Di Segni C, Raimondo S, Olivieri G, Silvestrini A, et al. (2016) Thyroid Hormones, Oxidative Stress, and Inflammation. Mediators Inflamm 2016: 6757154.
219. Yu D, Moore SC, Matthews CE, Xiang YB, Zhang X, et al. (2016) Plasma metabolomic profiles in association with type 2 diabetes risk and prevalence in Chinese adults. Metabolomics 12.
220. Xie W, Wood AR, Lyssenko V, Weedon MN, Knowles JW, et al. (2013) Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. Diabetes 62: 2141-2150.
221. Ziegler A, Mwambi H, Konig IR (2015) Mendelian Randomization versus Path Models: Making Causal Inferences in Genetic Epidemiology. Hum Hered 79: 194-204.
222. Schooneman MG, Vaz FM, Houten SM, Soeters MR (2013) Acylcarnitines: Reflecting or Inflicting Insulin Resistance? Diabetes 62: 1-8.

223. Soeters MR, Sauerwein HP, Duran M, Wanders RJ, Ackermans MT, et al. (2009) Muscle acylcarnitines during short-term fasting in lean healthy men. Clin Sci (Lond) 116: 585-592.
224. Costa CC, de Almeida IT, Jakobs C, Poll-The BT, Duran M (1999) Dynamic changes of plasma acylcarnitine levels induced by fasting and sunflower oil challenge test in children. Pediatr Res 46: 440-444.
225. Xu G, Hansen JS, Zhao XJ, Chen S, Hoene M, et al. (2016) Liver and Muscle Contribute Differently to the Plasma Acylcarnitine Pool During Fasting and Exercise in Humans. J Clin Endocrinol Metab 101: 5044-5052.
226. Kien CL, Everingham KI, R DS, Fukagawa NK, Muoio DM (2011) Short-term effects of dietary fatty acids on muscle lipid composition and serum acylcarnitine profile in human subjects. Obesity (Silver Spring) 19: 305-311.
227. Shrestha A, Mullner E, Poutanen K, Mykkanen H, Moazzami AA (2017) Metabolic changes in serum metabolome in response to a meal. Eur J Nutr 56: 671-681.
228. Brockman DA, Chen X, Gallaher DD (2013) Consumption of a high β-glucan barley flour improves glucose control and fatty liver and increases muscle acylcarnitines in the Zucker diabetic fatty rat. European Journal of Nutrition 52: 1743-1753.
229. Dehne LI, Klemm C, Henseler G, Hermann-Kunz E (1999) The German food code and nutrient data base (BLS II. 2). European journal of epidemiology 15: 355-358.
230. Bouchard-Mercier A, Rudkowska I, Lemieux S, Couture P, Vohl MC (2013) The metabolic signature associated with the Western dietary pattern: a cross-sectional study. Nutr J 12: 158.
231. Sichert-Hellert W, Kersting M, Chahda C, Schäfer R, Kroke A (2007) German food composition database for dietary evaluations in children and adolescents. Journal of Food Composition and Analysis 20: 63-70.
232. Nelson DL, Lehninger AL, Cox MM (2008) Lehninger principles of biochemistry: Macmillan.
233. Bene J, Márton M, Mohás M, Bagosi Z, Bujtor Z, et al. (2012) Similarities in serum acylcarnitine patterns in type 1 and type 2 diabetes mellitus and in metabolic syndrome. Annals of Nutrition and Metabolism 62: 80-85.
234. Pochini L, Oppedisano F, Indiveri C (2004) Reconstitution into liposomes and functional characterization of the carnitine transporter from renal cell plasma membrane. Biochim Biophys Acta 1661: 78-86.
235. Hoene M, Li J, Li Y, Runge H, Zhao X, et al. (2016) Muscle and liver-specific alterations in lipid and acylcarnitine metabolism after a single bout of exercise in mice. Sci Rep 6: 22218.

236. Ramsay RR, Gandour RD, van der Leij FR (2001) Molecular enzymology of carnitine transfer and transport. Biochim Biophys Acta 1546: 21-43.
237. Rinaldo P, Cowan TM, Matern D (2008) Acylcarnitine profile analysis. Genet Med 10: 151-156.
238. Rutkowsky JM, Knotts TA, Ono-Moore KD, McCoin CS, Huang S, et al. (2014) Acylcarnitines activate proinflammatory signaling pathways. Am J Physiol Endocrinol Metab 306: E1378-1387.
239. Adams SH, Hoppel CL, Lok KH, Zhao L, Wong SW, et al. (2009) Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid β-oxidation and altered tricarboxylic acid cycle activity in type 2 diabetic African-American women. The Journal of nutrition 139: 1073-1081.
240. Aguer C, McCoin CS, Knotts TA, Thrush AB, Ono-Moore K, et al. (2015) Acylcarnitines: potential implications for skeletal muscle insulin resistance. FASEB J 29: 336-345.
241. Liepinsh E, Makrecka-Kuka M, Makarova E, Volska K, Svalbe B, et al. (2016) Decreased acylcarnitine content improves insulin sensitivity in experimental mice models of insulin resistance. Pharmacol Res 113: 788-795.
242. Hotamisligil GS, Shargill NS, Spiegelman BM (1993) Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. Science 259: 87-91.
243. Nakadate T, Blumberg PM (1987) Modulation by palmitoylcarnitine of protein kinase C activation. Cancer Res 47: 6537-6542.
244. Wise BC, Glass DB, Chou CH, Raynor RL, Katoh N, et al. (1982) Phospholipid-sensitive Ca2+-dependent protein kinase from heart. II. Substrate specificity and inhibition by various agents. J Biol Chem 257: 8489-8495.
245. Sato T, Arita M, Kiyosue T (1993) Differential mechanism of block of palmitoyl lysophosphatidylcholine and of palmitoylcarnitine on inward rectifier K+ channels of guinea-pig ventricular myocytes. Cardiovasc Drugs Ther 7 Suppl 3: 575-584.
246. Wu J, Corr PB (1992) Influence of long-chain acylcarnitines on voltage-dependent calcium current in adult ventricular myocytes. Am J Physiol 263: H410-417.
247. McCoin CS, Knotts TA, Ono-Moore KD, Oort PJ, Adams SH (2015) Long-chain acylcarnitines activate cell stress and myokine release in C(2)C(12) myotubes: calcium-dependent and -independent effects. American Journal of Physiology - Endocrinology and Metabolism 308: E990-E1000.
248. Holland WL, Summers SA (2008) Sphingolipids, insulin resistance, and metabolic disease: new insights from in vivo manipulation of sphingolipid metabolism. Endocr Rev 29: 381-402.
249. Meikle PJ, Summers SA (2016) Sphingolipids and phospholipids in insulin resistance and related metabolic disorders. Nat Rev Endocrinol.

250. Williams P (2007) Nutritional composition of red meat. Nutrition & Dietetics 64: S113-S119.
251. Chaurasia B, Summers SA (2015) Ceramides - Lipotoxic Inducers of Metabolic Disorders. Trends Endocrinol Metab 26: 538-550.
252. Chavez JA, Siddique MM, Wang ST, Ching J, Shayman JA, et al. (2014) Ceramides and glucosylceramides are independent antagonists of insulin signaling. J Biol Chem 289: 723-734.
253. Chavez JA, Summers SA (2012) A ceramide-centric view of insulin resistance. Cell Metab 15: 585-594.
254. Park M, Kaddai V, Ching J, Fridianto KT, Sieli RJ, et al. (2016) A Role for Ceramides, but Not Sphingomyelins, as Antagonists of Insulin Signaling and Mitochondrial Metabolism in C2C12 Myotubes. J Biol Chem 291: 23978-23988.
255. Summers SA, Goodpaster BH (2016) CrossTalk proposal: Intramyocellular ceramide accumulation does modulate insulin resistance. J Physiol 594: 3167-3170.
256. Iwabuchi K, Nakayama H, Oizumi A, Suga Y, Ogawa H, et al. (2015) Role of Ceramide from Glycosphingolipids and Its Metabolites in Immunological and Inflammatory Responses in Humans. Mediators Inflamm 2015: 120748.
257. Presa N, Gomez-Larrauri A, Rivera IG, Ordonez M, Trueba M, et al. (2016) Regulation of cell migration and inflammation by ceramide 1-phosphate. Biochim Biophys Acta 1861: 402-409.
258. Tiper IV, East JE, Subrahmanyam PB, Webb TJ (2016) Sphingosine 1-phosphate signaling impacts lymphocyte migration, inflammation and infection. Pathog Dis 74.
259. Sugimoto M, Shimizu Y, Zhao S, Ukon N, Nishijima K, et al. (2016) Characterization of the role of sphingomyelin synthase 2 in glucose metabolism in whole-body and peripheral tissues in mice. Biochim Biophys Acta 1861: 688-702.
260. Pyne S, Adams DR, Pyne NJ (2016) Sphingosine 1-phosphate and sphingosine kinases in health and disease: Recent advances. Prog Lipid Res 62: 93-106.
261. Fayyaz S, Japtok L, Kleuser B (2014) Divergent role of sphingosine 1-phosphate on insulin resistance. Cell Physiol Biochem 34: 134-147.
262. Chen W, Lu H, Yang J, Xiang H, Peng H (2016) Sphingosine 1-phosphate in metabolic syndrome (Review). Int J Mol Med 38: 1030-1038.
263. Vance JE, Vance DE (2004) Phospholipid biosynthesis in mammalian cells. Biochem Cell Biol 82: 113-128.
264. Baeza G, Sarria B, Bravo L, Mateos R (2016) Exhaustive Qualitative LC-DAD-MSn Analysis of Arabica Green Coffee Beans: Cinnamoyl-glycosides and Cinnamoylshikimic Acids as New Polyphenols in Green Coffee. J Agric Food Chem 64: 9663-9674.

265. Cai L, Ma D, Zhang Y, Liu Z, Wang P (2012) The effect of coffee consumption on serum lipids: a meta-analysis of randomized controlled trials. Eur J Clin Nutr 66: 872-877.
266. Naidoo N, Chen C, Rebello SA, Speer K, Tai ES, et al. (2011) Cholesterol-raising diterpenes in types of coffee commonly consumed in Singapore, Indonesia and India and associations with blood lipids: a survey and cross sectional study. Nutr J 10: 48.
267. Cano-Marquina A, Tarin JJ, Cano A (2013) The impact of coffee on health. Maturitas 75: 7-21.
268. Maeba R, Maeda T, Kinoshita M, Takao K, Takenaka H, et al. (2007) Plasmalogens in human serum positively correlate with high- density lipoprotein and decrease with aging. J Atheroscler Thromb 14: 12-18.
269. Maeba R, Hara H, Ishikawa H, Hayashi S, Yoshimura N, et al. (2008) Myo-inositol treatment increases serum plasmalogens and decreases small dense LDL, particularly in hyperlipidemic subjects with metabolic syndrome. J Nutr Sci Vitaminol (Tokyo) 54: 196-202.
270. Yukawa GS, Mune M, Otani H, Tone Y, Liang X-M, et al. (2004) Effects of Coffee Consumption on Oxidative Susceptibility of Low-Density Lipoproteins and Serum Lipid Levels in Humans. Biochemistry (Moscow) 69: 70-74.
271. Mursu J, Voutilainen S, Nurmi T, Alfthan G, Virtanen JK, et al. (2005) The effects of coffee consumption on lipid peroxidation and plasma total homocysteine concentrations: a clinical trial. Free Radical Biology and Medicine 38: 527-534.
272. Agudelo-Ochoa GM, Pulgarin-Zapata IC, Velasquez-Rodriguez CM, Duque-Ramirez M, Naranjo-Cano M, et al. (2016) Coffee Consumption Increases the Antioxidant Capacity of Plasma and Has No Effect on the Lipid Profile or Vascular Function in Healthy Adults in a Randomized Controlled Trial. J Nutr 146: 524-531.
273. Watschinger K, Werner ER (2013) Orphan enzymes in ether lipid metabolism. Biochimie 95: 59-65.
274. Coleman RA, Lee DP (2004) Enzymes of triacylglycerol synthesis and their regulation. Prog Lipid Res 43: 134-176.
275. Prentki M, Madiraju SR (2008) Glycerolipid metabolism and signaling in health and disease. Endocr Rev 29: 647-676.
276. Perry RJ, Samuel VT, Petersen KF, Shulman GI (2014) The role of hepatic lipids in hepatic insulin resistance and type 2 diabetes. Nature 510: 84-91.
277. Finck BN, Hall AM (2015) Does Diacylglycerol Accumulation in Fatty Liver Disease Cause Hepatic Insulin Resistance? Biomed Res Int 2015: 104132.
278. Selathurai A, Kowalski GM, Burch ML, Sepulveda P, Risis S, et al. (2015) The CDP-Ethanolamine Pathway Regulates Skeletal

Muscle Diacylglycerol Content and Mitochondrial Biogenesis without Altering Insulin Sensitivity. Cell Metab 21: 718-730.
279. Meikle PJ, Wong G, Barlow CK, Weir JM, Greeve MA, et al. (2013) Plasma lipid profiling shows similar associations with prediabetes and type 2 diabetes. PLoS One 8: e74341.
280. Weir JM, Wong G, Barlow CK, Greeve MA, Kowalczyk A, et al. (2013) Plasma lipid profiling in a large population-based cohort. J Lipid Res 54: 2898-2908.
281. Ma DW, Arendt BM, Hillyer LM, Fung SK, McGilvray I, et al. (2016) Plasma phospholipids and fatty acid composition differ between liver biopsy-proven nonalcoholic fatty liver disease and healthy subjects. Nutr Diabetes 6: e220.
282. Arendt BM, Ma DW, Simons B, Noureldin SA, Therapondos G, et al. (2013) Nonalcoholic fatty liver disease is associated with lower hepatic and erythrocyte ratios of phosphatidylcholine to phosphatidylethanolamine. Appl Physiol Nutr Metab 38: 334-340.
283. Kroger J, Zietemann V, Enzenbach C, Weikert C, Jansen EH, et al. (2011) Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. Am J Clin Nutr 93: 127-142.
284. Kroger J, Jacobs S, Jansen EH, Fritsche A, Boeing H, et al. (2015) Erythrocyte membrane fatty acid fluidity and risk of type 2 diabetes in the EPIC-Potsdam study. Diabetologia 58: 282-289.
285. Zhang L, Xie C, Nichols RG, Chan SH, Jiang C, et al. (2016) Farnesoid X Receptor Signaling Shapes the Gut Microbiota and Controls Hepatic Lipid Metabolism. mSystems 1.
286. Sekita A, Okazaki Y, Katayama T (2016) Dietary phytic acid prevents fatty liver by reducing expression of hepatic lipogenic enzymes and modulates gut microflora in rats fed a high-sucrose diet. Nutrition 32: 720-722.
287. Lu Y, Fan C, Li P, Lu Y, Chang X, et al. (2016) Short Chain Fatty Acids Prevent High-fat-diet-induced Obesity in Mice by Regulating G Protein-coupled Receptors and Gut Microbiota. Sci Rep 6: 37589.
288. Weitkunat K, Schumann S, Petzke KJ, Blaut M, Loh G, et al. (2015) Effects of dietary inulin on bacterial growth, short-chain fatty acid production and hepatic lipid metabolism in gnotobiotic mice. J Nutr Biochem 26: 929-937.
289. Han S, Jiao J, Zhang W, Xu J, Wan Z, et al. (2015) Dietary fiber prevents obesity-related liver lipotoxicity by modulating sterol-regulatory element binding protein pathway in C57BL/6J mice fed a high-fat/cholesterol diet. Sci Rep 5: 15256.
290. Foerster J, Maskarinec G, Reichardt N, Tett A, Narbad A, et al. (2014) The Influence of Whole Grain Products and Red Meat on Intestinal Microbiota Composition in Normal Weight Adults: A

Randomized Crossover Intervention Trial. PLoS ONE 9: e109606.
291. Augustin LS, Kendall CW, Jenkins DJ, Willett WC, Astrup A, et al. (2015) Glycemic index, glycemic load and glycemic response: An International Scientific Consensus Summit from the International Carbohydrate Quality Consortium (ICQC). Nutr Metab Cardiovasc Dis 25: 795-815.
292. Wirstrom T, Hilding A, Gu HF, Ostenson CG, Bjorklund A (2013) Consumption of whole grain reduces risk of deteriorating glucose tolerance, including progression to prediabetes. Am J Clin Nutr 97: 179-187.
293. Fisher E, Boeing H, Fritsche A, Doering F, Joost HG, et al. (2009) Whole-grain consumption and transcription factor-7-like 2 (TCF7L2) rs7903146: gene-diet interaction in modulating type 2 diabetes risk. Br J Nutr 101: 478-481.
294. Hindy G, Sonestedt E, Ericson U, Jing XJ, Zhou Y, et al. (2012) Role of TCF7L2 risk variant and dietary fibre intake on incident type 2 diabetes. Diabetologia 55: 2646-2654.
295. Wang J, Hu F, Feng T, Zhao J, Yin L, et al. (2013) Meta-analysis of associations between TCF7L2 polymorphisms and risk of type 2 diabetes mellitus in the Chinese population. BMC Med Genet 14: 8.
296. Ma Y, Gao M, Liu D (2015) Chlorogenic Acid Improves High Fat Diet-Induced Hepatic Steatosis and Insulin Resistance in Mice. Pharmaceutical Research 32: 1200-1209.
297. Ong KW, Hsu A, Tan BKH (2013) Anti-diabetic and anti-lipidemic effects of chlorogenic acid are mediated by ampk activation. Biochemical Pharmacology 85: 1341-1351.
298. Murase T, Misawa K, Minegishi Y, Aoki M, Ominami H, et al. (2011) Coffee polyphenols suppress diet-induced body fat accumulation by downregulating SREBP-1c and related molecules in C57BL/6J mice. American Journal of Physiology - Endocrinology And Metabolism 300: E122-E133.
299. Salomone F, Galvano F, Li Volti G (2017) Molecular Bases Underlying the Hepatoprotective Effects of Coffee. Nutrients 9: 85.
300. Hammad SM (2011) Blood sphingolipids in homeostasis and pathobiology. Adv Exp Med Biol 721: 57-66.
301. van Dam RM, Willett WC, Rimm EB, Stampfer MJ, Hu FB (2002) Dietary fat and meat intake in relation to risk of type 2 diabetes in men. Diabetes Care 25: 417-424.
302. Ma W, Wu JH, Wang Q, Lemaitre RN, Mukamal KJ, et al. (2015) Prospective association of fatty acids in the de novo lipogenesis pathway with risk of type 2 diabetes: the Cardiovascular Health Study. Am J Clin Nutr 101: 153-163.

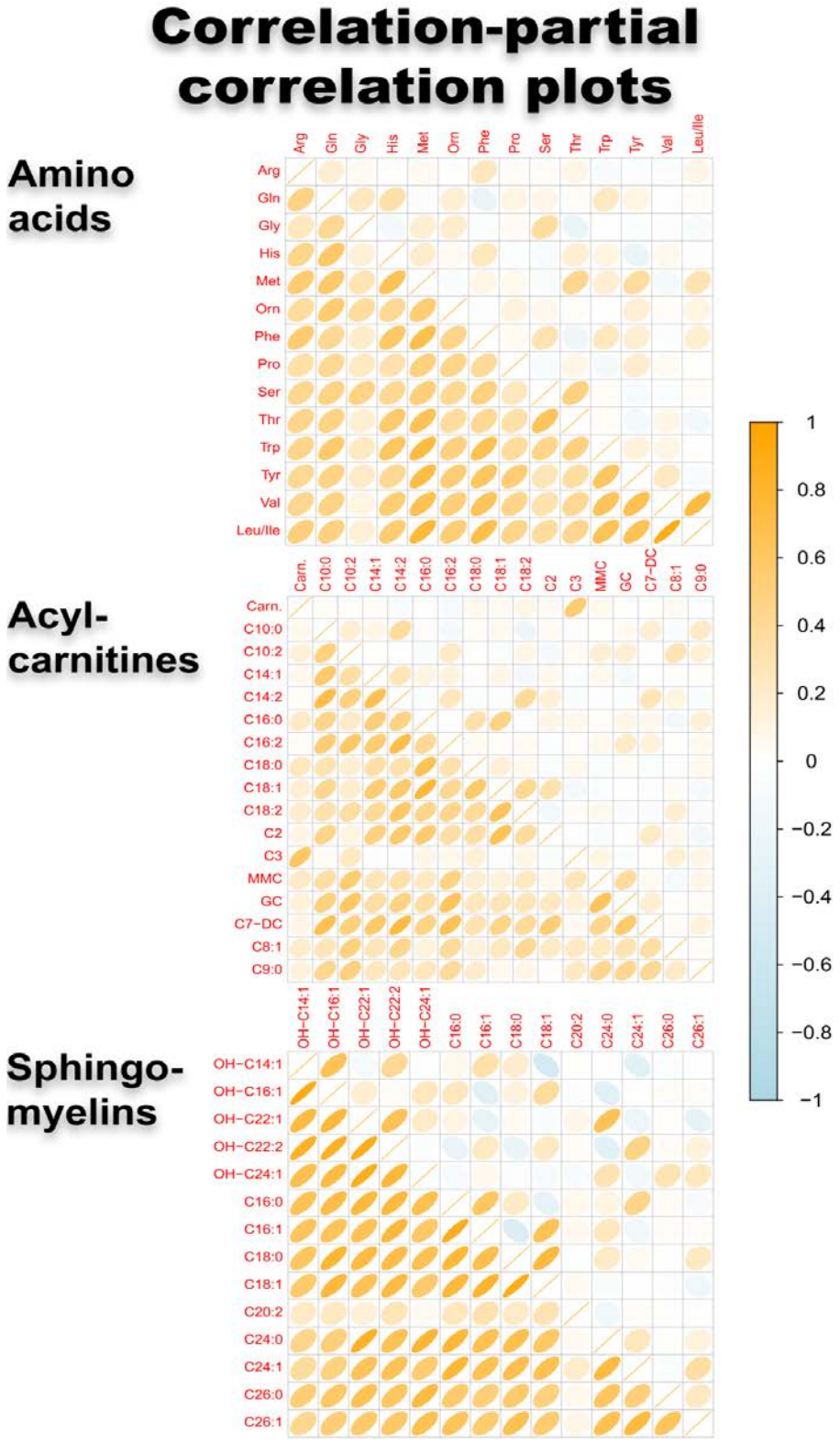
303. Hudgins LC, Hellerstein M, Seidman C, Neese R, Diakun J, et al. (1996) Human fatty acid synthesis is stimulated by a eucaloric low fat, high carbohydrate diet. J Clin Invest 97: 2081-2091.
304. van Meer G, Voelker DR, Feigenson GW (2008) Membrane lipids: where they are and how they behave. Nature reviews Molecular cell biology 9: 112-124.
305. Futerman AH, Riezman H (2005) The ins and outs of sphingolipid synthesis. Trends in cell biology 15: 312-318.
306. van Meer G (1989) Lipid traffic in animal cells. Annual review of cell biology 5: 247-275.
307. Simons K, Gruenberg J (2000) Jamming the endosomal system: lipid rafts and lysosomal storage diseases. Trends in cell biology 10: 459-462.
308. Shevchenko A, Simons K (2010) Lipidomics: coming to grips with lipid diversity. Nature reviews Molecular cell biology 11: 593-598.
309. Coskun Ü, Simons K (2011) Cell membranes: the lipid perspective. Structure 19: 1543-1548.
310. Tessaro FH, Ayala TS, Martins JO (2015) Lipid mediators are critical in resolving inflammation: a review of the emerging roles of eicosanoids in diabetes mellitus. Biomed Res Int 2015: 568408.
311. Khan SA, Ali A, Khan SA, Zahran SA, Damanhouri G, et al. (2014) Unraveling the complex relationship triad between lipids, obesity, and inflammation. Mediators Inflamm 2014: 502749.
312. Glass CK, Olefsky JM (2012) Inflammation and lipid signaling in the etiology of insulin resistance. Cell Metab 15: 635-645.
313. Imai Y, Dobrian AD, Morris MA, Taylor-Fishwick DA, Nadler JL (2016) Lipids and immunoinflammatory pathways of beta cell destruction. Diabetologia 59: 673-678.
314. Luo P, Wang MH (2011) Eicosanoids, beta-cell function, and diabetes. Prostaglandins Other Lipid Mediat 95: 1-10.
315. Robichaud PP, Surette ME (2015) Polyunsaturated fatty acid-phospholipid remodeling and inflammation. Curr Opin Endocrinol Diabetes Obes 22: 112-118.
316. Ricordi C, Garcia-Contreras M, Farnetti S (2015) Diet and Inflammation: Possible Effects on Immunity, Chronic Diseases, and Life Span. J Am Coll Nutr 34 Suppl 1: 10-13.
317. Huang CW, Chien YS, Chen YJ, Ajuwon KM, Mersmann HM, et al. (2016) Role of n-3 Polyunsaturated Fatty Acids in Ameliorating the Obesity-Induced Metabolic Syndrome in Animal Models and Humans. Int J Mol Sci 17.
318. Kapoor R, Huang YS (2006) Gamma linolenic acid: an antiinflammatory omega-6 fatty acid. Curr Pharm Biotechnol 7: 531-534.
319. Bhaswant M, Poudyal H, Brown L (2015) Mechanisms of enhanced insulin secretion and sensitivity with n-3 unsaturated fatty acids. J Nutr Biochem 26: 571-584.

320. Santilli F, Pignatelli P, Violi F, Davi G (2015) Aspirin for primary prevention in diabetes mellitus: from the calculation of cardiovascular risk and risk/benefit profile to personalised treatment. Thromb Haemost 114: 876-882.
321. Wymann MP, Schneiter R (2008) Lipid signalling in disease. Nat Rev Mol Cell Biol 9: 162-176.
322. Crowder MK, Seacrist CD, Blind RD (2017) Phospholipid regulation of the nuclear receptor superfamily. Adv Biol Regul 63: 6-14.
323. Ertunc ME, Hotamisligil GS (2016) Lipid signaling and lipotoxicity in metaflammation: indications for metabolic disease pathogenesis and treatment. J Lipid Res 57: 2099-2114.
324. Riserus U, Willett WC, Hu FB (2009) Dietary fats and prevention of type 2 diabetes. Prog Lipid Res 48: 44-51.
325. Maeba R, Nishimukai M, Sakasegawa S, Sugimori D, Hara H (2015) Plasma/Serum Plasmalogens: Methods of Analysis and Clinical Significance. Adv Clin Chem 70: 31-94.
326. Nishimukai M, Maeba R, Ikuta A, Asakawa N, Kamiya K, et al. (2014) Serum choline plasmalogens-those with oleic acid in sn-2- are biomarkers for coronary artery disease. Clin Chim Acta 437: 147-154.
327. Ly LD, Xu S, Choi SK, Ha CM, Thoudam T, et al. (2017) Oxidative stress and calcium dysregulation by palmitate in type 2 diabetes. Exp Mol Med 49: e291.
328. Hirsova P, Ibrabim SH, Gores GJ, Malhi H (2016) Lipotoxic lethal and sublethal stress signaling in hepatocytes: relevance to NASH pathogenesis. J Lipid Res 57: 1758-1770.
329. Rocha DM, Caldas AP, Oliveira LL, Bressan J, Hermsdorff HH (2016) Saturated fatty acids trigger TLR4-mediated inflammatory response. Atherosclerosis 244: 211-215.
330. Fritsche KL (2015) The science of fatty acids and inflammation. Adv Nutr 6: 293S-301S.
331. Hirsova P, Ibrahim SH, Krishnan A, Verma VK, Bronk SF, et al. (2016) Lipid-Induced Signaling Causes Release of Inflammatory Extracellular Vesicles From Hepatocytes. Gastroenterology 150: 956-967.
332. van Woudenbergh GJ, Kuijsten A, Tigcheler B, Sijbrands EJ, van Rooij FJ, et al. (2012) Meat consumption and its association with C-reactive protein and incident type 2 diabetes: the Rotterdam Study. Diabetes Care 35: 1499-1505.
333. Montonen J, Boeing H, Fritsche A, Schleicher E, Joost H-G, et al. (2013) Consumption of red meat and whole-grain bread in relation to biomarkers of obesity, inflammation, glucose metabolism and oxidative stress. European Journal of Nutrition 52: 337-345.

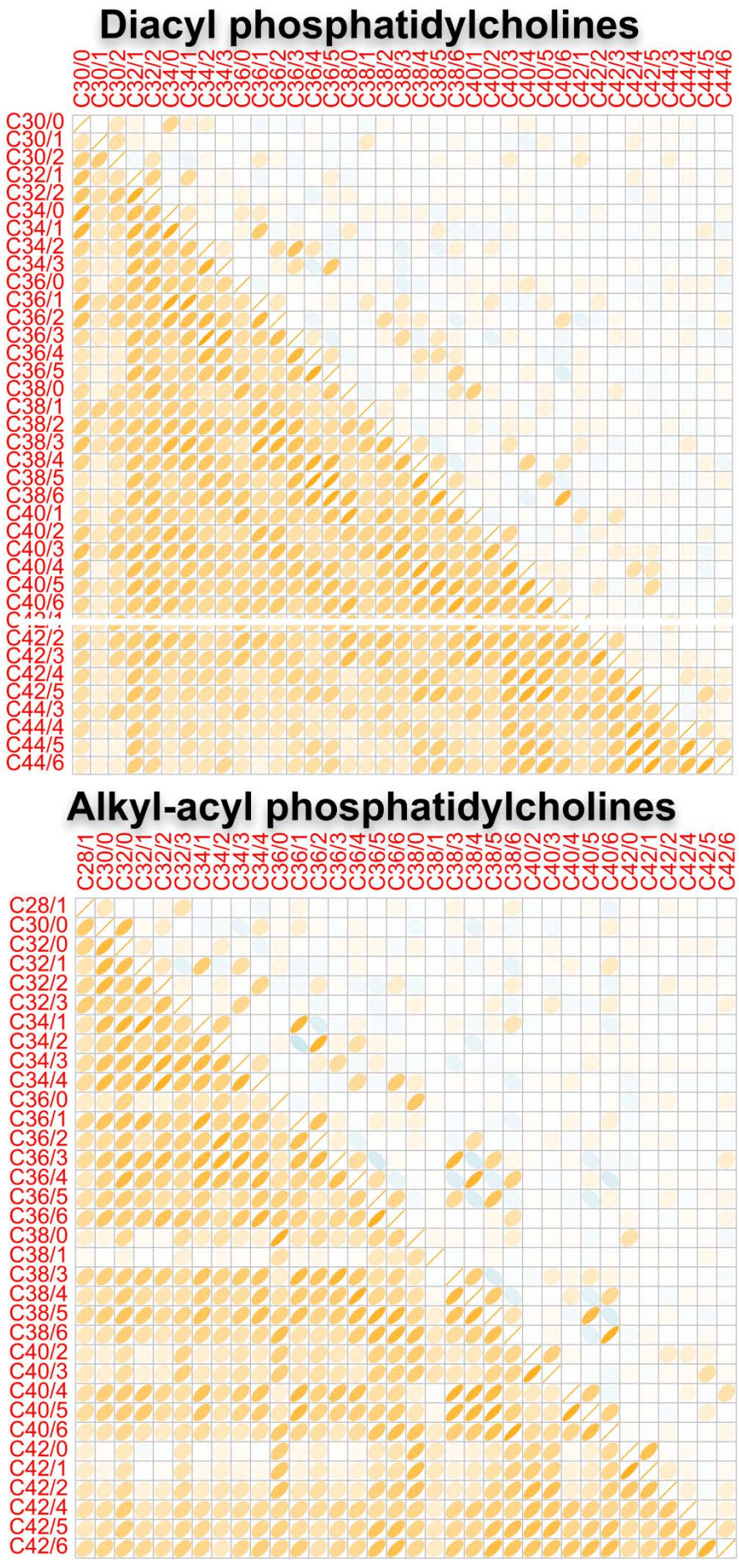
334. Masoodi M, Kuda O, Rossmeisl M, Flachs P, Kopecky J (2015) Lipid signaling in adipose tissue: Connecting inflammation & metabolism. Biochim Biophys Acta 1851: 503-518.
335. Varki A (2009) Multiple changes in sialic acid biology during human evolution. Glycoconj J 26: 231-245.
336. Alisson-Silva F, Kawanishi K, Varki A (2016) Human risk of diseases associated with red meat intake: Analysis of current theories and proposed role for metabolic incorporation of a non-human sialic acid. Mol Aspects Med 51: 16-30.
337. Samraj AN, Pearce OM, Laubli H, Crittenden AN, Bergfeld AK, et al. (2015) A red meat-derived glycan promotes inflammation and cancer progression. Proc Natl Acad Sci U S A 112: 542-547.
338. Nguyen DH, Tangvoranuntakul P, Varki A (2005) Effects of natural human antibodies against a nonhuman sialic acid that metabolically incorporates into activated and malignant immune cells. J Immunol 175: 228-236.
339. zur Hausen H (2012) Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. Int J Cancer 130: 2475-2483.
340. VanderWeele TJ (2013) Policy-relevant proportions for direct effects. Epidemiology 24: 175-176.
341. Hauser A, Bühlmann P (2015) Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77: 291-318.
342. Pearl J, Verma T (1991) A formal theory of inductive causation: University of California (Los Angeles). Computer Science Department.
343. Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin 2: 47-53.

8 Annex

8.1 Correlation-partial correlation plots of metabolite groups



Supplementary Figure 1: Correlation-partial correlation plots of amino acids, acylcarnitines, and sphingomyelins.



Supplementary Figure 2: Correlation-partial correlation plots of phosphatidylcholines

8.2 Factor analysis

Supplementary Table 1: Loadings of the metabolites in the first group-specific factor

Factor loadings					
Amino Acids					
Arg	0.61	Met	0.63	Thr	0.80
Gln	0.66	Orn	0.79	Trp	0.75
Gly	0.30	Phe	0.55	Tyr	0.81
His	0.71	Pro	0.55	Val	0.85
Leu/Ile	0.89	Ser	0.62		
Acylcarnitines					
C10:0	0.76	C18:0	0.45	C5-OH(C3-DC-M)	0.63
C10:2	0.60	C18:1	0.65	C7-DC	0.82
C14:1	0.72	C18:2	0.60	C8:1	0.45
C14:2	0.87	C2	0.60	C9:0	0.42
C16:0	0.63	C3	0.09	Carnitine	0.15
C16:2	0.77	C5-DC(C6-OH)	0.48		
Sphingomyelines					
OH-C14:1	0.79	C16:0	0.87	C24:0	0.80
OH-C16:1	0.86	C16:1	0.81	C24:1	0.74
OH-C22:1	0.91	C18:0	0.85	C26:0	0.70
OH-C22:2	0.92	C18:1	0.82	C26:1	0.70
OH-C24:1	0.85	C20:2	0.23		
Lysophosphatidylcholines					
C14:0	0.60	C18:0	0.80	C20:4	0.74
C16:0	0.86	C18:1	0.90	C28:1	0.26
C16:1	0.72	C18:2	0.70		
C17:0	0.60	C20:3	0.79		
Diacyl phosphatidylcholines					
C28/1	0.50	C36/2	0.63	C40/3	0.51
C30/0	0.70	C36/3	0.76	C40/4	0.75
C32/0	0.79	C36/4	0.75	C40/5	0.78
C32/1	0.70	C36/5	0.70	C40/6	0.66
C32/2	0.65	C36/6	0.78	C42/0	0.26
C32/3	0.63	C38/0	0.48	C42/1	0.35
C34/1	0.78	C38/1	0.15	C42/2	0.50
C34/2	0.62	C38/3	0.76	C42/4	0.62
C34/3	0.74	C38/4	0.73	C42/5	0.67
C34/4	0.78	C38/5	0.86	C42/6	0.7
C36/0	0.51	C38/6	0.68		
C36/1	0.76	C40/2	0.48		
Alkyl-acyl phosphatidylcholines					
C30/0	0.65	C36/4	0.66	C40/5	0.85
C30/1	0.30	C36/5	0.70	C40/6	0.75
C30/2	0.50	C38/0	0.62	C42/1	0.59
C32/1	0.81	C38/1	0.55	C42/2	0.74

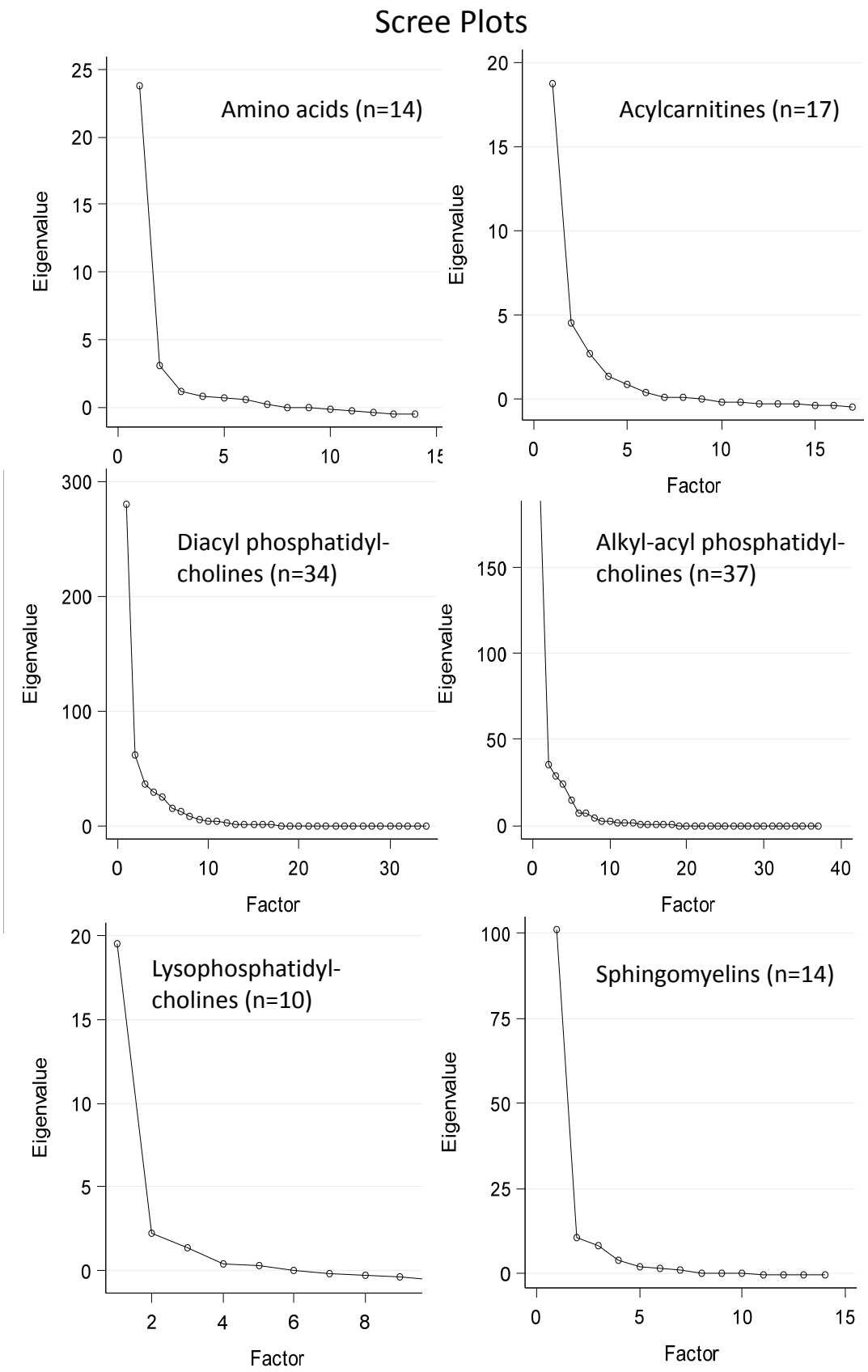
Factor loadings					
C32/2	0.81	C38/2	0.74	C42/3	0.74
C34/0	0.71	C38/3	0.73	C42/4	0.68
C34/1	0.76	C38/4	0.79	C42/5	0.71
C34/2	0.68	C38/5	0.74	C44/3	0.52
C34/3	0.65	C38/6	0.73	C44/4	0.56
C36/0	0.61	C40/1	0.69	C44/5	0.58
C36/1	0.68	C40/2	0.65	C44/6	0.58
C36/2	0.71	C40/3	0.83		
C36/3	0.71	C40/4	0.82		

As appropriate for the case-cohort design, the study sample to derive the factor loadings was restricted to the random subcohort (which is representative for the full cohort). Factor analysis was applied restricted to metabolite groups, i.e. separately among amino acids, acylcarnitines, sphingomyelins, lysophosphatidylcholines, diacyl phosphatidylcholines, and alkyl-acyl phosphatidylcholines, respectively. Analyses were based on the metabolite-residuals standardized for the participants' age, sex, BMI, and prevalence of hypertension.

Supplementary Table 2: Correlation between group-specific factors

	AA F1	DPC F1	AC F1	AEPC F1	LPC F1	SM F1
AA F1	1	0.20*	-0.09	0.21	0.25	0.06
DPC F1		1	0.09	0.71	0.40	0.33
AC F1			1	0.12	0.07	0.30
AEPC F1				1	0.33	0.52
LPC F1					1	0.20
SM F1						1

Pearson correlation coefficient between group-specific first factors of common variance (F1) in the subcohort (n=2,092). AA=amino acids; DPC = diacyl phosphatidylcholines; AC = acylcarnitines; AEPC = alkyl-acyl phosphatidylcholines; LPC = lysophosphatidylcholines; SM = sphingomyelins.



Supplementary Figure 3: Scree plots of group-specific factors

8.3 Theoretical background

8.3.1 Overview

This supplemental chapter on causal inference theory largely relies on the contributions by Judea Pearl as comprehensively summarized in his fundamental book *Causality* [129]. First, *causal models* and *causal diagrams* will be defined. Second, the terms *effect* and *effect identifiability* will be defined with respect to their use in causal inference literature. Third, criteria for the identifiability of effects from observational data will be derived including reflections on confounding and other sources of bias. Inference of causal structures from complex biological data will be subject of the last section.

8.3.2 Causal models and causal diagrams

Reiteration of the global biological research question of this work underscores its causal nature: *How does the habitual diet affect the metabolic pathway activities – and how does this translate into an altered risk of developing type 2 diabetes?* To answering such global causal questions from observational data, involves numerous variables to characterize exposure (dietary habits), potential mediators (metabolic activities), potential confounders (e.g. phenotypical characteristics and other diet & lifestyle factors) and outcome (time-to-type 2 diabetes incidence). Per definition, these variables are not randomly distributed with regard to the others, but show rather complex patterns of interdependencies. For metabolomics data this mutual dependency is of particular interest because of strong and biologically meaningful intercorrelation. Considering such high-dimensional intertwined information in large joint distribution functions rapidly reaches its limits – in terms of computation but even more so in terms of interpretability.

For most cases, however, each single variable directly depends on only a small subset of variables. An efficient and intuitive way to encode conditional independence information on multivariate distributions is the graphical representation as a network. Graphical approaches allow decomposing large joint distribution functions on complex observational data into several small distributions including only small subsets of variables [129]. The information from such small joint distributions can then be handled and interpreted separately but can be also coherently assembled again to generate a global picture. In other words, instead of

handling overly complex joint distributions, graphical approaches can be used to decompose global problems into "atomic" pieces [129].

Graphical models in this context can be used to encode the direct determinants of each variable in a joint model. Links encode data-generating mechanisms. From a graphical causal model of the data-generating process, sparse functions can be deduced to model the relation between any two variables in the system [140]. Thus a fully defined graphical causal model can be interpreted as a system of structured equations which in turn algebraically represents the data-generating mechanisms [195]. Partly specified causal models can still be of use to integrate incomplete knowledge on the data-generating process with observations on the conditional independency structure of the data [179,341]. Now the term *causal model* will be defined and *causal diagrams* will be specified as a subclass of graphical models.

In words, a causal model can be described as composite of three main ingredients. First, a set of variables U represents background factors that are not affected by variables inside the model. The set U can thus be seen as influential exogenous determinants that define the setting in which the actual mechanisms of interest are studied. Second, the set V represents variables within the model. Variables in V are sensitive to the setting U but might also be mutually dependent. The factors that determine the level of variable $V_i \in V$ are called parents PA_i with respect to V_i. The set of parents PA_i might comprise variables from V or U or both. Thus, V comprises variables of high interest for epidemiological models, for example the outcome but also mediators and the more "tricky" type of confounders that depend on the realization of other influential variables. The third ingredient of a causal model is a set of link functions F. These functions connect each set of parents to its child. Functions in F can be of any parametrical or non-parametrical form. In a causal model functions correspond to directed processes. This is in line with common understanding of epidemiological models in which exposure levels are used to estimate the occurrence of the outcome (perhaps via affecting some mediators) and not the other way round. The formal definition of a causal model reads as follows.

1 Causal Model[1]

A causal model is a triple

$M = \langle U, V, F \rangle$,

where:

i. *U is a set of background variables (also called exogenous), that are determined by factors outside this model;*

ii. *V is a set* $\{V_1, V_2, \ldots, V_n\}$ *of variables, called endogenous, that are determined by variables in the model – that is variables in* $U \cup V$*; and*

iii. *F is a set of functions* $\{f_1, f_2, \ldots, f_n\}$ *such that each* f_i *is mapping from (the respective domain of)* $U_i \cup PA_i$ *to* V_i*, where* $U_i \subseteq U$ *and* $PA_i \subseteq (U \cup V) \setminus V_i$ *and the entire set of F forms a mapping from U to V.*

In other words, each

$$f_i \text{ in } v_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n$$

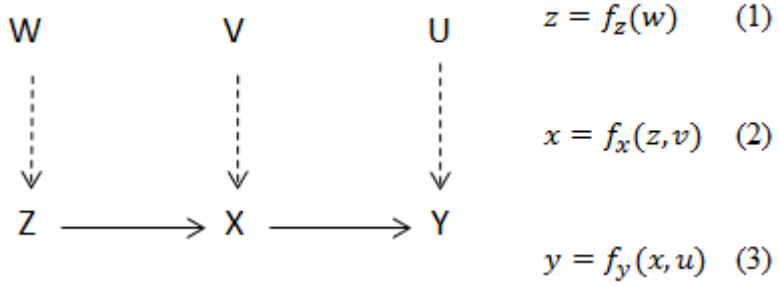
assigns a value to V_i *that depends on (the values of) a select set of variables in* $U \cup V$*, and the entire set F has a unique solution* $V(u)$.

Definition 1 specifies the causal model corresponding to an observed multivariate distribution as an unambiguous description of the structured process that generated these data. Each function corresponds to a direct mechanistical link, and no specified function implicates absence of a direct mechanistical link between two variables. The choice of the parents PA_i of V_i expresses the modelers' understanding of the variables that are mechanistically linked to V_i which will be subject to further elaborations below.

A causal model *M* can be associated with a directed graph G(*M*), named *causal diagram*. Nodes in such diagrams correspond to variables and directed edges (arrows) depict a directed causal link, i.e. a functional relation between a pair of variables. The encoded information depicted in *causal diagrams* is thus limited to qualitative assumptions on the direct influence of *endogenous* and *exogenous* variables on each V_i and does not specify the functional form of f_i. Graph G_1 (Supplementary Figure 4), e. g., encodes causal claims on the relation between the included variables. The graph contains *endogenous* variables $\{X, Y, Z\}$ corresponding to set V in definition (1) and *exogenous* variables $\{W, V, U\}$ corresponding to set U in definition (1). It should be noted that the absence of edges in

[1] Definition 1 literally quotes p.203 from 129. Pearl J (2009) Causality: Cambridge university press.

Supplementary Figure 4 claims the absence of direct influences (and non-correlated error terms for the *exogenous* variables $\{W, V, U\}$). The set of structural equations in algebraically expresses the causal claims encoded in G_1, and functions $\{f_X, f_Y, f_Z\}$ can generally be of any parametric or non-parametric form.
Translation from graphical into algebraic models and vice versa will be a key-characteristic of the methodological approach taken by this study. Causal modeling theory will inform this process in terms of variable selection and interpretation of thee obtained estimates.



Supplementary Figure 4: Causal graph G_1

8.3.3 Causal effects and effect identifiability

This work aims to identify metabolic mechanisms that connect diet to type 2 diabetes incidence. Therefore, a clear definition of effects (in distinction to associations) and criteria for effect identifiability from observational data is crucial. In plain language an effect refers to a mechanism that links exposure and outcome. The interpretation of statistical dependency as a causal effect cannot rely on observed associations alone. A causal interpretation of statistical estimates is only possible with reference to a causal model. If the causal model is wrong, insufficiently specified or not sufficiently covered by measurements a causal interpretation of estimates would be false and probably misleading. Causal inference thus depends on well-informed assumptions and cannot be driven by the data alone.

The reflections on causal inference will be limited to data-generating mechanisms that can be adequately described by structural equation models (*SEM*) and graphs that are Markovian (i.e. acyclic with uncorrelated error-terms of *exogenous* variables). In graphical models directed edges (i.e. arrows) will be used to represent specified causal or temporal relationships. The theoretical considerations treated here can be in principle extended to semi-Markovian models (i.e. acyclicity but

correlated error-terms of exogenous variables) but this will not be discussed specifically. Estimating effects within semi-Markovian models includes inference on latent (non-measured) variables and was not subject of this work.

In the last two decades, well-defined semantics and a well-founded logic have been developed for mathematically formalizing causal claims and encoding them in graphical models. Now the term *effect* as used in the causal inference literature will be defined and criteria for *effect identifiability* will be derived. It should be noted that an *effect* is a causal quantity, and is therefore defined relative to an underlying *causal model M*, unlike statistical parameters such as associations which are defined relative to a joint distribution $P_M(\dot{v})$ over a set of observed variables V.

2 Causal Effect [129][2]

Given two sets of disjoint variables, X and Y, the effect of X on Y, denoted either as $P(y|\hat{x})$ *or as* $P(y|do(x))$*, is a function of X to the space of probability distributions on Y. For each realization x of X,* $P(y|\hat{x})$ *gives the probability of Y = y induced by deleting all equations corresponding to variables in X and substituting X=x in the remaining equations.*

In other words, the effect of X on Y is defined as the variance in Y that is attributable to setting X to the level x (in contrary to the facts, X is assumed to being set to the alternative level without "touching" variables in the system others than X). Assuming that the level of Y is sensitive to such hypothetical manipulation of X entails the claim of a mechanistical link directed from X to Y. If this effect does not (necessarily) involve other variables in the system we say that X and Y have a parent-child-relation which will later be defined in a formal way. For now recall that X is considered as parent with respect to Y if X directly affects Y ; in turn Y is directly affected by X and Y is thus considered as child with respect to X; graphically such parent-child-relation is depicted by an arrow emanating from X with the arrow-head pointing into Y (X→Y). An *effect* is defined as the mechanism linking a cause to a consequence, and the graphical notation provides the mathematical formalism to express such simple and intuitive directionality assumption.

[2] Definition 2 literally quotes p.70 from 129. Pearl J (2009) Causality: Cambridge university press.

The information provided by observational data, however, is on the joint distribution $\mathrm{P_M}(\dot{\mathrm{v}})$ alone, and as explicated below in further detail the *equivalence class* of *causal models* generates the same distribution. Thus, it is generally not possible to infer the data-generating causal processes based on observed data alone and the causal quantity of interest might possibly not be unambiguously discernible – even assuming complete information on $\mathrm{P_M}(\dot{\mathrm{v}})$ over V in the source population. Identifiability applies if the prior knowledge suffices to specify the *causal model M* to an extent that it conveys the necessary assumptions to derive valid estimates on the sought quantity. This does not necessarily imply *M* being explicated in full detail.

3 Identifiability[3]

Let $Q(M)$ *be any computable quantity of a model* M. *We say that* Q *is identifiable in a class* M *of models if, for any pair of models* M_1 *and* M_2 *from* M, $Q(M_1) = Q(M_2)$ *whenever* $P_{M_1}(v) = P_{M_2}(v)$. *If our observations are limited and permit only a partial set* F_M *of features of* $P_M(v)$ *to be estimated, we define* Q *to be identifiable from* F_M *if* $Q(M_1) = Q(M_2)$ *whenever* $F_{M_1} = F_{M_2}$.

Identifiability is essential for integrating statistical data with incomplete causal knowledge. Particularly, identifiability enables to estimate quantities Q consistently from P(v) without necessarily specifying *M* in full detail; specifying the general characteristics of the class of *M* suffices. In this work the quantity Q of interest is the *effect* of X on Y, i.e. $P(y|\hat{x})$. Computation of this quantity from a fully specified model *M* is straightforward. More sophisticated is the task, whenever $P(y|\hat{x})$ needs to be computed without full specification of *M*; we might rather have an incomplete specification of *M* in form of some qualitative assumption, e. g. directionality assumption on the relation between (sets of) variables based on causal reasoning or temporal order of the data.

To this end, consider a class *M* of models sharing the same characteristics, i.e. the same parent-child families; and assume that all models induce a positive distribution on the observed variables ($P_M(v) > 0$). Relative to such classes *effect identifiability* from observational data is defined in the following.

[3] Definition 3 literally quotes p.82 from 129. Pearl J (2009) Causality: Cambridge university press.

4 Effect Identifiability[4]

The causal effect of X *on* Y *is identifiable from a graph* G *if the quantity* $P(y|\hat{x})$ *can be computed uniquely from any positive probability of the observed variables – that is, if* $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$ *for every pair of models* M_1 *and* M_2 *with* $P_{M_1}(v) = P_{M_2}(v) > 0$ *and* $G(M_1) = G(M_2) = G$.

Simply speaking, an effect is identifiable from observational data if assumptions on the class of causal models are specific enough to render the sought quantity which is assumed to be invariant to the non-specified causal relationships. It should be noted that even in the cases where the degree of model specification does not suffice to infer a unique estimate on the *causal effect* we might still be able to infer a limited set of possible *effects* based on available observations. In other words, in case of non-identifiability of a single *effect* it might still be informative to estimate the set of possible *effects* based on the correlation-derived *equivalence class* of causal models or a subset thereof, taking into account prior causal assumptions.

8.3.4 Deconfounding adjustment sets

Given information on unbiased samples of the source population, invariance of observation-based effect estimates generally depends upon adequate control of confounding in statistical models. Confounding itself is a causal concept, and as such inherits directionality assumptions. The causal nature of the concept of confounding becomes obvious when compared to the concept of mediation. Relative to the effect of X on Y, a confounder C is defined as affecting both X and Y (X←C→Y). Only directionality assumptions allow distinguishing between a confounder C and an intermediary factor I (aka mediator) that lies on a causal path from X to Y (X→I→Y). It is not possible to distinguish between confounders and mediators based on information on the joint distribution alone! Not adjusting for a confounder, however, would bias estimates for the effect of X on Y, whereas adjustment for mediators in contrary would result in biased estimates if the sought quantity is the total effect of X on Y. (If the sought quantity is the direct effect of X on Y, however, we have to adjust for mediators as well – *direct effects* will be

[4] Definition 4 literally quotes p.77 from 129. Pearl J (2009) Causality: Cambridge university press.

defined below.) Appropriate selection of adequate adjustment sets therefore necessarily relies on structural assumptions – be it implicitly or explicitly.

Graphical criteria to selecting sufficient (or *deconfounding*) adjustment sets to estimate valid *effects* from observational data will now be stated. An obvious choice for a sufficient adjustment set would be to estimate the effect of X on Y conditional on all (potential) predecessors of X. This set is often too large to be handled and information is often limited to a subset of (potential) predecessors. In fact, to efficiently control for potential confounding of any other variables, we need only to concern ourselves with the set of *Markovian parents* or shortly *parents* of X.

5 Markovian parents[5]

Let $V = \{X_1, \dots, X_n\}$ *be an ordered set of variables, and let* $P(v)$ *be the joint probability distribution on these variables. A set variables* PA_i *is said to be Markovian parents of* X_i *if* PA_i *is a minimal set of predecessors of* X_i *that renders* X_i *independent of all its other predecessors. In other words,* PA_i *is any subset of* $\{X_1, \dots, X_{i-1}\}$ *satisfying* $P(x_i|PA_i) = P(x_i|X_1, \dots, X_{i-1})$ *and such that no proper subset of* PA_i *satisfies the above-stated equation.*

A useful property of Markovian models is that the set of Markovian parents *d-separates* each variable X_i from all its nondescendants. In other words, "*each variable X_i is independent of all its nondescendants, given its parents PA_i in G"* (*Pearl and Verma [342]*). Every flow of information from distal predecessors of X_i has to be transmitted by a path involving the set of Markovian parents. In a graph G, the *adjacency set* of a node X_i denoted *$adj_i(G)$* is defined as the group of nodes directly connected to X_i by a single edge. Given G, the set of *Markovian parents* of X_i is a subset of the adjacency set ($\mathrm{pa_i} \subseteq adj_i$).

[5] Definition 5 literally quotes p.14 from 129. Pearl J (2009) Causality: Cambridge university press.

To define *d-separation* consider three disjoint sets of variables X, Y, and Z, which are represented as a directed acyclic graph G.

6 d-Separation[6]

A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if

i. p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node is in Z, or
ii. p contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that no descendant of m is in Z

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

Collider bias describes the phenomenon that observations on a common consequence of two independent causes provides information on the likelihood of both causes having occurred thereby introducing a dependency of the causes in every stratum of the consequence. This pattern of conditional dependencies was described by Joseph Berkson in 1946 [343] and is therefore known as *Berkson's paradox*. For illustration consider the following artificial example: A genetic variant has a high prevalence (say 20% heterozygote carriers of the rare variant). Now assume that heterozygote carriers have no disadvantages but that the homozygote genotype of the rare variant is embryonically lethal. Assuming this particular genetic variant not having any influence on selecting a partner and not taking into account evolution we would await to find an inverse association between the heterozygote genotype in parents of the same children, just because the prevalence of the heterozygote genotype in both parents clearly affects the probability to give birth to a healthy child. Stratifying on the effect of common causes renders the causes mutually dependent in the strata. For this simplified and artificial example the conclusion that the genotype of one parent cannot affect the genotype of the other parent can be unmistakable drawn based on well-established biological knowledge. In real-world biological problems, such conclusion would perhaps not be as obvious. The point is that it is not always safe to adjust for additional variables because adjusting for colliders would bias effect estimands.

[6] Definition 6 literally quotes p.16 from 129. Pearl J (2009) Causality: Cambridge university press.

Now criteria to select sufficient adjustment sets to control confounding will be deduced from the graphical definition of *d-separation* under the assumption that the underlying causal structure is known. Applying these criteria to the problem of estimating the effect of X on Y from a sample P(v) given the assumptions encoded in G, it can be tested whether a set of variables is sufficient for identifying $P(y|\hat{x})$ from observational data.

7 Back-Door criterion[7]

A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:

i. no node in Z is a descendant of X_i; and

ii. Z blocks every path between X_i and X_j that contains an arrow into X_i.

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

Thus, the *effect* of X on Y is identifiable from observational data whenever a *deconfounding* set Z of variables is available and the *back-door criterion* is applicable based on the causal assumption encoded in G. The *effect* is then identifiable by adjusting the relation between X and Y for Z.

$$P(y|\hat{x}) = \sum_z P(y|x,z)P(z)$$

Back-door refers to the fact that back-door paths contain variables with arrowheads pointing into the potential causal variable X in a considered cause-effect pair (X, Y). Analogously a front-door path contains arrows emanating from X.

8 Front-Door criterion[8]

A set of variables Z satisfies the front-door criterion relative to an ordered pair of variables (Xi, Xj) in a DAG G if:

i. Z intercepts all directed paths from X to Y;

ii. there is no unblocked back-door path from X to Z; and

iii. all back-door paths from Z to Y are blocked by X.

[7] Definition 7 literally quotes p.79 from 129. Pearl J (2009) Causality: Cambridge university press.

[8] Definition 8 literally quotes p.82 from 129. Pearl J (2009) Causality: Cambridge university press.

The front-door criterion implies that the effect of X on Y is also identifiable given information on a set of variables Z that block every causal path from X to Y but have no effect on X.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und ausschließlich die angegebenen Quellen genutzt habe. Weiterhin versichere ich, die Arbeit an keiner anderen Hochschule eingereicht zu haben.

Potsdam, den 26.04.2017 Clemens Wittenbecher

Danksagung

Zuerst möchte ich mich bei Matthias Schulze bedanken: Für das Vertrauen in die Ideen, die ich in meiner Arbeit verfolgen wollte. Für die Unterstützung dabei, die Rahmenbedingungen zu schaffen und die Kontakte herzustellen, durch die das möglich wurde. Und für die lehrreichen Kommentare und bereichernden Diskussionen. Ich hätte mir keine bessere Betreuung meiner Arbeit wünschen können.

Mein besonderer Dank gilt Jan Krumsiek. Er stand mir bei den bioinformatischen Aspekten der Arbeit zur Seite. Seine Sicht auf Netzwerke hat die Arbeit entscheidend beeinflusst.

Cornelia Weikert danke ich für ihre hilfreiche Rolle als zweite Betreuerin der Arbeit. Janine Kröger danke dafür, dass sie während der Arbeit als meine Mentorin zur Verfügung stand.

Ein herzlicher Dank geht auch an meine Kollegen und ehemaligen Kollegen in den Abteilungen für Molekulare Epidemiologie und Epidemiologie und am Deutschen Institut für Ernährungsforschung. Die anregenden Diskussionen zu Arbeitsthemen, alle anderen Gespräche und die gemeinsame Freizeit haben die letzten Jahre sehr bereichert. Ganz konkret möchte ich mich bei Cecilia Galbete, Catarina Schiborn, Fabian Eichelmann, Simone Jacobs, Kristin Mühlenbruch, Olga Kuxhaus und Elli Polemiti für die Unterstützung bei der Anfertigung der Dissertationsschrift bedanken.

Herzlicher Dank gilt meinen Eltern. Ohne ihre bedingungslose Unterstützung hätte ich Interessen und Beruf nicht so in Einklang bringen können, wie es mir bis jetzt möglich war. Mein verstorbener Vater hat mir viele Grundlagen mit auf den Weg gegeben, die mir bei der Realisierung dieser Arbeit geholfen haben. Meine Mutter auch. Darüber hinaus war sie auch im Alltag der vergangenen Jahre immer eine große Hilfe. Dank gilt auch meinem Bruder für den Austausch zu Inhalten der Arbeit und dem wissenschaftlichen Kontext.

Mein liebevoller Dank gilt meiner Familie, meiner Frau Galina und meinen Kindern Arthur, Irmelinde und Ferun. Euch widme ich diese Arbeit.