The Journal of Nutrition
Methodology and Mathematical Modeling

Gaussian Graphical Models Identify Networks of Dietary Intake in a German Adult Population[1–3]

Khalid Iqbal,[4]* Brian Buijsse,[4] Janine Wirth,[4] Matthias B Schulze,[5,6] Anna Floegel,[4,7] and Heiner Boeing[4,7]

Departments of [4]Epidemiology and [5]Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany; and [6]German Center for Diabetes Research, Neuherberg, Germany

Abstract

Background: Data-reduction methods such as principal component analysis are often used to derive dietary patterns. However, such methods do not assess how foods are consumed in relation to each other. Gaussian graphical models (GGMs) are a set of novel methods that can address this issue.

Objective: We sought to apply GGMs to derive sex-specific dietary intake networks representing consumption patterns in a German adult population.

Methods: Dietary intake data from 10,780 men and 16,340 women of the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort were cross-sectionally analyzed to construct dietary intake networks. Food intake for each participant was estimated using a 148-item food-frequency questionnaire that captured the intake of 49 food groups. GGMs were applied to log-transformed intakes (grams per day) of 49 food groups to construct sex-specific food networks. Semiparametric Gaussian copula graphical models (SGCGMs) were used to confirm GGM results.

Results: In men, GGMs identified 1 major dietary network that consisted of intakes of red meat, processed meat, cooked vegetables, sauces, potatoes, cabbage, poultry, legumes, mushrooms, soup, and whole-grain and refined breads. For women, a similar network was identified with the addition of fried potatoes. Other identified networks consisted of dairy products and sweet food groups. SGCGMs yielded results comparable to those of GGMs.

Conclusions: GGMs are a powerful exploratory method that can be used to construct dietary networks representing dietary intake patterns that reveal how foods are consumed in relation to each other. GGMs indicated an apparent major role of red meat intake in a consumption pattern in the studied population. In the future, identified networks might be transformed into pattern scores for investigating their associations with health outcomes. *J Nutr* 2016;146:646–52.

Keywords: Gaussian graphical models, GGMs, dietary pattern analysis, consumption pattern, dietary networks

Introduction

Dietary pattern analysis is a preferable method for characterizing dietary intake (1) and understanding eating behavior. An exploratory analysis based on data-reduction methods such as principal component analysis (PCA)[8] and cluster analysis are frequently used to derive dietary patterns (2). PCA has been of particular interest because it compresses food groups based on the correlation or covariance between original variables into a number of uncorrelated patterns called components or factors (3).

Although the correlation structure assessed by such methods helps to better understand data and identify the similarity pattern between food groups, it cannot completely unravel the understanding of the pairwise association between food variables. Pairwise correlations between food groups can be more informative if they are independent of the effect of other food groups (4). Such pairwise correlations between food groups that control for others identify a dependency of various food groups in the dietary data, which may be important in understanding how foods are consumed in relation to each other.

Moreover, the existing methods of dietary pattern analysis require several but crucial subjective choices during data analysis (5, 6), and the identified patterns are often difficult to

[1] Supported by the German Federal Ministry of Science (01 EA 9401), the European Union (SOC 95201408 05F02), and the Higher Education Commission of Pakistan (no. 50015451 to KI).

[2] Author disclosures: K Iqbal, B Buijsse, J Wirth, MB Schulze, A Floegel, and H Boeing, no conflicts of interest.

[3] Supplemental Figures 1–3 and Supplemental Tables 1 and 2 are available from the ''Online Supporting Material'' link in the online posting of the article and from the same link in the online table of contents at http://jn.nutrition.org.

[7] These authors contributed equally to this work.

[8] Abbreviations used: EPIC, European Prospective Investigation into Cancer and Nutrition; GGM, Gaussian graphical model; PCA, principal component analysis; SGCGM, semiparametric Gaussian copula graphical model.

*To whom correspondence should be addressed. E-mail: khalid.iqbal@dife.de.

Manuscript received July 29, 2015. Initial review completed September 4, 2015. Revision accepted December 17, 2015.
First published online January 27, 2016; doi:10.3945/jn.115.221135.

interpret (3, 7, 8). These limitations warrant the investigation of complementary approaches to characterize dietary intake patterns. Innovative methods that provide additional information might be advantageous over conventional ones and improve our understanding of the complexity of eating behaviors.

Gaussian graphical models (GGMs) form a promising class of methods for exploratory analysis (9). These are graphical methods that identify the conditional independence structure in the data set by assessing pairwise correlation between 2 variables controlling for others. GGMs assume multivariate normal distribution for underlying data and can infer a direct relation between variables in a given data set without prior knowledge (10). GGMs have been used to simplify and compress high-dimensional genetic (11, 12) and metabolomics (13, 14) data to explore respective underlying pathways.

Because dietary data are high-dimensional like genetic and metabolomics data, the application of GGMs for identifying conditional independence structures between food intake variables is an interesting approach. In dietary intake data, the pairwise correlation between 2 food groups controlling for others can identify both the internal structure (i.e., patterns) in the original data as well as the relation between the food groups consumed in the identified network. The latter characteristic is of particular interest because foods are consumed in specific combinations that reflect consumption patterns and may be helpful in providing insight into the eating behavior of the studied population. These networks may also identify key interrelated food groups that may be potential candidates for further investigation into confounder structures, thus leading to a much better understanding of the biological relations between diet and health status.

The objective of this study was therefore to apply GGMs to a well-studied set of food data by constructing sex-specific dietary intake networks in the EPIC (European Prospective Investigation into Cancer and Nutrition)-Potsdam study for further understanding the interrelation between food intakes. Moreover, because GGMs assume data normality, which may not be the case with certain dietary variables, GGM results were confirmed through semiparametric Gaussian copula graphical models (SGCGMs) (15), which do not require normal distribution of the data.

Methods

Study population

Participants of this cross-sectional study came from the EPIC-Potsdam cohort (16), which is part of the multicenter EPIC cohort study that was established to investigate associations between diet, lifestyle factors, and cancer as well as other chronic diseases (17). EPIC-Potsdam includes a total of 27,548 participants (59.9% women) who were mostly aged 35–65 y old at baseline, lived in Potsdam, Germany or the surrounding areas, and enrolled between 1994 and 1998. The ethics committee of the state of Brandenburg approved the study procedures. All participants provided written informed consent at baseline before examination. Data on age, sex, educational attainment, physical activity, smoking status, and anthropometric indicators, including weight and height, were collected at baseline. Anthropometric measurements were performed by trained staff while participants were in their underwear. Body weight was measured to the nearest 100 g and body height to the nearest 1 mm. Education attainment was defined in 3 categories: currently in training/no certificate or skill, professional school, and college or higher education. Physical activity was assessed using a short questionnaire and categorized as inactive, moderately inactive, moderately active, and active according to the Cambridge physical activity index (18). Smoking status was categorized as never, former, and current.

For this study, participants aged <35 y and those with missing data on dietary intake, educational attainment, and anthropometric indicators were excluded from the analysis. (See **Supplemental Figure 1** for a flowchart of participants selected for this study sample.) Thus, the final study sample comprised 27,120 subjects (10,780 men and 16,340 women).

Assessment of dietary intake

Habitual dietary intake was assessed at the time of enrollment with a validated, optical-readable, self-administered, semiquantitative FFQ (19). The questionnaire queried the frequency of consumption of 148 food items over the past 12 mo. Additional information regarding the fat content of dairy products consumed and type of fat used for food preparation was collected at the same time. Food intake portion sizes were estimated using photographs and standard portion sizes (when possible). Food intake frequency was assessed in 10 categories ranging from “never” to “5 times a day or more.” The intake of each food item was calculated from portion size and intake frequency. Single-food intakes were collapsed into 49 food groups (including alcoholic beverages) as described by Schulze et al. (20).

Statistical methods

GGMs. GGMs are probabilistic graphs used to analyze and visualize the dependency structures with the help of a graph that describe conditional independence between variables (21). These graphs present a set of nodes and edges, where nodes represent variables and edges represent conditional dependency relations. A missing edge between 2 variables in the dependence graph represents conditional independence between these variables given all other variables (22). Such conditional independence in a dependence graph is called a pairwise Markov property (9) and is quantified in terms of partial correlation. Model selection in GGMs results in a sparse graph that represents the underlying pattern of the associated variables.

Theoretical background. Suppose we have a data matrix X with n observations and p variables from a p-variate normal distribution that has a mean vector μ and covariance matrix Σ, which can also be expressed as Np (μ, Σ). Then from the inverse of this covariance matrix, which is also called the precision matrix, the conditional distribution of any 2 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 2 variables is called partial correlation (23). If the partial correlation between the 2 variables, i.e., p_1 and p_2, is 0, it is inferred that these 2 variables are conditionally independent given all other variables. The estimation of conditional independence in a precision matrix forms the basis of GGMs (24). In GGMs, the conditional independence relation between given variables is reflected in an undirected graph G (V, E), where V represents vertices (variables) and E represents edges (partial correlation between variables) of the graph G. From this, a GGM is defined as an undirected graph of p-variate normal distribution Np (μ, Σ) with a conditional independence restriction (i.e., 2 variables are independent given others) if the correlation in inverse covariance matrix between the 2 variables is 0, as defined by the pairwise Markov property (25).

In GGMs, conditional independence between variables is determined by identifying 0 entries in the inverse of the covariance matrix, known as the covariance selection problem (26) or model selection (also called structure learning) in the Gaussian concentration graph model (27). However, in a high-dimensional multivariate normally distributed data set, there may be no or few 0 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 because the aim of GGMs is to identify the topology (structure) of a graphical model that is an accurate and meaningful representation of the underlying data. The accuracy of such a model is assessed by the likelihood that the model explains the data (28). Such situations require adopting a regularization technique that enforces sparsity in the precision matrix for data representation. Although a number of methods exist (29, 30) for achieving sparsity in the precision matrix, graphical lasso (31) is a fast, efficient 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 0 and results in a sparse matrix of direct connections (31). In sum, rather than maximizing log likelihood, graphical lasso maximizes the regularized log likelihood for achieving sparsity. Regularization is achieved by penalizing log likelihood by the term $\lambda \times L_1$ norm, where L_1 norm is the absolute sum of the inverse covariance matrix and λ is a nonnegative-tuning shrinkage parameter. It is also called the regularization parameter. The value of λ depends on the research question (level of sparsity required) and is estimated from the best model fit (log likelihood) for different values of λ. 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 data set.

Analysis. The means ± SDs of food intakes, age, BMI (in kg/m^2), and percentages of participants in different smoking, physical activity, and educational categories were calculated for men and women separately with SAS version 6.1 (IBM). A *t* test was used for continuous variables, and a Chi-square test was used for categorical variables to assess statistical differences ($\alpha = 0.05$) between men and women.

GGMs were used to derive sex-specific networks of dietary intake. GGM analysis was conducted in R (version 3.0.3, R Foundation for Statistical Computing, Vienna, Austria) software as described by Højsgaard and Lauritzen (32). Gaussian assumption for GGM was visually assessed using a histogram and box plot in R. Because most of the dietary variables had skewed distributions, dietary data were log-transformed [ln (g/d + 1)] to improve normality. A 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” (31). The optimum value for the regularization parameter λ was assessed in the “huge” package by specifying a sequence of λ values (0.60–0.10) in a decreasing order for sparsity (33). The sequence of λ values was selected in such a way that the highest λ value (0.60) would result in an extremely sparse concentration matrix (no node connections) and the smallest λ value (0.10) would result in a less-sparse concentration matrix (very dense graph difficult to interpret). An optimum λ value of 0.25 was selected by the maximum likelihood estimate of the graphical models and used for all analyses. Estimated sparse concentration matrixes were exported to a yEd graph editor (34) and visualized as a dietary network separately for men and women. Network stability for the existing study sample was assessed by repeated bootstrapping 80% of the original sample with replacement.

To further evaluate the robustness of the results, dietary networks of the log-transformed intakes were reconstructed using SGCGMs (15), which do not require Gaussian distribution of the underlying data. SGCGMs transform the observed variables in the latent variables, and rank-based statistics, including Spearman’s ρ and Kendall’s τ, are exploited to estimate the correlation matrix, which is plugged into the parametric procedure to get a sparse precision matrix. Semiparametric analysis was conducted using the huge package in R.

Results

Table 1 shows the baseline characteristics of the study participants. In general, men were older ($P < 0.01$) and had a higher BMI ($P < 0.01$) than women. Moreover, men tended to have a higher level of education ($P < 0.01$) and were more frequently smokers compared with women ($P < 0.01$). **Table 2** shows the mean intakes of the food groups estimated from the FFQ in this population. **Supplemental Tables 1** and **2** show pairwise Spearman rank correlations among 49 food groups consumed by men and women, respectively.

TABLE 1 Baseline characteristics of the EPIC-Potsdam cohort participants included in this study[1]

Characteristics	Men (*n* = 10,780)	Women (*n* = 16,340)	*P*
Age at enrollment, y	52.6 ± 8.0	49.3 ± 9.2	<0.01[2]
BMI, kg/m^2	27.0 ± 3.7	25.9 ± 4.7	<0.01[2]
Physical activity, %			<0.01[3]
Inactive	22.1	22.0	
Moderately inactive	36.2	39.5	
Moderately active	24.3	23.5	
Active	17.4	15.1	
Education attainment, %			<0.01[3]
Currently in training/no certificate or skill	33.9	41.9	
Professional school	17.2	29.8	
≥College	48.9	28.3	
Smoking status, %			<0.01[3]
Never	30.4	58.0	
Former	44.7	24.3	
Smoker	24.9	17.7	

[1] Values are means ± SDs unless otherwise indicated. EPIC, European Prospective Investigation into Cancer and Nutrition.

[2] *t* test.

[3] Chi-square test.

GGM analysis identified 1 major dietary network that we termed the principal network and several smaller networks that consisted of similar food groups in men and women (**Figures 1 and 2**). In men, the principal dietary network consisted of the intake of 12 food groups, most of which clustered around red meat and cooked vegetables. Red meat intake was highly correlated with the intakes of poultry, processed meat, sauce, and potatoes, whereas the intake of cooked vegetables was highly correlated with the intake of mushrooms and cabbage. The network revealed that the intake of processed meat and poultry was conditionally dependent on red meat intake, whereas the intake of legumes and mushrooms was conditionally dependent on the intake of cooked vegetables in the identified pattern. In addition, there was a strong negative correlation between the intake of whole-grain and refined breads.

Other important networks identified in men consisted of dairy products defined by fat content, sweet foods, and fruits and vegetables. In the network of dairy products defined by fat content, there was a strong inverse correlation between the intakes of high- and low-fat food groups among men. On the other hand, in the network defined by the intake of sweet food groups, all food groups were positively correlated with each other. In the same network, the intakes of desserts as well as cakes and cookies were correlated with the intake of all other sweet foods. The network of fruits and vegetables showed that the intake of fresh fruits and vegetable fats were conditionally dependent on the intake of raw vegetables.

In women, the principal network consisted of the same food groups as identified in men, with the addition of fried potatoes (Figure 2). Similar to the intake network identified in men, the network in women revealed a central role of red meat and cooked vegetables intake 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.

As in men, other important networks identified in women consisted of intakes of dairy products defined by fat content, sweet foods, and fruits and vegetables. However, unlike in men, the network of dairy products defined by fat content additionally

TABLE 2 Dietary intakes of 49 food groups used to derive dietary networks among men and women of the EPIC-Potsdam cohort included in this study[1]

Food groups	Men (*n* = 10,780)	Women (*n* = 16,340)
Whole-grain bread, g/d	40.9 ± 57.3	48 ± 52.4
Refined bread, g/d	167 ± 88.0	106 ± 63
Grain flakes, grains, muesli, g/d	4.8 ± 15.4	5.9 ± 14.4
Cornflakes, crisps, g/d	1.4 ± 5.7	1.9 ± 6.7
Pasta, rice, g/d	16.5 ± 15.0	15.9 ± 14.4
Vegetarian dishes, g/d	1 ± 4.3	1.4 ± 5.7
Chips, g/d	2.6 ± 6.6	2 ± 5.0
Pizza, g/d	7.3 ± 11.0	6.8 ± 8.9
Cake, cookies, g/d	68.3 ± 71.8	59 ± 61.8
Confectionary, g/d	23.8 ± 27.6	20.9 ± 26.1
Sweet bread spreads, g/d	12.5 ± 13.7	11.2 ± 12.1
Eggs, g/d	19.4 ± 17.5	16.1 ± 14.6
Fresh fruit, g/d	122 ± 89.0	154 ± 99.0
Canned fruit, g/d	19.5 ± 26.0	17 ± 23.9
Raw vegetables, g/d	47.9 ± 39.8	61.7 ± 47.1
Cabbage, g/d	13.5 ± 13.9	14.2 ± 13.5
Cooked vegetables, g/d	27.5 ± 17.5	30.1 ± 18.6
Garlic, g/d	0.1 ± 0.4	0.1 ± 0.5
Mushrooms, g/d	2 ± 2.4	2 ± 2.4
Legumes, g/d	29.3 ± 24.0	19.3 ± 16.0
Potatoes, g/d	95.5 ± 52.2	75.2 ± 44.9
Fried potatoes, g/d	18.8 ± 17.3	10.8 ± 10.2
Nuts, g/d	3.7 ± 8.3	2.9 ± 7.9
Low-fat dairy products, g/d	83.1 ± 175	111 ± 194
High-fat dairy products, g/d	98.1 ± 170	101 ± 154
Low-fat cheese, g/d	6.2 ± 15.5	6.9 ± 14.3
High-fat cheese, g/d	30.6 ± 28.1	26.3 ± 23.6
Water, g/d	366 ± 404	470 ± 455
Coffee, g/d	440 ± 347	406 ± 297
Decaffeinated coffee, g/d	27.1 ± 120	31.1 ± 121
Tea, g/d	226 ± 328	294 ± 385
Fruit juice, g/d	186 ± 229	200 ± 223
Low-energy soft drinks, g/d	14.6 ± 92.4	8.5 ± 56.7
High-energy soft drinks, g/d	70 ± 180	27.4 ± 106
Beer, g/d	393 ± 525	46.3 ± 126
Wine, g/d	49.2 ± 102	51.9 ± 86.8
Spirits, g/d	5.1 ± 13.6	1.1 ± 5.1
Other alcoholic beverages, g/d	11.2 ± 21.3	14.4 ± 38.4
Butter, g/d	10.2 ± 14.6	7.7 ± 10.9
Margarine, g/d	18 ± 16.9	14.1 ± 13.5
Vegetable fat, g/d	3 ± 3.1	3.6 ± 3.5
Animal fat, g/d	0.3 ± 0.8	0.2 ± 0.6
Sauces, g/d	13.3 ± 12.4	11.2 ± 10.9
Desserts, g/d	16.7 ± 22.0	15.4 ± 22.6
Fish, g/d	28.2 ± 30.7	21.2 ± 21.9
Poultry, g/d	15 ± 14.3	11.4 ± 11.2
Meat, g/d	54 ± 35.6	34.4 ± 23.0
Processed meat, g/d	78.6 ± 54.8	48.1 ± 34.5
Soup, g/d	45.1 ± 41.9	38 ± 35.5

[1] Values are means ± SDs. Listed are 49 food groups derived from a 178-item FFQ by combining foods that are similar in nutrient composition. EPIC, European Prospective Investigation into Cancer and Nutrition.

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

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.

Dietary networks derived through SGCGMs to assess the robustness of the normality assumption of GGMs showed a strong resemblance to the GGM-derived networks (**Supplemental Figures 2** and **3**). The derived principal and smaller networks were similar compared to GGM and comprised similar food groups for both men and women.

Discussion

This study assessed GGMs, a complementary exploratory method for dietary pattern analysis that is an existing approach already in use in metabolomics (14), genetics (35), and climate research (36). GGMs, 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 between the 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 that are visualized as dietary networks.

A major advantage of GGMs is their ability to distinguish between direct and indirect associations between the food groups consumed. Data-reduction methods such as 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 2 food groups. The removal of indirect effects when assessing pairwise correlations between 2 food groups is crucial to understanding how different food groups are consumed in relation to each other. GGMs address the problem of indirect effects by calculating a measure of conditional independencies between the food groups (10). The resulting conditional independence measures reflect a pairwise correlation between 2 food groups independent of the linear effect of the other food groups. In other words, the partial correlation coefficients reflect the association between 2 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 relation to disease outcomes. Therefore, use of the identified networks may only partly be helpful for defining confounding during further analyses.

In this analysis, GGMs identified sex-specific networks consisting of a principal network and additional smaller networks. The principal networks among both sexes revealed that the 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 a PCA-derived dietary pattern in the same population (20). 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 GGMs. Similarly, in women, 10 of the 11 food groups with high-factor loading in the PCA-identified "plain-cooking" pattern were also part of the principal networks identified by GGMs. Moreover, high-fat and sweet PCA patterns were also comparable to the identified networks in both men and women. This comparison indicates that the identified food networks are not statistical artifacts but may reflect true patterns (1). There are several potential approaches for further investigating the identified networks. First, the conditional independence can be advantageously used for

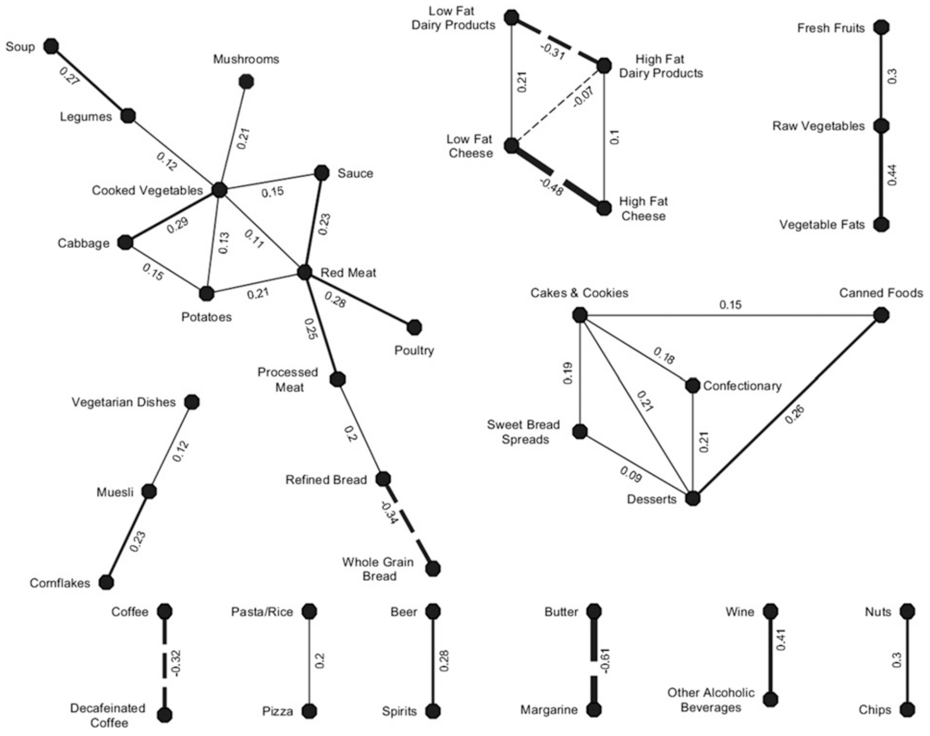

FIGURE 1 Dietary intake networks for men from the EPIC-Potsdam cohort included in this study derived by Gaussian graphical models. Vertices represent foods/food groups, and edges represent conditional dependencies (reflected by partial correlation coefficients) between foods/food groups. Continuous lines show positive partial correlations, whereas broken lines show negative partial correlations. Edge thickness is proportional to the strength of the correlation between connected food groups. The absence of an edge between 2 foods/food groups represents conditional independence between them in the network (*n* = 10,780). EPIC, European Prospective Investigation into Cancer and Nutrition.

identifying consumption probabilities of the foods/food groups identified in the network for each individual. Such probabilities would be helpful for modeling alternative intake patterns by modifying intake probabilities, which may help assess the impact of dietary behavior change or dietary recommendations. Second, algorithms could be developed to use these consumption probabilities to build innovative food network scores that could be used to investigate their relation to health outcomes

In addition, GGMs 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. Although high-factor loading of red meat in the PCA pattern

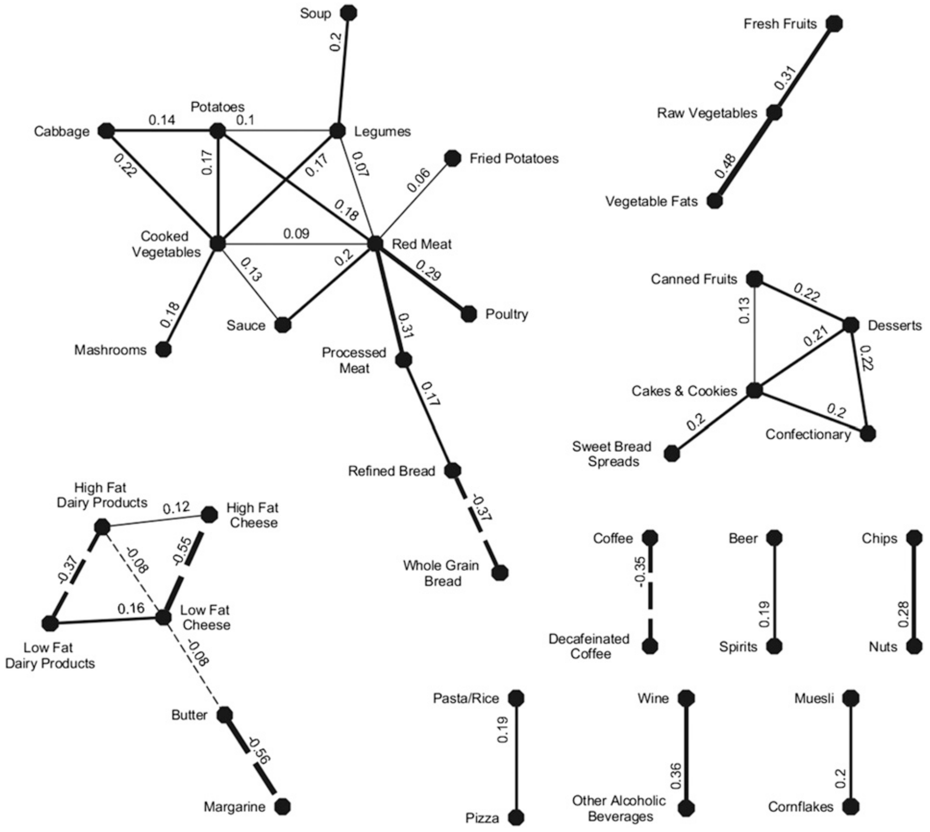

FIGURE 2 Dietary intake networks for women from the EPIC-Potsdam cohort included in this study derived by Gaussian graphical models. Vertices represent foods/food groups, and edges represent conditional dependencies (reflected by partial correlation coefficients) between foods/food groups. Continuous lines show positive partial correlations, whereas broken lines show negative partial correlations. Edge thickness is proportional to the strength of the correlation between connected food groups. The absence of an edge between 2 foods/food groups represents conditional independence between them in the network (*n* = 16,340). EPIC, European Prospective Investigation into Cancer and Nutrition.

underscores its importance, GGMs not only underscores its importance but 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 (37). This is interesting because the role of red meat for health outcomes is still unraveling in terms of causality (37–39), and its further investigation in relation to health outcomes is still a research agenda priority (40).

GGMs identified a separate network for fresh fruits, raw vegetables, and vegetable fats, reflecting a healthy pattern among both sexes. In addition, similar to PCA, GGMs identified separate networks for sweet foods and dairy products based on fat content. However, unlike PCA (41), GGMs identified independent networks of food groups in which each food group was part of only 1 network that facilitated their interpretation.

GGMs introduce sparsity (i.e., select only a few variables in the final model) and force other variables to 0 to explore data structure and facilitate interpretation. This advantage of GGMs is also shared by another data-reduction method called treelet transform, which was introduced in nutrition epidemiology in 2011 (42). Treelet transform 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 0 in each identified factor to achieve sparsity. This helps to identify factors that are easy to interpret. However, unlike GGMs, 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 GGMs are a robust method for dietary pattern analysis and that they revealed similar networks as SGCGMs. GGMs were the method of choice for this analysis because SGCGMs perform rank-based transformations of the original variables into new variables that have Gaussian distribution. After transforming the variables, SGCGMs use the same method to attain a graphical model as done in GGMs. Therefore, we preferred to keep log-transformed original variables and use GGMs rather than selecting a model on rank-based transformed variables.

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, we log-transformed all data, and although this does not always result in perfect normal distributions, our findings were robust compared with SGCGMs, which do not require the Gaussian assumption. Second, network sparsity depends on a regularization parameter that can be derived using different criteria (e.g., log likelihood or Aikake/Bayesian information criteria, cross-validation, etc). However, independent of the choice of method, the latent structure of the data remains the same and may be identified using any of the shrinkage parameter estimation methods—albeit with different sparsity levels. Third, changes in characteristics of the study sample may potentially yield a different network in the same manner as pattern analysis through PCA. This is true for all correlation-dependent methods and should be kept in mind for GGMs as well. Fourth, GGMs identify networks but do not assign individual scores to participants such as PCA or 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 the aim of this study. Nevertheless, advancing GGM methodology for possible calculation of quantitative scores or classification of individuals on the basis of an identified network could be of interest for future research. Furthermore, methods used for dietary pattern analysis, which assume sparsity, have also been criticized. It is argued that such methods reduce the pattern to several foods, although actual consumption comprises a large number of foods, all of which should be retained in the dietary pattern (1). Nevertheless, such arguments are challenged on several grounds. First, dietary pattern lacks a specific definition. The current definition is method-driven and can be defined operationally as data reduction (43). Because existing dietary pattern analysis tools have limitations, different methods may identify dietary patterns differently irrespective of the sparsity assumption (44) but still will be called dietary patterns. Second, use of sparsity for pattern recognition depends on the study question and may be advantageous in certain situations. For example, Assi et al. (45) used sparsity to identify a nutrient pattern associated with hormonal receptor-defined breast cancer. In this study, sparsity is advantageous because we could show not only the pattern but also how foods in the pattern are consumed in relation to each other.

Strengths of this study include the use of a population-based cohort with a large sample size. The large database allows sex-specific dietary networks to be calculated and their stability assessed. In addition, this analysis was based on an already published dietary data set that has been analyzed with other dietary pattern methods (e.g., PCA) that enable a direct comparison of previous methods to our results. Furthermore, the associations between the investigated food groups and chronic disease risk in this population are already known (46). Consequently, the identified networks may be further investigated for an association with health outcomes.

In conclusion, GGMs are a powerful exploratory method that can be used to construct dietary intake networks that represent 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. GGMs identified a central role of red meat consumption within the derived dietary pattern. Identified networks can be transformed to intake pattern scores for association with health outcome. Nevertheless, additional studies are required to validate this method in other populations.

Acknowledgments

We thank Manuela Bergman of the Human Study Centre for contributing to the EPIC-Potsdam study design. KI, JW, AF, and HB conceived the idea and designed the analysis plan; AF and HB were involved in developing the overall research plan and overseeing the study; KI conducted statistical analysis and wrote the manuscript; BB and AF provided support with statistical analyses; KI, AF, and HB had primary responsibility for final content; and BB, JW, MBS, AF, and HB contributed to data interpretation and provided critical comments on the manuscript. All authors read and approved the final manuscript.

References

1. Imamura F, Jacques PF. Invited commentary: dietary pattern analysis. Am J Epidemiol 2011;173:1105–8; discussion 9–10.
2. Varraso R, Garcia-Aymerich J, Monier F, Le Moual N, De Batlle J, Miranda G, Pison C, Romieu I, Kauffmann F, Maccario J. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. Am J Clin Nutr 2012;96:1079–92.

3. Michels KB, Schulze MB. Can dietary patterns help us detect diet-disease associations? Nutr Res Rev 2005;18:241–8.
4. Rasmussen MA, Bro R. A tutorial on the Lasso approach to sparse modeling. Chemom Intell Lab Syst 2012;119:21–31.
5. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, Krebs-Smith SM, Guenther PM. Dietary patterns: challenges and opportunities in dietary patterns research: an experimental biology workshop, April 1, 2006. J Am Diet Assoc 2007;107:1233–9.
6. Kant AK. Dietary patterns and health outcomes. J Am Diet Assoc 2004;104:615–35.
7. Martínez ME, Marshall JR, Sechrest L. Invited commentary: factor analysis and the search for objectivity. Am J Epidemiol 1998;148:17–9.
8. Schulze MB, Hoffmann K. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. Br J Nutr 2006;95:860–9.
9. Lauritzen S. Graphical models. New York: Clarendon Press; 1996.
10. 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. Stat Appl Genet Mol Biol 2008;7:14.
11. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. J Multivariate Anal 2004;90:196–212.
12. 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 Proc 2014;8(Suppl 1):S99.
13. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 2011;5:21.
14. Floegel A, Wientzek A, Bachlechner U, Jacobs S, Drogan D, Prehn C, Adamski J, Krumsiek J, Schulze MB, Pischon T, et al. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. Int J Obes 2005;2014:.
15. Han, L, Han, F, Yuan, M, Lafferty, J, Wasserman, L. High-dimensional semiparametric Gaussian copula graphical models. Ann Stat 2012;40:2293–326.
16. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43:205–15.
17. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992;3:783–91.
18. Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, Hennings S, Day NE. 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 Nutr 2003;6:407–13.
19. Kroke A, Klipstein-Grobusch K, Voss S, Moseneder J, Thielecke F, Noack R, Boeing H. Validation of a self-administered food-frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. Am J Clin Nutr 1999;70:439–47.
20. Schulze MB, Hoffmann K, Kroke A, Boeing H. Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. Br J Nutr 2001;85:363–73.
21. Strobl R, Grill E, Mansmann U. Graphical modeling of binary data using the LASSO: a simulation study. BMC Med Res Methodol 2012;12:16.
22. Krämer N, Schafer J, Boulesteix AL. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics 2009;10:384.
23. Edwards D. A brief introduction to graphical models [Internet]. [cited 2014 Mar 17]. Available from: https://djfextranet.agrsci.dk/sites/phd_course2_2010/public/Documents/ABriefIntro.pdf.
24. Mazumder R, Hastie T. The graphical lasso: new insights and alternatives. Electron J Stat 2012;6:2125–49.
25. Drton M, Perlman MD. Multiple testing and error control in Gaussian graphical model selection. Stat Sci 2007;22:430–49.
26. Dempster AP. Covariance selection. Biometrics 1972;28:157–75.
27. Wermuth DRC. Multivariate dependencies: models, analysis and interpretation. London: Chapman & Hall; 1996.
28. Honorio J, Rish DSI, Cecchi G. Variable selection for Gaussian graphical models. J Mach Learn Res 2012;22:538–46.
29. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika 2007;94:19–35.
30. Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Stat 2006;34:1436–62.
31. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9:432–41.
32. Højsgaard SED, Lauritzen S. Graphical models with R. New York: Springer-Verlag, 2012.
33. 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;13:1059–62.
34. yWorks. yEd software version 3.13.4[Internet]. c2015 [cited 2015 Jul 12]. Available from: http://www.yworks.com.
35. Ma S, Gong Q, Bohnert HJ. An Arabidopsis gene network based on the graphical Gaussian model. Genome Res 2007;17:1614–25.
36. Zerenner T, Friederichs P, Lehnertz K, Hense A. A Gaussian graphical model approach to climate networks. Chaos 2014;24:023103.
37. 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. Int J Clin Exp Med 2014;7:2100–10.
38. 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. Br J Nutr 2014;112:762–75.
39. Larsson SC, Orsini N. Red meat and processed meat consumption and all-cause mortality: a meta-analysis. Am J Epidemiol 2014;179:282–9.
40. WHO International Agency for Research on Cancer. IARC monographs on the evaluation of carcinogenic risks to humans: report of the advisory group to recommend priorities for IARC monographs during 2015–2019 [monograph on the Internet]. Lyon (France): WHO; 2014 [cited 2015 Mar 15]. Available from: http://monographs.iarc.fr/ENG/Publications/internrep/14-002.pdf.
41. 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. J Diabetes Metab Disord 2014;13:71.
42. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K. Exploring dietary patterns by using the treelet transform. Am J Epidemiol 2011;173:1097–104.
43. Slattery ML. Defining dietary consumption: is the sum greater than its parts? Am J Clin Nutr 2008;88:14–5.
44. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K. Gorst-Rasmussen et al. respond to “Dietary Pattern Analysis.” Am J Epidemiol 2011;173:1109–10.
45. Assi N, Moskal A, Slimani N, Viallon V, Chajes V, Freisling H, Monni S, Knueppel S, Forster J, Weiderpass E, 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 Nutr 2015 Feb 23 (Epub ahead of print; DOI: 10.1017/S1368980015000294).
46. 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. Eur J Clin Nutr 2013;67:412–9.