Application of Gaussian Graphical Models to Derive Dietary Intake Networks and Association of the Identified Networks with Risk of Major Chronic Diseases in European Prospective Investigation into Cancer and Nutrition-Potsdam Cohort

vorgelegt von

MSc. (Hons)

Khalid Iqbal

geb. in Swabi, Pakistan

von der Fakultät VII–Wirtschaft und Management

der Technischen Universität Berlin

zur Erlangung des akademischen Grades

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

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Tobias Kurth

Gutachter: Prof. Dr. Heiner Boeing

Gutachter: Prof. Dr. Reinhard Busse

Tag der wissenschaftlichen Aussprache: 25 July 2017

Berlin 2017

... to my parents who sent me to school,

they didn't go themselves

and

to my wife and children

who kept me moving ahead.

Contents

List of tables

List of figures

Glossary of abbreviations

EPIC	European Prospective Investigation into Cancer and Nutrition
FFQ	Food frequency questionnaire
GGM	Gaussian graphical models
SGCGM	Semiparametric Gaussian copula graphical models
BMI	Body mass index
T2D	Type 2 diabetes
MI	Myocardial infarction
CVD	Cardiovascular diseases
PCA	Principal Component Analysis
PAL	Physical activity level
MET	Metabolic equivalent
SAS	Statistical Analysis System
SD	Standard deviation
CI	Confidence interval
UI	Uncertainty interval
RRR	Reduced rank regression
HR	Hazard ratio
IARC	International agency for research on cancer
MET	metabolic equivalent
BMR	Basal metabolic rate

1. Introduction

1.1 Background

Human kind has endured an overwhelming shift in dietary and physical activity patterns since the appearance of Paleolithic man on earth (1). Changes in dietary patterns ["*the quantities, proportions, variety or combinations of different foods and beverages in diets, and the frequency with which they are habitually consumed*" (2)] and lifestyle, as humans drifted from classic dietary sources through hunting and forest foraging to farming, might have profoundly influenced their health status. As early as 370 BC, Hippocrates recognized such an influence by surmising that environmental factors, including diet, influence imbalance of four body fluids or Humor responsible for human diseases (3, 4). Similarly, in 1799, classic work of Easton, who compiled biographical data on longevity of 1712 people, living for 100 years or more, from AD 66 to 1799 (5), revealed that many centenarian shared common characteristic of sparing consumption (6), implying an implicit association between a certain dietary pattern and longevity. Since last century, substantial data regarding the role of diet in health has further strengthened this conviction. However, evidence regarding explicit relationship between dietary intake and health outcomes is still inconclusive.

Given the earlier indications concerning the importance of diet in human health and advances in food biochemistry, there was an immense interest to explore the role of diet in health and disease. Nevertheless, most of the earlier studies, were primarily limited to investigate the role of single food or food constituent (nutrient and other dietary component) in relation to health outcomes (7). The investigations to identify effect of single nutrient or food on health outcomes such as the role of vitamin C in curing scurvy and Vitamin A in vision and immunity, though important, have certain limitations. First, complex health conditions and diseases like obesity, type 2 diabetes (T2D), cardiovascular diseases (CVD) and cancer are a consequence of complex interplay between multiple dietary factors and other environmental factors and hence the role of diet can best be investigated by taking into account the multidimensional approaches rather than single nutrient studies. Second, single nutrient studies are hindered by interaction of nutrients with each other and with other food complexes (8). Therefore, it was proposed that dietary pattern analysis rather than studies of single nutrient may provide important insights into the role of diet in health outcomes (9).

This is a pragmatic approach since food is consumed in combination and the correlated nature of foods and the food components can be of advantage if studied together. Moreover, eating pattern differ across cultures, religions and regions. Thus, pattern analysis may identify unique dietary patterns reflecting different traditions that may also be related to health outcomes (10).

Exploratory analysis based on data-reduction methods, like Principal Component Analysis (PCA) and cluster analysis, is frequently used to derive dietary patterns (11). PCA has been of particular interest, as it compresses food groups, based on correlation or covariance among original variables, into a number of uncorrelated patterns called components or factors (12). The identified patterns characterize dietary intake and are used as exposure to investigate health outcomes.

Although the correlation structure assessed by such methods helps to better understand data and identify the similarity pattern among food groups, it, however, cannot completely unravel understanding of the pairwise association among food variables. Pairwise correlations among food groups can be more informative if it is independent of the effect of other food groups (13). Such pairwise correlations among food groups controlling for others identify dependency of various food groups in the dietary data, which may be important to understand how different foods are consumed in relation to each other(14).

Moreover, the existing methods of dietary pattern analysis require several but crucial subjective choices during data analysis (15, 16). Besides, the identified patterns are often difficult to interpret (12, 17, 18) and are frequently not associated with health outcomes (19). These limitations warrant investigation of complementary approaches to characterize dietary intake patterns. Innovative methods that provide additional insights into intake patterns and minimize these concerns might be advantageous over conventional ones. Such methods can improve understanding the complexity of eating behaviors and diet-disease relationship.

Gaussian graphical models (GGM) form a promising class of methods for exploratory analysis (20). These are graphical methods that identify conditional independence structure in the dataset by assessing pairwise correlation among two variables controlling for others. GGM assume multivariate normal distribution for underlying data and can infer direct relationship among variables in a given data set without prior knowledge

(21). GGM have been used to simplify and compress high dimensional genetic (22, 23) and metabolomics (24, 25) data to explore respective underlying pathways.

As dietary data are high dimensional like genetic and metabolomics data, the application of GGM to identify conditional independence structures among food intake variables is an interesting approach. In dietary intake data, the pairwise correlation among two food groups controlling for others can identify both the internal structure (i.e. patterns) in the original data as well as the relationship among the food groups consumed in the identified network. The latter characteristic is of particular interest as foods are consumed in specific combinations that reflects consumption patterns and may be helpful in providing insight into eating behavior of the studied population. In addition, these networks may also identify key interrelated food groups that may be potential candidates for further investigation into confounder structures and in-depth understanding of biological relationships between diet and health status. This feature of GGM may be of higher significance to identify patterns and food groups that may be associated with major chronic diseases like T2D, myocardial infarction (MI), stroke and cancer.

1.2 Objectives of the study

The primary aim of the study was to introduce and apply GGM as an innovative approach of dietary pattern analysis for further understanding the interrelationship between food groups consumed and investigate association of the identified dietary patterns with risk of major chronic diseases in the EPIC-Potsdam cohort study.

Specific objectives of the study were to:

i. Apply GGM to construct and validate the networks of dietary intake representing dietary patterns of the participants of the EPIC-Potsdam cohort.

ii. Investigate the adherence to the identified food intake pattern (food intake networks) and risk of T2D, MI, stroke, CVD, cancer, cardiometabolic diseases (T2D and CVD combined), and overall chronic diseases (T2D, MI, stroke, and cancer combined).

iii. Investigate the associations of the foods/foods groups constituting GGM derived dietary intake networks with risk of T2D, MI, stroke, CVD, cancer, cardiometabolic diseases, and overall chronic diseases.

iv. Compare the results of GGM identified pattern with that of PCA by reconstructing PCA dietary patterns in EPIC-Potsdam as done previously by Schulze et al (26) and investigate association of the derived dietary patterns with T2D, MI, stroke, CVD, cancer, cardiometabolic diseases, and overall chronic diseases.

1.3 Significance of the research question

Relevance to the field of nutrition epidemiology:

Existing statistical methods of dietary pattern analysis identify the independent vectors called patterns, in covariance or correlation matrix and assign scores to individuals, which are used to assess adherence to the identified patterns (27). Introduction of such methods (e.g. PCA) to nutrition epidemiology helped overcome limitations associated with single nutrient/food studies and facilitated characterization of population specific dietary patterns. Adoption of such methods has also helped to recognize health benefits of following a certain type of dietary pattern. Moreover, these statistical methods have yielded consistent results from studies in certain populations, strengthening evidence to define optimum diet, which have been translated into dietary guidelines (28). Nevertheless, since foods are consumed in specific combinations; hence dietary recommendations for a population require understanding of food intake in context of how these foods are consumed in relation to each other. Moreover, the existing understanding of optimal diet is incomplete, requiring further evidence to establish what constitute a healthy dietary pattern. Current methods may be limited in providing such evidence. For example, existing methods do not provide information on how foods are consumed in relation to each other. Likewise, in dietary patterns identified with methods such as PCA, several foods may contribute to more than one pattern, rendering it difficult to conclude its role in relation to health outcome (29).

Introduction of GGM to identify consumption patterns attempts to fill this gap in Nutrition Epidemiology. This approach explores conditional independence relationship among food groups that are consumed in a population. The conditional independence network of dietary intake best represents the underlying structure of the dataset, where each food is part of only one network at a time, in contrast to the existing methods. These conditional independence networks represent dietary patterns and can be scored to investigate adherence to these patterns and risk of major chronic diseases. The method also offers the opportunity to study the effect of single food groups in the identified food networks.

Relevance to public health

Chronic diseases including T2D, CVD, and cancer are a major cause of deaths and loss of disability-adjusted life years (30). Global burden of disease analyses showed a significant increase in global mortality estimates related to diabetes [1.5 million deaths (95% uncertainty interval (UI): 1.5 million to 1.6 million)], CVD [17.9 million (95% UI: 17.6 million to 18.3)], and cancer [8·8 million (95% UI: 8·6 million to 8.9 million)] from the years 2005-2015 (31). These figures represent an increase in total deaths attributed to diabetes by 32.1% (95% UI: 27.7-36.3), that of CVD by 12.5% (95% UI: 10.6-14.4), and that of cancers by 17·0% (95% UI: 14·8-19·3). Diabetes also contributed to 418,000 (95% UI: 389,000-441,000) deaths, representing an increase of 39·5% (95% UI: 35.4-43·5) from 2005 to 2015. Diabetes and cancer also caused increased years of life lost (YLL) from 2005 to 2015. By year 2015, YLLs due to diabetes rose by 25·4% (95% UI: 20·4-30·0), advancing it ranking from 18th to 15th place. Moreover, the age-standardized prevalence of diabetes nearly doubled in the last 25 years, rising from 4.7% to 8.5% in the adult population (32).

These statistics show that urgent actions are warranted to tackle the high burden of these non-communicable chronic diseases to prevent early death, YLLs and higher proportion of DALYs attributed to these diseases. Diet is considered an important exposure in development of these major chronic diseases (33). Several dietary patterns (prudent, western etc.) have been consistently linked with risk of T2D, CVD and certain types of cancer (34). However, the notion of what constitute a true pattern, how different foods are consumed in relation to each other, and how it relates to these diseases in different populations is still elusive. In absence of a gold standard to assess dietary patterns, it is hard to conclude what constitute a true pattern; nevertheless, the application of GGM in different populations can help identify unique dietary intake patterns that best represent underlying data and can also be investigated in relation to risk of chronic diseases.

In brief, this approach identifies the interrelationship of foods consumed in a population. Understanding such relationships among foods help identify intervention space in the existing patterns of intake that may help formulate practical approaches for dietary modification to prevent diet related chronic diseases.

2. Background of approaches to dietary pattern analysis in nutrition epidemiology

Dietary pattern analysis has emerged as a favorable approach to characterize dietary intake (35) and understand eating behavior. The underlying concept of pattern analysis is to compress and summarize dietary variables in a meaningful representation of overall diet that can be easily analyzed and compared in contrast to individual foods (36).

In nutrition epidemiology, two approaches are usually applied to describe dietary patterns. First is called a priori approach and is based on prior knowledge of the effect of a food or food component in relation to a health outcome. In this method, the key foods, food groups or food components are transformed into a score, which is then used to investigate its association with health outcomes. Second approach is called a-posteriori and is based on the analysis of the empirical data. In this data-driven approach, multivariate structure of the data is explored to identify meaningful components, which are later used to explore its association with health outcomes.

A posteriori approach is of particular interest, as it relies on empirical data and requires fewer subjective decisions in determining the components of the identified patterns. These approaches are mostly used for exploratory analysis, which yield important information about eating patterns of the studied population.

2.1 Existing data-driven methods

Several data-driven statistical methods are used for dietary pattern analysis. Three major statistical methods used in nutritional epidemiology are i) factor analysis ii) cluster analysis, and iii) reduced rand regression (RRR).

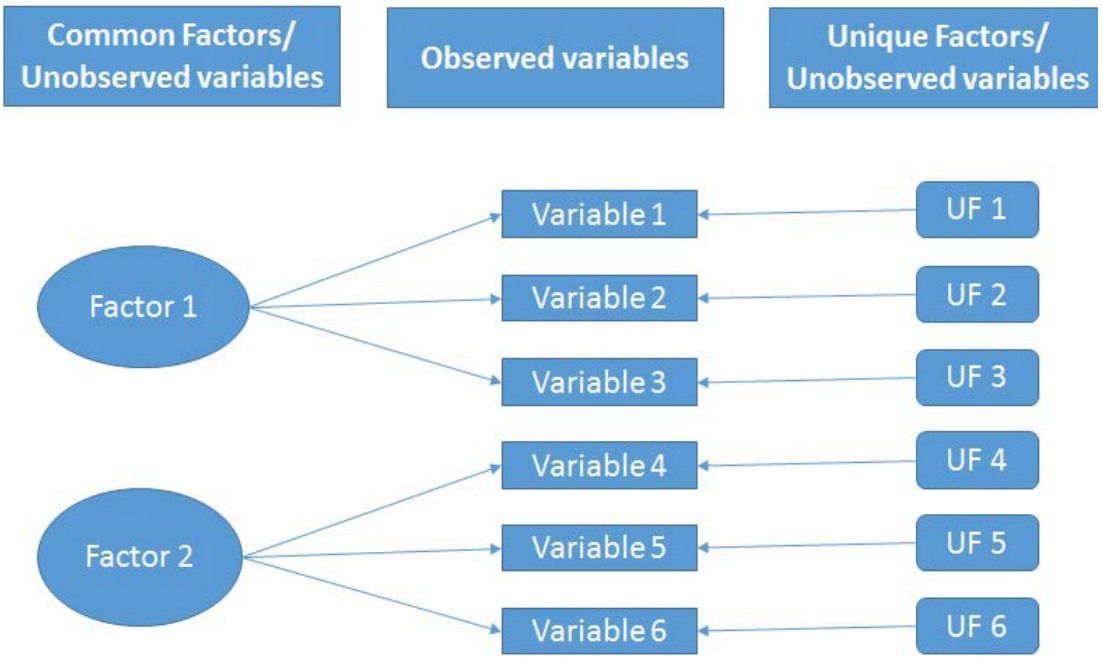
i) Factor analysis:

Factor analysis is a multivariate statistical technique that evaluates correlation or covariance matrix of the food groups under investigation to identify some traits (variables), which explain maximum variance in the dataset. Thus, the method reduces the actual number of observed variables into several variables or factors that characterize dietary intake of a population (37). Two types of factors extraction methods are commonly used:

a. Common factor analysis:

Often simply referred as factor analysis, common factor analysis includes exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis, which identifies minimum number of factors that explain the common variance (correlation) between variables, is the most commonly used factor analysis approach. This multivariate method assumes that a common underlying variable influences two or more observed variables and is responsible for the correlation between/among these variables. Thus, in effect the method explores the common variance among the variables. The common variable responsible for the observed correlation is called common factor. As the common variables are not directly measured, such variables are also referred as latent variables or underlying factors (38). In common factor analysis, the common factors are not the only variables affecting the observed variables but there also exists a unique unobserved variable for each observed variable that influence the observed variable as shown in the figure.

Figure 1: Common factor model with six variables. The common factor i.e. Factor 1 influences the first three correlated observed variables i.e. Variable 1 to Variable 3 and the second common factor i.e. Factor 2 influences the three observed variables i.e. Variable 4 to Variable 6. Each unique Factor i.e. UF1 to UF6 influences one specific observed variable i.e. Variable 1 to Variable 6, respectively. Adapted from (39).



Mathematical model of the factor analysis can be expressed using an example. Suppose there are 'n' observed variables (X), where n = 1, 2, 3, …, n, and 'm' denotes the number of latent factors (F), where m=1,2, 3, …, m. Then the observed variables X_j, where $_j$ =1,2,3, …, j, are linear combination of the latent factors F_j as expressed in Equation 1.

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + e_j \quad \text{----------} \quad \text{Equation 1}$$

In the above equation $a_{j1}, a_{j2}, \ldots, a_{jn}$, are factor loadings, where "*j*" represent the observation number i.e. 'a_{j1}' mean factor loading of j^{th} observation in the first common factor (40).

Model in equation 1 assumes that there are 'm' underlying factors (F), in which each observation is the linear combination of these factors along with its residual variate (e_j). The common factor F in the model is the shared variance and the residual variate represents the unique variance of the observed variables. In exploratory factor analysis the unique factors are assumed to be orthogonal, therefore, they do not contribute to the covariance among the variables (41). Hence, the covariance between variables is attributed to the common factors only. The value of factor loadings quantify the contribution of each observation to the common factor (42). Similar to the regression weights these factor loadings show strength of correlation between the observed variable and the latent factor (43).

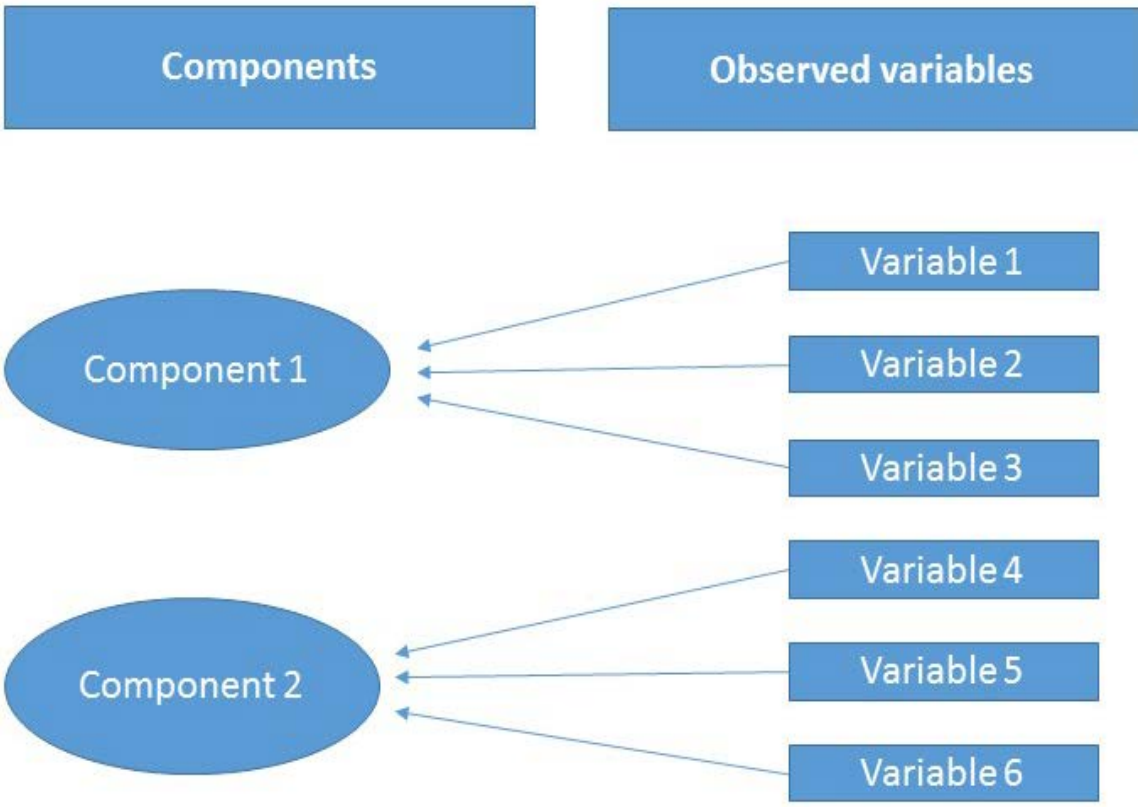
In factor analysis, the first extracted factor explains the largest proportion of shared variance in the variables. The second extracted pattern explain the largest proportion of the remaining variance and so on, unless all the of the variance is explained (44). The number of factors are retained using Kaisers criteria (45), scree plot (46) and/or interpretability depending on objectives of the study.

b. Component Analysis:

This approach use total variance in the initial extraction and is considered as a data-reduction method. Principal Component Analysis (PCA) is the most commonly applied component analysis method, which is used to reduce the number of original variables into a few representative components or patterns (47). Thus, PCA is a linear

transformation of the observed variables into a set of smaller uncorrelated variables also called patterns or components. PCA includes total variance in the initial extraction in contrast to the common factor analysis, which includes only the shared variance (48).

Figure 2: Principal component analysis with six variables. Component 1 is the linear weighted combination of the first three variables (Variable 1 to Variable 3) and component 2 is the linear weighted combination of the last three variables (Variable 4 to Variable 6). Adapted from (39).



Thus, principal components are the linear combination of the weighted observations and can be mathematically expressed using an example.

Suppose there are 'n' observed variables (X), where n = 1,2,3, …, n, and 'm' denotes the number of extracted patterns (P), where m=1,2, 3, …, m. Then the component score (P) of the participants is the linear transformation of the observed variables, which is mathematically expressed as:

$$P_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jm}X_m \quad \text{----------} \quad \text{Equation 2}$$

In the above equation a_{j1}, a_{j2}, ..., a_{jm}, are factor loadings, where "*j*" represent the observation variable i.e. 'a_{j1}' mean factor loading of j^{th} variable in the first component (40). In PCA, the first extracted factor explains the largest proportion of total variance in the variables. The second extracted pattern explains the largest proportion of the remaining variance and so on, unless all of the variance is explained (44). The number of factors are retained using Kaisers criteria (45), scree plot (46) and/or interpretability depending on objectives of the study.

ii) **Cluster Analysis:**

Cluster analyses are a set of statistical techniques that classify individuals with same characteristics into a finite number of groups or clusters based on two or more observed variables. Subjects in the same cluster share common characteristics but are statistically different from members of the other groups. Various clustering methods (logarithms) are applied for clustering of observations depending on study objectives. These methods primarily measure the statistical distance between individuals and the groups of observations. The measured distances can be considered as Euclidean distance, or another conceptualization (49).

Among the clustering approaches, Ward's cluster analysis and K-means cluster analyses are commonly used in nutrition epidemiology (50). K-means clustering is the most commonly used method that aggregates individuals in a multidimensional space using the squared Euclidean distances between observations to position individuals (51). In this method of clustering the number of clusters i.e. *K* are predetermined. The algorithm assigns the individuals to the cluster with smallest distances from its mean. Initially, the individuals are assigned to K clusters and cluster means are assessed. Then, the distances of individuals from the means are reassessed and individuals with smallest distance to a cluster means are reassigned. The process continues until there is little successive change in the clusters' means. In the last step, the individuals with the smallest distance to a particular cluster mean are permanently assigned to that cluster.

Sum of squared errors (SSE), also called scatter, is used to assess quality of clustering. Euclidean distance i.e. error of each data point (distance from the cluster mean) are summed to give a total sum of squared errors. Clusters with smallest total SSEs are preferred as an actual representation of the data points in the cluster (52).

In analysis of dietary intake data, cluster analysis aggregate individuals in several clusters based on their food intake differences. Individuals are placed in common, non-overlapping clusters based on similar dietary intake (16). The identified clusters are later used as exposure/independent variable in statistical analysis.

***iii)* Reduced Rank Regression (RRR):**

RRR or maximum redundancy analysis is a statistical method that identifies patterns, which explain the maximum variance in the response variables. This method depends on the prior knowledge about phenomenon under investigation e.g. diet-disease association. RRR combines both a priori knowledge to define response variables (related to a disease etc.) and the empirical data to derive patterns, to answer the question at hand (53).

Mathematical base of the RRR method is similar to PCA. If X_1, X_2, X_3, …, X_n, are the number of predictor variables and Y_1, Y_2, Y_3, …, Y_m, are the number of response variables, then C_j, where j=1,2,3, …, j, is linear combination of predictor variables that explains maximum variance in the response variable Y_m. The first factor in RRR explains maximum variance in response variable, in contrast to PCA, which explains maximum variance in predictor variables. In this method, initially covariance matrix of response variables is used to derive a linear function to form a response score. The response score is then projected on the predictor space to form a factor score. As the method maximize variance in the response variables, the response scores rather than factor scores are used in evaluation of factors.

iv) Other Methods:

Several other methods are also less frequently used for dietary pattern analysis. These include partial least squares (53), treelet transform (54), and Artificial Neural networks (26) etc.

2.2 Gaussian graphical models – a novel approach for dietary pattern analysis

GGM are probabilistic graphs used to analyze and visualize the dependency structures with the help of a graph that describe conditional independence among variables (55). These graphs present a set of nodes and edges, where nodes represent variables and

edges represent conditional dependency relations. A missing edge between two variables, in the dependence graph, represents conditional independence between these variables given all other variables (56). Such conditional independence in a dependence graph is called pair-wise Markov property (57) and is quantified in terms of partial correlation. Model selection in GGM results in a sparse graph that represents the underlying pattern of the associated variables.

Theoretical background:

Suppose a data matrix X with n observations and p variables from p-variate normal distribution having mean vector μ and covariance matrix ∑, which can also be expressed as N*p* (μ, ∑). Then from the inverse of this covariance matrix, which is also called precision matrix, the conditional distribution of any two random variables given other variables can be obtained e.g. P_1 and P_2, given all other variables, and the correlation coefficient in this distribution between the two variables (e.g. P_1, P_2) is called partial correlation (58). If the partial correlation between the two variables e.g. P_1 and P_2 is zero, it is inferred that the two variables i.e. P_1 and P_2 are conditionally independent given all other variables. Estimation of conditional independence in a precision matrix forms basis of GGM (59). In GGM, the conditional independence relationship between given variables is reflected in an undirected graph *G* (*V, E*), where *V* represent vertices (variables) and *E* represent edges (partial correlation among variables) of the graph *G*. From this, GGM is defined as, an undirected graph of p-variate normal distribution N*p* (μ, ∑) with conditional independence restriction i.e. two variables are independent given others, if the correlation in inverse covariance matrix between the two variables is zero, defined by the pairwise Markov Property (60).

In GGM, conditional independence among variables is determined by identifying zero entries in the inverse of the covariance matrix, known as the covariance selection problem (61) or model selection (also called structure learning) in the Gaussian concentration graph model (62). However, in a high dimensional multivariate normally distributed dataset there may be no or few zero entries in the precision matrix, which may result in a dense concentration graph, with each node connected to other nodes in the graph. Such graphs are less informative as aim of GGM is to identify topology (structure) of a graphical model, which is accurate and meaningful representation of the underlying data. Accuracy of such a model is assessed by the likelihood that the model explains the data (63). Such situations

require adoption of regularization technique that enforces sparsity in the precision matrix for data representation. While a number of methods exist (64, 65) for achieving sparsity in the precision matrix, graphical lasso (66) is an efficient and fast approach for structure learning in graphical models. It is a regularized (penalized) likelihood optimization method that puts a penalty on the off-diagonal elements of the inverse covariance matrix, shrinking the estimated values of pairwise partial correlations, which forces small or noisy values to zero and results in a sparse matrix of direct connections (67). In sum, rather than maximizing log-likelihood, graphical lasso maximizes the regularized log-likelihood to achieve sparsity. Regularization is achieved by penalizing log-likelihood by term [*Lambda* * L_1 norm], where $L1$ norm is the absolute sum of the inverse-covariance matrix and *Lambda* is a non-negative tuning shrinkage parameter. It is also called regularization parameter. The value of *Lambda* depends on the research question (level of sparsity required) and is estimated from the best model fit (log likelihood) for different values of *Lambda*. This model for continuous data assumes multivariate Gaussian distribution and the estimated sparse concentration matrix represents the graphical model that is visualized as the underlying structure or pattern in the given dataset.

2.3 Semiparametric Gaussian copula graphical models

Suppose a data matrix X with n observations and p variables from p-variate normal distribution having mean vector μ and covariance matrix ∑, which can also be expressed as Np (μ, ∑). An undirected graph in the inverse of the covariance matrix (∑) of this data matrix can be obtained using GGM. However, one of the constraints of this approach is Gaussian assumption. In dietary intake dataset, all intake variables may or may not be normally distributed and thus the assumption of multivariate normality may not be fulfilled. An alternative approach to learn the underlying pattern in the dataset is to use a SGCGM (68). In this approach the p variables i.e. p = (p_1, p_2, p_3, …, p_n) are transformed to a function of the original variable i.e. f (p) = (f p_1, f p_2, f p_3, …, f p_n), where f (p) is assumed to follow multi-variate Gaussian distribution. This results in the nonparametric extension of the normal called paranormal distribution. This model depends on function *f(j),* mean μ and the covariance matrix ∑. The conditional independence in the model is encoded in the precision matrix. A non-parametric estimate of the function *f(j)* is employed for graphical learning using graphical lasso (69).

3. Methods

3.1 Study design and study population

Study Design:

European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam is part of the ongoing multi-center EPIC cohort study conducted in 10 European countries, which was established with the aim to investigate relation of diet with lifestyle factors, cancer and other chronic diseases in the studied population (70, 71). EPIC study is coordinated by International Agency for Research on Cancer (IARC) of the World Health Organization. Study center of EPIC-Potsdam cohort is located in German Institute of Human Nutrition Potsdam-Rehbrucke.

Study population:

Target population of the EPIC-Potsdam cohort was general population, including both men and women, mainly aged 35-64 years living in Potsdam and the surrounding areas adjacent to the study institute. Thus, study region comprised of the Potsdam city and the surrounding communities.

Recruitment procedure:

For recruitment of the study participants addresses of the potential participants from general population registries were obtained. Potential participants meeting the study criteria (age) were invited by mail to the study center. In case of non-response within two weeks, participants were reminded by mail and telephone. Voluntary participation from the uninvited participants were also allowed if they met the study criteria. During recruitment period from August 1994 to September 1998 a total of 27,548 participants agreed to participate in the cohort study and underwent examination. Participation rate (response rate) in the study was 27.2%.

3.2 Ethical approval and participants consent

Ethics committee of the state of Brandenburg approved the study procedures. Before enrollment in the study, all participants were informed about the study objectives, study participation procedures including data collection methods, data protection and their

right to withdraw anytime during study. All participants provided written informed consent, at baseline, before examination.

3.3 Study variables

Following variables were used to meet objectives of the study

i. *Outcome variables (End points)*: Outcome variables included verified incident cases of T2D, MI, stroke, CVD (as a sum of MI and stroke), cancer, cardiometabolic diseases (as sum of T2D, MI and stroke), and overall chronic diseases (as sum of T2D, MI, stroke, and cancer).

ii. *Exposure Variables (dietary patterns)*: Scores of dietary patterns identified by GGM and PCA were used as exposure variables. In addition, food groups of the GGM pattern were also used as exposure variable in separate models to assess its contribution to risk of outcomes of interest. GGM identified sex-specific pattern scores included principal pattern score, high fat dairy pattern score, fruit and vegetable pattern score, sweets pattern score, and breakfast cereals pattern score. Standardized intakes of food groups of each GGM pattern was used as exposure for association of single foods with risk of major chronic diseases. PCA identified pattern scores included plain cooking pattern score, cereal pattern score, sweet pattern score, and high fat dairy score in men and plain cooking pattern score, sweet pattern score, bread and sausage pattern score, and low fat diary product pattern score in women. For association of intake network score, networks with three or more foods were considered as exposure. All networks (with 3 or more nodes) and the PCA patterns were adjusted for each other in the cox-regression models.

iii. *Covariates*: Covariates included age, education level, smoking status, physical activity level, educational attainment, BMI, total energy intake, alcohol consumption, vitamin supplementation, history of hypertension and under-reporting of energy. Analysis for association of food groups with risk of all outcomes were further adjusted for sex.

3.4 Data collection

Study procedures and data collection tools were developed according to core EPIC protocols (72). A number of tools were used to collect baseline information (Table 1).

Table 1: Baseline assessment instruments and variables (used in this study)[1]

Instruments	Exposure variables
Self-administered lifestyle questionnaires[2]	
Food frequency questionnaires	Frequency and quantity of food consumption, use of sauces and fat, regular use of supplements
Lifestyle questionnaire	Family status, education, occupational position, history of alcohol consumption, etc.
PC interviews	
PC-guided menu-controlled interactive interview	Occupation, smoking history, physical activity in winter and summer, weight history, subjective health situation, medical anamnesis, use of medication during the previous four weeks, etc.
EPIC-SOFT 24-hour recall	Food consumption within the last 24 h
Physical examinations	Anthropometry (height, weight, waist- and hip circumference, skinfold measurements), blood pressure, heart rate, blood withdrawal, bone density

*1 A*dapted from Boeing et al, (73)

2 The returned and completed questionnaires were read by optical scanning at the examination center and immediately checked for reading errors and missing information by a computer program. All unclear information were immediately clarified with the participant via a PC-guided menu-controlled interview.

3.4.1 Assessment of dietary intake

Habitual dietary intake was assessed at the time of enrollment with a validated, optical readable, self-administered, semi-quantitative food frequency questionnaire (FFQ) (74). Details of the questionnaire development, reliability, and validation are provided elsewhere (75). In brief, a 148-item FFQ was developed for the EPIC-Potsdam cohort to assess intake of the study participants. Reproducibility of the FFQ was assessed at interval of six months and it showed good reproducibility among the same participants. When compared with intake assessed through 24-hour recalls, spearman correlation between the respective methods ranged from 0.49 for legumes to 0.90 for alcohol beverages.

The FFQ queried the frequency of consumption of 148 food items over the past 12 months. Additional information regarding fat content of dairy products consumed and type of fat used for food preparation were also collected at the same time. Portion sizes of food intake were estimated using photographs and standard portion sizes (e.g. one cup, one piece, one teaspoon etc.). Frequency of food intake was assessed in ten categories ranging from 'Never' to '5 times a day or more'. Participants described their intakes in terms of the standard portion sizes e.g. one third, half, equal, double of the portion size etc. or described intakes in absolute sizes of the pictures i.e. small, medium or large servings of the shown foods/dishes. Intake of each food item was calculated from portion size and intake frequency. The questionnaire also assessed intake of nutritional supplements.

Intakes of the single foods were collapsed into 49 food groups including alcoholic beverages (Supplement S3) based on their nutrient composition using food codes from a specialized software called EPIC-SOFT. This software is designed for assessment of 24-hour recalls in all EPIC study centers for data standardization (76). The software automatically assigned similar codes to the consumed food in all centers making it easier to merge food dataset from other EPIC centers. The software also prevented outliers and missingness by implementing quality checks.

The German Nutrient Database, BLS version-III, was used to calculate energy/nutrient composition of the foods. BLS is a standard data-base of foods and dishes consumed in Federal Republic of Germany (77).

3.4.2 Anthropometric assessment

Trained staff conducted anthropometric assessment in the study center as per protocol (73). Body weight of all participants was measured with Seca 876 digital scale to the nearest 0.1 Kg, in light underwear and without wearing shoes. Similarly, body height was measured with Seca 222 Stadiometer to the nearest 1 mm. Body mass index (Quetelet Index) was calculated using the formula:

$$BMI = \text{weight (Kg)} / \text{height (Meter)}^2$$

3.4.3 Assessment of lifestyle and other covariates

Data on age, sex, smoking status and educational attainment were assessed through a self-administered questionnaire at baseline in EPIC-Potsdam cohort. Due to differences in educational qualification among participants (having old and new type of qualification) that were equivalent but not directly comparable a new variable educational attainment was used to assess educational level. Educational attainment was defined as i) currently in training/no certificate or skill ii) professional school (vocational training) iii) and college or higher education.

Physical activity was assessed using a short questionnaire. Physical activity level (PAL) over the previous year was determined and expressed in metabolic equivalent [MET] (78). Smoking status of the participants was assessed and categorized as never smoker, former smoker, current smoker, and smoking ≥ 20 cigarettes/day.

Prevalent health conditions including diagnosed T2D, MI, stroke, and cancer was assessed at baseline. The assessment also included recording of information regarding intake of any medication during the last four weeks.

Energy misreporting was assessed as using the approach of Mendez et al as reported earlier in the EPIC-Potsdam validation sub-study. Energy intake was assessed from FFQ using BLS version-III. Energy requirement was estimated as basal metabolic rate (BMR) x PAL. BMR was estimated using Mifflin equation (79).

3.4.4 Follow-up and ascertainment of cases (outcome variables)

Participants, in the EPIC-Potsdam cohort, were followed-up every three years for approximately 15 years (median 11.4 years). Follow-ups were conducted using special questionnaires that enquired about health status, modification in diet and intake of any

medication. Follow-up information were used to record incidence of chronic diseases including T2D, MI, stroke and all form of cancers. Several procedures including reminders, computer program to detect missingness, tracing of non-respondents etc. were adopted to ensure completeness of follow-up (80). Focus on follow-up of the participants resulted in high response rate, validity and completeness of the follow-up data. All incidence cases were coded according to International Classification of Diseases (ICD) codes (81). These include ICD-10 codes: E11 for T2D, I21 for myocardial infarction, I60, I61, I63, I64 for stroke, and C00-97 for cancer (except C44: non-melanoma skin cancer).

In the initial follow-up, participants were asked if they had ever been diagnosed with a certain disease. If the answer was yes, then age of the diagnosis was asked to identify prevalent or incident chronic diseases. In the later follow-ups, incident cases were identified by asking participants directly if they were diagnosed with any disease since last visit. This question was complemented by questions related to intake of any medication. Questions related to medication included names and doses of all medications consumed in the last four weeks. For incident case verification, dates of diagnosis, contact and postal address of treating physician and/or hospital and other related information were collected as well. As a further verification of health status reasons for changes of diet, if any, were also recorded in all follow-ups.

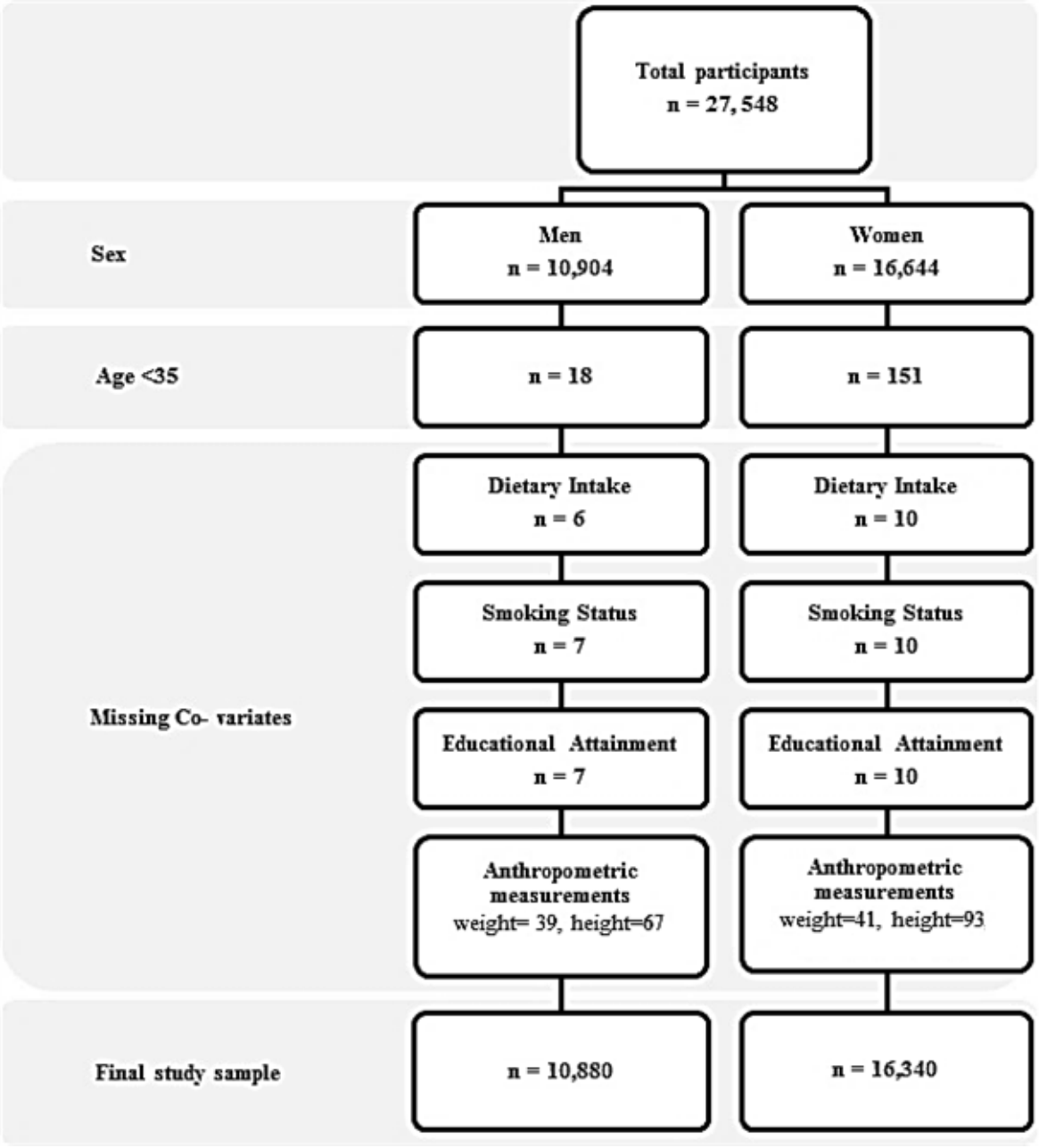
Self-reported incident cases were duly verified by either contacting the treating physician or the hospital. The physicians were asked to provide information regarding type of diagnosis, date, means of diagnosis and medication prescribed, on a standardized reporting proforma. A mix of active and passive procedures was adopted for better case ascertainment. Therefore, active follow-up by questionnaires were complemented by passive follow-ups using cancer registries, statistics of the State of Brandenburg for incidence cases, and death certificates from local health offices to identify and verify incidence cases.

In this study two end points i.e. cardiometabolic diseases and overall chronic diseases was also considered as health outcomes. The endpoint, cardiometabolic diseases was created by combining cases of T2D, MI and stroke. This outcome was created as the three endpoints i.e. T2D, MI and stroke share common risk factors assessed in this study; therefore, the combined endpoint may provide enough power to evaluate the association with the exposure. The overall chronic disease outcome variable was created by combining cases of T2D, MI, stroke and cancer.

3.5 Inclusion and exclusion criteria

In the current study, maximum sample was used to construct GGM networks. Therefore, all participants aged 35 years or over with complete data available on dietary intake (FFQ information), anthropometric measurements, physical activity level, smoking status, and educational attainment were included in the analysis (Figure 1).

Figure 3: Flow chart for selection of study participants (EPIC*-Potsdam) for dietary pattern analysis



*EPIC (European Prospective Investigation into Cancer and Nutrition)

After exclusion of the participants with missing data a total of 27,360 participants including 10,880 men and 16,340 women were included in the study to achieve specific objective 1 of the study. To achieve specific objectives 2 and 3, prevalent cases of T2D, MI, stroke and cancer were further excluded from the sample. Incidence cases were defined as T2D, MI, stroke or cancer depending on which ever was diagnosed first. Unverified incidence cases for all outcomes were retained in the dataset. Nevertheless, these were excluded from analysis of their respective outcomes. For example, unverified T2D cases were excluded from analysis when T2D was used as outcome but included in analysis when MI, stroke or cancer were used as outcomes in the model.

3.6 Statistical analysis

3.6.1 Descriptive Statistics

Mean ± standard deviations (SD) were used to describe age, BMI, PAL, energy, alcohol consumption, and report intakes of the 49 food groups. Percentages were used to report proportion of participants in categories of smoking, educational attainment, history of hypertension, vitamin and mineral supplementation. All analyses were stratified by sex. Statistical analyses were conducted using SAS Enterprise, Version 6.1 (SAS Institute Inc., Cary, NC, USA) and R Software Version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria) and yEd graph editor (Version 3.10.1, yWorks GmbH, Tübingan, Germany).

3.6.2 Dietary pattern analysis using GGM

GGM were used to derive sex-specific "networks" of dietary intake. GGM analyses were conducted in R Software as described by Højsgaard et al. (82). Gaussian assumption for GGM was visually assessed using histogram and boxplot in R. As most of the dietary variables had skewed distributions, dietary data was log-transformed (ln [g/day+1]) to improve normality. Sparse inverse covariance (precision) matrix was estimated from the log-transformed data using graphical lasso (least absolute shrinkage and selection operator) , in R package "glasso" (67). Optimum value for regularization parameter (lambda) was assessed in package "huge" by specifying a sequence of lambda values (0.60 to 0.10) in a decreasing order for sparsity(83). Sequence of the lambda values was selected in such a way that

highest value of lambda i.e. 0.60 would result in extremely sparse concentration matrix (no node connections) and the smallest lambda value i.e. 0.10 would result in a less sparse concentration matrix (very dense graph difficult to interpret). Optimum lambda value of 0.25 was selected by maximum likelihood estimate of the graphical models and used for all analysis. Estimated sparse concentration matrices were exported to yEd graph editor (84) and visualized as a dietary network separately for men and women.

Stability of the networks for the existing study sample was assessed by repeated bootstrapping 80% of the original sample with replacement. A total of 100 replications were made.

3.6.3 Dietary pattern analysis using SGCGM

To further evaluate the robustness of the results, dietary networks of the log transformed intakes were reconstructed using SGCGM (68), which did not require Gaussian distribution of the underlying data. Semi-parametric analysis was conducted using the "huge" package in R.

3.6.4 Scoring the network of dietary intake

All major networks including the principal network, dairy products network, fruit & vegetable network, sweets network and breakfast cereals network (identified only in men) were scored to evaluate adherence with the pattern and risk of major chronic diseases. In the first step, dietary intake variables (log-transformed) included in the food networks (Figure 4 and Figure 5) were standardized to the same mean and 1 standard deviation. In the second step, standardized intakes of the food groups in the same networks were combined by either adding or subtracting it from the rest depending on the direction of the correlation in their respective networks (Supplement 5.1). Sex specific scores were created using the food intake networks. Principal network was scored and named as principal pattern. Network of high fat dairy product was named high fat dairy pattern; network consisting of fresh fruit, fresh vegetables and vegetable fat/oil was named fruit & vegetables pattern; and the network comprising of sweet foods was named sweet pattern. Food intake networks in men comprised of an additional network consisting of muesli, cornflakes and vegetarian dishes, which was also scored and named breakfast cereal pattern. In summary, the sex-specific network patterns scores comprised of plain cooking, dairy, fruit & vegetables, and sweet in both men

and women. Breakfast cereal pattern score was created only for men. The network scores were categorized in quintiles. Both continuous scores and quintiles were used as exposure variables, although separately, in further analysis. All the patterns scores were adjusted for each other in the analysis.

3.6.5 Association of dietary intake networks with chronic disease risk in EPIC-Potsdam

Cox Proportional Hazard Model

Adherence to the identified patterns reflected by food intake networks was assessed using sex-specific network scores. Participants were ranked according to the food network scores and divided into quintiles (5 groups). Cox proportional hazard models (85) were used to assess relation between food network scores (continuous)/categories of food network score and risk of developing T2D, MI, stroke, CVD, cancer, cardiometabolic diseases, and chronic diseases. Hazard Ratios (HR) with 95% confidence intervals for each quintile of the food network score was estimated with lowest quintile as reference category. Age was entered as a time-scale in the cox model with age on the date of filling the questionnaire as entry in the risk set and age at date of either incidence (diagnosis) of one of the outcomes of interest, loss to follow-up (date of last questionnaire returned), death, or censoring at the 5th follow-up as exit age from the risk set. Age is an important risk factor for chronic diseases under study; therefore, to avoid violation of proportional hazard assumptions all analysis was stratified by age, using strata command in SAS. This procedure treats each year as a stratum and allowed each year having its own baseline hazard. Follow-up time or time contributed by each participant to the study was calculated from difference of age of exit and age of entry to the risk set.

In addition to the dietary patterns (network) scores, individual food/food groups of each network were also investigated for association with risk of chronic diseases. For this purpose, each GGM pattern score was replaced with its food group components, simultaneously in subsequent models. Food groups from one pattern were adjusted for other pattern scores in all models.

Two models were developed to assess risk of developing chronic diseases under

study. For all outcomes, basic model (Model 1) was adjusted for age only while multivariable model (Model 2) was adjusted for age, educational attainment, smoking status, alcohol intake, PAL, BMI, total energy intake, history of hypertension, vitamin supplementation and energy under-reporting. Energy intake were adjusted using residual method (86). Model 2 was not adjusted for history of hypertension when outcome of interest was cancer. Associations of patterns with major chronic diseases were assessed using both quintiles and per SD increase of the score.

Assessment of model Assumptions

Cox proportional hazard model is the regression of log of hazards on the variables of interest. In this model, the baseline hazard is the time varying intercept term. In addition, the covariates act multiplicatively on hazard at any given time (87). This leads to some key assumptions of the model that needs to be addressed. Therefore, four assumptions for estimation and assessment (88) for cox regression model were considered as described below:

i. *Covariates have multiplicative effect on hazard function*. This means that each unit increase in a covariate results in a constant change in hazard rate. As this is related to proportional hazard; therefore, this assumption was not assessed.

ii. *Covariates have linear effect on natural logarithm of the hazard function*. This assumption was assessed visually using Martingale's residuals and through likelihood ratio test using restricted cubic splines. For visual assessment, Martingales' residuals were assessed without the variable(s) of interest in the model and the residuals obtained were plotted against the variables. A smooth line showed linear association. Restricted cubic splines assess all variables on continuous scales. Four knots at 10^{th}, 25^{th}, 75^{th}, and 95^{th} percentiles were used for assessment. Associations were considered non-linear by visual assessment and using p-value. A p-value less than 0.05 was considered significant evidence against linearity of the association.

iii. *Proportional Hazard Assumptions:* Proportional hazard assumptions for variables of interest were assessed by adding time dependent variables to the model. Time dependent variable were defined in the model as follows:

X(*i*)Time= X(*i*) x follow-up time;

X(i) represent variables of interest for which proportionality assumption were assessed. Proportionality assumptions of the cox model were fulfilled for the assessed variables.

iv. *There are no tied events:* Tied events were handled using Efron method. This method effectively handles the tied events in the cox regression models.

Assessment of collinearity

In multivariable analysis, it is assumed that independent variables are uncorrelated. Consequently, if two or more independent variables in such a model strongly correlated the coefficient estimates can be biased (89). This leads to biased standard errors and may result in wider confidence intervals (90). Therefore, multicollinearity in the final model was assessed using tolerance (1/ variance inflation factor), and conditional index using linear regression diagnostics in SAS. Variance inflation factor (VIF) for all covariates were lower than 3 and condition index of the model was lower than 100 showing no evidence of considerable collinearity.

Effect modifications

Effect modification by smoking status (never smoker, former smoker, current smoker, and smoking ≥ 20 cigarettes/day), educational attainment (currently in training/no certificate or skill, professional school/vocational training, and college or higher education), and BMI (BMI<25, 25≤ BMI<30, and BMI≥30) were investigated by adding a combination of interaction terms in the cox model.

Test for linear trend

Tests for linear trends in risk (hazard ratios) across the score quintiles were estimated using medians of the quintile score as continuous variables in the model. Results were reported as *p- for linear trend*.

Sensitivity Analysis

Sensitivity analyses were conducted by excluding those who had experienced outcome (incident cases) in the first two years.

3.6.6 PCA derived dietary patterns and risk of chronic diseases risk in EPIC-Potsdam cohort

Schulze *et al* (91) earlier identified sex-specific dietary patterns in EPIC-Potsdam and investigated its association with lifestyle factors and nutrient intake. However, associations of these patterns with risk of major chronic disease were not investigated. Therefore, the seven patterns earlier reported were recreated in the same population using the same procedures (including the inclusion and exclusion criteria) as described by Schulze *et al*. In brief, dietary patterns were recreated using PROC FACTOR procedure in SAS. Seven factors were retained in two steps using Kaiser criterion and Scree plot. Varimax rotation was applied for interpretability. Food groups with factor loadings greater than 0.20 were considered to contribute significantly to the pattern. Patterns were named according to characteristics of the food groups. Participants were ranked according to the factor scores and categorized into quintiles for further analysis.

For association with health outcomes, five patterns including plain cooking pattern, cereal pattern, sweet pattern, fruit and vegetable pattern, and high fat pattern in men, and plain cooking pattern, sweet pattern, bread & sausage pattern, fruit & vegetable pattern, and low fat dairy pattern in women, were retained. Two of the seven patterns were not included in the analysis due to lack of their interpretability. They also explained less variance as compared to the patterns included in the analysis.

Cox proportional hazard models were used to assess risk of chronic diseases both as per SD increase in score and in the categories of the pattern scores as described in section 2.6.5.

4. Results

4.1 Application of GGM to derive dietary intake networks

4.1.1 Characteristics of the study sample

Table 2 shows the baseline characteristics of the study participants. Men were on average older and had higher alcohol consumption than women. Moreover, men tended to have a higher level of education and were more frequently smokers as compared to women.

Table 2: Baseline characteristics of the EPIC-Potsdam cohort participants included in the study

Characteristics	Men (n=339)	Women (n=325)
Age (years)	51.3±7.6	48.4±8.9
Total energy intake (Kcal)	2460.8±690.1	1914.8±566.9
Alcohol Consumption (g/day)	22.8±22.4	8.5±10.9
Physical Activity Level (Met)	1.5±0.3	1.5±0.2
BMI (Kg/m^2)	26.8±3.5	25.6±4.5
Educational Attainment (%)		
No vocational training	33.1	40.39
Technical college	16.4	29.89
University	50.6	29.72
Smoking Status (%)		
Never- smoker	32.43	59.31
Former smoker	41.95	22.54
Current smoker	15.54	14.96
Smoker >=20 units/day	10.07	3.18
History of Hypertension		
Yes	44.9	60.3
No	55.1	39.7
Vitamin Supplement		
No	85.7	81.9
Yes	14.3	18.1

* Continuous variables are presented as mean ± standard deviation and categorical variables as percentage. n = sample size, BMI = body mass index, Kg= kilogram

Mean intakes of the food groups estimated from the FFQ in this population (both men and women) are shown in Table 3. Both men and women had higher intakes of refined (other) bread as compared to whole grain bread in contrast to recommendation of German Nutrition Society for preference of whole grain products (92). Similarly, the average intakes of both fruit and vegetables, in both sexes, were considerably lower as compared to DGE recommendations. On the other hand, both sexes had considerably higher proportion of red and processed meat intakes. Consumption of red and processed meat intake was higher in men as compared to women. Intakes of alcohol beverages revealed that on average men consumed higher quantities of beer against recommendation of moderate intake.

Table 3: Dietary intake of 49 food groups used to derive dietary networks among men and women of EPIC-Potsdam cohort, included in the study.

Food Groups[1] (g/d)	**Men[2] (n=10,780)**	**Women[2] (*n*=16,340)**
Whole grain bread	40.9 ± 57.3	48 ± 52.4
Refined (other) bread	167 ± 88.0	106 ± 63
Grain flakes, grains, muesli	4.8 ± 15.4	5.9 ± 14.4
Cornflakes, crisps	1.4 ± 5.7	1.9 ± 6.7
Pasta, rice	16.5 ± 15.0	15.9 ± 14.4
Vegetarian dishes	1 ± 4.3	1.4 ± 5.7
Chips	2.6 ± 6.6	2 ± 5.0
Pizza	7.3 ± 11.0	6.8 ± 8.9
Cake, cookies	68.3 ± 71.8	59 ± 61.8
Confectionary	23.8 ± 27.6	20.9 ± 26.1
Sweet bread spreads	12.5 ± 13.7	11.2 ± 12.1
Eggs	19.4 ± 17.5	16.1 ± 14.6
Fresh fruit	122 ± 89.0	154 ± 99.0
Canned fruit	19.5 ± 26.0	17 ± 23.9
Raw vegetables	47.9 ± 39.8	61.7 ± 47.1
Cabbage	13.5 ± 13.9	14.2 ± 13.5
Cooked vegetables	27.5 ± 17.5	30.1 ± 18.6
Garlic	0.1 ± 0.4	0.1 ± 0.5
Mushrooms	2 ± 2.4	2 ± 2.4
Legumes	29.3 ± 24.0	19.3 ± 16.0
Potatoes	95.5 ± 52.2	75.2 ± 44.9
Fried potatoes	18.8 ± 17.3	10.8 ± 10.2
Nuts	3.7 ± 8.3	2.9 ± 7.9

Food Groups[1] (g/d)	Men[2] (n=10,780)	Women[2] (*n*=16,340)
Low-fat dairy products	83.1 ± 175	111 ± 194
High-fat dairy products	98.1 ± 170	101 ± 154
Low-fat cheese	6.2 ± 15.5	6.9 ± 14.3
High-fat cheese	30.6 ± 28.1	26.3 ± 23.6
Water	366 ± 404	470 ± 455
Coffee	440 ± 347	406 ± 297
De-caffeinated coffee	27.1 ± 120	31.1 ± 121
Tea	226 ± 328	294 ± 385
Fruit juice	186 ± 229	200 ± 223
Low-energy soft drinks	14.6 ± 92.4	8.5 ± 56.7
High-energy soft drinks	70 ± 180	27.4 ± 106
Beer	393 ± 525	46.3 ± 126
Wine	49.2 ± 102	51.9 ± 86.8
Spirits	5.1 ± 13.6	1.1 ± 5.1
Other alcoholic beverages	11.2 ± 21.3	14.4 ± 38.4
Butter	10.2 ± 14.6	7.7 ± 10.9
Margarine	18 ± 16.9	14.1 ± 13.5
Vegetable fat	3 ± 3.1	3.6 ± 3.5
Animal fats	0.3 ± 0.8	0.2 ± 0.6
Sauce	13.3 ± 12.4	11.2 ± 10.9
Desserts	16.7 ± 22.0	15.4 ± 22.6
Fish	28.2 ± 30.7	21.2 ± 21.9
Poultry	15 ± 14.3	11.4 ± 11.2
Meat	54 ± 35.6	34.4 ± 23.0
Processed meat	78.6 ± 54.8	48.1 ± 34.5
Soup	45.1 ± 41.9	38 ± 35.5

[1]Listed are 49 food groups derived from 178 item FFQ by combining foods having similarity in nutrient composition

[2]Presented are mean values ± SD

4.1.2 Dietary pattern analysis using GGM

GGM analysis identified one major dietary network that was termed "principal network" and several smaller networks consisting of similar food groups in men and women (Figure 4 and Figure 5). In men, the principal dietary network consisted of intake of 12 food groups that grouped around red meat and cooked vegetables. Intake of red meat was highly correlated with the intakes of poultry, processed meat, sauce and potatoes, while intake of cooked vegetables was highly correlated with intake of mushrooms and cabbage. The network revealed that intakes of processed meat and poultry were conditionally dependent on red meat intake while intakes of legumes and mushrooms were conditionally dependent on intake of cooked vegetables in the identified pattern. Besides, there was a strong negative correlation between intakes of whole grain bread and refined bread.

Other important networks identified in men, consisted of dairy products defined by fat content, sweet foods, fruit & vegetables, and breakfast cereals. In the network of dairy product defined by fat content there was a strong inverse correlation between intakes of high and low fat food groups among men. On the other hand, in the network defined by intake of sweet food groups, all food groups were positively correlated with each other. In the same network, intakes of desserts as well as cakes and cookies were correlated with intake of all other sweet foods. The fruit & vegetables intake network showed that intake of fresh fruit and vegetable fats were conditionally dependent on intake of raw vegetables. The breakfast cereal network had vegetarian dishes as one of the food group; however, the stability of the connection of vegetarian dishes was less stable as compared to other edges in all networks in men (78%).

In women, the principal network consisted of the same food groups as identified in men, with addition of fried potatoes as shown in Figure 5. Similar to the principal network identified in men, the principal network in women revealed a central role of red meat and cooked vegetables intakes but showed more conditional dependencies between intakes of food groups compared to the network in men. In addition, legumes and potatoes were also central to the intake network.

Like in men, other important networks identified in women, consisted of intakes of dairy products defined by fat content, sweet foods, and fruit & vegetables. However, unlike in men, the network of dairy products defined by fat content additionally included butter and

margarine in women. While in the network comprising sweet foods, the intake of cakes and cookies was connected to the intakes of all other food groups.

A major difference between patterns in men and women was the intake relationship between food groups in the identified networks. For example, intake of red meat was related with intakes of five food groups in men but seven food groups in women.

Figure 4: Dietary intake networks for men of EPIC Potsdam cohort, included in the study, derived by Gaussian Graphical Models. Vertices represent foods/food groups and edges represent conditional dependencies (reflected by partial correlation coefficients) among foods/food groups. Continuous black lines show positive and red broken lines show negative partial correlations. Thicknesses of the edges are proportional to the strength of the correlation among connected food groups. Absence of an edge between two foods/food groups represents conditional independence between them in the network (n=10,780).

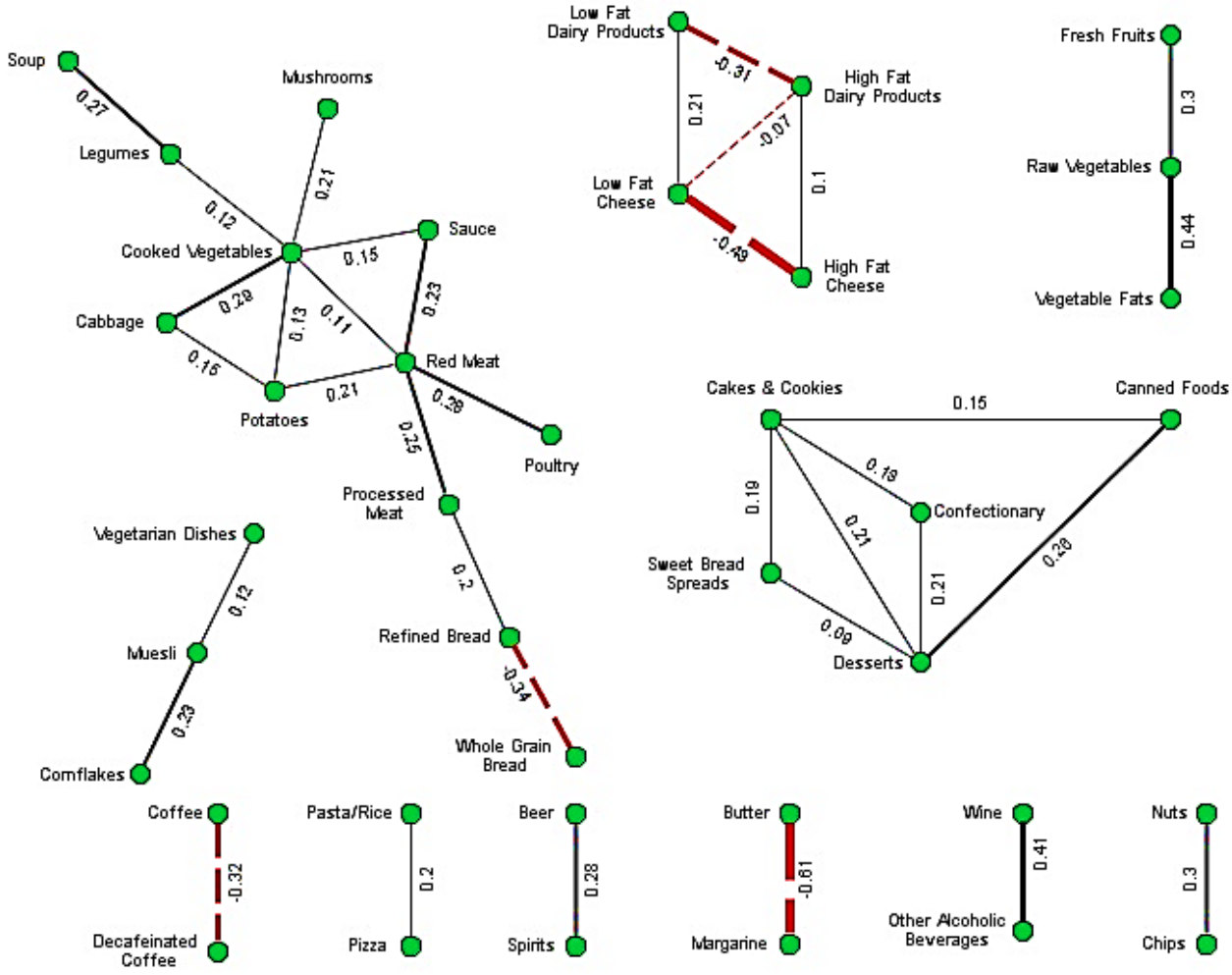
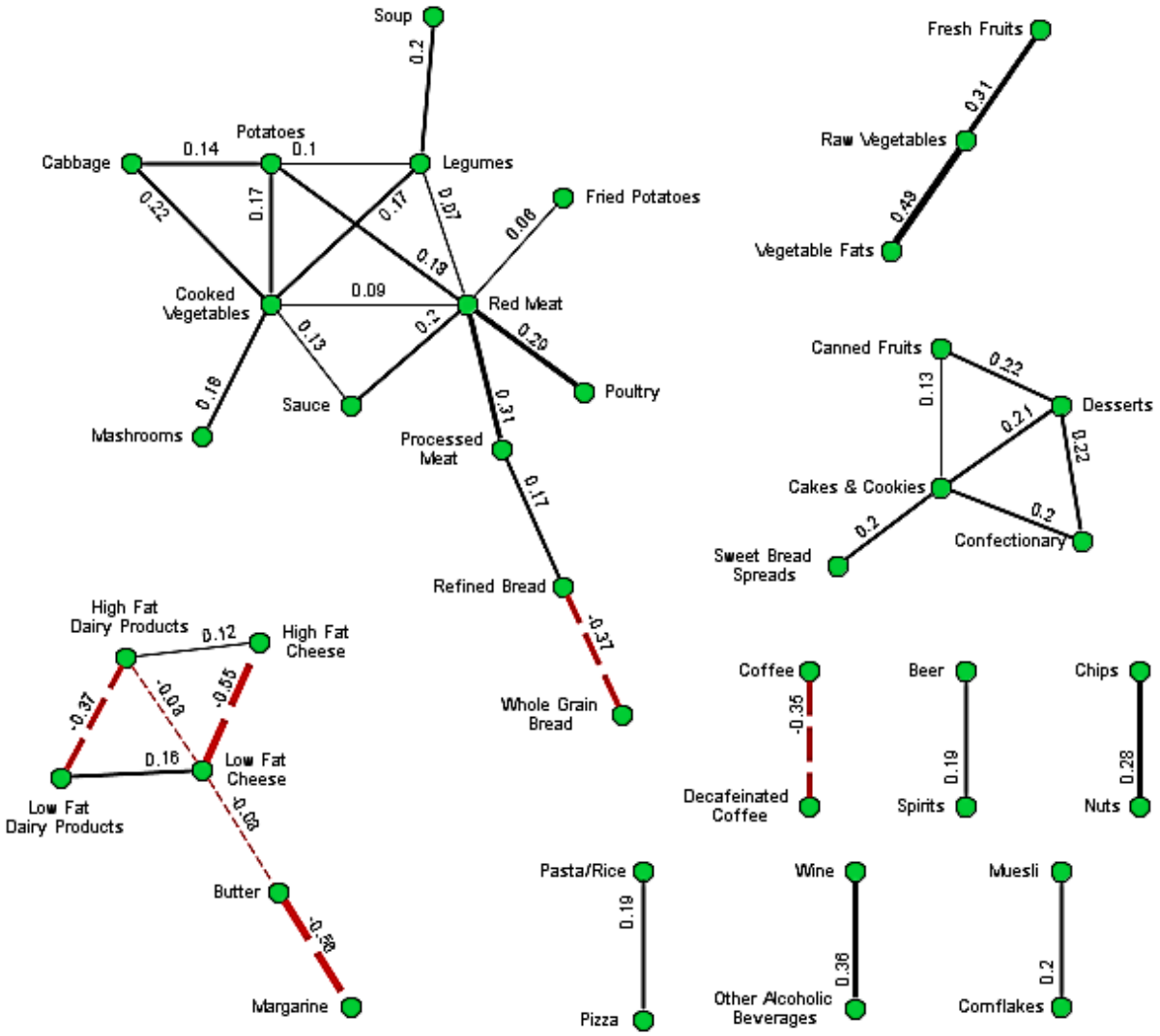


Figure 5: Dietary intake networks for women of EPIC-Potsdam cohort, included in the study, derived by Gaussian Graphical Models. Vertices represent foods/food groups and edges represent conditional dependencies (reflected by partial correlation coefficients) among foods/food groups. Continuous black lines show positive and broken red lines show negative partial correlations. Thicknesses of the edges are proportional to the strength of the correlation among connected food groups. Absence of an edge between two foods/food groups represents conditional independence between them in the network (n=16,340).



Stability analysis by bootstrap sampling revealed that the identified networks were stable in the current population. No structural variations in major networks were observed in both men and women when bootstrap sampling was applied (Tables 4 & 5).

Table 4: Reproducibility (in percent) of edges in the Gaussian Graphical Models identified dietary intake networks among men (n=10,880)

S.No.	Node 1	Node 2	Reproducibility (%)
1	Whole grain bread	Refined bread	100
2	Refined bread	Processed meat	100
3	Muesli	Corn flakes	100
4	Muesli	Vegetarian dishes	78
5	Pasta & rice	Pizza	100
6	Chips	Nuts	100
7	Cakes & cookies	Confectionary	100
8	Cakes & cookies	Sweet bread spreads	100
9	Cakes & cookies	Canned fruit	100
10	Cakes & cookies	Desserts	100
11	Confectionary	Desserts	100
12	Sweet bread spreads	Desserts	98
13	Fresh fruit	Raw vegetables	100
14	Canned fruit	Desserts	100
15	Raw vegetables	Vegetable fats	100
16	Cabbage	Cooked vegetables	100
17	Cabbage	Potatoes	100
18	Cooked vegetables	Mushrooms	100
19	Cooked vegetables	Legumes	100
20	Cooked vegetables	Potatoes	100
21	Cooked vegetables	Sauce	100
22	Cooked vegetables	Red meat	100
23	Legumes	Soup	100
24	Potatoes	Red meat	100
25	Low fat dairy products	High fat dairy products	92
26	Low fat dairy products	Low fat cheese	100
27	High fat dairy products	Low fat cheese	100
28	High fat dairy products	High fat cheese	100
29	Low fat cheese	High fat cheese	100
30	Coffee	Decaffeinated coffee	100
31	Beer	Spirits	100
32	Wine	Other alcoholic beverages	100
33	Butter	Margarine	98
34	Sauce	Red meat	100
35	Poultry	Red meat	100
36	Red meat	Processed meat	100

Table 5: Reproducibility (in percent) of edges in the Gaussian Graphical Models identified dietary intake networks among women (n=16,340)

S.No.	Node 1	Node 2	Reproducibility (%)
1	Whole grain bread	Refined bread	100
2	Refined bread	Processed meat	100
3	Muesli	Corn flakes	100
4	Muesli	Vegetarian dishes	78
5	Pasta & rice	Pizza	100
6	Chips	Nuts	100
7	Cakes & cookies	Confectionary	100
8	Cakes & cookies	Sweet bread spreads	100
9	Cakes & cookies	Canned fruit	100
10	Cakes & cookies	Desserts	100
11	Confectionary	Desserts	100
12	Sweet bread spreads	Desserts	98
13	Fresh fruit	Raw vegetables	100
14	Canned fruit	Desserts	100
15	Raw vegetables	Vegetable fats	100
16	Cabbage	Cooked vegetables	100
17	Cabbage	Potatoes	100
18	Cooked vegetables	Mushrooms	100
19	Cooked vegetables	Legumes	100
20	Cooked vegetables	Potatoes	100
21	Cooked vegetables	Sauce	100
22	Cooked vegetables	Red meat	100
23	Legumes	Soup	100
24	Potatoes	Red meat	100
25	Low fat dairy products	High fat dairy products	92
26	Low fat dairy products	Low fat cheese	100
27	High fat dairy products	Low fat cheese	100
28	High fat dairy products	High fat cheese	100
29	Low fat cheese	High fat cheese	100
30*	Butter	High fat dairy products	68
31	Coffee	Decoffee	100
32	Beer	Spirits	100
33	Wine	Other alcoholic beverages	100
34*	Wine	Beer	65
35	Butter	Margarine	98
36	Sauce	Red meat	100
37	Poultry	Red meat	100
38	Red meat	Processed meat	100

*Pair of nodes not observed in the full (final) network

4.1.3 Dietary pattern analysis using SGCGM

Dietary networks derived through SGCGM, to assess robustness of the normality assumption of GGM, showed a strong resemblance to the GGM derived networks (Figure 6 and Figure 7). The SGCGM identified principal networks for both sexes had an additional dependence relationship between whole grain bread and muesli. Moreover, the sweet food network and high fat dairy network formed a single network, in men, though exhibiting similarity in conditional independence relationship to that of networks identified by GGM. Overall, the derived principal and smaller networks were similar compared to GGM networks and comprised of similar food groups for both men and women. In both the sexes, red meat, cooked vegetable and potatoes were central to the intake in the principal networks.

Figure 6: Dietary intake networks for men of EPIC-Potsdam cohort, included in the study, derived by Semiparametric Gaussian Copula Graphical Models. Vertices represent foods/food groups and edges represent conditional dependencies (reflected by partial correlation coefficients) among foods/food groups. Absence of an edge between two foods/food groups represents conditional independence between them in the networks (n=10,880).

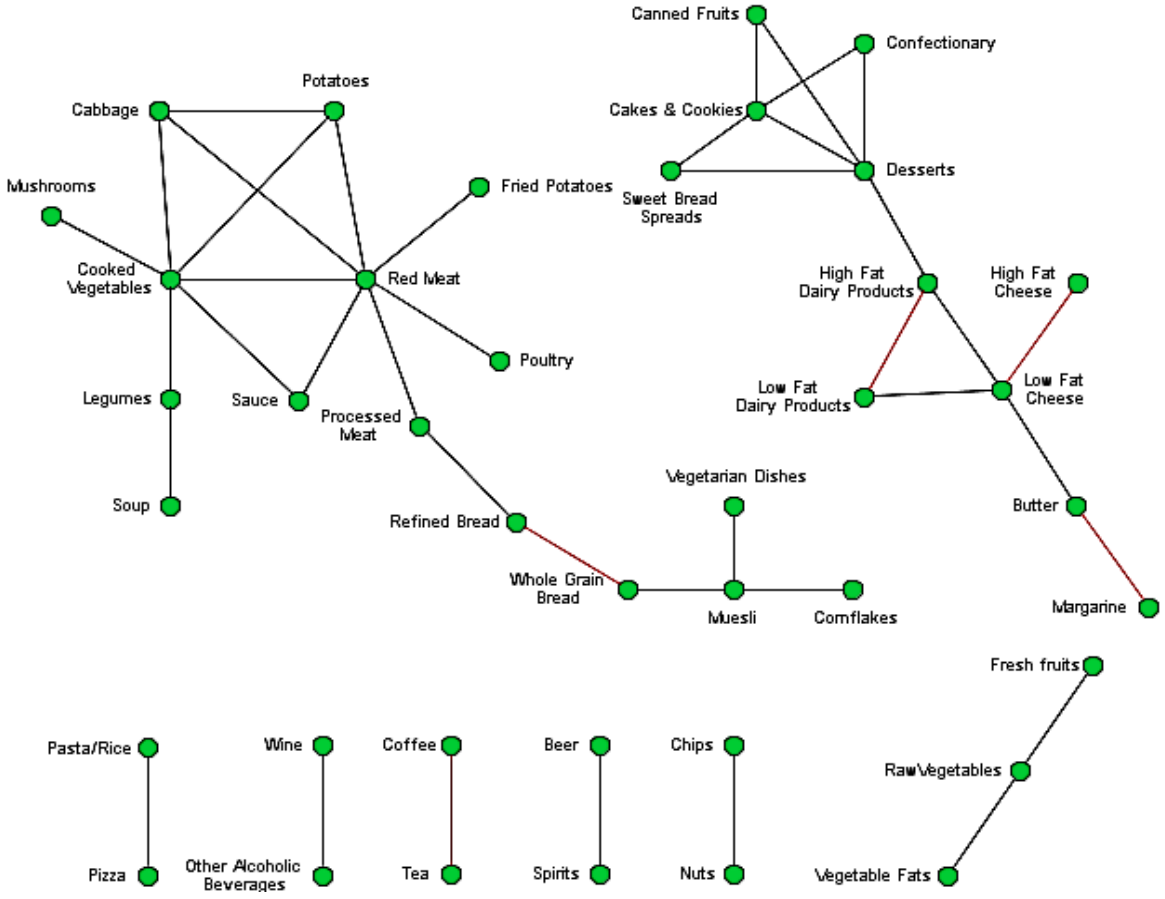
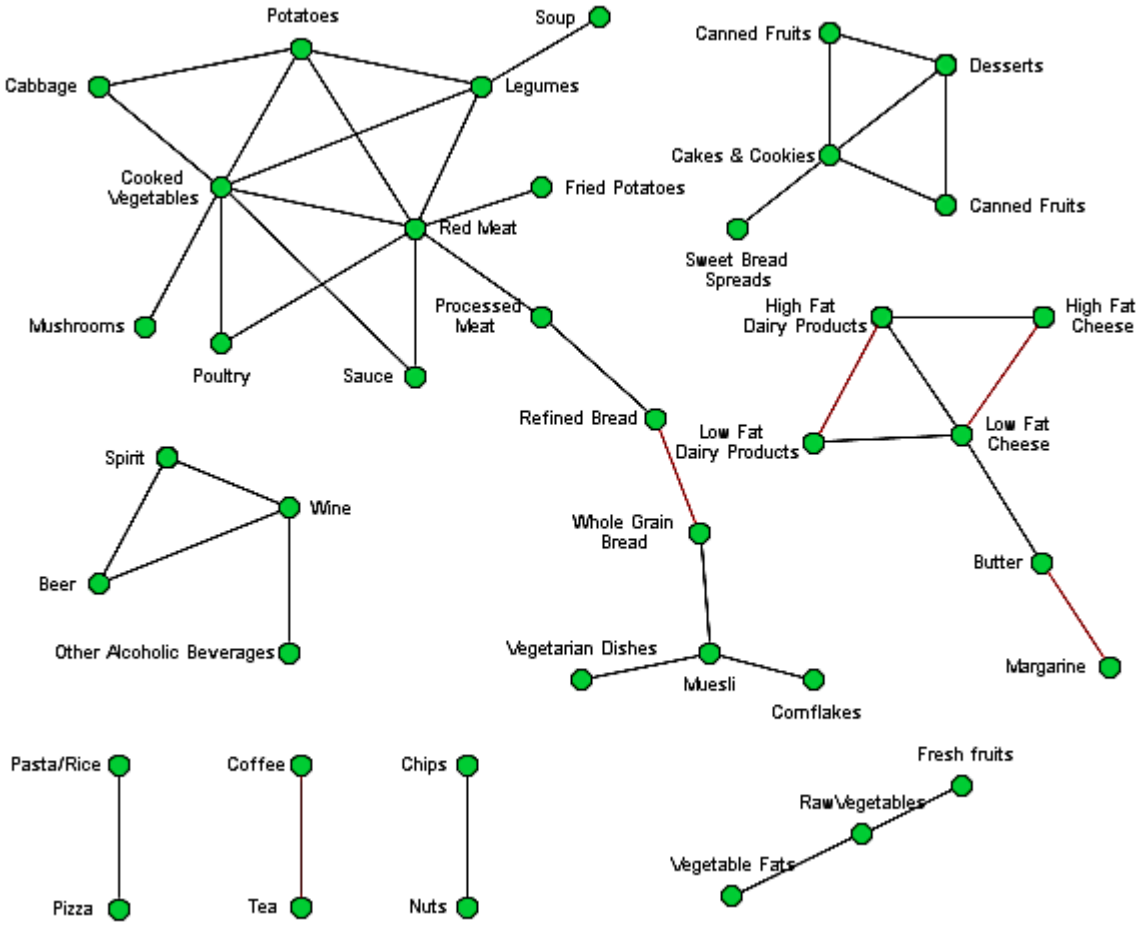


Figure 7: Dietary intake networks for women of EPIC-Potsdam cohort, included in the study, derived by Semiparametric Gaussian Copula Graphical Models. Vertices represent foods/food groups and edges represent conditional dependencies (reflected by partial correlation coefficients) among foods/food groups. Absence of an edge between two foods/food groups represents conditional independence between them in the networks (n=16,340).



4.1.4 Dietary pattern analysis using PCA

Seven sex-specific PCA patterns were identified by Shulze *et. al.* and were reconstructed in this analysis for comparison. The seven dietary patterns explained 30.3% of total variance in both sexes. For comparison with GGM identified dietary patterns (food networks), only five patterns were retained for both sexes.

In men (Table 6), the first PCA component loaded high (factor loadings > 0.20) on meat, cooked vegetables, sauce, potatoes, cabbage, poultry, legumes, mushrooms, animal fat, and soup and it was named *Plain Cooking pattern*. The second component loaded high on cereal based food groups and was named *Cereals pattern*. The third component loaded high on sweet foods and was named *Sweet pattern*. The fourth pattern loaded high on fruit and raw vegetables and was named *Fruit & Vegetables pattern*. The fifth component loaded high on high fat dairy products and was named *High Fat Dairy pattern*.

In women (Table 7), the first PCA component loaded high on foods similar to that of *Plain Cooking pattern* and was given the same name. The second pattern loaded high on sweet foods and was named *Sweet pattern*. The third pattern loaded high on bread and sausage and named *Bread & Sausage pattern*. The fourth component loaded high on raw vegetables, fruit, vegetable fat, and tea. This component was named *Fruit & Vegetables pattern*. The fifth pattern loaded high on low fat dairy product and was named *Low Fat Dairy pattern*.

Table 6: Factor-loading matrix of the principal component analysis identified dietary intake patterns for men in the European Prospective Investigation into Cancer and Nutrition-Potsdam study (n =8679)

Good Groups	*Plain Cooking*	*Cereals*	*Sweet*	*Fruit & Vegetables*	*High Fat Dairy*
Meat	**0.73**	-0.08	-0.03	-0.09	0.13
Cooked vegetables	**0.68**	0.11	0.03	0.16	-0.02
Sauce	**0.60**	-0.04	0.08	0.05	0.11
Potatoes	**0.58**	-0.22	0.07	-0.06	-0.05
Cabbage	**0.52**	0.01	0.04	0.15	0.03
Poultry	**0.50**	0.10	-0.07	0.05	-0.03
Legumes	**0.41**	-0.10	0.04	-0.01	-0.21
Mushrooms	**0.38**	0.09	-0.07	0.14	0.01
Animal fat	**0.22**	0.03	0.05	-0.07	0.01
Pasta, rice	0.20	**0.55**	0.02	-0.05	0.06
Pizza	0.03	**0.53**	0.00	-0.14	0.16
Vegetarian dishes	0.04	**0.45**	0.04	0.11	-0.04
Grain flakes, grains, muesli	-0.07	**0.42**	0.15	0.11	-0.02
Whole grain bread	-0.04	**0.38**	0.04	0.27	-0.20
Wine	-0.01	**0.36**	-0.14	0.07	0.01
Cornflakes, crisps	-0.05	**0.26**	0.11	0.01	0.03
Garlic	0.03	**0.22**	-0.08	0.13	-0.07
Refined bread	-0.04	-0.45	0.18	0.03	**0.36**
Cake, cookies	0.08	-0.02	**0.60**	-0.03	0.01
Deserts	0.11	0.11	**0.60**	-0.07	-0.04
Canned fruit	0.19	-0.13	**0.50**	0.07	-0.19
Sweet bread spreads	-0.03	-0.02	**0.48**	0.03	-0.01
Confectionary	0.01	0.02	**0.46**	-0.03	0.17
High-fat dairy products	-0.03	0.17	**0.31**	0.02	**0.27**
Tea	-0.04	0.21	**0.25**	0.17	0.05
Fruit juice	-0.03	0.00	**0.23**	0.15	-0.06
Spirits	0.07	-0.14	-0.28	0.03	0.00
Beer	0.11	-0.25	-0.41	-0.06	0.07
Raw vegetables	0.10	0.05	-0.01	**0.80**	0.01
Vegetable fat	0.16	0.20	-0.04	**0.70**	0.01
Fresh fruit	0.00	-0.03	0.18	**0.63**	0.00
Water	-0.02	0.17	-0.09	0.24	-0.10
Eggs	0.10	-0.12	0.01	0.00	0.14
Chips	-0.04	0.06	0.05	-0.03	0.01
Nuts	-0.03	0.10	0.08	0.07	-0.04
Other alcoholic beverages	0.03	0.14	-0.02	0.00	-0.06
Fried potatoes	**0.28**	0.01	0.05	-0.13	0.03
Fish	0.09	0.01	0.06	**0.24**	-0.10
Soup	**0.24**	-0.01	0.18	0.06	-0.22
Coffee	0.07	-0.15	-0.01	-0.07	0.12

Good Groups	*Plain Cooking*	*Cereals*	*Sweet*	*Fruit & Vegetables*	*High Fat Dairy*
High-fat cheese	0.03	0.05	0.07	0.11	**0.62**
Decaffeinated coffee	0.05	0.00	0.08	-0.03	-0.10
Low-fat dairy products	-0.03	0.04	0.11	0.08	-0.44
Low-fat cheese	-0.05	-0.01	0.01	0.16	-0.59
Margarine	0.01	-0.08	0.05	0.02	-0.07
Processed meat	0.10	-0.31	0.05	0.11	**0.22**
Butter	0.05	-0.20	0.15	0.02	**0.36**

Table 7: Factor-loading matrix of the principal component analysis identified dietary intake patterns for women in the European Prospective Investigation into Cancer and Nutrition-Potsdam study (n = 13,373) *

Good Groups	*Plain Cooking*	*Sweet*	*Bread & Sausage*	*Fruit & Vegetables*	*Low fat dairy*
Cooked vegetables	**0.65**	-0.03	-0.16	0.18	-0.02
Meat	**0.64**	0.02	**0.33**	-0.07	-0.06
Potatoes	**0.61**	0.05	0.13	-0.10	-0.02
Legumes	**0.53**	0.07	-0.02	-0.03	0.04
Cabbage	**0.49**	-0.02	-0.08	0.09	-0.04
Sauce	**0.48**	0.12	**0.22**	0.05	-0.11
Poultry	**0.47**	-0.03	0.07	0.08	0.13
Mushrooms	**0.35**	-0.12	-0.06	0.20	-0.05
Soup	**0.33**	0.23	-0.16	0.03	0.05
Fried potatoes	**0.30**	0.15	0.11	-0.18	-0.06
Animal fat	**0.21**	0.04	-0.01	-0.04	0.03
Cake, cookies	0.07	**0.63**	0.08	-0.02	-0.02
Deserts	0.10	**0.60**	-0.05	-0.09	0.01
Confectionary	-0.02	**0.59**	0.06	-0.02	-0.08
Canned fruit	0.24	**0.48**	0.07	0.02	0.10
Sweet bread spreads	0.01	**0.46**	-0.03	0.04	-0.12
Fruit juice	-0.01	**0.26**	0.08	0.12	-0.04
Cornflakes, crisps	-0.09	0.18	-0.14	0.01	0.06
Other bread	-0.01	0.15	**0.61**	0.02	-0.23
Processed meat	0.11	0.14	**0.55**	0.12	0.01
Margarine	-0.09	0.09	**0.51**	0.07	**0.40**
Garlic	0.08	-0.01	-0.27	0.13	-0.02
Whole grain bread	-0.04	0.03	-0.40	0.23	0.15
Vegetarian dishes	0.04	0.00	-0.40	0.11	-0.11
Grain flakes, grains, muesli	-0.10	0.12	-0.42	0.11	0.06
Raw vegetables	0.03	0.00	-0.03	**0.83**	0.03
Vegetable fat	0.10	-0.05	-0.13	**0.75**	-0.02
Fresh fruit	0.01	0.15	-0.03	**0.62**	0.06

Good Groups	*Plain Cooking*	*Sweet*	*Bread & Sausage*	*Fruit & Vegetables*	*Low fat dairy*
Tea	0.01	0.20	-0.23	**0.24**	0.03
Low-fat cheese	0.04	-0.01	-0.14	0.10	**0.60**
Low-fat dairy products	-0.03	0.13	-0.10	0.12	**0.49**
Water	0.04	-0.16	-0.22	0.23	**0.24**
Decaffeinated coffee	0.05	0.06	-0.02	-0.04	0.18
High-fat dairy products	-0.01	0.23	-0.09	0.04	-0.36
High-fat cheese	-0.05	0.11	**0.25**	0.20	-0.48
Butter	0.13	0.17	-0.13	-0.05	-0.62
Other alcoholic beverages	0.00	0.06	0.00	0.01	0.06
Wine	-0.02	-0.12	-0.19	0.09	-0.04
Spirits	0.11	-0.12	0.03	-0.02	0.00
Beer	0.07	-0.14	0.04	-0.05	-0.12
Chips	-0.06	0.21	0.08	0.00	-0.02
Eggs	0.08	0.15	0.13	-0.03	-0.04
Coffee	0.05	-0.01	0.18	-0.02	-0.02
Fish	0.21	0.18	-0.08	0.19	0.16
Nuts	-0.03	**0.22**	-0.12	0.10	-0.03
Pizza	-0.01	0.03	-0.13	-0.06	-0.10
Pasta, rice	0.19	0.03	-0.23	0.01	-0.06

4.2 Association of dietary intake patterns with risk of major chronic diseases

4.2.1 Association of pattern scores with background characteristics

Associations of the GGM patterns with background characteristics for men are shown in Table 8. Most noticeably, participants with higher principal pattern score were more likely to be current smokers and more likely to have higher BMI, higher alcohol and energy intakes, lower educational attainment and lower physical activity level. Participants with higher score of high fat dairy pattern were more likely to be current smokers and more likely to have higher energy intakes, lower BMI and lower physical activity levels. Participants with higher fruit & vegetable pattern were less likely to be current smokers, alcohol consumers, and more likely to have higher educational attainment. Participants with higher sweet pattern were less likely to be current smokers, alcohol consumers, and physically active and more likely to have higher educational attainment. Participants with higher score of cereal pattern were less likely to smoke and more likely to have higher educational attainment and physically activity.

Table 8: Association (regression coefficient with 95% confidence intervals*) of Gaussian Graphical Models identified pattern with background characteristics in men (n=8,679)

Characteristics (Unit)	Principal	High fat dairy	Fruit & vegetables	Sweet	Breakfast Cereals
Age (Years)	**0.08** *(0.07,0.1)*	**-0.03** *(-0.04, -0.03)*	0.01 *(0,0.01)*	**0.05** *(0.04,0.05)*	**-0.03** *(-0.04, -0.03)*
Current smoking**	**0.54** *(0.29,0.79)*	**0.2** *(0.07,0.33)*	**-0.69** *(-0.81, -0.58)*	**-0.68** *(-0.82, -0.55)*	**-0.45** *(-0.54, -0.36)*
Education**	**-0.64** *(-0.88, -0.4)*	**0.16** *(0.04,0.28)*	**0.5** *(0.4,0.61)*	**0.46** *(0.33,0.59)*	**0.54** *(0.45,0.63)*
BMI (Kg/meter^2)	**0.12** *(0.09,0.15)*	**-0.08** *(-0.1, -0.07)*	**0.02** *(0.01,0.04)*	**-0.08** *(-0.09, -0.06)*	**-0.05** *(-0.07, -0.04)*
Physical activity (MET)	**-1.12** *(-1.82, -0.42)*	**-0.53** *(-0.88, -0.17)*	**0.46** *(0.14,0.77)*	**-0.49** *(-0.88, -0.1)*	-0.12 *(-0.38,0.13)*
Alcohol intake (g/day)	**0.18** *(0.08,0.29)*	0.04 (-0.02,0.09)	**-0.12** (-0.17, -0.07)	**-0.87** (-0.93, -0.81)	**-0.2** (-0.24, -0.17)
Energy (Kcal)	**8.52** (8.11,8.92)	**2** (1.79,2.2)	**1.96** (1.78,2.14)	**6.91** (6.69,7.14)	**0.51** (0.36,0.66)
Vitamin Supplementation**	**-0.99** (-1.3, -0.69)	**-0.23** (-0.38, -0.07)	**0.27** (0.13,0.41)	**-0.05** (-0.22,0.12)	**0.53** (0.42,0.64)

*Parameter estimates of multiple linear regression analysis with 'network score (pattern score)' as dependent variable and age, smoking status, BMI, physical activity level, education, alcohol intake, energy intake and supplementation intake during the last one month). Important associations are shown in bold.

**Educational attainment was defined as i) currently in training/no certificate or skill ii) professional school (vocational training) and college or higher education. Smoking status of the participants was assessed and categorized as non-smoker vs smoker. Vitamin Supplementation was assessed as yes/no.

Associations of the GGM patterns with background characteristics for women are shown in Table 9. Most noticeably, participants with higher principal pattern score were more likely to be current smokers, more likely to have higher BMI, higher alcohol and energy intakes, lower educational attainment and lower physical activity level. Participants with higher score of high fat dairy pattern were more likely to be current smokers, more likely to have higher energy intakes, lower BMI and lower physical activity levels. Participants with higher fruit & vegetable pattern were less likely to be current smokers and consume alcohol; participants with sweet intake were less likely to be current smokers and consume alcohol; and participants with higher score of cereal pattern were less likely to be current smokers, more likely to have higher educational attainment and higher physical activity.

Table 9: Association (regression coefficient with 95% confidence interval*) of Gaussian Graphical Models identified pattern with background characteristics in women (n=13,574)

Characteristics (Unit)	**Principal**	**High fat dairy**	**Fruit & vegetables**	**Sweet**	**Breakfast Cereals**
Age (Years)	**0.06** *(0.05,0.07)*	<0.001 *(-0.01,0.01)*	0.01 *(0,0.01)*	0.02 *(0.01,0.02)*	**-0.04** *(-0.05, -0.04)*
Current smoking**	**0.38** *(0.14,0.63)*	**0.19** *(0.03,0.35)*	**-0.59** *(-0.69, -0.5)*	**-0.71** *(-0.82, -0.59)*	**-0.45** *(-0.54, -0.36)*
Education**	**-0.62** *(-0.81, -0.42)*	0.08 *(-0.04,0.21)*	**0.44** *(0.37,0.52)*	0.01 *(-0.08,0.1)*	**0.51** *(0.44,0.58)*
BMI (Kg/meter2)	**0.12** *(0.1,0.14)*	**-0.16** *(-0.17, -0.15)*	0.01 *(0,0.02)*	**-0.02** *(-0.03, -0.01)*	**-0.03** *(-0.04, -0.02)*
Physical activity (MET)	**-1.16** *(-1.88, -0.43)*	**-0.62** *(-1.09, -0.16)*	**1.06** *(0.77,1.34)*	**0.16** *(-0.18,0.5)*	**0.41** *(0.14,0.68)*
Alcohol intake (g/day)	**0.35** *(0.25,0.45)*	**0.12** *(0.06,0.19)*	0.07 *(0.03,0.11)*	**-0.58** *(-0.63, -0.54)*	**-0.07** *(-0.11, -0.03)*
Energy (Kcal)	**8.66** *(8.33,9)*	**2.55** *(2.34,2.77)*	**1.67** *(1.54,1.8)*	**7.03** *(6.88,7.19)*	**0.74** *(0.61,0.86)*
Vitamin Supplementation**	**-0.82** *(-1.06, -0.58)*	**-0.31** *(-0.47, -0.15)*	**0.25** *(0.16,0.35)*	**-0.13** *(-0.24, -0.02)*	**0.44** *(0.36,0.53)*

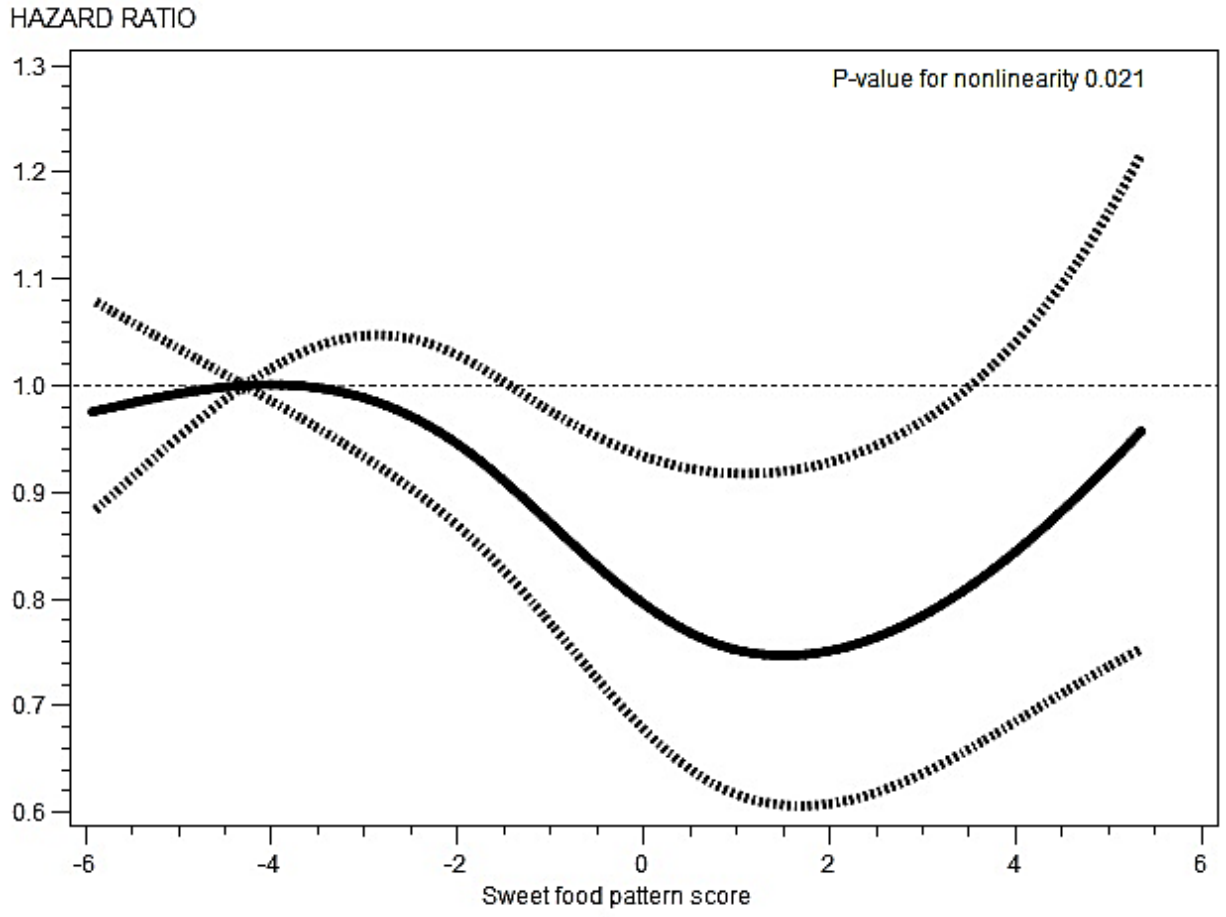
*Parameter estimates of multiple linear regression analysis with 'network score (pattern score)' as dependent variable and age, smoking status, BMI, physical activity level, education (no schooling vs higher schooling, smoking status (current vs no smoking), physical activity, alcohol intake, energy intake and supplementation intake during the last one month). Important associations are shown in bold.

**Educational attainment was defined as i) currently in training/no certificate or skill ii) professional school (vocational training) and college or higher education. Smoking status of the participants was assessed and categorized as non-smoker vs smoker. Vitamin Supplementation was assessed as yes/no.

4.2.2 Association of GGM identified patterns with risk of chronic diseases

After a median follow-up of 11.4 years (both in men and women) 701 verified cases of T2D, 199 verified cases of MI, 156 verified cases of stroke, and 729 verified cases of cancer were identified in men and 554 verified cases of T2D, 75 verified cases of MI, 132 verified cases of stroke, and 778 verified cases of cancer were identified in women. Assessment of cox-proportional hazard model assumptions showed no violations except for linearity of GGM identified sweet food pattern with risk of T2D in men as shown in Figure 8.

Figure 8: Non-linear relation of sweet food pattern with risk of T2D in men.



4.2.2.1 GGM identified pattern and risk of chronic diseases in men

Principal network score was characterized by higher intakes of refined bread, red meat, processed meat, and lower intake of whole grain bread. Adherence to the principal pattern (per SD increase) was weakly associated with risk of T2D but not associated with risk of MI, stroke, CVD, cardiometabolic diseases or overall chronic diseases in men (Table 10). This pattern was not associated with MI, stroke, CVD or overall chronic diseases in multi-variable adjusted models. No linear trend could be observed across quintile of the principal pattern score.

High fat dairy intake network represented intake pattern of high fat dairy products. High fat dairy pattern was inversely associated with risk of T2D and cardiometabolic diseases in men (Table 11). In multivariable adjusted model, participants in highest as compared to the lowest quintile had 28% lower risk of T2D [Hazard Ratio (HR): 0.72; Confidence Interval (CI): 0.56, 0.92; p for trend = 0.001]. The same dairy pattern was also associated with decreased risk of cardiometabolic diseases. Men in highest quintile of the dairy pattern had 23% lower risk of cardiometabolic diseases as compared to those in the lowest quintile (HR: 0.77; CI: 0.63,0.96; *p* for trend 0.004). Association of the pattern with T2D and cardiometabolic diseases was consistent both in age and multi-variable adjusted models. High fat dairy pattern was not associated with risk other major chronic diseases in multivariable adjusted models.

Pattern representing fruit & vegetables was associated with lower risk of cardiometabolic diseases (Table 12). As compared to the highest quintiles, participants in the lowest quintile had 29% lower risk of cardiometabolic diseases in multivariable adjusted model (HR: 0.71; CI: 0.57, 0.89; *p* for trend 0.013). The pattern was also weakly associated with risk of stroke, though no trend across quintiles was observed. Sweet foods and breakfast cereals patterns were not associated with risk of major chronic diseases in men (Table 13).

4.2.2.2 GGM identified pattern and risk of chronic diseases in women

Higher adherence to the principal pattern was associated with increased risk of T2D, cardiometabolic diseases and chronic diseases in women (Table 15). Multivariable adjusted model showed that participants in highest than in the lowest quintile of the principal pattern had 69% higher risk of T2D (HR: 1.69; CI: 1.24,2.32, p for trend = <0.0001), 58% higher

risk of cardiometabolic diseases (HR: 1.58; CI: 1.21,2.06, p for trend = <0.0001) and 22% higher risk of chronic diseases (HR: 1.22; CI: 1.02,1.46, p for trend = 0.01). Principal pattern was not associated with risk of MI, stroke, CVD or cancer.

High fat dairy product pattern was weakly inversely associated with risk of T2D (Table 16). Association of the pattern with T2D was consistent both in age and multi-variable adjusted models. Women in highest quintile of the dairy pattern had 32% lower risk of T2D as compared to those in the lowest quintile (HR: 0.68; CI: 0.51,0.92; *p* for trend 0.184). No association between high fat dairy pattern and risk of CVD, MI, stroke or cancer were observed either in age or multi-variable adjusted models. Although a linear trend across quintile was not observed; however, the per SD increase of the score showed some evidence of linear association with T2D.

Fruit & vegetables pattern was not associated with risk of any outcome of interest (Table 17). In age-adjusted model, higher compliance to the fruit & vegetables pattern was associated with lower risk of T2D. However, the association was attenuated after adjustment for other relevant risk factors in multivariable model.

Sweet pattern was also not associated with risk of major chronic diseases (Table 18). Inverse association of the pattern with risk of T2D and cardiometabolic diseases was observed in age-adjusted models. However, these associations were attenuated in multivariate adjusted models.

Table 10: Association between Gaussian Graphical Models identified principal pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	1.18(0.91,1.52)	1.33(1.04,1.71)	1.34(1.04,1.72)	1.61(1.25,2.07)		1.03(1.02,1.05)
	Model 2		(Ref.)	1.11(0.86,1.44)	1.11(0.86,1.44)	1.12(0.86,1.47)	1.31(0.98,1.75)	0.086	1.01(1,1.03)†
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	1(0.64,1.57)	0.79(0.5,1.27)	0.81(0.51,1.3)	1.31(0.85,2.02)		1.01(0.99,1.04)
	Model 2		(Ref.)	0.91(0.58,1.44)	0.73(0.45,1.19)	0.71(0.43,1.17)	1.06(0.63,1.78)	0.985	1.01(0.97,1.04)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.94(0.56,1.58)	1.04(0.62,1.73)	1.27(0.78,2.09)	1.04(0.61,1.77)		1.01(0.98,1.04)
	Model 2		(Ref.)	0.92(0.54,1.56)	0.95(0.56,1.62)	1.17(0.68,2.01)	0.93(0.5,1.73)	0.873	1(0.97,1.04)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.98(0.69,1.39)	0.91(0.64,1.3)	0.99(0.7,1.41)	1.25(0.89,1.76)		1.02(0.99,1.04)
	Model 2		(Ref.)	0.93(0.65,1.33)	0.84(0.58,1.21)	0.89(0.61,1.29)	1.04(0.7,1.55)	0.924	1(0.98,1.03)
Cancer	*Model 1*	8550(729)	(Ref.)	0.88(0.7,1.12)	1.01(0.81,1.27)	0.82(0.64,1.04)	0.92(0.72,1.16)		1(0.98,1.01)
	Model 2		(Ref.)	0.88(0.69,1.11)	1.01(0.8,1.28)	0.8(0.62,1.04)	0.88(0.67,1.17)	0.281	0.99(0.98,1.01)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	1.14(0.93,1.41)	1.18(0.96,1.46)	1.2(0.97,1.48)	1.49(1.21,1.84)		1.03(1.01,1.04)
	Model 2		(Ref.)	1.09(0.88,1.36)	1.03(0.83,1.29)	1.03(0.82,1.29)	1.23(0.96,1.57)	0.192	1.01(0.99,1.03)
Chronic Diseases	*Model 1*	8403(1593)	(Ref.)	1.06(0.9,1.25)	1.12(0.95,1.31)	1.07(0.91,1.26)	1.29(1.09,1.51)		1.02(1.01,1.03)
	Model 2		(Ref.)	1.03(0.87,1.21)	1.02(0.87,1.21)	0.95(0.8,1.14)	1.1(0.91,1.33)	0.538	1(0.99,1.02)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.
†p-value: 0.046

Table 11: Association between Gaussian Graphical Models identified high fat diary pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.78(0.63,0.96)	0.6(0.48,0.76)	0.64(0.51,0.8)	0.58(0.46,0.74)		0.93(0.9,0.95)
	Model 2		(Ref.)	0.86(0.69,1.07)	0.74(0.58,0.94)	0.75(0.59,0.96)	0.72(0.56,0.92)	0.001	0.95(0.93,0.98)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	1.2(0.8,1.81)	0.61(0.37,1)	1.05(0.69,1.62)	1.01(0.65,1.57)		1(0.95,1.05)
	Model 2		(Ref.)	1.22(0.81,1.86)	0.63(0.37,1.05)	1.06(0.68,1.67)	1(0.63,1.59)	0.784	1.01(0.95,1.07)
Stroke	*Model 1*	8593 (156)	(Ref.)	1.22(0.78,1.91)	0.88(0.53,1.46)	0.98(0.6,1.6)	0.75(0.43,1.29)		0.98(0.92,1.03)
	Model 2		(Ref.)	1.2(0.76,1.9)	0.92(0.55,1.53)	1(0.6,1.65)	0.79(0.45,1.4)	0.467	0.98(0.92,1.04)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	1.18(0.87,1.6)	0.71(0.49,1.02)	0.99(0.71,1.38)	0.87(0.61,1.22)		0.99(0.95,1.02)
	Model 2		(Ref.)	1.19(0.87,1.62)	0.74(0.51,1.06)	1.01(0.71,1.42)	0.89(0.62,1.28)	0.371	0.99(0.95,1.03)
Cancer	*Model 1*	8550(729)	(Ref.)	1.01(0.81,1.27)	1.03(0.82,1.29)	1.07(0.85,1.34)	1.03(0.82,1.3)		1.01(0.98,1.04)
	Model 2		(Ref.)	1.04(0.83,1.31)	1.07(0.84,1.35)	1.1(0.87,1.4)	1.05(0.82,1.34)	0.490	1.01(0.98,1.04)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.87(0.72,1.04)	0.64(0.53,0.79)	0.74(0.61,0.9)	0.67(0.55,0.82)		0.95(0.93,0.97)
	Model 2		(Ref.)	0.93(0.77,1.12)	0.75(0.61,0.92)	0.84(0.68,1.03)	0.77(0.63,0.96)	0.004	0.97(0.95,0.99)
Chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.91(0.78,1.05)	0.79(0.68,0.92)	0.86(0.74,1)	0.8(0.69,0.94)		0.97(0.96,0.99)
	Model 2		(Ref.)	0.95(0.82,1.1)	0.86(0.74,1.01)	0.93(0.8,1.1)	0.88(0.74,1.03)	0.087	0.99(0.97,1.01)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 12: Association between Gaussian Graphical Models identified fruit & vegetables pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.93(0.73,1.17)	1(0.79,1.26)	0.94(0.74,1.19)	0.79(0.62,1.01)		0.97(0.94,1)
	Model 2		(Ref.)	0.96(0.76,1.22)	1.02(0.81,1.29)	0.98(0.77,1.25)	0.8(0.62,1.04)	0.1655	0.97(0.94,1)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.89(0.58,1.36)	0.82(0.53,1.27)	0.86(0.56,1.32)	0.73(0.47,1.16)		0.97(0.91,1.03)
	Model 2		(Ref.)	1(0.65,1.54)	0.94(0.6,1.47)	0.99(0.63,1.55)	0.89(0.55,1.43)	0.869	1(0.94,1.07)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.62(0.38,1.01)	0.66(0.41,1.08)	0.89(0.57,1.39)	0.52(0.31,0.89)		0.93(0.87,0.99)
	Model 2		(Ref.)	0.61(0.37,1)	0.68(0.41,1.11)	0.9(0.57,1.43)	0.55(0.31,0.95)	0.140	0.93(0.86,1)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.77(0.55,1.06)	0.75(0.54,1.05)	0.87(0.63,1.2)	0.62(0.44,0.88)		0.95(0.9,0.99)
	Model 2		(Ref.)	0.81(0.58,1.12)	0.82(0.59,1.15)	0.95(0.68,1.31)	0.7(0.48,1.01)	0.188	0.97(0.92,1.01)
Cancer	*Model 1*	8550(729)	(Ref.)	1.06(0.84,1.34)	0.92(0.72,1.17)	1.12(0.88,1.41)	1.11(0.87,1.4)		1.01(0.98,1.04)
	Model 2		(Ref.)	1.1(0.87,1.4)	0.97(0.76,1.25)	1.19(0.94,1.51)	1.18(0.92,1.51)	0.166	1.02(0.99,1.06)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.87(0.71,1.06)	0.91(0.75,1.11)	0.89(0.73,1.08)	0.69(0.56,0.85)		0.96(0.93,0.98)
	Model 2		(Ref.)	0.9(0.74,1.1)	0.95(0.78,1.16)	0.94(0.77,1.15)	0.71(0.57,0.89)	0.013	0.96(0.93,0.99)
Chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.92(0.79,1.08)	0.92(0.79,1.08)	0.95(0.81,1.1)	0.84(0.71,0.98)		0.98(0.95,1)
	Model 2		(Ref.)	0.96(0.82,1.12)	0.96(0.82,1.12)	0.99(0.85,1.16)	0.87(0.74,1.03)	0.195	0.98(0.96,1)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 13: Association between Gaussian Graphical Models identified sweet food pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.78(0.62,0.97)	0.69(0.55,0.87)	0.49(0.39,0.63)	0.67(0.53,0.85)		0.95(0.93,0.97)
	Model 2		(Ref.)	0.85(0.67,1.07)	0.82(0.64,1.06)	0.66(0.49,0.89)	0.93(0.66,1.3)	0.255	0.98(0.95,1.02)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.79(0.49,1.26)	1.08(0.7,1.66)	0.92(0.59,1.44)	0.99(0.63,1.55)		1.01(0.96,1.05)
	Model 2		(Ref.)	0.82(0.5,1.34)	1.19(0.73,1.93)	1.08(0.63,1.85)	1.08(0.57,2.02)	0.571	1.03(0.97,1.11)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.73(0.45,1.18)	0.66(0.4,1.08)	0.69(0.42,1.13)	0.65(0.39,1.06)		0.96(0.91,1)
	Model 2		(Ref.)	0.73(0.44,1.23)	0.69(0.4,1.21)	0.74(0.41,1.35)	0.7(0.34,1.44)	0.349	0.96(0.89,1.03)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.76(0.53,1.07)	0.89(0.64,1.25)	0.81(0.58,1.14)	0.86(0.61,1.2)		0.99(0.96,1.02)
	Model 2		(Ref.)	0.77(0.53,1.1)	0.96(0.66,1.39)	0.92(0.61,1.38)	0.93(0.58,1.5)	0.938	1(0.95,1.06)
Cancer	*Model 1*	8550(729)	(Ref.)	1.07(0.85,1.35)	1.01(0.8,1.28)	0.94(0.74,1.2)	0.88(0.69,1.12)		0.98(0.96,1)
	Model 2		(Ref.)	1.1(0.85,1.4)	1.06(0.81,1.37)	1.01(0.76,1.35)	0.93(0.66,1.3)	0.503	0.99(0.96,1.03)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.76(0.63,0.92)	0.75(0.62,0.91)	0.59(0.48,0.73)	0.71(0.58,0.86)		0.96(0.94,0.98)
	Model 2		(Ref.)	0.82(0.67,1.01)	0.86(0.69,1.07)	0.74(0.58,0.96)	0.9(0.68,1.2)	0.279	0.99(0.96,1.02)
Chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.88(0.76,1.03)	0.84(0.72,0.98)	0.71(0.6,0.83)	0.75(0.64,0.88)		0.97(0.95,0.98)
	Model 2		(Ref.)	0.92(0.78,1.08)	0.92(0.78,1.1)	0.82(0.68,1)	0.86(0.69,1.08)	0.090	0.98(0.96,1.01)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 14: Association between Gaussian Graphical Models identified breakfast cereals pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.94(0.74,1.19)	0.8(0.62,1.05)	0.92(0.72,1.17)	0.68(0.51,0.92)		0.94(0.89,0.98)
	Model 2		(Ref.)	0.99(0.78,1.26)	0.99(0.76,1.3)	1.12(0.88,1.43)	0.96(0.71,1.3)	0.9302	0.99(0.95,1.04)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	1.19(0.79,1.79)	0.63(0.36,1.1)	0.89(0.56,1.41)	0.8(0.47,1.35)		0.95(0.87,1.04)
	Model 2		(Ref.)	1.32(0.87,2)	0.69(0.39,1.21)	1.04(0.65,1.67)	1.05(0.61,1.79)	0.862	0.99(0.91,1.09)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.95(0.57,1.59)	0.78(0.43,1.4)	0.99(0.6,1.64)	0.65(0.33,1.26)		0.94(0.85,1.04)
	Model 2		(Ref.)	1(0.6,1.67)	0.83(0.46,1.51)	1.1(0.66,1.84)	0.76(0.39,1.5)	0.519	0.97(0.87,1.08)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	1.08(0.78,1.5)	0.72(0.48,1.08)	0.95(0.67,1.34)	0.77(0.51,1.16)		0.95(0.89,1.02)
	Model 2		(Ref.)	1.17(0.84,1.62)	0.79(0.52,1.18)	1.09(0.77,1.55)	0.96(0.63,1.46)	0.745	0.99(0.93,1.06)
Cancer	*Model 1*	8550(729)	(Ref.)	0.82(0.64,1.06)	0.81(0.62,1.05)	0.97(0.77,1.22)	0.91(0.7,1.18)		0.99(0.95,1.03)
	Model 2		(Ref.)	0.85(0.66,1.09)	0.85(0.65,1.11)	1.02(0.8,1.28)	0.97(0.75,1.27)	0.975	1(0.96,1.04)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	1.01(0.83,1.23)	0.75(0.6,0.95)	0.94(0.77,1.16)	0.73(0.57,0.93)		0.95(0.91,0.98)
	Model 2		(Ref.)	1.07(0.87,1.3)	0.89(0.7,1.13)	1.14(0.92,1.4)	0.99(0.77,1.28)	0.898	1(0.96,1.04)
Chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.91(0.77,1.07)	0.78(0.65,0.93)	0.95(0.81,1.12)	0.81(0.67,0.97)		0.96(0.93,0.99)
	Model 2		(Ref.)	0.95(0.81,1.12)	0.87(0.73,1.05)	1.09(0.93,1.28)	0.98(0.82,1.19)	0.878	1(0.97,1.03)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 15: Association between Gaussian Graphical Models identified principal pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	1.38(1,1.91)	1.91(1.4,2.6)	2(1.47,2.72)	2.32(1.71,3.15)		1.06(1.04,1.07)
	Model 2		(Ref.)	1.25(0.9,1.73)	1.62(1.19,2.21)	1.62(1.18,2.22)	1.69(1.24,2.32)	<.0001	1.04(1.02,1.06)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	0.77(0.37,1.61)	0.54(0.24,1.21)	0.72(0.34,1.52)	1.14(0.57,2.25)		1.03(1.02,1.05)
	Model 2		(Ref.)	0.78(0.37,1.65)	0.56(0.24,1.27)	0.77(0.36,1.65)	1.07(0.53,2.16)	0.698	1.01(0.97,1.06)
Stroke	*Model 1*	13471 (132)	(Ref.)	0.97(0.52,1.79)	1.34(0.75,2.39)	1.17(0.64,2.13)	1.71(0.97,3)		1.02(0.98,1.06)
	Model 2		(Ref.)	0.96(0.52,1.79)	1.26(0.7,2.26)	1.08(0.59,1.98)	1.48(0.83,2.62)	0.146	1.03(0.99,1.06)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	0.9(0.55,1.46)	1.01(0.62,1.62)	1.05(0.66,1.69)	1.49(0.95,2.33)		1.02(0.98,1.06)
	Model 2		(Ref.)	0.91(0.56,1.48)	0.98(0.6,1.59)	1.02(0.63,1.66)	1.33(0.84,2.1)	0.263	1.02(0.99,1.05)
Cancer	*Model 1*	13397(778)	(Ref.)	0.9(0.71,1.13)	1(0.79,1.25)	0.9(0.71,1.13)	1.03(0.82,1.3)		1.03(1,1.07)
	Model 2		(Ref.)	0.91(0.72,1.15)	1.02(0.82,1.29)	0.93(0.73,1.17)	1.05(0.84,1.33)	0.542	1(0.99,1.02)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	1.22(0.93,1.61)	1.59(1.22,2.07)	1.71(1.31,2.22)	2.05(1.59,2.66)		1.05(1.04,1.06)
	Model 2		(Ref.)	1.13(0.86,1.49)	1.4(1.07,1.82)	1.46(1.12,1.91)	1.58(1.21,2.06)	<.0001	1.04(1.02,1.05)
Chronic Diseases	*Model 1*	13268(1448)	(Ref.)	1.02(0.85,1.22)	1.23(1.03,1.46)	1.18(0.99,1.41)	1.39(1.17,1.65)		1.02(1.01,1.03)
	Model 2		(Ref.)	1(0.83,1.19)	1.17(0.98,1.39)	1.12(0.94,1.34)	1.22(1.03,1.46)	0.007	1.02(1,1.03)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 16: Association between Gaussian Graphical Models identified high fat diary pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	0.88(0.7,1.12)	0.78(0.61,1.01)	0.7(0.54,0.91)	0.49(0.37,0.65)		0.93(0.91,0.95)
	Model 2		(Ref.)	1.03(0.81,1.31)	1(0.77,1.28)	1.05(0.8,1.36)	0.68(0.51,0.92)	0.184	0.98(0.95,1)†
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	1.63(0.8,3.33)	0.8(0.34,1.89)	1.24(0.58,2.68)	1.81(0.88,3.7)		0.97(0.95,0.99)
	Model 2		(Ref.)	1.79(0.87,3.68)	0.9(0.38,2.14)	1.51(0.69,3.28)	1.88(0.9,3.93)	0.297	1.04(0.97,1.11)
Stroke	*Model 1*	13471 (132)	(Ref.)	1.15(0.68,1.93)	0.9(0.52,1.58)	0.74(0.41,1.33)	1.28(0.76,2.16)		1.04(0.97,1.11)
	Model 2		(Ref.)	1.12(0.66,1.88)	0.92(0.53,1.62)	0.78(0.43,1.41)	1.36(0.8,2.3)	0.494	1.02(0.97,1.07)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	1.36(0.88,2.09)	0.89(0.55,1.44)	0.97(0.6,1.55)	1.52(0.98,2.34)		1.04(0.98,1.11)
	Model 2		(Ref.)	1.38(0.9,2.13)	0.93(0.57,1.51)	1.07(0.67,1.73)	1.64(1.05,2.55)	0.142	1.03(0.99,1.08)
Cancer	*Model 1*	13397(778)	(Ref.)	0.88(0.71,1.1)	0.91(0.73,1.14)	0.92(0.73,1.14)	0.97(0.78,1.22)		1.01(0.96,1.06)
	Model 2		(Ref.)	0.88(0.71,1.1)	0.92(0.74,1.15)	0.93(0.74,1.16)	0.97(0.78,1.22)	0.928	1.01(0.99,1.03)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	1.01(0.81,1.25)	0.81(0.64,1.02)	0.75(0.59,0.94)	0.69(0.55,0.88)		0.96(0.94,0.97)
	Model 2		(Ref.)	1.13(0.91,1.4)	0.98(0.78,1.24)	1.04(0.82,1.32)	0.93(0.73,1.19)	0.767	0.99(0.97,1.01)
Chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.94(0.81,1.1)	0.86(0.73,1.01)	0.83(0.71,0.98)	0.81(0.69,0.96)		0.98(0.96,0.99)
	Model 2		(Ref.)	1.01(0.86,1.18)	0.97(0.83,1.14)	1(0.85,1.18)	0.98(0.83,1.16)	0.979	1(0.98,1.01)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

†P-value: 0.044

Table 17: Association between Gaussian Graphical Models identified fruit & vegetables pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	0.79(0.61,1.03)	0.76(0.58,0.99)	0.83(0.64,1.07)	0.75(0.57,0.97)		0.94(0.91,0.98)
	Model 2		(Ref.)	0.91(0.7,1.19)	0.8(0.61,1.05)	0.94(0.72,1.22)	0.82(0.63,1.07)	0.589	0.98(0.93,1.02)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	2.05(0.97,4.35)	1.55(0.7,3.42)	1.18(0.51,2.7)	1.31(0.58,2.96)		0.96(0.92,1)
	Model 2		(Ref.)	2.48(1.16,5.31)	2.14(0.95,4.83)	1.61(0.69,3.76)	1.84(0.8,4.26)	0.865	1.03(0.92,1.15)
Stroke	*Model 1*	13471 (132)	(Ref.)	1.23(0.72,2.1)	1.19(0.69,2.04)	0.86(0.48,1.53)	0.95(0.53,1.67)		0.99(0.89,1.1)
	Model 2		(Ref.)	1.23(0.72,2.11)	1.17(0.68,2.02)	0.84(0.47,1.51)	0.94(0.52,1.67)	0.641	0.99(0.9,1.08)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	1.42(0.91,2.21)	1.22(0.77,1.93)	0.99(0.61,1.59)	1.06(0.66,1.7)		1.05(0.95,1.16)
	Model 2		(Ref.)	1.52(0.97,2.37)	1.33(0.83,2.11)	1.08(0.67,1.75)	1.19(0.73,1.93)	0.948	1.01(0.94,1.09)
Cancer	*Model 1*	13397(778)	(Ref.)	0.92(0.73,1.15)	0.93(0.74,1.16)	0.92(0.74,1.15)	0.9(0.72,1.13)		0.97(0.9,1.05)
	Model 2		(Ref.)	0.94(0.75,1.17)	0.97(0.77,1.21)	0.96(0.77,1.21)	0.94(0.75,1.18)	0.867	1(0.97,1.04)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	0.91(0.72,1.15)	0.88(0.7,1.11)	0.89(0.71,1.13)	0.84(0.66,1.06)		0.96(0.93,0.99)
	Model 2		(Ref.)	1.03(0.81,1.3)	0.95(0.75,1.2)	1.01(0.8,1.28)	0.95(0.74,1.2)	0.941	0.99(0.95,1.03)
Chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.91(0.77,1.07)	0.91(0.78,1.07)	0.9(0.77,1.06)	0.84(0.71,1)		0.97(0.95,1)
	Model 2		(Ref.)	0.95(0.81,1.12)	0.95(0.81,1.12)	0.98(0.83,1.16)	0.9(0.76,1.07)	0.545	0.99(0.96,1.02)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 18: Association between Gaussian Graphical Models identified sweet pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518(554)	(Ref.)	0.95(0.73,1.24)	0.85(0.65,1.11)	1.02(0.79,1.32)	0.74(0.56,0.98)		0.97(0.95,1)
	Model 2		(Ref.)	0.97(0.74,1.27)	0.88(0.66,1.15)	1.14(0.87,1.48)	0.85(0.64,1.12)	0.3923	1.02(0.97,1.06)
Myocardial Infarction	*Model 1*	13556(75)	(Ref.)	1.25(0.59,2.65)	1.22(0.57,2.63)	1.17(0.55,2.5)	1.09(0.5,2.37)		0.99(0.96,1.01)
	Model 2		(Ref.)	1.29(0.6,2.78)	1.43(0.65,3.13)	1.42(0.65,3.1)	1.38(0.62,3.06)	0.4777	1.01(0.9,1.13)
Stroke	*Model 1*	13471(132)	(Ref.)	1.24(0.74,2.09)	0.82(0.46,1.45)	0.9(0.52,1.57)	0.71(0.39,1.27)		1.03(0.95,1.1)
	Model 2		(Ref.)	1.37(0.81,2.31)	0.95(0.53,1.7)	1.04(0.6,1.83)	0.84(0.47,1.53)	0.8709	1(0.91,1.09)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	1.22(0.79,1.87)	0.91(0.57,1.44)	0.86(0.55,1.37)	0.78(0.49,1.25)		1.05(0.97,1.13)
	Model 2		(Ref.)	1.31(0.85,2.02)	1.04(0.65,1.66)	1.01(0.63,1.61)	0.95(0.59,1.54)	0.485	1(0.93,1.08)
Cancer	*Model 1*	13397(778)	(Ref.)	0.94(0.76,1.18)	0.9(0.72,1.13)	0.82(0.65,1.04)	1.02(0.82,1.28)		0.96(0.91,1.01)
	Model 2		(Ref.)	0.95(0.76,1.19)	0.92(0.73,1.16)	0.85(0.67,1.07)	1.06(0.85,1.34)	0.6065	1.01(0.97,1.05)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	1.02(0.81,1.29)	0.87(0.69,1.11)	1.01(0.8,1.27)	0.76(0.59,0.97)		0.97(0.95,1)
	Model 2		(Ref.)	1.05(0.83,1.33)	0.92(0.72,1.17)	1.12(0.88,1.42)	0.87(0.68,1.12)	0.7896	1.01(0.97,1.05)
Chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.97(0.83,1.14)	0.88(0.74,1.04)	0.92(0.78,1.08)	0.89(0.75,1.05)		0.99(0.97,1)
	Model 2		(Ref.)	0.98(0.84,1.16)	0.92(0.78,1.09)	0.98(0.83,1.16)	0.97(0.82,1.15)	0.7118	1.01(0.98,1.04)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

4.2.3 Association of PCA identified dietary patterns with risk of chronic diseases

4.2.3.1 Dietary pattern and risk of chronic diseases in men

PCA identified plain cooking pattern was not associated with any of the outcomes of interest (Table 19). There was some evidence of association between the plain cooking pattern and T2D in age adjusted model (HR: 1.27; CI: 1.01, 1.6). However, the association was attenuated after adjustment of relevant covariates in multivariable model.

Per SD increase of cereal pattern was inversely associated with risk of T2D, cardiometabolic diseases and chronic diseases (Table 20). Per SD increase of cereal pattern score was associated with 11% lower risk of diabetes (HR: 89; CI: 0.81, 0.98), 8% lower risk of both cardiometabolic diseases (HR 0.92; CI: 0.85, 0.99; p-value 0.0265), and overall chronic diseases (HR: 0.92; CI: 0.87, 0.98; p-value 0.007). No association of cereal pattern with MI, stroke or CVD was observed.

Sweet pattern was inversely associated with risk of stroke in men (Table 21). Participants in highest than in the lowest quintile had 62% lower risk of stroke in multivariable adjusted model (HR: 0.38; CI: 0.19, 0.79, p for trend 0.014). A trend for decrease risk of overall chronic diseases was observed across sweet pattern score (p for trend 0.033); however, no association was observed between per SD increase of the sweet pattern score and risk of overall chronic diseases (p-value 0.095).

Adherence to the fruit & vegetable pattern was associated with lower risk of stroke in men (Table 22). However, no trend could be observed across quintiles of the scores.

High fat dairy pattern was inversely associated with risk of T2D (Table 23). Comparison of the extreme quintile showed that participants in the highest quintile had 21% lower risk of developing T2D (HR: 0.79; CI: 0.61, 1.01; p for trend 0.006; p for continuous exposure 0.013). No association of the pattern was observed with other chronic diseases in multivariable adjusted model.

4.2.3.2 Dietary pattern and risk of chronic diseases in women

In women, plain cooking pattern was not associated with risk of any chronic disease (Table 24). Similarly, no associations were observed between sweet pattern and risk of T2D, MI, stroke, and cancer or any of the combined outcomes (Table 26).

Adherence with bread & sausage pattern was associated with higher risk of T2D, cardiometabolic diseases and chronic diseases (Table 25). Multivariable adjusted model showed that women in highest quintile as compared to the lowest quintile had 103% higher risk of T2D (HR 2.01; CI: 1.47,2.81; p for trend <0.0001), 84% higher risk of cardiometabolic diseases (HR 1.84: CI: 1.39, 2.44; p for trend <0.0001), and 46% higher risk of overall chronic diseases (HR 1.46; CI: 1.21,1.75; p for trend 0.0003). No associations between compliance to the pattern and risk of MI, stroke and CVD were observed.

No associations were observed for fruit & vegetable pattern (Table 26) as well as low fat dairy intake pattern (Table 27) and risk of T2D, MI, stroke, Cancer, cardiometabolic or overall chronic diseases.

Table 19: Principal component analysis derived plain cooking pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.87(0.68,1.12)	0.98(0.77,1.25)	0.93(0.73,1.19)	1.27(1.01,1.6)		1.11(1.04,1.2)
	Model 2		(Ref.)	0.86(0.66,1.1)	0.92(0.72,1.18)	0.9(0.69,1.17)	1.18(0.91,1.54)	0.101	1.09(1,1.19)†
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.84(0.53,1.34)	0.86(0.55,1.36)	0.97(0.62,1.51)	0.99(0.63,1.54)		1.05(0.92,1.22)
	Model 2		(Ref.)	0.8(0.5,1.29)	0.78(0.49,1.26)	0.91(0.56,1.46)	0.83(0.5,1.39)	0.833	1.04(0.88,1.23)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.65(0.39,1.1)	0.86(0.53,1.39)	0.65(0.39,1.08)	0.86(0.53,1.4)		0.99(0.84,1.17)
	Model 2		(Ref.)	0.64(0.38,1.1)	0.79(0.48,1.31)	0.61(0.35,1.06)	0.76(0.44,1.33)	0.588	0.93(0.76,1.13)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.73(0.51,1.04)	0.84(0.6,1.17)	0.82(0.58,1.15)	0.94(0.67,1.3)		1.03(0.93,1.15)
	Model 2		(Ref.)	0.71(0.49,1.01)	0.77(0.54,1.1)	0.76(0.53,1.09)	0.8(0.55,1.17)	0.636	0.99(0.87,1.13)
Cancer	*Model 1*	8550(729)	(Ref.)	1.03(0.81,1.3)	1.1(0.87,1.38)	0.78(0.61,1)	0.96(0.76,1.22)		0.98(0.91,1.06)
	Model 2		(Ref.)	1.05(0.82,1.33)	1.13(0.89,1.43)	0.8(0.62,1.04)	0.98(0.75,1.28)	0.360	0.99(0.9,1.09)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.79(0.64,0.98)	0.9(0.73,1.1)	0.89(0.72,1.09)	1.12(0.92,1.36)		1.09(1.02,1.16)
	Model 2		(Ref.)	0.78(0.63,0.97)	0.85(0.69,1.05)	0.85(0.69,1.06)	1.01(0.81,1.27)	0.378	1.05(0.97,1.14)
Overall chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.87(0.74,1.02)	1.03(0.88,1.2)	0.87(0.74,1.02)	1.09(0.93,1.28)		1.06(1.01,1.12)
	Model 2		(Ref.)	0.87(0.74,1.03)	0.99(0.84,1.17)	0.86(0.72,1.02)	1.01(0.85,1.21)	0.671	1.03(0.97,1.1)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.
P-value: 0.051

Table 20: Principal component analysis derived cereal pattern and risk (hazard ratio with 95% confidence interval) of major chronic diseases among men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.98(0.79,1.22)	0.87(0.69,1.09)	0.84(0.67,1.06)	0.59(0.46,0.77)		0.79(0.73,0.87)
	Model 2		(Ref.)	1(0.8,1.26)	0.94(0.74,1.19)	0.96(0.75,1.23)	0.78(0.6,1.03)	0.083	0.89(0.81,0.98)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.82(0.55,1.23)	0.58(0.37,0.91)	0.88(0.59,1.33)	0.61(0.39,0.97)		0.84(0.72,0.99)
	Model 2		(Ref.)	0.94(0.62,1.42)	0.7(0.44,1.12)	1.17(0.76,1.8)	0.88(0.54,1.43)	0.983	0.97(0.82,1.15)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.87(0.55,1.38)	0.89(0.56,1.41)	0.5(0.28,0.86)	0.86(0.52,1.41)		0.9(0.75,1.08)
	Model 2		(Ref.)	0.85(0.53,1.37)	0.89(0.55,1.45)	0.52(0.29,0.93)	1(0.59,1.69)	0.650	0.97(0.8,1.18)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.85(0.62,1.16)	0.74(0.53,1.02)	0.71(0.51,0.99)	0.74(0.52,1.04)		0.87(0.77,0.98)
	Model 2		(Ref.)	0.91(0.66,1.25)	0.83(0.59,1.16)	0.87(0.61,1.23)	0.97(0.67,1.39)	0.935	0.97(0.86,1.11)
Cancer	*Model 1*	8550(729)	(Ref.)	0.96(0.77,1.19)	0.88(0.71,1.1)	0.86(0.68,1.08)	0.85(0.66,1.08)		0.91(0.84,0.99)
	Model 2		(Ref.)	1(0.81,1.25)	0.94(0.75,1.18)	0.93(0.74,1.19)	0.93(0.72,1.19)	0.507	0.94(0.87,1.03)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.93(0.77,1.12)	0.84(0.69,1.02)	0.78(0.64,0.95)	0.64(0.52,0.8)		0.81(0.76,0.88)
	Model 2		(Ref.)	0.97(0.8,1.18)	0.92(0.75,1.12)	0.92(0.75,1.14)	0.85(0.68,1.07)	0.191	0.92(0.85,0.99)
Overall chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.9(0.77,1.04)	0.84(0.72,0.97)	0.79(0.68,0.92)	0.7(0.59,0.82)		0.85(0.8,0.9)
	Model 2		(Ref.)	0.93(0.8,1.08)	0.91(0.78,1.06)	0.91(0.78,1.07)	0.86(0.72,1.02)	0.117	0.92(0.87,0.98)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 21: Principal component analysis derived sweet pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases in men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.8(0.64,1)	0.67(0.53,0.84)	0.61(0.49,0.77)	0.59(0.47,0.75)		0.83(0.77,0.9)
	Model 2		(Ref.)	0.95(0.75,1.2)	0.89(0.69,1.16)	0.83(0.63,1.1)	0.92(0.66,1.3)	0.515	0.96(0.85,1.08)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.91(0.58,1.44)	1.02(0.65,1.58)	0.8(0.51,1.27)	1.03(0.67,1.6)		1.07(0.93,1.23)
	Model 2		(Ref.)	1.01(0.62,1.62)	1.08(0.66,1.76)	0.85(0.49,1.46)	1.01(0.54,1.88)	0.963	1.15(0.92,1.43)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.64(0.4,1.02)	0.47(0.28,0.77)	0.56(0.35,0.9)	0.46(0.28,0.75)		0.76(0.64,0.9)
	Model 2		(Ref.)	0.59(0.36,0.96)	0.41(0.24,0.72)	0.48(0.27,0.86)	0.38(0.19,0.79)	0.014	0.67(0.51,0.88)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.78(0.56,1.08)	0.69(0.5,0.97)	0.69(0.49,0.96)	0.76(0.55,1.04)		0.94(0.85,1.05)
	Model 2		(Ref.)	0.8(0.56,1.13)	0.69(0.47,1)	0.67(0.45,1.01)	0.7(0.43,1.12)	0.153	0.94(0.79,1.12)
Cancer	*Model 1*	8550(729)	(Ref.)	1.05(0.83,1.33)	1.02(0.81,1.29)	0.76(0.59,0.97)	0.96(0.76,1.21)		0.95(0.88,1.03)
	Model 2		(Ref.)	1.09(0.85,1.4)	1.05(0.8,1.36)	0.78(0.58,1.05)	0.97(0.69,1.35)	0.404	0.97(0.85,1.09)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.79(0.65,0.95)	0.65(0.53,0.79)	0.6(0.49,0.73)	0.63(0.52,0.77)		0.86(0.8,0.92)
	Model 2		(Ref.)	0.89(0.73,1.09)	0.79(0.63,0.98)	0.74(0.58,0.94)	0.83(0.62,1.11)	0.137	0.94(0.85,1.04)
Overall chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.89(0.76,1.03)	0.78(0.67,0.91)	0.64(0.55,0.75)	0.73(0.63,0.85)		0.89(0.85,0.94)
	Model 2		(Ref.)	0.96(0.81,1.12)	0.86(0.73,1.03)	0.72(0.6,0.88)	0.83(0.67,1.04)	0.033	0.93(0.86,1.01)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 22: Principal component analysis derived fruit & vegetable pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases in men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.96(0.75,1.22)	1.18(0.94,1.49)	1.05(0.83,1.33)	0.99(0.77,1.26)		0.98(0.91,1.06)
	Model 2		(Ref.)	0.94(0.73,1.2)	1.1(0.87,1.4)	1(0.78,1.27)	0.91(0.7,1.17)	0.714	0.96(0.88,1.04)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.68(0.44,1.06)	0.62(0.4,0.97)	0.72(0.47,1.1)	0.77(0.51,1.17)		0.95(0.81,1.1)
	Model 2		(Ref.)	0.76(0.49,1.18)	0.7(0.44,1.1)	0.84(0.54,1.29)	0.88(0.57,1.36)	0.969	1.02(0.87,1.18)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.74(0.46,1.2)	0.95(0.61,1.5)	0.58(0.35,0.97)	0.64(0.38,1.06)		0.82(0.68,0.99)
	Model 2		(Ref.)	0.71(0.44,1.16)	0.94(0.59,1.48)	0.57(0.34,0.97)	0.62(0.37,1.06)	0.066	0.8(0.66,0.98)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.71(0.51,0.99)	0.77(0.56,1.06)	0.68(0.49,0.94)	0.7(0.5,0.97)		0.9(0.79,1.01)
	Model 2		(Ref.)	0.73(0.52,1.02)	0.81(0.58,1.12)	0.72(0.51,1.01)	0.73(0.52,1.03)	0.145	0.92(0.81,1.04)
Cancer	*Model 1*	8550(729)	(Ref.)	0.95(0.74,1.21)	1.04(0.82,1.31)	1.01(0.8,1.27)	1.19(0.94,1.49)		1.05(0.97,1.12)
	Model 2		(Ref.)	0.99(0.77,1.26)	1.07(0.84,1.36)	1.05(0.83,1.34)	1.23(0.97,1.56)	0.091	1.05(0.98,1.14)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.87(0.71,1.07)	1.01(0.83,1.23)	0.88(0.72,1.07)	0.84(0.68,1.03)		0.94(0.88,1.01)
	Model 2		(Ref.)	0.87(0.71,1.07)	0.97(0.8,1.19)	0.86(0.7,1.05)	0.8(0.65,0.99)	0.099	0.93(0.87,1)
Overall chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.9(0.77,1.05)	1.02(0.87,1.19)	0.93(0.79,1.09)	0.96(0.82,1.13)		0.99(0.94,1.04)
	Model 2		(Ref.)	0.9(0.77,1.06)	1.01(0.86,1.18)	0.92(0.79,1.08)	0.95(0.81,1.12)	0.737	0.98(0.93,1.04)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 23: Principal component analysis derived high fat dairy pattern and risk (hazard ratios with 95% confidence interval) of chronic diseases in men in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	8618 (701)	(Ref.)	0.74(0.6,0.92)	0.6(0.48,0.75)	0.61(0.49,0.77)	0.62(0.49,0.79)		0.83(0.77,0.89)
	Model 2		(Ref.)	0.79(0.64,0.99)	0.67(0.53,0.84)	0.73(0.58,0.93)	0.79(0.61,1.01)	0.006	0.91(0.84,0.98)
Myocardial Infarction	*Model 1*	8608 (199)	(Ref.)	0.82(0.53,1.28)	0.92(0.6,1.41)	1.05(0.69,1.62)	1.05(0.67,1.64)		1.06(0.92,1.22)
	Model 2		(Ref.)	0.8(0.51,1.25)	0.88(0.57,1.36)	0.99(0.64,1.53)	0.96(0.6,1.53)	0.957	1.05(0.9,1.22)
Stroke	*Model 1*	8593 (156)	(Ref.)	0.94(0.58,1.52)	1.34(0.85,2.12)	0.9(0.54,1.51)	0.93(0.54,1.62)		1(0.84,1.19)
	Model 2		(Ref.)	0.92(0.57,1.5)	1.33(0.84,2.11)	0.93(0.55,1.57)	1.04(0.58,1.85)	0.806	1.01(0.84,1.21)
Cardiovascular diseases	*Model 1*	8524(342)	(Ref.)	0.88(0.63,1.22)	1.08(0.79,1.48)	0.98(0.7,1.37)	0.99(0.69,1.4)		1.03(0.92,1.15)
	Model 2		(Ref.)	0.87(0.62,1.21)	1.05(0.76,1.46)	0.96(0.68,1.36)	0.97(0.67,1.41)	0.964	1.02(0.91,1.14)
Cancer	*Model 1*	8550(729)	(Ref.)	1.03(0.83,1.29)	1.08(0.86,1.34)	1.04(0.82,1.3)	1.02(0.8,1.3)		1.02(0.95,1.1)
	Model 2		(Ref.)	1.06(0.85,1.32)	1.1(0.87,1.37)	1.05(0.83,1.32)	1(0.78,1.29)	0.921	1.03(0.95,1.11)
Cardiometabolic Diseases	*Model 1*	8466(960)	(Ref.)	0.78(0.65,0.94)	0.73(0.6,0.88)	0.73(0.6,0.88)	0.73(0.6,0.9)		0.89(0.83,0.95)
	Model 2		(Ref.)	0.82(0.68,0.99)	0.78(0.64,0.95)	0.81(0.67,1)	0.85(0.69,1.06)	0.047	0.95(0.89,1.01)
Overall chronic Diseases	*Model 1*	8403(1593)	(Ref.)	0.9(0.78,1.05)	0.87(0.75,1.01)	0.84(0.72,0.98)	0.85(0.72,1)		0.94(0.9,0.99)
	Model 2		(Ref.)	0.94(0.81,1.09)	0.91(0.78,1.06)	0.89(0.76,1.05)	0.92(0.78,1.09)	0.159	0.98(0.93,1.03)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 24: Principal component analysis derived plain cooking pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	1.1(0.81,1.51)	1.33(0.98,1.79)	1.41(1.05,1.89)	1.54(1.16,2.05)		1.15(1.07,1.25)
	Model 2		(Ref.)	1.09(0.8,1.5)	1.3(0.95,1.77)	1.27(0.93,1.74)	1.31(0.96,1.8)	0.085	1.08(0.99,1.19)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	0.86(0.38,1.92)	0.74(0.32,1.69)	1.1(0.52,2.32)	1.12(0.54,2.34)		1.19(0.95,1.48)
	Model 2		(Ref.)	0.81(0.36,1.85)	0.72(0.31,1.7)	1.08(0.49,2.37)	0.93(0.43,2.05)	0.807	1.13(0.89,1.45)
Stroke	*Model 1*	13471 (132)	(Ref.)	1.09(0.59,2.01)	1.34(0.74,2.4)	1.14(0.63,2.08)	1.28(0.72,2.3)		1.1(0.93,1.3)
	Model 2		(Ref.)	1.1(0.59,2.05)	1.35(0.73,2.49)	1.16(0.61,2.19)	1.2(0.63,2.3)	0.793	1.07(0.88,1.31)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	0.96(0.58,1.59)	1.13(0.7,1.83)	1.18(0.74,1.89)	1.24(0.78,1.97)		1.13(0.99,1.3)
	Model 2		(Ref.)	0.95(0.57,1.58)	1.13(0.68,1.85)	1.19(0.72,1.96)	1.11(0.67,1.85)	0.608	1.11(0.95,1.3)
Cancer	*Model 1*	13397(778)	(Ref.)	0.83(0.66,1.04)	0.88(0.7,1.1)	0.75(0.59,0.94)	0.91(0.73,1.13)		0.99(0.92,1.06)
	Model 2		(Ref.)	0.85(0.67,1.07)	0.91(0.72,1.15)	0.77(0.61,0.99)	0.94(0.74,1.2)	0.711	1(0.92,1.09)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	1.06(0.81,1.39)	1.27(0.98,1.65)	1.35(1.05,1.74)	1.47(1.14,1.88)		1.15(1.07,1.23)
	Model 2		(Ref.)	1.04(0.79,1.37)	1.23(0.94,1.61)	1.22(0.93,1.61)	1.23(0.94,1.62)	0.117	1.08(0.99,1.18)
Overall chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.91(0.76,1.09)	1.02(0.86,1.21)	0.96(0.81,1.14)	1.1(0.93,1.3)		1.06(1.01,1.12)
	Model 2		(Ref.)	0.91(0.76,1.09)	1.01(0.85,1.21)	0.92(0.77,1.1)	1.01(0.84,1.22)	0.671	1.03(0.97,1.09)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 25: Principal component analysis derived sweet pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	0.9(0.7,1.17)	0.76(0.58,0.99)	0.94(0.73,1.21)	0.75(0.57,0.98)		0.91(0.83,0.99)
	Model 2		(Ref.)	0.94(0.72,1.23)	0.85(0.63,1.15)	1.17(0.86,1.59)	0.94(0.63,1.4)	0.954	1.02(0.87,1.19)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	0.78(0.37,1.66)	0.81(0.39,1.7)	0.66(0.31,1.44)	1.4(0.74,2.67)		1.23(1.04,1.46)
	Model 2		(Ref.)	0.72(0.33,1.56)	0.71(0.33,1.55)	0.55(0.24,1.28)	0.9(0.37,2.19)	0.916	1.17(0.83,1.67)
Stroke	*Model 1*	13471 (132)	(Ref.)	0.9(0.51,1.6)	1.53(0.93,2.54)	0.82(0.47,1.46)	0.75(0.42,1.34)		0.92(0.77,1.11)
	Model 2		(Ref.)	1.02(0.56,1.84)	1.75(1,3.09)	0.95(0.48,1.86)	0.82(0.35,1.88)	0.480	0.96(0.69,1.33)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	0.88(0.56,1.39)	1.25(0.82,1.9)	0.77(0.48,1.22)	0.91(0.58,1.42)		1.01(0.88,1.16)
	Model 2		(Ref.)	0.93(0.58,1.5)	1.33(0.84,2.11)	0.8(0.47,1.38)	0.88(0.46,1.66)	0.539	1.08(0.83,1.4)
Cancer	*Model 1*	13397(778)	(Ref.)	0.92(0.73,1.15)	1.02(0.82,1.27)	0.86(0.69,1.08)	1(0.8,1.24)		1.01(0.94,1.08)
	Model 2		(Ref.)	0.93(0.74,1.18)	1.05(0.83,1.34)	0.91(0.69,1.18)	1.07(0.78,1.48)	0.702	1.03(0.9,1.18)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	0.92(0.73,1.16)	0.92(0.73,1.15)	0.94(0.75,1.19)	0.79(0.62,1)		0.94(0.87,1.02)
	Model 2		(Ref.)	0.96(0.76,1.22)	1.01(0.79,1.31)	1.11(0.84,1.46)	0.91(0.64,1.29)	0.750	1.03(0.89,1.18)
Overall chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.89(0.76,1.05)	0.95(0.81,1.12)	0.89(0.76,1.05)	0.9(0.77,1.06)		0.98(0.93,1.03)
	Model 2		(Ref.)	0.91(0.77,1.08)	1.01(0.85,1.21)	0.98(0.81,1.19)	0.99(0.78,1.25)	0.940	1.01(0.92,1.11)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake energy+ under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 26: Principal component analysis derived bread & sausage pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	1.68(1.23,2.29)	1.79(1.32,2.45)	2.06(1.52,2.8)	2.6(1.92,3.52)		1.33(1.22,1.45)
	Model 2		(Ref.)	1.47(1.08,2.02)	1.6(1.17,2.19)	1.67(1.22,2.29)	2.03(1.47,2.81)	<.0001	1.22(1.1,1.35)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	1.4(0.57,3.45)	2.12(0.92,4.88)	2.29(0.99,5.27)	2.24(0.96,5.25)		1.33(1.04,1.7)
	Model 2		(Ref.)	1.23(0.5,3.06)	1.84(0.79,4.28)	1.77(0.76,4.14)	1.42(0.58,3.45)	0.315	1.17(0.89,1.54)
Stroke	*Model 1*	13471 (132)	(Ref.)	1.28(0.74,2.22)	1.14(0.65,2.02)	0.88(0.48,1.63)	1.62(0.93,2.81)		1.12(0.93,1.35)
	Model 2		(Ref.)	1.22(0.7,2.13)	1.11(0.62,1.98)	0.81(0.43,1.51)	1.34(0.74,2.41)	0.760	1.05(0.86,1.28)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	1.37(0.84,2.21)	1.4(0.87,2.27)	1.29(0.79,2.12)	1.8(1.12,2.89)		1.19(1.02,1.38)
	Model 2		(Ref.)	1.26(0.78,2.06)	1.29(0.79,2.09)	1.1(0.66,1.82)	1.35(0.81,2.23)	0.408	1.1(0.93,1.3)
Cancer	*Model 1*	13397(778)	(Ref.)	1.16(0.92,1.46)	1.09(0.86,1.37)	1.07(0.85,1.36)	1.32(1.05,1.65)		1.06(0.99,1.14)
	Model 2		(Ref.)	1.17(0.93,1.48)	1.1(0.87,1.39)	1.07(0.84,1.37)	1.31(1.03,1.68)	0.092	1.06(0.97,1.15)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	1.54(1.17,2.02)	1.75(1.34,2.29)	1.92(1.47,2.5)	2.4(1.85,3.12)		1.31(1.21,1.41)
	Model 2		(Ref.)	1.37(1.04,1.8)	1.57(1.2,2.06)	1.57(1.2,2.07)	1.84(1.39,2.44)	<.0001	1.2(1.09,1.31)
Overall chronic Diseases	*Model 1*	13268(1448)	(Ref.)	1.27(1.07,1.52)	1.27(1.06,1.51)	1.35(1.14,1.61)	1.67(1.4,1.98)		1.17(1.1,1.23)
	Model 2		(Ref.)	1.2(1,1.43)	1.19(1,1.42)	1.22(1.02,1.46)	1.46(1.21,1.75)	0.0003	1.11(1.04,1.18)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 27: Principal component analysis derived fruit & vegetable pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases in women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	0.88(0.68,1.14)	0.8(0.61,1.04)	0.88(0.68,1.15)	0.9(0.69,1.18)		0.96(0.88,1.05)
	Model 2		(Ref.)	0.95(0.73,1.23)	0.8(0.61,1.05)	0.96(0.73,1.26)	0.93(0.69,1.24)	0.673	0.98(0.89,1.08)
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	1.08(0.53,2.2)	1.11(0.55,2.23)	0.95(0.46,1.96)	0.78(0.36,1.69)		0.83(0.65,1.08)
	Model 2		(Ref.)	1.19(0.58,2.44)	1.26(0.62,2.57)	1.1(0.52,2.33)	0.79(0.35,1.8)	0.389	0.88(0.67,1.15)
Stroke	*Model 1*	13471 (132)	(Ref.)	1.12(0.65,1.94)	1.09(0.63,1.88)	1(0.57,1.74)	1.01(0.57,1.79)		0.96(0.8,1.14)
	Model 2		(Ref.)	1.09(0.63,1.89)	1.05(0.6,1.84)	0.96(0.54,1.7)	0.95(0.51,1.76)	0.668	0.94(0.76,1.15)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	1.17(0.75,1.82)	1.14(0.73,1.78)	0.99(0.63,1.57)	1.01(0.64,1.62)		0.93(0.8,1.08)
	Model 2		(Ref.)	1.17(0.75,1.83)	1.18(0.75,1.86)	1.02(0.64,1.64)	1.01(0.61,1.67)	0.695	0.95(0.81,1.12)
Cancer**	*Model 1*	13397(778)	(Ref.)	0.95(0.76,1.19)	0.97(0.77,1.21)	0.95(0.76,1.18)	0.97(0.77,1.22)		0.98(0.91,1.06)
	Model 2		(Ref.)	0.97(0.77,1.21)	1(0.8,1.25)	0.98(0.78,1.24)	1.03(0.81,1.31)	0.633	0.99(0.92,1.08)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	0.94(0.74,1.18)	0.87(0.69,1.1)	0.93(0.74,1.17)	0.97(0.77,1.23)		0.97(0.9,1.04)
	Model 2		(Ref.)	0.98(0.77,1.24)	0.88(0.69,1.12)	0.99(0.77,1.25)	0.99(0.76,1.28)	0.890	0.99(0.91,1.07)
Overall chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.95(0.81,1.12)	0.92(0.78,1.09)	0.92(0.78,1.08)	0.94(0.8,1.11)		0.97(0.92,1.02)
	Model 2		(Ref.)	0.97(0.82,1.14)	0.93(0.79,1.1)	0.95(0.8,1.13)	0.96(0.81,1.15)	0.813	0.98(0.92,1.04)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake+ energy under-reporting. ** Model for cancer was not adjusted for history of hypertension.

Table 28: Principal component analysis derived low fat dairy pattern and risk (hazard ratios with 95% confidence interval) of major chronic diseases among women in EPIC-Potsdam cohort

Exposure/ Chronic diseases	Model*	n (cases)	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	P-value for Linear Trend	Hazard ratio per SD increase
Type 2 diabetes	*Model 1*	13518 (554)	(Ref.)	0.97(0.72,1.32)	1.31(0.99,1.75)	1.19(0.89,1.59)	1.93(1.47,2.52)		1.28(1.18,1.38)
	Model 2		(Ref.)	0.86(0.63,1.18)	1.1(0.82,1.47)	0.87(0.65,1.18)	1.19(0.9,1.57)	0.143	1.09(1,1.19)†
Myocardial Infarction	*Model 1*	13556 (75)	(Ref.)	0.97(0.48,1.97)	0.67(0.31,1.47)	1.02(0.51,2.03)	0.9(0.44,1.85)		0.91(0.72,1.15)
	Model 2		(Ref.)	1.29(0.62,2.69)	0.84(0.38,1.89)	1.29(0.63,2.66)	1(0.47,2.14)	0.997	0.93(0.74,1.17)
Stroke	*Model 1*	13471 (132)	(Ref.)	0.65(0.37,1.16)	0.78(0.45,1.34)	0.87(0.51,1.47)	0.95(0.57,1.59)		0.98(0.82,1.17)
	Model 2		(Ref.)	0.69(0.38,1.23)	0.76(0.43,1.33)	0.86(0.5,1.49)	0.89(0.52,1.51)	0.978	0.94(0.79,1.13)
Cardiovascular diseases	*Model 1*	13453(198)	(Ref.)	0.8(0.51,1.25)	0.77(0.49,1.22)	0.93(0.61,1.43)	0.88(0.57,1.36)		0.93(0.81,1.08)
	Model 2		(Ref.)	0.89(0.56,1.41)	0.8(0.5,1.27)	0.97(0.62,1.52)	0.85(0.54,1.33)	0.604	0.9(0.78,1.05)
Cancer**	*Model 1*	13397(778)	(Ref.)	0.78(0.62,0.99)	0.88(0.71,1.1)	0.92(0.74,1.14)	0.88(0.7,1.09)		0.97(0.91,1.05)
	Model 2		(Ref.)	0.79(0.63,1)	0.88(0.7,1.11)	0.92(0.73,1.15)	0.86(0.68,1.08)	0.469	0.97(0.9,1.04)
Cardiometabolic Diseases	*Model 1*	13397(715)	(Ref.)	0.97(0.75,1.25)	1.18(0.92,1.51)	1.14(0.89,1.46)	1.63(1.29,2.06)		1.19(1.1,1.28)
	Model 2		(Ref.)	0.9(0.69,1.17)	1.03(0.8,1.33)	0.91(0.7,1.17)	1.12(0.87,1.42)	0.285	1.04(0.96,1.12)
Overall chronic Diseases	*Model 1*	13268(1448)	(Ref.)	0.88(0.74,1.05)	1.03(0.87,1.22)	1.02(0.86,1.2)	1.22(1.04,1.43)		1.08(1.03,1.14)
	Model 2		(Ref.)	0.85(0.72,1.02)	0.95(0.8,1.13)	0.89(0.75,1.06)	0.97(0.82,1.15)	0.966	1(0.95,1.05)

* Model 1: Adjusted for age; and Model 2: Adjusted for model 1 + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake + energy under-reporting. **For cancer model was not adjusted for history of hypertension.
†P-value: 0.061

Table 29: Per serving increase in intake of components of Gaussian Graphical Models networks and risk (Hazard ratios with 95% confidence intervals*; significant results shown in bold) of type 2 diabetes (T2D), MI, stroke, cardiovascular diseases (CVD), cancer, cardiometabolic diseases and overall chronic diseases in EPIC-Potsdam (n=22,136)†

Food/Food Groups (Portion Size in grams)	T2D	MI	Stroke	CVD	Cancer	Cardiometabolic Diseases	Overall chronic Diseases
Whole grain bread (50)	**0.53(0.4,0.71)**	0.72(0.39,1.3)	**0.5(0.28,0.91)**	**0.57(0.37,0.88)**	1.02(0.78,1.34)	**0.57(0.44,0.73)**	**0.73(0.6,0.88)**
Refined bread (50)	1.21(0.86,1.7)	0.9(0.43,1.88)	0.83(0.41,1.66)	0.87(0.52,1.46)	1.26(0.93,1.7)	1.13(0.84,1.52)	1.16(0.93,1.44)
Processed meat (100)	**2.31(1.6,3.33)**	0.77(0.37,1.6)	1.05(0.51,2.17)	0.95(0.57,1.6)	1.22(0.9,1.66)	**1.82(1.33,2.48)**	**1.49(1.19,1.87)**
Red meat (100)	**2.44(1.64,3.62)**	1.61(0.72,3.61)	2.06(0.92,4.62)	1.68(0.94,3)	0.96(0.69,1.34)	**2.11(1.5,2.96)**	**1.46(1.15,1.87)**
Poultry (100)	1.15(0.87,1.53)	0.86(0.48,1.53)	0.93(0.52,1.66)	0.98(0.64,1.5)	1.17(0.9,1.51)	1.04(0.81,1.33)	1.06(0.88,1.28)
Sauce (5)	**1.37(1,1.86)**	1.05(0.55,1.99)	0.87(0.46,1.62)	0.95(0.6,1.51)	0.96(0.73,1.26)	**1.34(1.02,1.74)**	1.15(0.94,1.39)
Potatoes (100)	1.17(0.85,1.61)	1.42(0.72,2.8)	0.54(0.29,1.01)	0.85(0.53,1.37)	0.83(0.62,1.1)	1.05(0.8,1.38)	1(0.82,1.23)
Fried potatoes (100)	1.23(0.93,1.63)	0.88(0.5,1.57)	1.18(0.66,2.11)	1.03(0.68,1.56)	1.04(0.8,1.34)	1.14(0.9,1.45)	1.07(0.9,1.29)
Cabbage (100)	1.12(0.84,1.5)	0.59(0.32,1.08)	1.36(0.74,2.49)	0.94(0.61,1.47)	0.87(0.67,1.14)	1.09(0.85,1.4)	0.95(0.78,1.14)
Cooked vegetables (100)	1.15(0.88,1.52)	1.72(0.96,3.1)	0.74(0.42,1.31)	1.1(0.72,1.66)	1.07(0.83,1.38)	1.13(0.89,1.43)	1.07(0.9,1.28)
Mushrooms (10)	**1.21(1.06,1.38)**	0.95(0.71,1.27)	1.03(0.78,1.37)	1.01(0.82,1.24)	1.02(0.9,1.16)	1.13(1,1.27)	1.09(1,1.19)
Legumes (100)	1.2(0.93,1.54)	0.74(0.44,1.25)	1.37(0.81,2.32)	1.05(0.72,1.54)	0.88(0.7,1.11)	1.18(0.95,1.47)	1.04(0.89,1.23)
Soup (250)	**1.71(1.21,2.41)**	0.96(0.47,1.97)	0.94(0.47,1.9)	0.96(0.58,1.61)	1.01(0.74,1.38)	1.42(1.06,1.91)	1.21(0.97,1.51)
Low fat cheese (30)	0.96(0.87,1.06)	0.94(0.74,1.18)	0.82(0.65,1.02)	0.87(0.74,1.03)	1.02(0.92,1.12)	0.94(0.85,1.02)	0.98(0.91,1.04)
High fat cheese (30)	1(0.91,1.1)	0.98(0.79,1.21)	**0.8(0.65,0.99)**	0.88(0.75,1.02)	1.02(0.92,1.12)	0.97(0.89,1.06)	0.99(0.93,1.06)
Low fat dairy products (100)	0.87(0.64,1.19)	0.98(0.51,1.88)	0.93(0.49,1.78)	0.93(0.58,1.49)	1.06(0.8,1.4)	0.85(0.65,1.11)	0.9(0.74,1.1)
High fat dairy products (100)	**0.69(0.51,0.92)**	0.89(0.47,1.68)	0.75(0.41,1.37)	0.83(0.53,1.3)	1.15(0.88,1.51)	**0.73(0.57,0.94)**	0.84(0.7,1.02)

Food/Food Groups (Portion Size in grams)	T2D	MI	Stroke	CVD	Cancer	Cardiometabolic Diseases	Overall chronic Diseases
Butter (5)	0.92(0.82,1.03)	1.15(0.9,1.47)	1.05(0.82,1.34)	1.1(0.92,1.31)	1.06(0.95,1.18)	0.97(0.88,1.07)	1.01(0.93,1.08)
Margarine (5)	0.95(0.85,1.07)	0.91(0.71,1.15)	0.96(0.76,1.23)	0.94(0.79,1.12)	0.98(0.88,1.09)	0.97(0.87,1.07)	0.97(0.91,1.05)
Raw vegetables (100)	**0.7(0.51,0.97)**	0.89(0.45,1.77)	**0.51(0.26,1)**	0.65(0.4,1.06)	1.18(0.87,1.6)	**0.7(0.53,0.93)**	0.88(0.72,1.09)
Fresh fruit (100)	1.31(0.99,1.74)	1.47(0.81,2.66)	0.93(0.52,1.66)	1.25(0.82,1.92)	0.88(0.67,1.15)	**1.3(1.02,1.66)**	1.06(0.88,1.27)
Vegetable fats (5)	1.23(1.1,1.37)	1.01(0.8,1.28)	1.09(0.86,1.38)	1.05(0.89,1.24)	1.02(0.93,1.13)	**1.16(1.05,1.27)**	**1.09(1.01,1.17)**
Sweet spreads (10)	1.07(0.93,1.24)	**1.47(1.07,2)**	**0.73(0.54,0.98)**	1.04(0.83,1.29)	1.08(0.95,1.24)	1.07(0.95,1.22)	1.05(0.96,1.16)
Cakes and cookies (50)	1.08(0.83,1.42)	0.99(0.56,1.74)	1.25(0.71,2.2)	1.1(0.73,1.65)	0.96(0.74,1.23)	1.1(0.87,1.38)	1.06(0.89,1.26)
Desserts (50)	**2(1.53,2.61)**	1.23(0.71,2.13)	0.9(0.52,1.56)	1.12(0.75,1.67)	0.84(0.66,1.08)	**1.65(1.31,2.08)**	**1.19(1,1.41)**
Confectionary (50)	**1.7(1.31,2.21)**	0.85(0.5,1.45)	1.53(0.89,2.63)	1.16(0.79,1.72)	1.06(0.84,1.34)	**1.49(1.19,1.86)**	**1.22(1.03,1.44)**
Canned fruit (50)	**3.02(2.24,4.09)**	1.27(0.67,2.41)	**0.51(0.28,0.96)**	0.84(0.53,1.32)	1.11(0.84,1.46)	**2.09(1.61,2.72)**	**1.56(1.29,1.9)**
Muesli (50)	0.89(0.68,1.17)	0.64(0.35,1.17)	**0.54(0.3,0.98)**	**0.57(0.37,0.88)**	1.12(0.9,1.39)	**0.76(0.59,0.96)**	0.96(0.81,1.13)
Cornflakes (50)	1.17(0.91,1.51)	1.45(0.85,2.46)	0.67(0.36,1.24)	1(0.67,1.51)	1.08(0.86,1.34)	1.17(0.94,1.46)	1.1(0.94,1.29)
Vegetarian dishes (100)	0.75(0.52,1.08)	0.56(0.24,1.32)	1.26(0.64,2.49)	0.95(0.56,1.62)	0.79(0.59,1.05)	0.79(0.58,1.08)	0.83(0.67,1.03)

* Models adjusted for age + educational attainment + smoking status + history of hypertension + physical activity level + alcohol intake + vitamin supplementation + energy intake + energy under-reporting. For cancer model was not adjusted for history of hypertension.

4.2.4 Association of single food groups identified in GGM networks with risk of major chronic diseases in EPIC-Potsdam

Associations of single food groups as a component of GGM networks with major chronic diseases are shown in Table 29. Food groups are presented as components of their respective patterns.

1. *Food components of principal pattern and risk of major chronic diseases.*

As a component of the principal pattern represented by principal network, per serving (50g) increase in the intake of whole grain bread was inversely associated with risk of T2D (HR: 0.53; CI: 0.4,0.71), stroke (HR:0,5; CI: 0.28,0.91), CVD (HR: 0.57; CI: 0.37,0.88), cardiometabolic diseases (HR: 0.57; CI: 0.44, 0.73), and overall chronic diseases (HR: 0.73; CI: 0.6, 0.88). No association for intake of whole grain bread and MI was observed. Red meat and processed meat intake (per serving of 100g) were associated with increased risk of T2D, cardiometabolic diseases, and overall chronic diseases. Intakes of red and processed meats were not associated with risk of MI, stroke or CVD in this analysis. Weak association was observed for intake of sauces (per serving of 5g) with risk of T2D (HR: 1.37; CI: 1, 1.86). Sauce intake was also positively associated with higher risk of cardiometabolic diseases (HR: 1.34; CI: 1.02, 1.74). Intake of mushrooms was related to higher risk of T2D and cardiometabolic diseases but not with other major chronic diseases. Intake of soup (per serving 250 g) was also related to higher risk of T2D (HR: 1.71; CI: 1.21, 2.41) and cardiometabolic diseases (HR: 1.42; CI: 1.06, 1.91).

2. *Component of dairy intake network and risk of major chronic diseases*

Intake of high fat dairy products showed inverse association with risk of T2D and cardiometabolic diseases. Interestingly, inverse association was also observed between intake of high fat cheese and the risk of stroke in the current analyses.

3. *Component of fruit & vegetable pattern and risk of major chronic disease*

Raw vegetable intake was inversely associated with risk of T2D and cardiometabolic diseases. In addition, a weak but inverse association was also observed between the intake of fresh fruit and vegetables and risk of stroke. Fresh fruit were unexpectedly positively associated with higher risk of cardiometabolic diseases. However,

no associations were observed between fresh fruit and other health outcomes in the study. Intake of vegetable fats was positively associated with risk of T2D, cardiometabolic diseases and overall chronic diseases.

4. *Component of sweet food pattern and risk of major chronic diseases*

Among the sweet food groups, intakes of canned fruit and confectionary were positively associated with risk of developing T2D, cardiometabolic diseases and overall chronic diseases. Interestingly, higher intakes of canned fruit were inversely associated with risk of stroke. Intakes of desserts showed a positive association with risk of T2D and cardiometabolic diseases. Association of desserts intake with risk of overall chronic diseases was positive but weaker as compared to other outcomes. Intakes of sweet spreads showed positive association with risk of MI but negative with stroke.

5. *Component of cereal pattern and risk of major chronic diseases*

Cereal pattern included muesli, cornflakes and vegetarian dishes. Muesli intake was inversely associated with risk of stroke and cardiometabolic diseases. No associations were observed for other food groups with risk of major chronic diseases.

4.2.5 Effect modification

There was no significant effect modification by educational attainment, smoking status, blood pressure (above or below median value), and BMI categories (<25, 25.1-29.9, 30 or above) in both sexes.

4.2.6 Sensitivity analysis

In both men and women when the first two years of cases were removed from the analysis the magnitude and direction of the effect did not change.

5. Discussion

GGM are a set of promising approach that was used to construct dietary intake networks representing dietary intake patterns. These conditional independence networks provided an insight into food consumption patterns and identified food groups that were central to the network structure. Moreover, the identified pattern and the component food groups of the patterns were related to major chronic diseases. Adherence to principal pattern, which was high in foods like red meat, processed meat, and refined bread and low in whole-grain bread, was associated with higher risk of T2D, cardiometabolic diseases and overall chronic diseases in women. In contrast, high fat dairy pattern was related with a lower risk of T2D both in men and women. Analysis of individual food groups in relation to risk of major chronic diseases showed that red meat, processed meat, sauce, and soup were positively and whole grain-bread, raw vegetables, high fat dairy products and muesli were inversely associated with risk of major chronic diseases.

5.1 Strengths and limitations of the study design

EPIC-Potsdam cohort is part of the pan-European prospective cohort study (EPIC) to investigate influence of diet on chronic disease outcomes. Under the auspice of the Diet and Cancer Unit of IARC, the cancer research institute of World Health Organization, EPIC is a well-designed cohort study implemented with standardized protocols. Advantage of the prospective studies are the temporal relation of exposure and diseases, where exposure is followed by disease/events providing temporal framework to assess causality and strong scientific evidence of associations (93).

Like other cohort studies this study had advantage to investigate several chronic diseases simultaneously, which allowed exploring association of dietary patterns and risk of T2D, MI, Stroke and Cancer in the current analysis. As each participant of the cohort was in the risk set for all chronic diseases until censored; therefore, only first incident-events were considered as outcome of interest for this study. Second, third and the following events as well as recurrent events of the first incident-event were not considered in analysis. A limitation of the current study was its inability to investigate rare diseases like site specific cancer due to low number of such cases (94).

Target population of EPIC-Potsdam study was general population, therefore, exposure distribution among the subjects can be assumed as random. Nevertheless, external validity of this study may be limited as the study participation rate in EPIC-Potsdam was 22.7%. Moreover, due to exhaustive examination schedule people with severe illness and disability were excluded. Therefore, the final sample might have been healthier than general population. Nevertheless, it is difficult to avoid selection bias in such studies as healthy people are more frequently willing to participate in health related studies than unhealthy ones in the target population (95, 96).

This cohort had a rigorous approach to maintain high response-rate at all follow-ups. For each follow-up, participants were mailed follow-up questionnaire. In case of non-response, they were reminded by mail and telephone. Strong follow-up in this cohort resulted in a follow-up response rate of 93-96% (97). Similarly, stringent criteria for case definition and ascertainment were followed. Cases were defined through standard clinical diagnosis and coded according to ICD-10. Both active and passive follow-up were used for case identifications. Participants were actively asked to provide information about any incident diseases. In case of any incident disease, the concerned physician who diagnosed the case or treatment center where the case was diagnosed, were contacted for verification. All information including diagnosis procedures were recoded to define verified cases. Moreover, to ensure identification of all cases, cancer registries, statistics record of relevant departments and other incident data were actively assessed. As additional source of information, next-of-kin were also accessed for identification of cases.

Overall EPIC-cohort study contributes to generate a high quality of scientific evidence owing to the nature of its design, size, and standardized procedures that it follows to ensure achieve the study goals.

5.2 Strengths and limitations of the study methods

Methods of data collections:

A major strength of the methods used in the EPIC-cohort was standardized procedures for data collection. Standard and uniform procedures were followed to record relevant information including assessment of anthropometric measurements. In addition, all self-administered questionnaire including FFQ were machine readable that minimized

human error while transferring recorded information from the questionnaires to computer for tabulation.

For exposure assessment, validated food frequency questionnaires were used that captured commonly consumed foods during the last 12 months. Since, assessment of all food intakes inclusive of rarely consumed food could result in a long FFQ putting high physical and recall burden on the participants, the FFQ used in EPIC-Potsdam queried only commonly consumed 148 food items. Reproducibility and reliability of the questionnaire were assessed in a validation study (74). The FFQ had good reproducibility and acceptable relative validity. In addition, the same validation study also found that participants in EPIC-Potsdam underreported energy intake by 22% on average and the levels of under-reporting increased with higher energy intakes. Consequently, all analyses in the current study were adjusted for under-reporting of energy intake to address the issue of energy under-reporting. Nevertheless, residual confounding in the adjusted model cannot be ruled out due to measurement errors.

Recall bias was an important limitation of the exposure and other co-variate assessment. Assessment of dietary intake depended on participants' memory, which entails inherent problem of recall bias. Similarly, portion size estimation is subject to misclassification and can result in under or over reporting of the foods consumed. Although this bias cannot be removed, however, list of commonly consumed foods and pictures of the portion sizes were available to facilitated recall and estimation of portion sizes in the FFQ.

Statistical methods:

This study applied GGM, a novel approach of dietary pattern analysis, which is already in use in other fields. GGM is a data-reduction method that not only identifies sex-specific dietary networks representing patterns but also explores relationship among food groups. GGM assume Gaussian distribution of the data, which is not often the case. Therefore, the method is validated using an alternative approach called SGCGM, which does not require Gaussian assumption. Further, the food intake networks from GGM were compared with dietary patterns from PCA, an established and conventional method of dietary pattern analysis in nutrition epidemiology. Validation of the GGM derived pattern and comparison with the establish methods strengthen confidence in result of the

new dietary pattern analysis approach. Comparison with PCA also helped in determining complimentary advantage of the GGM approach for dietary pattern analysis.

There are also some potential limitations of the GGM method. First, it requires data to be Gaussian distributed which is not the case for all dietary variables. However, dietary intake data were log-transformed, and even though this do not always result in perfect normal distributions findings from GGM analysis were robust when compared against SGCGM, which do not require the Gaussian assumption. Second, network sparsity depends on regularization parameter that can be derived using different criteria e.g. log-likelihood or AIC/BIC, cross-validation etc. However, independent of the choice of method, latent structure of the data remains the same and may be identified using any of the shrinkage parameter estimation methods, though with different sparsity levels. Third, changes in characteristics of the study sample may potentially yield a different network in a same way as pattern analysis through other methods like PCA will do. This is true for all correlation dependent methods and should be kept in mind for GGM as well. Fourth, GGM identifies networks but neither does it assign individual scores to participants like PCA nor classifies individuals in groups as done by cluster analysis. It is important to note that a major aim of dietary pattern analysis is to classify individuals based on a pattern variable, which was not aim of the current study. Nevertheless, advancement of GGM methodology for possible calculation of quantitative scores or classification of individuals on the basis of identified network could be of interest, which is done in the second part of this work and can be done in future studies as well in more innovative ways. Furthermore, methods used for dietary pattern analysis, which assume sparsity, have also been criticized. It is argued that such methods reduce pattern to several foods although actual consumption comprise of large number of foods, all of which should be retained in the dietary pattern (35). Nevertheless, such arguments are challenged on several grounds. First dietary pattern lack a specific definition. Current definition of dietary pattern is method driven and operationally can be defined as data reduction (98). As existing dietary pattern analysis tools have limitations, different methods may identify dietary pattern differently, irrespective of sparsity assumption (99) but still will be called dietary patterns. Second, use of sparsity for pattern recognition depends on the study question and may be of advantage in certain situations. For example, Assi et al (100) has used sparsity to identify nutrient pattern associated with hormonal receptor-defined breast cancer. In the present study sparsity is of advantage

because it showed not only the pattern but also how foods in the pattern are consumed in relation to each other. Moreover, the sparsity also allowed investigating component food groups of the patterns in relation to risk of major chronic diseases.

Another important aspect of this study was the novel approach of creating patterns (networks) scores. Scores were created using standardized intakes and direction of correlation. Although this approach considers the actual intake combinations of different food groups; however, it does not include a priori information related to health outcomes. This feature of the score can be considered a strengthened-on one hand as it reduces subjectivity; and a weakness on another hand as food groups in such combinations of intakes may not be necessarily related to health outcomes in the same direction, which may result in cancelling out of the effect in the pattern score. For example, two food groups with positive correlation in a network will contribute to higher score showing greater adherence to the pattern but may be individually related to risk of a disease in opposite direction. Nevertheless, individual food component of the patterns was also investigated in relation to risk major chronic diseases, to help explain such instances.

EPIC-Potsdam dietary data-set is well studied and analyzed with other dietary pattern methods, e.g. PCA, which enabled direct comparison of previous methods to these results. Furthermore, the associations between the investigated food groups and chronic disease risk in this population are already published (101). Therefore, outcomes of association of the GGM identified patterns and its food components could be compared with these results to assess advantage of the new approach.

5.3 Discussion of Results

5.3.1 Application of GGM for dietary pattern analysis:

This study assessed a complementary exploratory method for dietary pattern analysis, called GGM, an existing exploratory approach already in use in metabolomics (25), genetics (102), and climate research (103). GGM, a novel approach for dietary pattern analysis, help to identify latent structures in the dietary intake data by constructing dietary intake networks based on conditional independence among intake of food groups. Moreover, this approach, when applied to dietary data, minimizes subjective choices during data analysis and identifies easy to interpret internal structures in the dietary data,

visualized as dietary networks.

Major advantage of GGM is its ability to distinguish between direct and indirect associations among the food groups consumed. Data reduction methods like PCA depend upon the correlation matrix of the food groups that does not control for the indirect effect of other foods in the pairwise correlation between two food groups. Removal of indirect effects when assessing pairwise correlation among two food groups is crucial to understand how different food groups are consumed in relation to each other. GGM is addressing the problem of indirect effects by calculating a measure of conditional independencies among the food groups (21). The resulting conditional independence measures reflect pairwise correlation between two food groups independent of the linear effect of the other food groups. In other words, the partial correlation coefficients reflect the association between two food groups independent of the effect of other food groups. However, it is pertinent to note that the conditional independence measures do not provide any information concerning the relationship to disease outcomes. Therefore, use of the identified networks may only partly be helpful for defining confounding during further analyses.

In the current analysis, GGM identified sex-specific networks consisting of a principal network and additional smaller networks. Among both sexes, the principal networks revealed that consumption of red meat and cooked vegetables were independent of any specific food group intake underlining their potential key role in determining dietary behavior.

The findings of this study are consistent with PCA derived dietary pattern in the same population (91). For example, in men, 8 of the 12 food groups with high factor loading, in the PCA identified “Plain Cooking” pattern, were also part of the principal networks identified by GGM. Similarly, in women, 10 of the 11 food groups with high factor loadings, in the PCA identified “Plain Cooking” pattern, were also part of the principal networks identified by GGM. Moreover, “high fat dairy” in men and “sweet” PCA patterns in both men and women were also comparable with the identified networks. This comparison indicates that the identified food networks are not statistical artefacts but may reflect true patterns (35). However, in addition to the common patterns PCA also identified a number of other patterns like alcohol pattern in both men and women and bread & sausage pattern in women alone. This could be expected as GGM and PCA as

well as other data-reduction methods are statistical approaches with different mathematical basis as described in Sections 1.4 & 1.5.

In addition, GGM showed that red meat consumption is central to the dietary intake in the studied population, a finding that cannot be derived explicitly from the PCA pattern. Though high factor loading of red meat in the PCA pattern underscore its importance, GGM not only underlines its importance, it also reveals the pattern of its consumption i.e. how it is consumed in relation to other foods, in a given population. Moreover, the networks show a strong positive association between red meat and processed meat intake, a finding also observed in other populations (104). This is interesting since the role of red meat for health outcomes is still unraveled in terms of causality (104-106) and its further investigation in relation to health outcomes is still a priority on the research agenda (107).

GGM identified a separate network for fresh fruit, raw vegetables and vegetable fats, reflecting a healthy pattern among both sexes. In addition, like PCA, GGM identified separate networks for sweet foods and dairy products based on fat content. However, unlike PCA (108), GGM identified independent networks of food groups in which each food group was part of only one network that facilitated their interpretation.

GGM introduces sparsity i.e. select only few variables in the final model, forcing other variables to zero, to explore data structure and facilitate interpretation. This advantage of GGM is also shared by another data reduction method called Treelet Transform (TT), recently introduced in nutrition epidemiology (54). TT combines data reduction features of PCA and the interpretability advantage of cluster analysis to identify sparse latent structures called factors. Low or noisy factor loadings are forced to zero in each identified factor to achieve sparsity. This helps to identify factors that are easy to interpret. However, unlike GGM, it does not estimate a single pattern of individual foods as a unique solution for the estimated model. Moreover, the food groups in each factor are not independent of the effect of each other.

This study also showed that GGM is a robust method for dietary pattern analysis. It revealed similar networks as SGCGM. For current analysis GGM was method of choice because SGCGM perform rank based transformation of the original variables into new variables having Gaussian distribution. After transformation of the variables SGCGM use the same method to get graphical model as done in GGM. Therefore, it was preferred to

keep log-transformed original variables and use GGM than perform model selection on rank-based transformed variables.

This study showed that GGM is a powerful exploratory method that can be used to construct dietary intake networks representing dietary intake patterns. These conditional independence networks provide an insight into dietary intake patterns of a population and identify food groups that are central to the network structure.

5.3.2 Discussion of association between intake patterns and risk of major chronic diseases

GGM is a powerful exploratory approach for dietary pattern analysis and the identified pattern can be related with risk of diet related chronic diseases. In this analysis, GGM patterns could identify risk of major chronic diseases and the patterns could further be exploited to investigate association of component food groups of the patterns with risk of chronic diseases. Likewise, GGM patterns were easier to interpret in relation to health outcomes as each food was part of one pattern and the patterns consisted mostly of similar food groups.

In the current study GGM identified principal pattern was strongly linked with risk of developing T2D and combined endpoints of major chronic diseases in women, independent of other risk factors. In contrast, a similar pattern identified by PCA was not associated with risk of major chronic diseases. Due to novelty of this approach, this intake pattern cannot be compared directly with patterns identified by other methods; nevertheless, these results are comparable with findings of studies reporting patterns comprising similar food groups. For example a prospective study of women aged 38-64 years showed that western diet pattern including red and processed meat, refined grains and French fries beside other foods was associated with increased risk of T2D (109). Likewise, in a British birth cohort, a dietary pattern that included lower intake of whole grains and higher intakes of processed meat was associated with higher risk of T2D in women but not in men (110). This is interesting to note that many studies have reported in general similar pattern to that of principal pattern identified by GGM; however, these patterns had remarkably different components (as food groups) (111, 112) and may not provide a realistic comparison. Therefore, it may be important to consider the components of the identified patterns as well to understand the relation between the pattern and the

observed health outcome. Such analysis in this study showed that whole grain bread was inversely associated with risk of chronic diseases while red mead, processed meat, sauce, mushrooms, and soup were positively associated with risk of chronic diseases. These findings are consistent with earlier studies. For example, two of the food groups i.e. whole grain bread and red meat were associated with risk of T2D and overall chronic diseases in a previous study in the same population (101). Similarly, other studies also showed that higher intakes of red meat (109), processed meat (113), refined grain and fried potatoes (114) were positively associated with risk of T2D while higher intake of whole grains [including from whole grain bread] was inversely associated with the risk of T2D (115, 116). It is also necessary to underline that some foods are consumed in combination and the observed independent effect may have a synergetic or antagonistic effect in combination with other foods. For example, sauce is not consumed in isolation and may have a varying role in relation to risk of diseases depending on its composition and the combination of foods with which it is consumed.

The inverse association observed between high fat dairy pattern and T2D in this study is consistent with earlier studies (117-119), though the scientific evidence from earlier studies regarding role of high fat dairy in cardiometabolic diseases is conflicting (120, 121). Nevertheless, these results are in line with several meta-analyses that reported higher intakes of high fat dairy products lowers (122) rather than increase risk of metabolic diseases (120, 123). One meta-analysis attributed the positive influence of high fat dairy product on risk of T2D to its fat content (117). It was reported that fat from high fat dairy products but not from meat were associated with risk of T2D (117). In addition, it is suggested that dairy products may alter risk of T2D, CVD and other chronic diseases by providing an array of lipids, proteins and micronutrients (120, 124). Investigation of components of this pattern with risk of major diseases showed that high fat dairy products but not high fat cheese was inversely associated with T2D and CVD. Interestingly, high fat dairy pattern was not associated with stroke but high fat cheese as component of this pattern was inversely associated with risk of stroke in this cohort, an observation also reported in a recent meta-analysis (125) but not in earlier studies (126) .

GGM identified fruit & vegetable pattern was inversely associated with risk of cardiometabolic diseases in men but not in women. A similar pattern identified by PCA was inversely associated with stroke and cardiometabolic diseases in men but not with any

outcomes in women. The results are consistent with earlier findings that fresh fruit and vegetables are inversely associated with risk of T2D and CVD (127-130). It must be noted that use of cardiometabolic diseases as endpoint in this study limits its direct comparison with other studies. Nevertheless, as the three diseases i.e. T2D, MI and stroke share common risk factors assessed in this study, it may provide enough power to evaluate the association with the exposure. Association of the pattern score with background characteristics showed inverse relation with smoking and alcohol consumption but positive association with physical activity and education, which shows a positive health behavior among high consumers of this pattern. Assessment of individual components of fruit & vegetables pattern showed that raw vegetables are inversely related with risk of T2D, stroke and cardiometabolic diseases. These findings are consistent with an earlier study from this cohort (101). Moreover, vegetable fat in this pattern was positively associated with risk of T2D, which might be one factor related to lack of association of the fruit & vegetable pattern with risk of T2D.

Another finding related to inverse association of PCA identified sweet pattern with stroke in men are similar to earlier observations from the same cohort. Earlier studies have attributed the inverse association of sweet foods intake with chronic disease outcome to under-reporting (74, 131). Although all analysis was adjusted for energy-under reporting in this study, nevertheless, the observed associations remained unchanged. One reason could be use of subjectively measured physical activity levels i.e. from FFQ for estimation of energy requirement in this study as compared to objectively measured physical activity levels used in the studies reporting energy under-reporting in this cohort.

Association of PCA identified cereals pattern with lower risk of stroke, cardiometabolic diseases, and overall chronic diseases in this analysis are also in line with other studies (132-134). Whole grains contain bran, germ and endosperm, which are important sources of fiber, nutrients and energy respectively (135). Higher intakes of cereals were also inversely related with smoking, BMI, and cholesterol (91) in men, reflecting healthier behavior of the participants. This pattern also loaded high on wine, tea and vegetables (factor loading: 20). Wine in moderate consumption is associated with lower risk of CVD (136, 137). Likewise, higher intakes of tea and vegetable intake are also attributed with lower risk of CVD (138, 139).

Another PCA pattern, bread & sausage, identified in women was positively related

to risk of T2D and the combined endpoints of major chronic diseases. Dietary patterns including similar foods identified in two US cohorts were also related to increased risk of T2D and CVD (140, 141). Similar observations were also reported from a large cross-sectional survey in Taiwanese population (142).

These results showed that GGM identified patterns can not only characterize dietary intake but also identify risk of diet related major chronic diseases. In the current study, GGM approach was compared with a commonly used food pattern analysis i.e. PCA. GGM and PCA pattern were differently related to risk of major chronic diseases. These results are unsurprising because GGM is mathematically a different approach than PCA and other statistical methods used for dietary pattern analysis. Hence, a direct comparison of the methods is not possible. In absence of gold standard, it cannot be concluded which methods is best for dietary pattern analysis; however, depending on the study question, methods with minimum limitations may be preferred over others. As a complimentary exploratory analysis approach, GGM approach showed several advantages when relating the identified patterns with health outcomes. For example, the principal pattern of GGM was related to risk of chronic diseases whereas the PCA identified first pattern that explained much of the variance, in contrast, was not related to risk of chronic diseases, an observation also reported in earlier studies (29, 143-145). Likewise, interpretations of GGM patterns were easier as compared to PCA patterns in which many food groups have non-zero loadings. For example, high fat dairy pattern identified by GGM was composed of dairy foods only; however, PCA high fat dairy pattern also comprised of processed meat and refined bread. Moreover, it is interesting to note that people consume a number of foods in different combinations that are related to each other and may be associated with risk of chronic diseases. GGM could capture such combinations (multiple networks including different combinations) of food consumed and were related to health outcomes. On the other hand, PCA identified large number of independent patterns (uncorrelated) assuming that each pattern could be followed in isolation, which is contrary to the reality. Likewise, the GGM could also be used to evaluate association of the individual food groups, in contrast to the PCA, which is an advantage that can be used in exploratory analysis for hypothesis testing.

No single method is best for dietary pattern analysis that identify pattern, which are related to chronic diseases as well. One method, Reduced Rank Regression (53) can be

preferred over other methods for dietary pattern analysis when associations with chronic diseases are also desired beside characterization of the dietary intake. However, this approach relies on prior knowledge of diet and disease, rendering it unsuitable for exploratory analysis in absence of evidence regarding exposure and disease outcome. The addition of GGM as exploratory analysis with its ability to identify pattern associated with risk of chronic diseases expands the choices of methods available for pattern analysis.

6. Conclusions and outlook

GGM are a set of powerful exploratory methods that can be used to construct dietary intake networks representing dietary intake patterns. These conditional independence networks provide an insight into food consumption patterns of a population and identify food groups that are central to the network structure. Applying this approach to EPIC-Potsdam data, sex-specific intake patterns were identified, which highlighted a central role of red meat and cooked vegetables consumption within the derived dietary patterns.

The GGM identified networks can also be scored to rank intakes of the participants across the patterns in relation to risk of chronic diseases. Using this approach, analysis in this study revealed higher risk of T2D, cardiometabolic diseases and overall chronic diseases in women following a pattern high in foods like red meat, processed meat, and refined bread and low in whole-grain bread. In contrast, lower risk of chronic diseases was observed for intake of high fat dairy products both in men and women. Additionally, the GGM approach gives leverage to investigate component food groups of a pattern without the need to include all the foods simultaneously in the model. Using this advantage, important positive role of red meat, processed meat, sauce, and soup and an inverse role of whole grain-bread, raw vegetables, high fat dairy products and muesli could be identified in relation to risk of chronic diseases.

GGM is a promising approach for dietary pattern analysis. Nevertheless, additional studies are required to validate this method in other populations. In addition, there are several other potential applications of the identified networks. First, the conditional independence can be advantageously used to identify consumption probabilities of the foods/food groups identified in the network for each individual. Such probabilities will be helpful to model alternative intake patterns by modifying intake probabilities, which may be helpful to assess impact of dietary behavior change or dietary recommendations. Partial correlations matrix of the GGM model can also be used, as an alternative approach, to achieve the same objective. Moreover, there is a rapid development in statistical techniques of graphical structure estimation for GGM e.g. FastGGM (146) and Fractional Marginal Pseudo-likelihood method (147) that can be used for efficient model selection in GGM.

7. Public health implication

Chronic diseases including T2D, CVD, and cancer are major contributors to the global burden of non-communicable diseases. The global burden of diseases analysis showed a significant increase in global morbidity and mortality related to these diseases. Diet is considered as an important risk factor for these chronic diseases (33). Therefore, urgent actions are warranted to reduce the health and economic cost of these chronic diseases.

Introduction of GGM to nutrition epidemiology expand the methodological arsenal of nutrition epidemiologist to investigate dietary patterns with a new dimension. The GGM approach adds the new aspect of exploring relationship of food intake in consumptions data. Using this approach, population-specific patterns can be identified in different regions, which help in understanding how foods are consumed in relation to each other. This information can be used to plan dietary recommendations keeping in view how change in one food will affect intake of others. Similarly, the possibility to investigate individual foods provides an opportunity to apply this approach for identification of foods associated with different health outcomes. These networks may be further investigated for planning dietary recommendations. Likewise, in future studies, these networks may be exploited to model healthy dietary patterns such as Mediterranean diet, low energy or a diet with some other desirable healthy characteristics to prevent risk of diet related chronic diseases.

8. Summary

Dietary pattern analysis is a preferable approach to characterize dietary intake and understand eating behaviors. Existing data reduction methods like principal component analysis (PCA) although identify the similarity patterns of food intake in a population; it does not provide insight into how foods are consumed in relation to each other. Moreover, these methods require several subjective but important decisions during analysis and the identified patterns are often difficult to interpret. Gaussian graphical models (GGM) are a set of powerful exploratory approaches that can identify dietary intake networks resembling dietary pattern, which may be easier to interpret and may reveal important insight into eating patterns. Therefore, aims of the current study were to apply GGM as a novel approach of dietary pattern analysis and investigate association of the identified patterns with risk of major chronic diseases in European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam.

In this study, dietary intake data from 10,780 men and 16,340 women from EPIC-Potsdam were used to construct dietary intake networks representing dietary patterns. In the first step, GGM were applied to log-transformed intakes of 49 food groups to construct sex-specific dietary intake networks. Semiparametric Gaussian copula graphical models (SGCGM) were used to confirm GGM results. Stability of the networks were assessed using 100 bootstrap samples. In a second step, dietary intake networks were scored, which were used as exposure variable to evaluate its association with risk of major chronic diseases including type 2 diabetes (T2D), myocardial infarction (MI), stroke, cardiovascular diseases (MI and stroke), cancer, cardiometabolic diseases (T2D, MI, and stroke), and overall chronic diseases (T2D, MI, stroke, and cancer). As comparative approach to GGM, Principal Component Analysis (PCA) patterns were recreated and its respective scores were used to evaluate adherence to the patterns and risk of major chronic diseases. Cox-proportional hazard analysis models were used to investigate associations of the patterns' scores with risk of major chronic diseases.

In both sexes, GGM identified one major network, named principal intake network, and several smaller networks. In men, the principal network comprised 12 food groups, which grouped around red meat and cooked vegetables, and several smaller networks including dairy network, sweet food network, fresh fruit & vegetable network, and breakfast cereals network. In women, the principal network consisted of the same food

Summary groups as identified in men, with addition of fried potatoes. A major difference between patterns in men and women was the intake relationship between food groups in the identified networks. For example, intake of red meat was related with intakes of five food groups in men but seven food groups in women. CGGM results showed that GGM is a robust approach for patterns identification and bootstrap results revealed that GGM identified networks are stable.

In relation to risk of chronic diseases, principal pattern, which included higher intakes of foods like red meat, processed meat, sauce and refined bread and low intakes of whole-grain bread, was positively associated with risk of T2D, cardiometabolic diseases and overall chronic disease in women but not in men. However, a similar pattern identified by PCA was not related with risk of any major chronic disease in both sexes. GGM identified high fat dairy pattern was inversely related to risk of T2D, cardiometabolic diseases, and overall chronic diseases in men; and risk of T2D in women. A similar pattern identified by PCA was also related to lower risk of T2D but only in men. GGM identified fruit & vegetable pattern was inversely related with risk of cardiometabolic diseases in men whereas a similar pattern identified by PCA, in men, was inversely associated with risk of stroke and cardiometabolic diseases. GGM identified breakfast cereal pattern was not related with risk of chronic diseases in both sexes. However, PCA cereal pattern (including more food than GGM' pattern) was inversely related with risk of T2D, cardiometabolic diseases and overall chronic diseases, in men. PCA also identified a "bread & sausage" pattern in women, which was positively related with risk of T2D, cardiometabolic diseases, and overall chronic diseases. Evaluation of the individual food groups of the networks showed that red meat, processed meat, sauce, and soup were positively and whole grain-bread, raw vegetables, high fat dairy products and muesli were negatively associated with risk of major chronic diseases.

In brief, GGM is a complimentary approach of dietary pattern analysis that can provide insight into how foods are consumed in relation to each other in different populations. It identifies easy to interpret patterns, where each food is part of only one specific pattern. These patterns may also be related with risk of chronic diseases and can be used to investigate association of the component food groups with health outcomes. Further studies are required to confirm validity of this approach in other populations. Information related to food intake and risk of chronic diseases revealed from this approach may be useful for dietary recommendations to prevent risk of chronic diseases.

9. Zusammen Fassung

Ernährungsmusteranalysen sind eine bevorzugte Methode um die Ernährungsaufnahme zu charakterisieren und um das Essverhalten zu verstehen. Existierende Datenreduktionsmethoden wie Hauptkomponentenanalysen (PCA) identifizieren ähnliche Muster der Nahrungsaufnahme. Jedoch sind sie nicht geeignet um zu erkennen, welche Lebensmittel in Verbindung mit anderen konsumiert werden. Darüber hinaus erfordern diese Methoden mehrere subjektive, aber wichtige Entscheidungen während der Analyse und die identifizierten Muster sind nur schwer zu interpretieren. Gaußsche grafische Modelle (GGM) sind leistungsfähige explorative Ansätze, die, als komplementärer Ansatz der Ernährungsmusteranalyse verwendet werden können, um leicht zu interpretierende Muster zu identifizieren, die wichtige Einsichten in Essgewohnheiten offenbaren. Die Ziele dieser Arbeit waren daher Gaußsche grafische Modelle als neue Methode der Ernährungsmusteranalyse anzuwenden und die Assoziationen der identifizierten Muster zum Risiko für schwere chronische Krankheiten in der European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam-Studie zu untersuchen.

In dieser Studie wurden die Ernährungsdaten von 10,780 Männern und 16,340 Frauen aus EPIC-Potsdam genutzt um daraus Nahrungsmittelnetzwerke zu abzuleiten, die Lebensmittelmuster darstellen. Im ersten Schritt wurden GGM auf die log-transformierten Verzehrsmengen von 49 Lebensmittelgruppen angewendet um geschlechtsspezifische Nahrungsmittelnetzwerke zu bilden. Semiparametrische Gauß-Copula grafische Modelle (SGCGM) wurden genutzt um die Ergebnisse der GGM zu bestätigen. Die Stabilität der Netzwerke wurde mittels 100 Bootstrap-Stichproben bewertet. In einem zweiten Schritt wurden den Nahrungsmittelnetzwerken ein Score zugewiesen, welcher als Expositionsvariable genutzt wurde um die Assoziationen mit schweren chronischen Krankheiten einschließlich Diabetes mellitus Typ 2 (T2D), Herzinfarkt, Schlaganfall, kardiovaskulären Erkrankungen (Herzinfarkt und Schlaganfall), Krebs, kardiometabolischen Erkrankungen (T2D, Herzinfarkt und Schlaganfall) und chronischen Gesamterkrankungen (T2D, Herzinfarkt, Schlaganfall und Krebs). Zum Vergleich mit den GGM-Ergebnissen wurden zudem Muster mittels einer Hauptkomponentenanalyse gebildet und deren Scores genutzt um die Adhärenz zu den Mustern und dem Risiko für schwere chronische Erkrankungen zu bewerten. Proportionale Hazardmodelle nach Cox

wurden genutzt um die Assoziationen der Muster-Scores mit Zusammenfassung dem Risiko für schwere chronische Erkrankungen zu untersuchen.

In beiden Geschlechtern identifizierte GGM ein Hauptnetz, genannt Haupteinnahme-Netzwerk, und mehrere kleinere Netzwerke. Das Netzwerk für Männer beinhaltete 12 Nahrungsmittelgruppen, welche um rotes Fleisch und gekochtes Gemüse gruppiert sind und einige kleinere Netzwerke wie ein Molkereinetzwerk, süßes Lebensmittelnetzwerk, frisches Obst u. Gemüsenetzwerk und Frühstücksgetreidenetzwerk. Bei den Frauen bestand das Hauptnetzwerk aus den gleichen Nahrungsmittelgruppen, wie bei den Männern sowie Bratkartoffeln. Ein wesentlicher Unterschied zwischen den Mustern bei Männern und Frauen war die Aufnahmebeziehung zwischen Nahrungsmittelgruppen in den identifizierten Netzwerken. Zum Beispiel war die Aufnahme von rotem Fleisch bei Männern mit der Einnahme von fünf Nahrungsmittelgruppen verbunden, aber bei den Frauen mit sieben Nahrungsmittelgruppen.

In Bezug auf das Risiko von chronischen Krankheiten war das Hauptmuster, welches eine höhere Aufnahme von Lebensmitteln wie rotem Fleisch, verarbeitetem Fleisch, Sauce und Weißbrot und niedrigen Einnahmen von Vollkornbrot beinhaltete, mit dem Risiko von T2D, kardiometabolischen Erkrankungen und chronischen Gesamterkrankungen bei Frauen assoziiert, aber nicht bei Männern. Jedoch war ein ähnliches Muster, welches mittels PCA identifiziert wurde, nicht mit dem Risiko einer größeren chronischen Erkrankung bei beiden Geschlechtern verbunden. Ein durch GGM identifiziertes Muster, bestehend aus hohem Molkereifettanteil, war invers verbunden mit dem Risiko von T2D, kardiometabolischen Krankheiten und chronischen Gesamterkrankungen bei Männern und dem Risiko von T2D bei Frauen. Ein ähnliches Muster, das durch PCA identifiziert wurde, war auch mit einem niedrigeren Risiko von T2D verbunden, aber nur bei Männern. Ein durch GGM identifiziertes Obst- und Gemüsemuster war invers verbunden mit dem Risiko von kardiometabolischen Erkrankungen bei Männern, während ein ähnliches Muster identifiziert mittels PCA, bei Männern invers assoziiert war mit dem Risiko von Schlaganfall und kardiometabolischen Krankheiten.

Ein durch GGM identifiziertes Muster aus Cerealien zum Frühstück war nicht verbunden mit dem Risiko von chronischen Erkrankungen in beiden Geschlechtern.

Jedoch ein mittels PCA identifiziertes Cerealien Muster (welches mehr Lebensmittel Zusammenfassung beinhaltete als das GGM Muster) war invers verbunden mit dem Risiko von T2D, kardiometabolischen Erkrankungen und chronischen Gesamterkrankungen bei Männern. Jedoch war das PCA-Getreidemuster (einschließlich mehr Nahrung als das GGM-Muster) umgekehrt mit dem Risiko von T2D, kardiometabolischen Erkrankungen und chronischen Gesamterkrankungen bei Männern assoziiert. PCA identifizierte auch ein "Brot & Wurst" -Muster bei Frauen, welches positiv mit dem Risiko von T2D, kardiometabolischen Erkrankungen und chronischen Gesamterkrankungen verbunden war. Die Bewertung der einzelnen Lebensmittelgruppen in den Netzwerken zeigte, dass rotes Fleisch, verarbeitetes Fleisch, Sauce und Suppe positiv sowie Vollkornbrot, rohes Gemüse, fettreiche Milchprodukte und Müsli negativ mit dem Risiko für schwere chronische Erkrankungen assoziiert waren.

Zusammenfassend kann gesagt werden, dass GGM ein vielversprechendes Konzept der Ernährungs-Muster-Analyse darstellt, um zu untersuchen wie Lebensmittel in Beziehung zueinander in verschiedenen Bevölkerungsgruppen konsumiert werden. Es kann leicht zu interpretieren Muster identifizieren, wo jedes Nahrungsmittel Teil von nur einem bestimmten Muster ist. Diese Muster können auch mit dem Risiko von chronischen Krankheiten in Verbindung gebracht werden und können verwendet werden, um die Assoziation von Nahrungsmittelgruppen mit Erkrankungsendpunkten zu untersuchen. Weitere Studien sind erforderlich, um die Gültigkeit dieses Ansatzes in anderen Populationen zu bestätigen. Informationen in Bezug auf die Nahrungsaufnahme und das Risiko von chronischen Krankheiten, die mit diesem Ansatz gewonnen werden, können für Ernährungsempfehlungen nützlich sein, um das Risiko von chronischen Erkrankungen zu senken.

Bibliography

1. Popkin BM. Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases. The American journal of clinical nutrition. 2006 Aug;84(2):289-98. PubMed PMID: 16895874. Epub 2006/08/10. Eng.
2. U.S. Department of Health and Human Services and U.S. Department of AgricultureDietary Guidelines for Americans 2015-2020. 8th Edition. Published 2016. http://health.gov/dietaryguidelines/2015/guidelines/. Accessed January 2016: .
3. Cardenas D. Let not thy food be confused with thy medicine: The Hippocratic misquotation. e-SPEN Journal. 2013 12//;8(6):e260-e2.
4. Jouanna J. Hippocrate, Paris, Fayard, 1992. et V Langholf,«Nachrichten bei Platon über die Kommunikation zwischen Aerzten und Patienten», in R Wittern et P Pellegrin, Hippokratische Medizin und antike Philosophie, Hildesheim, Olms. 1996:113-42.
5. Easton J. Human Longevity: Recording the Name, Age, Place of Residence, and Year, of the Decease of 1712 Persons, who Attained a Century, & Upwards, from AD 66 to 1799,... By James Easton: James Easton; sold also by John White, London; 1799.
6. Diet, Nutrition, and Health: McGill-Queen's University Press; 1989.
7. Edelstein S. Nutrition in public health: Jones & Bartlett Publishers; 2010.
8. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Current opinion in lipidology. 2002 Feb;13(1):3-9. PubMed PMID: 11790957. Epub 2002/01/16. Eng.
9. Wirfalt AK, Jeffery RW. Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. Journal of the American Dietetic Association. 1997 Mar;97(3):272-9. PubMed PMID: 9060944. Epub 1997/03/01. Eng.
10. Willett WC. Diet and health: what should we eat? Science (New York, NY). 1994 Apr 22;264(5158):532-7. PubMed PMID: 8160011. Epub 1994/04/22. Eng.
11. Varraso R, Garcia-Aymerich J, Monier F, Le Moual N, De Batlle J, Miranda G, et al. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. The American journal of clinical nutrition. 2012 November 1, 2012;96(5):1079-92.

12. Michels KB, Schulze MB. Can dietary patterns help us detect diet-disease associations? Nutrition research reviews. 2005 Dec;18(2):241-8. PubMed PMID: 19079908. Epub 2005/12/01. eng.
13. Rasmussen MA, Bro R. A tutorial on the Lasso approach to sparse modeling. Chemometrics and Intelligent Laboratory Systems. 2012 10/1/;119(0):21-31.
14. Iqbal K, Buijsse B, Wirth J, Schulze MB, Floegel A, Boeing H. Gaussian Graphical Models Identify Networks of Dietary Intake in a German Adult Population. The Journal of nutrition. 2016;146(3):646-52.
15. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, et al. Dietary Patterns: Challenges and Opportunities in Dietary Patterns Research: An Experimental Biology Workshop, April 1, 2006. Journal of the American Dietetic Association. 2007 7//;107(7):1233-9.
16. Kant AK. Dietary patterns and health outcomes. Journal of the American Dietetic Association. 2004 Apr;104(4):615-35. PubMed PMID: 15054348. Epub 2004/04/01. Eng.
17. Martinez ME, Marshall JR, Sechrest L. Invited commentary: Factor analysis and the search for objectivity. American journal of epidemiology. 1998 Jul 1;148(1):17-9. PubMed PMID: 9663398. Epub 1998/07/15. eng.
18. Schulze MB, Hoffmann K. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. The British journal of nutrition. 2006 May;95(5):860-9. PubMed PMID: 16611375. Epub 2006/04/14. eng.
19. Hodge A, Bassett J. What can we learn from dietary pattern analysis? Public health nutrition. 2016 Feb;19(2):191-4. PubMed PMID: 26784585. Epub 2016/01/20. eng.
20. Lauritzen S. Graphical Models New York Clarendon Press, Oxford University Press; 1996.
21. Villers F, Schaeffer B, Bertin C, Huet S. Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. Statistical applications in genetics and molecular biology. 2008;7(1):Article 14. PubMed PMID: 18976229. Epub 2008/11/04. eng.
22. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004 7//;90(1):196-212.
23. Talluri R, Shete S. Gaussian graphical models for phenotypes using pedigree data

and exploratory analysis using networks with genetic and nongenetic factors based on Genetic Analysis Workshop 18 data. BMC Proceedings. 2014;8(Suppl 1):S99. PubMed PMID: doi:10.1186/1753-6561-8-S1-S99.

24. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC systems biology. 2011;5:21. PubMed PMID: 21281499. Pubmed Central PMCID: PMC3224437. Epub 2011/02/02. eng.
25. Floegel A, Wientzek A, Bachlechner U, Jacobs S, Drogan D, Prehn C, et al. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. International journal of obesity (2005). 2014 Mar 10. PubMed PMID: 24608922. Epub 2014/03/13. Eng.
26. Hearty AP, Gibney MJ. Analysis of meal patterns with the use of supervised data mining techniques--artificial neural networks and decision trees. The American journal of clinical nutrition. 2008 Dec;88(6):1632-42. PubMed PMID: 19064525. Epub 2008/12/10. Eng.
27. Jacques PF, Tucker KL. Are dietary patterns useful for understanding the role of diet in chronic disease? The American journal of clinical nutrition. 2001 Jan;73(1):1-2. PubMed PMID: 11124739. Epub 2000/12/22. Eng.
28. Cespedes EM, Hu FB. Dietary patterns: from nutritional epidemiologic analysis to national guidelines. The American journal of clinical nutrition. 2015 May;101(5):899-900. PubMed PMID: 25832336. Pubmed Central PMCID: PMC4409695. Epub 2015/04/03. Eng.
29. Osler M, Helms Andreasen A, Heitmann B, Hoidrup S, Gerdes U, Morch Jorgensen L, et al. Food intake patterns and risk of coronary heart disease: a prospective cohort study examining the use of traditional scoring techniques. European journal of clinical nutrition. 2002 Jul;56(7):568-74. PubMed PMID: 12080395. Epub 2002/06/25. eng.
30. Cardiovascular disease, chronic kidney disease, and diabetes mortality burden of cardiometabolic risk factors from 1980 to 2010: a comparative risk assessment. The lancet Diabetes & endocrinology. 2014 Aug;2(8):634-47. PubMed PMID: 24842598. Pubmed Central PMCID: PMC4572741. Epub 2014/05/21. Eng.
31. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global

Burden of Disease Study 2015. Lancet (London, England). 2016 Oct 8;388(10053):1459-544. PubMed PMID: 27733281. Epub 2016/10/14. Eng.

32. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet (London, England). 2016 Apr 9;387(10027):1513-30. PubMed PMID: 27061677. Epub 2016/04/12. eng.
33. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet (London, England). 2012 Dec 15;380(9859):2224-60. PubMed PMID: 23245609. Pubmed Central PMCID: PMC4156511. Epub 2012/12/19. eng.
34. Forouzanfar MH, Alexander L, Anderson HR, Bachman VF, Biryukov S, Brauer M, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet (London, England). 2015 Dec 05;386(10010):2287-323. PubMed PMID: 26364544. Pubmed Central PMCID: PMC4685753. Epub 2015/09/15. eng.
35. Imamura F, Jacques PF. Invited commentary: dietary pattern analysis. American journal of epidemiology. 2011 May 15;173(10):1105-8; discussion 9-10. PubMed PMID: 21474588. Epub 2011/04/09. eng.
36. Slattery ML. Defining dietary consumption: is the sum greater than its parts? The American journal of clinical nutrition. 2008 Jul;88(1):14-5. PubMed PMID: 18614718. Pubmed Central PMCID: PMC2925519. Epub 2008/07/11. Eng.
37. Thurstone LL. Multiple factor analysis. Psychological Review. 1931;38(5):406.
38. Gorsuch RL. Factor Analysis. Second ed. Hillsdale, NJ:Lawrence Erlbaum Associates1983.
39. O'Rourke N, Hatcher L. A step-by-step approach to using SAS for factor analysis and structural equation modeling: Sas Institute; 2013.
40. Yong AG, Pearce S. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in Quantitative Methods for Psychology. 2013;9(2):79-94.
41. Kim J-O, Mueller CW. Factor analysis: Statistical methods and practical issues: Sage; 1978.
42. Harman HH. Modern factor analysis: University of Chicago Press; 1976.

43. Kline P. An Easy Guide to Factor Analysis: Psychology Press; 1994.
44. Suhr DD. Exploratory or confirmatory factor analysis?: SAS Institute Cary; 2006.
45. Kaiser HF. The application of electronic computers to factor analysis. Educational and psychological measurement. 1960.
46. Cattell RB. The scree test for the number of factors. Multivariate behavioral research. 1966;1(2):245-76.
47. Tabachnick B, Fidell L. Using Multivariate Statistics., 5th edn.(Pearson: Boston, MA.). 2007.
48. Beavers AS, Lounsbury JW, Richards JK, Huck SW, Skolits GJ, Esquivel SL. Practical considerations for using exploratory factor analysis in educational research. Practical assessment, research & evaluation. 2013;18(6):1-13.
49. Johnson RA, Wichern DW. Applied multivariate statistical analysis: Prentice hall Upper Saddle River, NJ; 2002.
50. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, et al. Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. Journal of the American Dietetic Association. 2007 Jul;107(7):1233-9. PubMed PMID: 17604756. Epub 2007/07/03. eng.
51. Reedy J, Wirfält E, Flood A, Mitrou PN, Krebs-Smith SM, Kipnis V, et al. Comparing 3 Dietary Pattern Methods—Cluster Analysis, Factor Analysis, and Index Analysis—With Colorectal Cancer Risk: The NIH–AARP Diet and Health Study. American journal of epidemiology. 2010 12/21 06/15/received 11/04/accepted;171(4):479-87. PubMed PMID: PMC2842201.
52. Tan P-N, Steinbach M, Kumar V. Intro. to Data Mining. Michigan State University and University of Minnesota. 2006:207-23.
53. Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. American journal of epidemiology. 2004 May 15, 2004;159(10):935-44.
54. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K. Exploring dietary patterns by using the treelet transform. American journal of epidemiology. 2011 May 15;173(10):1097-104. PubMed PMID: 21474587. Epub 2011/04/09. eng.
55. Strobl R, Grill E, Mansmann U. Graphical modeling of binary data using the LASSO: a simulation study. BMC medical research methodology. 2012;12:16. PubMed PMID: 22353192. Pubmed Central PMCID: PMC3305667. Epub

2012/02/23. eng.

56. Kramer N, Schafer J, Boulesteix AL. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC bioinformatics. 2009;10:384. PubMed PMID: 19930695. Pubmed Central PMCID: PMC2808166. Epub 2009/11/26. eng.
57. Lauritzen sl. Graphical Models. Oxford University Press; 1996.
58. Edwards D. A Brief Introduction to Graphical Models. Internet: https://djfextranet.agrsci.dk/sites/phd_course2_2010/public/Documents/ABriefIntro.pdf. (Accessed 17.03.2014). 2011.
59. Mazumder R, Hastie T. The graphical lasso: New insights and alternatives. 2012 2012:2125-49. en.
60. Drton M, Perlman MD. Multiple Testing and Error Control in Gaussian Graphical Model Selection. 2007 2007/08:430-49. en.
61. Dempster AP. Covariance selection. Biometrics. 1972:157-75.
62. Wermuth DRCaN. Multivariate Dependencies. Models, Analysis and Interpretation: Chapman & Hall, London; 1996. 272. p.
63. J. Honorio DS, I. Rish, and G. Cecchi. Variable Selection for Gaussian Graphical Models. Journal of Machine Learning Research - Workshop and Conference Proceedings. 2012;22:538-46.
64. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007 March 1, 2007;94(1):19-35.
65. Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the Lasso. 2006 2006/06:1436-62. en.
66. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England). 2008 July 1, 2008;9(3):432-41.
67. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England). 2008 Jul;9(3):432-41. PubMed PMID: 18079126. Pubmed Central PMCID: PMC3019769. Epub 2007/12/15. eng.
68. LIU H, HAN ,F.,YUAN ,M., AFFERTY ,J.D.& WASSERMAN ,L.A. High-dimensional semiparametric Gaussian copula graphical models. Ann Statistics. 2012;40, 2293–326.
69. Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research.

2009;10(Oct):2295-328.

70. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. International journal of epidemiology. 1997;26 Suppl 1:S6-14. PubMed PMID: 9126529. Epub 1997/01/01. Eng.
71. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Annals of nutrition & metabolism. 1999;43(4):205-15. PubMed PMID: 10592369. Epub 1999/12/11. Eng.
72. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Annals of oncology : official journal of the European Society for Medical Oncology. 1992 Dec;3(10):783-91. PubMed PMID: 1286041. Epub 1992/12/01. Eng.
73. Boeing H, Wahrendorf J, Becker N. EPIC-Germany--A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. Annals of nutrition & metabolism. 1999;43(4):195-204. PubMed PMID: 10592368. Epub 1999/12/11. Eng.
74. Kroke A, Klipstein-Grobusch K, Voss S, Moseneder J, Thielecke F, Noack R, et al. Validation of a self-administered food-frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. The American journal of clinical nutrition. 1999 Oct;70(4):439-47. PubMed PMID: 10500011. Epub 1999/09/29. eng.
75. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J. Reproducibility and relative validity of food group intake in a food frequency questionnaire developed for the German part of the EPIC project. European Prospective Investigation into Cancer and Nutrition. International journal of epidemiology. 1997;26 Suppl 1:S59-70. PubMed PMID: 9126534. Epub 1997/01/01. Eng.
76. Voss S, Charrondiere UR, Slimani N, Kroke A, Riboli E, Wahrendorf J, et al. [EPIC-SOFT a European computer program for 24-hour dietary protocols]. Zeitschrift fur Ernahrungswissenschaft. 1998 Sep;37(3):227-33. PubMed PMID: 9800313. Epub 1998/11/04. EPIC-SOFT ein europaisches Computerprogramm fur 24-Stunden-Erinnerungsprotokolle. Ger.

77. Dehne LI, Klemm C, Henseler G, Hermann-Kunz E. The German Food Code and Nutrient Data Base (BLS II.2). European journal of epidemiology. 1999 Apr;15(4):355-9. PubMed PMID: 10414376. Epub 1999/07/22. Eng.
78. Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, Hennings S, et al. Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public health nutrition. 2003 Jun;6(4):407-13. PubMed PMID: 12795830. Epub 2003/06/11. eng.
79. Mifflin MD, St Jeor ST, Hill LA, Scott BJ, Daugherty SA, Koh YO. A new predictive equation for resting energy expenditure in healthy individuals. The American journal of clinical nutrition. 1990 Feb;51(2):241-7. PubMed PMID: 2305711. Epub 1990/02/01. eng.
80. Bergmann MM, Bussas U, Boeing H. Follow-up procedures in EPIC-Germany--data quality aspects. European Prospective Investigation into Cancer and Nutrition. Annals of nutrition & metabolism. 1999;43(4):225-34. PubMed PMID: 10592371. Epub 1999/12/11. Eng.
81. Organisation T. International Classification of Diseases (ICD)-10. 2010.
82. Højsgaard S ED, Lauritzen S. Graphical Models with R. Gentleman R HK, Parmigiani GG, editor. New York: Springer-Verlag; 2012.
83. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge package for high-dimensional undirected graph estimation in R. J Mach Learn Res. 2012;13(1):1059-62.
84. GmbH y. yEd Software. http://www.yworks.com. 3.10.1 ed2000-2015.
85. David CR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society. 1972;34:187-220.
86. Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. American journal of epidemiology. 1986 Jul;124(1):17-27. PubMed PMID: 3521261. Epub 1986/07/01. Eng.
87. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer. 2003 07/2912/06/received 04/30/accepted;89(3):431-6. PubMed PMID: PMC2394368.
88. Borucka J. Assessment of Cox Proportional Hazard Model adequacy using Proc

Phreg and Proc Gplot. SAS Conference Proceedings: Pharmaceutical Users Software Exchange 2010; October 14-17, 2010; Berlin, Germany2010.

89. Schisterman EF, Perkins NJ, Mumford SL, Ahrens KA, Mitchell EM. Collinearity and causal diagrams - a lesson on the importance of model specification. Epidemiology (Cambridge, Mass). 2016 Aug 26. PubMed PMID: 27676260. Epub 2016/09/28. Eng.
90. Yoo W, Mayberry R, Bae S, Singh K, Peter He Q, Lillard JW, Jr. A Study of Effects of MultiCollinearity in the Multivariable Analysis. International journal of applied science and technology. 2014 Oct;4(5):9-19. PubMed PMID: 25664257. Pubmed Central PMCID: PMC4318006. Epub 2015/02/11. Eng.
91. Schulze MB, Hoffmann K, Kroke A, Boeing H. Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. The British journal of nutrition. 2001 Mar;85(3):363-73. PubMed PMID: 11299082. Epub 2001/04/12. eng.
92. Oberritter H, Schäbethal K, von Ruesten A, Boeing H. The DGE nutrition circle—Presentation and basis of the food-related recommendations from the German Nutrition Society (DGE). Ernährungs-Umschau Int. 2013;2:24-9.
93. Everitt BS, Palmer C. Encyclopaedic companion to medical statistics: John Wiley & Sons; 2011.
94. Müller MJ, Trautwein EA. Gesundheit und Ernährung-public health nutrition: Ulmer; 2005.
95. Silva Junior SH, Santos SM, Coeli CM, Carvalho MS. Assessment of participation bias in cohort studies: systematic review and meta-regression analysis. Cadernos de saude publica. 2015 Nov;31(11):2259-74. PubMed PMID: 26840808. Epub 2016/02/04. Eng.
96. Schneider R. Welche Methoden gibt es Ernährungsinformationen zu ermitteln. Vom Umgang mit Zahlen und Daten: Eine praxisnahe Einführung in Statistik und Ernährungsepidemiologie Frankfurt a Main, Umschau Zeitschriftenverlag. 1997:101-32.
97. von Ruesten A, Steffen A, Floegel A, van der AD, Masala G, Tjonneland A, et al. Trend in obesity prevalence in European adult cohort populations during follow-up since 1996 and their predictions to 2015. PloS one. 2011;6(11):e27455. PubMed PMID: 22102897. Pubmed Central PMCID: PMC3213129. Epub 2011/11/22. Eng.

98. Slattery ML. Defining dietary consumption: is the sum greater than its parts? The American journal of clinical nutrition. 2008 July 1, 2008;88(1):14-5.
99. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K. Gorst-Rasmussen et al. Respond to "Dietary Pattern Analysis". American journal of epidemiology. 2011 May 15, 2011;173(10):1109-10.
100. Assi N, Moskal A, Slimani N, Viallon V, Chajes V, Freisling H, et al. A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). Public health nutrition. 2015 Feb 23:1-13. PubMed PMID: 25702596. Epub 2015/02/24. Eng.
101. von Ruesten A, Feller S, Bergmann MM, Boeing H. Diet and risk of chronic diseases: results from the first 8 years of follow-up in the EPIC-Potsdam study. European journal of clinical nutrition. 2013 Apr;67(4):412-9. PubMed PMID: 23388667. Epub 2013/02/08. eng.
102. Ma S, Gong Q, Bohnert HJ. An Arabidopsis gene network based on the graphical Gaussian model. Genome research. 2007 Nov;17(11):1614-25. PubMed PMID: 17921353. Pubmed Central PMCID: PMC2045144. Epub 2007/10/09. eng.
103. Zerenner T, Friederichs P, Lehnertz K, Hense A. A Gaussian graphical model approach to climate networks. Chaos (Woodbury, NY). 2014 Jun;24(2):023103. PubMed PMID: 24985417. Epub 2014/07/06. eng.
104. Li F, An S, Hou L, Chen P, Lei C, Tan W. Red and processed meat intake and risk of bladder cancer: a meta-analysis. International journal of clinical and experimental medicine. 2014;7(8):2100-10. PubMed PMID: 25232394. Pubmed Central PMCID: PMC4161554. Epub 2014/09/19. eng.
105. Abete I, Romaguera D, Vieira AR, Lopez de Munain A, Norat T. Association between total, processed, red and white meat consumption and all-cause, CVD and IHD mortality: a meta-analysis of cohort studies. The British journal of nutrition. 2014 Sep 14;112(5):762-75. PubMed PMID: 24932617. Epub 2014/06/17. eng.
106. Larsson SC, Orsini N. Red meat and processed meat consumption and all-cause mortality: a meta-analysis. American journal of epidemiology. 2014 Feb 1;179(3):282-9. PubMed PMID: 24148709. Epub 2013/10/24. eng.
107. CANCER IAFRO. Report of the Advisory Group to Recommend Priorities for IARC Monographs during 2015–2019. LYON, FRANCE: 2014 18–19, April 2014. Report

No.

108. Shadman Z, Akhoundan M, Poorsoltan N, Larijani B, Qorbani M, Nikoo MK. New challenges in dietary pattern analysis: combined dietary patterns and calorie adjusted factor analysis in type 2 diabetic patients. Journal of diabetes and metabolic disorders. 2014;13:71. PubMed PMID: 25032128. Pubmed Central PMCID: PMC4100028. Epub 2014/07/18. eng.
109. Fung TT, Schulze M, Manson JE, Willett WC, Hu FB. Dietary patterns, meat intake, and the risk of type 2 diabetes in women. Archives of internal medicine. 2004 Nov 08;164(20):2235-40. PubMed PMID: 15534160. Epub 2004/11/10. eng.
110. Pastorino S, Richards M, Pierce M, Ambrosini GL. A high-fat, high-glycaemic index, low-fibre dietary pattern is prospectively associated with type 2 diabetes in a British birth cohort. The British journal of nutrition. 2016 May;115(9):1632-42. PubMed PMID: 27245103. Pubmed Central PMCID: PMC4907349. Epub 2016/06/02. eng.
111. van Dam RM, Rimm EB, Willett WC, Stampfer MJ, Hu FB. Dietary patterns and risk for type 2 diabetes mellitus in U.S. men. Annals of internal medicine. 2002 Feb 05;136(3):201-9. PubMed PMID: 11827496. Epub 2002/02/06. eng.
112. Montonen J, Knekt P, Härkänen T, Järvinen R, Heliövaara M, Aromaa A, et al. Dietary Patterns and the Incidence of Type 2 Diabetes. American journal of epidemiology. 2005 February 1, 2005;161(3):219-27.
113. Micha R, Wallace SK, Mozaffarian D. Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, Stroke, and Diabetes Mellitus. A Systematic Review and Meta-Analysis. 2010;121(21):2271-83.
114. Halton TL, Willett WC, Liu S, Manson JE, Stampfer MJ, Hu FB. Potato and french fry consumption and risk of type 2 diabetes in women. The American journal of clinical nutrition. 2006 Feb;83(2):284-90. PubMed PMID: 16469985. Epub 2006/02/14. eng.
115. de Munter JSL, Hu FB, Spiegelman D, Franz M, van Dam RM. Whole Grain, Bran, and Germ Intake and Risk of Type 2 Diabetes: A Prospective Cohort Study and Systematic Review. PLoS Medicine. 2007 08/28/received 07/17/accepted;4(8):e261. PubMed PMID: PMC1952203.
116. Aune D, Norat T, Romundstad P, Vatten LJ. Whole grain and refined grain consumption and the risk of type 2 diabetes: a systematic review and dose-response

meta-analysis of cohort studies. European journal of epidemiology. 2013 Nov;28(11):845-58. PubMed PMID: 24158434. Epub 2013/10/26. eng.

117. Ericson U, Hellstrand S, Brunkwall L, Schulz CA, Sonestedt E, Wallstrom P, et al. Food sources of fat may clarify the inconsistent role of dietary fat intake for incidence of type 2 diabetes. The American journal of clinical nutrition. 2015 May;101(5):1065-80. PubMed PMID: 25832335. Epub 2015/04/03. eng.
118. Diaz-Lopez A, Bullo M, Martinez-Gonzalez MA, Corella D, Estruch R, Fito M, et al. Dairy product consumption and risk of type 2 diabetes in an elderly Spanish Mediterranean population at high cardiovascular risk. European journal of nutrition. 2016 Feb;55(1):349-60. PubMed PMID: 25663611. Epub 2015/02/11. eng.
119. Kirii K, Mizoue T, Iso H, Takahashi Y, Kato M, Inoue M, et al. Calcium, vitamin D and dairy intake in relation to type 2 diabetes risk in a Japanese cohort. Diabetologia. 2009 Dec;52(12):2542-50. PubMed PMID: 19823801. Epub 2009/10/14. eng.
120. Eussen SJPM, van Dongen MCJM, Wijckmans N, den Biggelaar L, Oude Elferink SJWH, Singh-Povel CM, et al. Consumption of dairy foods in relation to impaired glucose metabolism and type 2 diabetes mellitus: the Maastricht Study. British Journal of Nutrition. 2016 2016/004/28;115(8):1453-61.
121. Tong X, Dong JY, Wu ZW, Li W, Qin LQ. Dairy consumption and risk of type 2 diabetes mellitus: a meta-analysis of cohort studies. European journal of clinical nutrition. 2011 Sep;65(9):1027-31. PubMed PMID: 21559046. Epub 2011/05/12. eng.
122. Elwood PC, Pickering JE, Givens DI, Gallacher JE. The consumption of milk and dairy foods and the incidence of vascular disease and diabetes: an overview of the evidence. Lipids. 2010 Oct;45(10):925-39. PubMed PMID: 20397059. Pubmed Central PMCID: PMC2950929. Epub 2010/04/17. eng.
123. Elwood PC, Givens DI, Beswick AD, Fehily AM, Pickering JE, Gallacher J. The survival advantage of milk and dairy consumption: an overview of evidence from cohort studies of vascular diseases, diabetes and cancer. Journal of the American College of Nutrition. 2008 Dec;27(6):723s-34s. PubMed PMID: 19155432. Epub 2009/01/22. eng.
124. Schwab U, Lauritzen L, Tholstrup T, Haldorssoni T, Riserus U, Uusitupa M, et al. Effect of the amount and type of dietary fat on cardiometabolic risk factors and risk of developing type 2 diabetes, cardiovascular diseases, and cancer: a systematic

review. Food & nutrition research. 2014;58. PubMed PMID: 25045347. Pubmed Central PMCID: PMC4095759. Epub 2014/07/22. eng.

125. de Goede J, Soedamah-Muthu SS, Pan A, Gijsbers L, Geleijnse JM. Dairy Consumption and Risk of Stroke: A Systematic Review and Updated Dose-Response Meta-Analysis of Prospective Cohort Studies. Journal of the American Heart Association. 2016 May 20;5(5). PubMed PMID: 27207960. Pubmed Central PMCID: PMC4889169. Epub 2016/05/22. eng.
126. Larsson SC, Männistö S, Virtanen MJ, Kontto J, Albanes D, Virtamo J. Dairy Foods and Risk of Stroke. Epidemiology (Cambridge, Mass). 2009;20(3):355-60. PubMed PMID: PMC3498757.
127. Panagiotakos D, Pitsavos C, Chrysohoou C, Palliou K, Lentzas I, Skoumas I, et al. Dietary patterns and 5-year incidence of cardiovascular disease: a multivariate analysis of the ATTICA study. Nutrition, metabolism, and cardiovascular diseases : NMCD. 2009 May;19(4):253-63. PubMed PMID: 18722096. Epub 2008/08/30. eng.
128. Brunner EJ, Mosdol A, Witte DR, Martikainen P, Stafford M, Shipley MJ, et al. Dietary patterns and 15-y risks of major coronary events, diabetes, and mortality. The American journal of clinical nutrition. 2008 May;87(5):1414-21. PubMed PMID: 18469266. Epub 2008/05/13. eng.
129. Steffen LM, Van Horn L, Daviglus ML, Zhou X, Reis JP, Loria CM, et al. A modified Mediterranean diet score is associated with a lower risk of incident metabolic syndrome over 25 years among young adults: the CARDIA (Coronary Artery Risk Development in Young Adults) study. The British journal of nutrition. 2014 Nov 28;112(10):1654-61. PubMed PMID: 25234439. Epub 2014/09/23. eng.
130. Bazzano LA, He J, Ogden LG, Loria CM, Vupputuri S, Myers L, et al. Fruit and vegetable intake and risk of cardiovascular disease in US adults: the first National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. The American journal of clinical nutrition. 2002 Jul;76(1):93-9. PubMed PMID: 12081821. Epub 2002/06/26. eng.
131. Gottschald M, Knuppel S, Boeing H, Buijsse B. The influence of adjustment for energy misreporting on relations of cake and cookie intake with cardiometabolic disease risk factors. European journal of clinical nutrition. 2016 Nov;70(11):1318-24. PubMed PMID: 27460264. Epub 2016/11/03. eng.
132. Flight I, Clifton P. Cereal grains and legumes in the prevention of coronary heart

disease and stroke: a review of the literature. European journal of clinical nutrition. 2006 Oct;60(10):1145-59. PubMed PMID: 16670693. Epub 2006/05/04. eng.

133. Aune D, Keum N, Giovannucci E, Fadnes LT, Boffetta P, Greenwood DC, et al. Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: systematic review and dose-response meta-analysis of prospective studies. BMJ (Clinical research ed). 2016 Jun 14;353:i2716. PubMed PMID: 27301975. Pubmed Central PMCID: PMC4908315. Epub 2016/06/16. eng.
134. Wu H, Flint AJ, Qi Q, van Dam RM, Sampson LA, Rimm EB, et al. Association between dietary whole grain intake and risk of mortality: two large prospective studies in US men and women. JAMA internal medicine. 2015 Mar;175(3):373-84. PubMed PMID: 25559238. Pubmed Central PMCID: PMC4429593. Epub 2015/01/07. eng.
135. Steffen LM, Jacobs DR, Jr., Stevens J, Shahar E, Carithers T, Folsom AR. Associations of whole-grain, refined-grain, and fruit and vegetable consumption with risks of all-cause mortality and incident coronary artery disease and ischemic stroke: the Atherosclerosis Risk in Communities (ARIC) Study. The American journal of clinical nutrition. 2003 Sep;78(3):383-90. PubMed PMID: 12936919. Epub 2003/08/26. eng.
136. Wollin SD, Jones PJH. Alcohol, Red Wine and Cardiovascular Disease. The Journal of nutrition. 2001 May 1, 2001;131(5):1401-4.
137. Chiva-Blanch G, Arranz S, Lamuela-Raventos RM, Estruch R. Effects of wine, alcohol and polyphenols on cardiovascular disease risk factors: evidences from human studies. Alcohol and alcoholism (Oxford, Oxfordshire). 2013 May-Jun;48(3):270-7. PubMed PMID: 23408240. Epub 2013/02/15. eng.
138. Hartley L, Flowers N, Holmes J, Clarke A, Stranges S, Hooper L, et al. PP10 Green and Black Tea for the Primary Prevention of Cardiovascular Disease (CVD): A Cochrane Systematic Review. Journal of Epidemiology and Community Health. 2013;67(Suppl 1):A52-A3.
139. Nancy Santesso, Eric Manheimer. A Summary of a Cochrane Review: Green and Black Tea for the Primary Prevention of Cardiovascular Disease. Global Advances in Health and Medicine. 2014;3(2):66-7.
140. McNaughton SA, Mishra GD, Brunner EJ. Dietary patterns, insulin resistance, and incidence of type 2 diabetes in the Whitehall II Study. Diabetes care. 2008

Jul;31(7):1343-8. PubMed PMID: 18390803. Pubmed Central PMCID: PMC2453656. Epub 2008/04/09. eng.

141. Fung TT, Willett WC, Stampfer MJ, Manson JE, Hu FB. Dietary patterns and the risk of coronary heart disease in women. Archives of internal medicine. 2001 Aug 13-27;161(15):1857-62. PubMed PMID: 11493127. Epub 2001/08/30. eng.
142. Muga MA, Owili PO, Hsu C-Y, Rau H-H, Chao JCJ. Association between Dietary Patterns and Cardiovascular Risk Factors among Middle-Aged and Elderly Adults in Taiwan: A Population-Based Study from 2003 to 2012. PloS one. 2016 07/01 12/07/received 06/04/accepted;11(7):e0157745. PubMed PMID: PMC4930186.
143. Osler M, Heitmann BL, Gerdes LU, Jorgensen LM, Schroll M. Dietary patterns and mortality in Danish men and women: a prospective observational study. The British journal of nutrition. 2001 Feb;85(2):219-25. PubMed PMID: 11242490. Epub 2001/03/10. eng.
144. Schulze MB, Hoffmann K, Kroke A, Boeing H. Risk of Hypertension among Women in the EPIC-Potsdam Study: Comparison of Relative Risk Estimates for Exploratory and Hypothesis-oriented Dietary Patterns. American journal of epidemiology. 2003 August 15, 2003;158(4):365-73.
145. Fung T, Hu FB, Fuchs C, Giovannucci E, Hunter DJ, Stampfer MJ, et al. Major dietary patterns and the risk of colorectal cancer in women. Archives of internal medicine. 2003 Feb 10;163(3):309-14. PubMed PMID: 12578511. Epub 2003/02/13. eng.
146. Wang T, Ren Z, Ding Y, Fang Z, Sun Z, MacDonald ML, et al. FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. PLoS Comput Biol. 2016;12(2):e1004755.
147. Leppä-aho J, Pensar J, Roos T, Corander J. Learning Gaussian graphical models with fractional marginal pseudo-likelihood. arXiv preprint arXiv:160207863. 2016.

Supplement

S1: Scoring of food networks

1. **Principal Network**

Men

Principal pattern = (**0**-whole grain bread) + refined bread + processed meat + red meat + poultry + sauce + potatoes + cooked vegetables + cabbage + mushrooms + legumes + soup;

Women

Principal pattern = (**0**-whole grain bread) + refined bread + processed meat + red meat + poultry + sauce + potatoes + cooked vegetables + cabbage + mushrooms + legumes + soup + fried potatoes;

2. **High Fat Dairy Network**

Men

Dairy pattern = (**0**-low fat cheese) + high fat cheese + (**0**-low fat dairy products) + high fat dairy products;

Women

Dairy pattern = (**0**-low fat cheese) + high fat cheese + (**0**-low fat dairy products) + high fat dairy products + butter + (**0**-margarine);

3. **Fresh Fruit & Vegetables Networks**

Fruit & vegetable pattern = Raw vegetables + Fresh vegetables + vegetable fat;

4. **Sweet Network**

Sweet pattern =Sweet bread spread + cakes & cookies + desserts + confectionary + canned fruits;

5. **Breakfast Cereals Network**

Breakfast cereal pattern = muesli + corn flakes + vegetarian dishes;

S2: Scree plots for selection of number of patterns in principal component analysis

Figure 1: Scree plots: a) Plot of Eigenvalue against number of factors. Right) Proportion of variance explained against number of factors (Men)

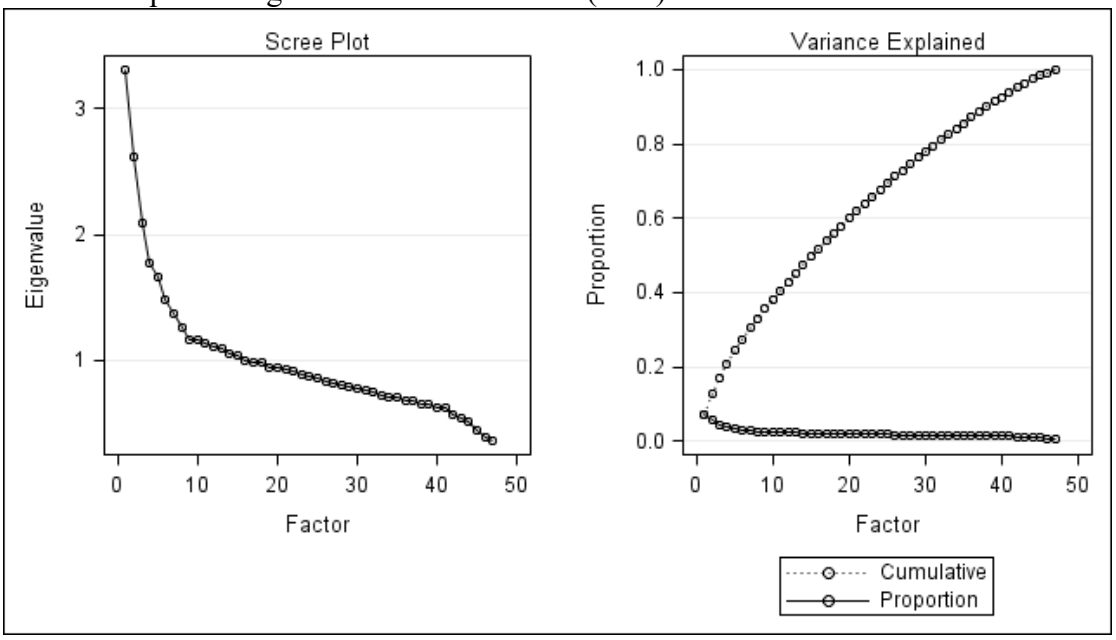
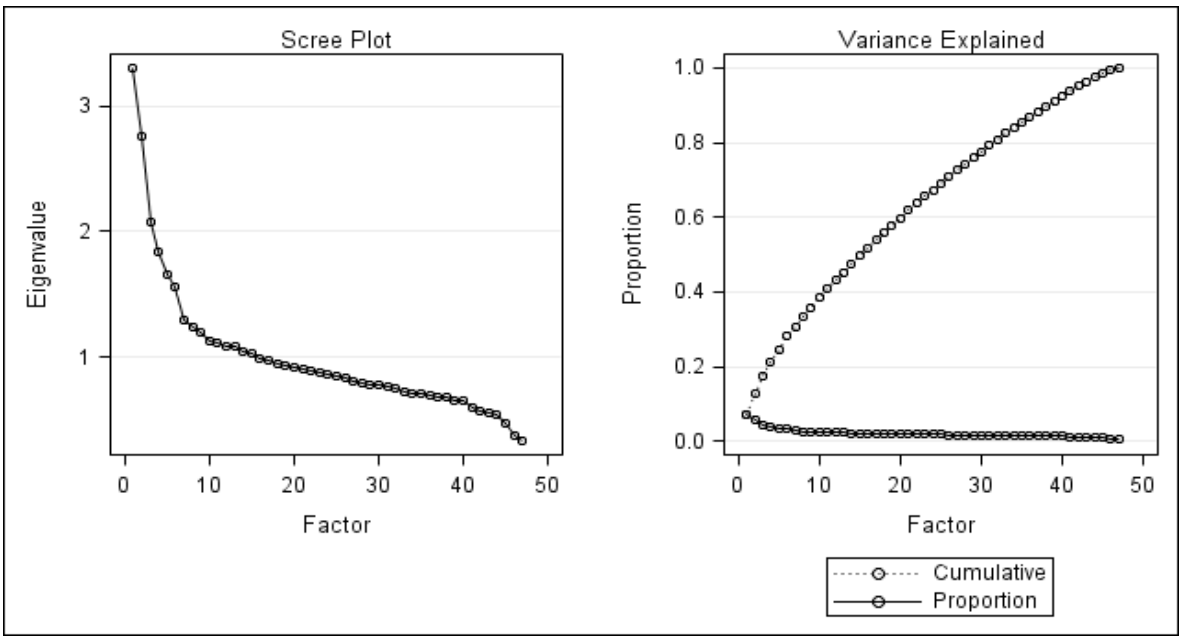


Figure 2: Scree plots: a) Plot of Eigenvalue against number of factors. Right) Proportion of variance explained against number of factors (Women)



S3: Food groups used to construct dietary intake patterns (networks)

Food or food group	Food items included in group
Whole-grain bread	Whole-grain bread, dark and whole-grain rolls
White (Other) bread	Rye bread, wheat bread, mixed bread, pale rolls, crisp bread, croissants
Muesli	Whole-grain, breakfast cereal, muesli
Corn flakes	Corn flakes, crisps
Pasta, rice	Cooked pasta, cooked rice
Vegetarian dishes	Vegetarian dishes, vegetarian spreads
Chips, salt sticks	Chips, flips, salt sticks, crackers
Pizza	Pizza, onion tart, quiche
Cake, cookies	Fruit cake, pound cake, sponge cake, cream cake, flan, brioches, pastries, sweet particles, biscuits, cookies, pancakes
Confectionary	Chocolate, candy bars, pralines, sugar in coffee and tea, ice cream
Sweet bread spreads	Jam, honey, chocolate spread, peanut butter
Eggs	Boiled eggs, fried eggs, omelettes
Fresh fruit	Apple, pear, peach, nectarine, cherry, plum, plums, grapes, strawberries, currants, raspberries, blackberries, bananas, kiwi, mango, fresh pineapple, orange, Grape fruit, mandarin
Canned fruit	Fruit compote, Canned fruit
Raw vegetables	Cucumber, radish, cabbage, carrots, seeds, sprouts, peppers, chilli pepper, tomato, raw onion, lettuce, endive, 107hinese cabbage, mixed salad
cabbage	Cauliflower, red cabbage, white cabbage, kohlrabi, broccoli, other cabbage
Cooked vegetables	Tomatoes, tomato sauce, sweet peppers, zucchini, eggplant (aubergine), spinach, carrots, asparagus, pea-carrot vegetable mix, leeks, and celery, broccoli, cauliflower, red and white cabbage, and kohlrabi, green peas, green beans, and pea/bean/lentil stew
Garlic	Raw or fried/cooked garlic (YES/NO answer)
Mushrooms	Fresh mushrooms, mushroom dishes
legumes	Green peas, green beans, lentil soup, pea soup, bean stew
Cooked potatoes	Salted potatoes, jacket potatoes, mashed potatoes, potato salad, dumplings
Fried potatoes	French fries, potato fritters, fried potatoes
Nuts	Nuts

Low-fat dairy products	Milk, dairy drinks, yoghurt, fruit yoghurt, sour milk, kefir, quark, herb quark (fat<=1.5%)
High-fat dairy products	Milk, dairy drinks, yoghurt, fruit yoghurt, quark, herb quark (fat>3.5% or no matter answer or fat>=20%), whipped cream
Low-fat cheese	Cream cheese, Gouda, Emmental, Tilsiter, Camembert, Brie, Gorgonzola (reduced-fat or skim stage)
high-fat cheese	Cream cheese, Gouda, Emmental, Tilsiter, Camembert, Brie, Gorgonzola, cheese (normal/creamfat, double fat / no matter answer)
Meat	Beef, veal, pork, lamb, venison, mixed ground meat, German beef roulade, beef goulash, roast pork, pork goulash, gyros, shashlik/meat skewer, hamburger/meatball
Poultry	Poultry, chicken/turkey ragout
Processed meat	Liverwurst, salami, mettwurst, cabanossi, bologna/polony, ham sausage, cooked ham, raw ham, poultry sausage, aspic, collared pork, blood sausage, frankfurter/wiener/hot dog, bratwurst, Bavarian veal sausage, Bavarian meat loaf, meat salad, other
Offals	Liver, other offals
Fish	Salmon, mackerel, herring, herring salad, salted herring, fried herring, rolled pickled herring, hot smoked herring, sprat, eel, redfish, trout, tuna, saithe/pollock, codfish, fish sticks, fish bake, calamari, craps/shellfish, other
Butter	Butter
Margarine	Margarine
Eggs	Egg salad, boiled egg, fried/scrambled eggs, other
Water	Tap water, mineral water
Coffee	Coffee with caffeine (black, with milk, with condensed milk, milk with sweetener)
Decaffeinated coffee	Decaffeinated coffee (black, with milk, with condensed milk, with sweetener)

Other non-alcoholic beverages	Mineral water/drinking water, herbal tea, fruit tea
Coffee	Coffee/espresso, coffee/espresso with (concentrated) milk, coffee without caffeine, coffee without caffeine with (concentrated) milk, cappuccino/caffè latte
Juice	Multi-vitamin juice, apple juice, orange juice, grape juice, grapefruit juice, elder juice, lemon juice, other fruit juice, tomato juice, other vegetable juice
Soft drinks	Lemonade, diet lemonade, cola, diet cola, non-alcoholic beer, malt beer, other
Tea	Black tea, black tea with (concentrated) milk, green tea, other hot drinks
Wine	White wine, red wine, rosé wine
Beer	Beer, strong beer/malt liquor
Spirits	Spirits
Other alcoholic beverages	Sparkling wine, beer shandy, liqueur, hot wine punch, wine spritzer, other mixed beverages, other
Sauces	Gravy with vegetables, gravy with meat/fish, gravy with side dish, salad dressing, mayonnaise, ketchup, other
Soup	Vegetable stew, stew with meat, clear soup, crème of vegetable soup, other

Acknowledgments

In the name of Allah who is the Most Benevolent, the Most Merciful and Who taught man by pen (Al-Quran). All praises to Almighty Allah who blessed me with the courage and strength to complete this work.

I am highly indebted to my supervisor, Professor Dr. Heiner Boeing, for giving me the opportunity to work on this interesting topic. He has been a source of knowledge and motivation since start of my studies and will continue through years. I learnt from him to be patient, keen, and innovative. He instilled in me the confidence to think beyond the borders of what we know or what we could know without fear and consequences. My contribution to science in this piece of writing would not have been possible without his guidance.

I am also very thankful to Professor Dr. Reinhard Busse for accepting me to Faculty VII and giving me the opportunity to submit this work.

My special thanks to Dr. Brian Buijsse who was very patient with my mistakes in early days of interaction with SAS. He showed me the way to epidemiology. Though he left and I could not get the benefit of his immense scholarship in later years of my PhD. His earlier support and guidance has shown me the path to follow and reach where I am today.

I also want to extend my sincere thanks to my new mentor Lukas Schwingshackl, who is an amazing person. I appreciate his passionate discussions, feedback on the texts and continued support to complete my work. I extend my sincere thanks to Dr Marta Stelmach-Mardas as well, who kept me on toe on various occasions during analysis and writing of my thesis. Her jests and lively conversations were a great source of motivation and refreshment. Also, thanks to my other many colleagues and friends who were always available to help out on anything they could do.

I am highly indebted to my brothers, who didn’t let me worry about anything while I was away from home. Their love and support were instrumental to complete my work uninterrupted. Lastly, I thank my beloved wife and two lovely daughters who always supported me and gave me everything - that which I wished and that which I did not. They are loving, beautiful and a perfect family.

Khalid Iqbal