British Journal of Nutrition (2003), **89**, 409–418
DOI: 10.1079/BJN2002778

An approach to construct simplified measures of dietary patterns from exploratory factor analysis

Matthias B. Schulze*, Kurt Hoffmann, Anja Kroke and Heiner Boeing

German Institute of Human Nutrition, Department of Epidemiology, Arthur-Scheunert-Allee 114–116, 14558 Bergholz-Rehbruecke, Germany

(Received 2 January 2002 – Revised 29 August 2002 – Accepted 3 October 2002)

Exploratory factor analysis might work well in elucidating the major dietary patterns prevailing in specific study populations. However, patterns extracted in one study population and their associations with disease risk cannot be reproduced with this data-specific method in other study populations. To construct less population-dependent pattern variables of similar content as original exploratory patterns, we proposed to derive so-called simplified pattern variables. They represent the sum of the unweighted standardised food variables which loaded high at the pattern of interest. Data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study suggest that these simplified pattern variables might adequately approximate factor analysis-based dietary patterns. A simplified pattern variable based on the six highest loading food variables showed a correlation $>0{\cdot}95$ with the originally derived factor score, which consisted of forty-seven food variables. Moreover, simplified pattern variables might adequately approximate patterns across different study populations. A simplified pattern variable showed similar factor loadings, ranging from 0·34 to 0·52, as well as similar associations with nutrient intake as a 'western' pattern originally reported from an US study population. These simplified pattern variables can subsequently be used to study pattern associations with disease risk, especially in multi-centre studies. It is therefore an approach that might overcome one of the most frequently claimed limitations of factor analyses applied in epidemiology: their non-comparable risk estimates.

Dietary patterns: Food habits: Factor analysis: Reproducibility of results

Dietary patterns are of considerable interest in nutritional epidemiology to reflect the complexity of dietary intake in relation to diseases. The dominating epidemiological approach of examining single nutrients or foods is fraught with the complexity of dietary intake in relation to diseases. The high degree of intercorrelation among nutrients as well as among foods makes it difficult to attribute effects to single dietary components. Separation of effects by adjustment in ordinary and logistic regression is hard to accomplish, because of the high intercorrelation results in unstable models, as well as large CI (Hoffmann *et al.* 2002). More important, however, might be the fact that by trying to separate effects one might miss associations between diet and disease. On the other hand, summary variables reflecting dietary patterns might take interactions between single nutrients and foods into account and might allow to estimate overall effects of diet.

Two general approaches have been used to define these summary variables (Trichopoulos & Lagiou, 2001). The so-called '*a posteriori*' approach builds on statistical exploratory methods. One method predominantly used in this context is exploratory factor analysis. This method works well for identifying the major dietary patterns of a particular study population (Jacques & Tucker, 2001; Trichopoulos & Lagiou, 2001), but independent from their relevance for any disease. On the other hand, the so-called '*a priori*' approach focuses on the construction of pattern variables that reflect hypothesis-oriented patterns based on available scientific evidence for specific diseases. *A priori* pattern scores were, for example, constructed on the basis of dietary recommendations. The Healthy Eating Index (Kennedy *et al.* 1995), as one of these recommendation-based pattern scores, measures how well

Abbreviation: EPIC, European Prospective Investigation into Cancer and Nutrition.

* **Corresponding author:** Dr Matthias B. Schulze, fax +49 33200 88 444, email mschulze@www.dife.de

diets conform to the recommendations of the US Department of Agriculture. The index consists of ten equally weighted components measuring adherence to serving recommendations for grains, vegetables, fruits, milk and meat, as well as measuring intake of total fat, saturated fat, cholesterol and Na and diet diversity.

A major criticism of the *a posteriori* approach is that the patterns extracted in one study population cannot be reproduced with the data-specific exploratory methods in other study populations (Martinez *et al.* 1998; Jacques & Tucker, 2001). Not surprisingly, nutritional studies using exploratory factor analysis reported generally quite different patterns (Gex-Fabry *et al.* 1988; Randall *et al.* 1990; Whichelow & Prevost 1996; Slattery *et al.* 1998; Hu *et al.* 2000; Maskarinec *et al.* 2000; Tseng *et al.* 2000; Williams *et al.* 2000; Schulze *et al.* 2001; Osler *et al.* 2001). Even though patterns have been successfully linked to disease risk (Slattery *et al.* 1998; Hu *et al.* 2000; Fung *et al.* 2001) and mortality (Osler *et al.* 2001), study specific estimates of relative risks are consequently not reproducible and comparable as well. Overall measures of effect across studies, e.g. determined by meta-analysis, require unified pattern variables, which limits the significance of the *a posteriori* pattern analysis approach in epidemiological research. Moreover, in multi-centre studies, like the European Prospective Investigation into Cancer and Nutrition (EPIC), a common data analysis will also rely on the same pattern variables for all centres that cannot be retained with exploratory factor analysis within centres.

On the other hand, the *a priori* approach offers the possibility of constructing pattern variables based on scientific evidence, replacing study-specificity by disease-specificity. In contrast to the *a posteriori* approach, different pattern variables can be applied for different diseases. The data to construct these patterns might come from observational studies of various dietary habits that appear to be associated with the specific diseases (Trichopoulos & Lagiou, 2001). This concept of constructing an *a priori* pattern variable has been demonstrated to be successful in studying, for example, whether a diet score reflecting key elements of a Mediterranean diet (originally described by Trichopoulou *et al.* (1995)) is related to mortality in an Anglo-Celtic Australian population (Kouris-Blazos *et al.* 1999). Similarly, *a priori* pattern variables might be constructed based on observational studies that determined the dietary habits by exploratory factor analysis and proved their significance for a specific disease. However, so far no attempts have been made to construct *a priori* pattern variables that are of similar content as originally identified exploratory patterns. The present study proposes a method that might be appropriate to do so.

Subjects and methods

Study population and data collection

The study population was selected from participants of the EPIC-Potsdam study, which contributes a general population sample of 27 548 men and women to the EPIC multi-centre cohort study (Riboli & Kaaks, 1997; Boeing *et al.* 1999). The analysis was restricted to men only (*n* 10 904). Men with missing information on dietary intake, smoking status, educational attainment, anthropometric measurements, and men reporting a change of their diet within the year before the assessment, were excluded from this study, retaining a total of 8975 study subjects.

Assessment of the study population was carried out between August 1994 and September 1998. Study participants filled out a self-administered food-frequency questionnaire. The food-frequency questionnaire assessed the usual food and nutrient intake of individuals during the 12 months prior to the examination and included 148 single food items. Photographs and, if available, standard portion sizes supported the estimation of portion sizes. The frequency of intake was measured using ten categories, ranging from 'never' to 'once per month or less' to 'five times per day or more'. The information on portion sizes and frequency of food intake was used to calculate the amount of each food item (g) consumed on average per d. The food items in the food-frequency questionnaire were aggregated into forty-nine separate food groups (Table 1). The grouping was based on the schemes of the German Food Code (Dehne *et al.* 1999) and EUROCODE (Kohlmeier, 1992) and on experience from other studies as well (Slattery *et al.* 1998; Hu *et al.* 1999). Nutrient intake per d was estimated from the consumed food items using the German Food Code (Dehne *et al.* 1999). Since nutrient intake is usually highly correlated with energy intake, we calculated energy-adjusted nutrient intakes, using the regression residual method (Willett & Stampfer, 1986).

Construction of pattern variables

Exploratory factor analysis aims to compress information on many variables into a few underlying factors by analysing their covariance structure. Details on the applied factor analytic methods have been reported elsewhere (Schulze *et al.* 2001). Briefly, the factor analysis started with a principal component solution, commonly used for the purpose of extracting dietary patterns (Hu, 2002). Thus, the pattern variables, called factor scores, were optimised linear combinations of the standardised food variables and were constructed to account for as much total variance of the food variables as possible. The retained factor scores were perfectly uncorrelated with each other and remained so after the applied subsequent rotation procedure *varimax*. We used an eigenvalue $>1{\cdot}25$ criterion, which is in agreement with Slattery *et al.* (1998), and finally retained seven factors. The commonly applied eigenvalue $>1{\cdot}00$ criterion yielded to many patterns (sixteen) for further analysis and no clear break between eigenvalues were observed in a scree plot. Food items with absolute factor loadings $>0{\cdot}20$ were considered as significantly contributing to a pattern. We excluded low-energy and high-energy soft drinks from the final analysis, because they did not load on any factor retained.

Table 1. Food groupings used in the dietary pattern analysis

Foods or food groups	Food items
Cooked vegetables	Tomatoes, tomato sauce, sweet pepper, courgette, aubergine, spinach, carrots, asparagus, pea–carrot vegetable mix, leek, celery (all cooked)
Cabbage family	Broccoli, cauliflower, red and white cabbage, kohlrabi (all cooked)
Legumes	Green peas, green beans, pea–bean–lentil stew
Cooked potatoes	Salted potatoes, jacket potatoes, mashed potatoes, potato salad, dumplings
Mushrooms	Fresh mushrooms, mushroom dishes
Sauce	Ketchup, brown and white sauce, salad dressing, sauce to vegetables
Poultry	Fried, grilled or roasted chicken or turkey
Meat except fish and poultry	Pork, beef, hamburger, minced meat, liver, lamb, roast hare
Animal fat except butter	Animal fat used for food preparation
Dessert	Pudding, sweet soufflé
Cake, cookies	Cake, tart, cookies
Confectionery, ice cream	Chocolate, candy bars, pralines, sugar, ice cream
Jam, honey, chocolate spread	Jam, honey, chocolate spread, peanut butter
Canned fruit	Canned fruit
Fruit juice	Citrus, apple, orange, grapefruit, grape, cherry, pineapple juice, multi-vitamin drinks
Tea	Black tea, green tea, fruit and herbal teas
Muesli	Wholegrain breakfast cereal, muesli
Cornflakes	Cornflakes, other refined grain breakfast cereal
Pasta, rice	Cooked pasta, cooked rice
Pizza	Pizza, quiche
Vegetarian dishes	Vegetarian dishes
Garlic	Raw, or fried or cooked garlic
Wholemeal bread	Wholemeal bread, dark and wholemeal rolls
Other bread	Rye bread, wheat bread, mixed bread, pale rolls, crispbread, croissants
Olive oil	Olive oil used for food preparation
Fresh fruits	Apple, pear, peach, cherry, grape, strawberry, blackberry, raspberry, kiwi, pineapple, mango, banana
Raw vegetables	Cucumber, carrot, sprouts, paprika, tomato, onion, radish
Other vegetable oils and fats	Vegetable fat used for food preparation (frying, dressing etc.)
Water	Tap water, mineral water
Fish	Fish, canned fish, smoked fish
Nuts	Nuts
Chips, salt sticks	Chips, salt sticks, cracker
Fried potatoes	French fries, potato fritters, fried potatoes
Beer	Beer
Spirits	Spirits
Wine	Wine, fruit wine, champagne
Other alcoholic beverages	Dessert wine, liqueur, aperitif
Eggs	Boiled eggs, fried eggs, omelette
Coffee	Coffee
Soup	Vegetable and potato stew, vegetable soup, meat and fish soup, broth, thickened soup
Processed meat	Salami, cold cuts sausage, ham, fried sausage
Low-fat dairy products	Milk and yogurt (≤150 g fat/kg), soured milk, low-fat curd cheese
High-fat dairy products	other milk and yogurt, or curd cheese, cream
Low-fat cheese	Low-fat fromage, low-fat cheese
High-fat cheese	Other fromage or cheese
Butter	Butter as bread spread and for food preparation
Margarine	Margarine as bread spread and for food preparation

A pattern variable from exploratory factor analysis can be described by the following equation:

$$\text{Pattern variable} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_m X_m, \quad (1)$$

where X_i, $i = 1, \ldots, m$, are the standardised food variables with zero mean and unit standard deviation. The parameters β_i, $i = 1, \ldots, m$, can be considered as weights that are equivalent to the factor loadings. Factor loadings are interpreted as the correlation coefficients between food variables and orthogonal factors (Hatcher, 1994).

It is, however, well recognised that other strategies of pattern variable determination, that are more simple than the optimal solution from equation 1, lead to pattern variables of the same tenor. For example, Comrey (1988) proposed to sum the unweighted individual values of the original variables which load most highly on the pattern, an approach that has not been implemented into pattern analysis in nutritional epidemiology so far. The corresponding simplified pattern variable applied in the present study, deduced from equation 1, is given by:

$$\text{Simplified pattern variable} = X_1 + X_2 + X_3 + \ldots + X_r \ (r < m), \quad (2)$$

where the omitted m–r variables are characterised by low weights β_i. In the case of negative factor loadings, a negative algebraic sign was assigned to the corresponding food groups in the calculation of the simplified pattern variable.

Pearson correlation coefficients between the simplified pattern variable and its entered food variables generally tend to increase with increasing intercorrelation among food variables (see Table 2 and Appendix). The sum of a relatively small number of standardised variables, as in the ideal case of a simplified pattern variable (equation 2), will even be markedly correlated with the original variables x_r if intercorrelations among x_r are small. For example, in the case of zero correlation between food variables, the simplified pattern variable based on four food items will have an average Pearson correlation 0·50 with the food items, while one based on ten items will have a correlation of only 0·32. This result is based on a mathematical formula proven in the Appendix. Consequently, simplified pattern variables can be uniformly applied to represent patterns from factor analysis carried out in different populations, even if the original food variables do not demonstrate similarly strong intercorrelations, but rather weak intercorrelations. As well as a simplified pattern variable as derived reflecting a pattern from factor analysis of the EPIC-Potsdam study dietary data, a second simplified pattern variable was constructed similar to a pattern reported to be associated with cardiovascular disease in an US study population (Hu *et al.* 2000). The latter pattern, originally labelled 'western', reflects a diet relatively high in meat, processed meat, refined grains, chips and popcorn, sweets and deserts, French fries, high-fat dairy products, high-energy soft drinks and eggs. These food groups were used to calculate a corresponding simplified pattern variable in the EPIC-Potsdam study.

Statistical analysis

In a first step, agreement between the first identified original factor analysis-based pattern variable (equation 1) and a deduced simplified pattern variable (equation 2) based on the same data set was assessed by calculating Pearson correlation coefficients and by comparing associations with original food variables.

Second, agreement between the US pattern originally reported and the simplified pattern variable was assessed by comparing food and nutrient profiles between both patterns. Here, Pearson correlation coefficients were calculated between the simplified pattern variable and food as well as nutrient intake, and these coefficients were compared with the corresponding correlation coefficients of the 'western' pattern that was originally reported by Hu *et al.* (2000). All analyses were performed with the SAS System® for Windows™, release 8.00 (SAS Institute Inc., Cary, NC, USA).

Results

Constructing simplified pattern variables from the same study populations

That the pattern variables defined by the simplified equation 2 and the original equation 1 were highly correlated is demonstrated in Fig. 1. Here, Pearson correlation coefficients between both pattern variables for varying degrees of simplification are shown. The simplified pattern variables were based on different numbers of standardised food variables, starting with the one that showed the highest loading at the original pattern, and adding consecutively food variables with decreasing loadings. A simplified pattern variable based on the six highest-loading food variables (six food items had loadings $>0{\cdot}5$ at the original pattern) showed already a correlation $>0{\cdot}95$ with the original factor score. Note that the correlation decreased when food variables with small loadings were added. The simplified pattern variable demonstrated similar correlations with food variables as the original factor score (Table 3).

Applying simplified pattern variables derived from different populations

A simplified pattern variable was applied that has been constructed correspondingly to the 'western' pattern reported from an US population by Hu *et al.* (2000; Table 4). Intercorrelations between food groups, reported to be correlated with the 'western' pattern (Hu *et al.* 2000), were generally low ($\leq 0{\cdot}30$) with a mean value

Table 2. Relationships between simplified pattern variables and original variables for varying intercorrelations between original variables and for varying numbers of original variables combined in the simplified pattern variable*

(Mean Pearson correlation coefficients for 8975 men)

	Mean Pearson correlation coefficients between original variables									
No. of original variables	0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
2	0·71	0·74	0·77	0·81	0·84	0·87	0·89	0·92	0·95	0·97
3	0·58	0·63	0·68	0·73	0·77	0·82	0·86	0·89	0·93	0·97
4	0·50	0·57	0·63	0·69	0·74	0·79	0·84	0·88	0·92	0·96
5	0·45	0·53	0·60	0·66	0·72	0·77	0·82	0·87	0·92	0·96
6	0·41	0·50	0·58	0·65	0·71	0·76	0·82	0·87	0·91	0·96
7	0·38	0·48	0·56	0·63	0·70	0·76	0·81	0·86	0·91	0·96
8	0·35	0·46	0·55	0·62	0·69	0·75	0·81	0·86	0·91	0·96
9	0·33	0·45	0·54	0·61	0·68	0·75	0·80	0·86	0·91	0·95
10	0·32	0·44	0·53	0·61	0·68	0·74	0·80	0·85	0·91	0·95

* Data were from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study (Riboli & Kaaks 1997; Boeing *et al.* 1999); for details of calculations and procedures, see p. 410 and Appendix.

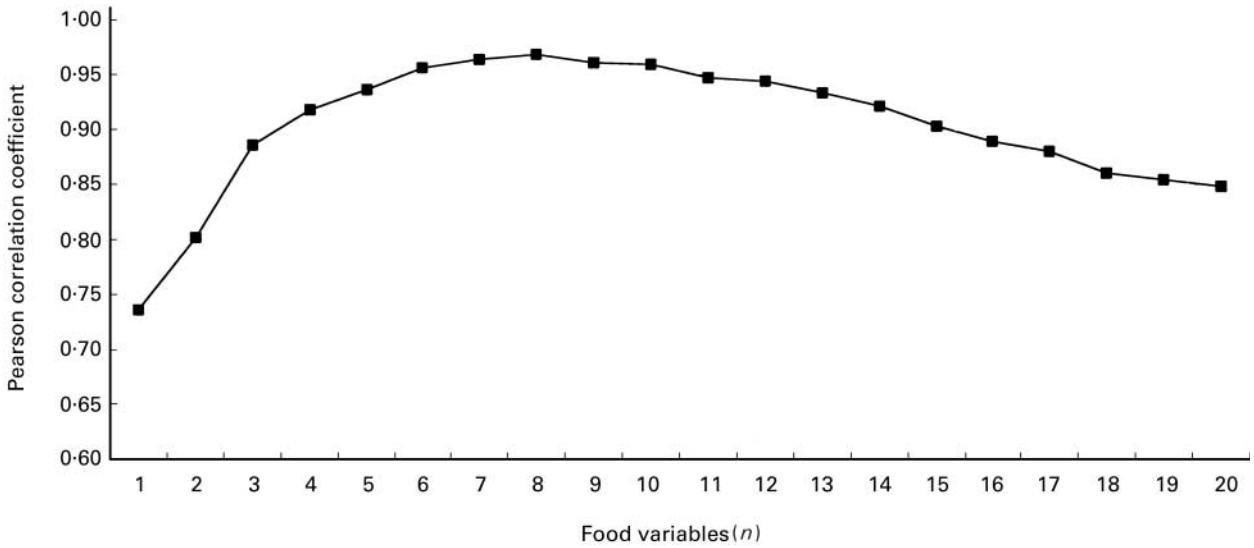

Fig. 1. Pearson correlation between a factor-analysis-based pattern variable (sum of forty-seven optimally weighted standardised food variables) and deduced simplified pattern variables (sum of unweighted standardised food variables, starting with the one that showed the highest correlation with the original factor-analysis-based pattern and adding consecutively food variables with decreasing correlation coefficients). The number (*n*) of food variables that were assigned to the simplified pattern variable are displayed on the *x*-axis; the corresponding factor loadings of the consecutively added single food variables from the original exploratory factor analysis were: 1 0·74, 2 0·67, 3 0·66, 4 0·55, 5 0·52, 6 0·52, 7 0·38, 8 0·38, 9 0·24, 10 0·23, 11 0·21, 12 0·20, 13 0·18, 14 0·16, 15 0·10, 16 0·09, 17 0·09, 18 0·09, 19 0·09, 20 0·09. Data were from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study (Riboli & Kaaks, 1997; Boeing *et al.* 1999) for 8975 men. For details of calculations and procedures, see p. 410 and Appendix.

0·09. The constructed simplified pattern variable had, however, markedly higher correlations with these food variables, namely 0·44 on average, which can also be calculated by using the formula shown in the Appendix. Furthermore, the simplified pattern variable was positively correlated with the energy-adjusted intakes of protein, saturated fat, monounsaturated fat and cholesterol (Pearson correlation coefficients 0·20, 0·13, 0·21 and 0·32 respectively), but negatively correlated with fibre and folate intake ($-0·14$ and $-0·18$ respectively). Energy-adjusted

Table 3. Relationships between the intake of single food groups and an original factor score and a deduced simplified pattern variable*

(Pearson correlation coefficients for 8975 men)

Food group	Original factor score†	Simplified pattern variable‡
Meat except fish and poultry	0·74	0·73
Sauce	0·67	0·69
Cooked vegetables	0·66	0·66
Potatoes	0·55	0·60
Poultry	0·52	0·57
Cabbage family	0·52	0·57
Mushrooms	0·38	0·27
Legumes	0·38	0·25
Fried potatoes	0·24	0·16
Animal fat, except butter	0·23	0·15
Soup	0·21	0·14
Pasta	0·20	0·10
Canned fruits	0·18	0·15
Vegetable oils and fats except olive oil	0·16	0·16
Dessert	0·10	0·09
Processed meat	0·09	0·12
Coffee	0·09	0·08
Fish	0·09	0·11
Remaining twenty-nine food groups	≤0·10	≤0·10

* Data were from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study (Riboli & Kaaks, 1997; Boeing *et al.* 1999); for details of calculations and procedures, see p. 410 and Appendix.

† Original factor score = sum of forty-seven optimally-weighted standardised food variables derived from factor analysis.

‡ Simplified pattern variable = sum of unweighted standardised intake of meat, sauce, cooked vegetables, potatoes, poultry, and cabbage.

Table 4. Relationships between single food groups and a simplified pattern variable that corresponds to the 'western' dietary pattern*†‡
(Pearson correlation coefficients for 8975 men)

	Meat	Processed meat	Refined grains	Chips, popcorn	Sweets and desserts	French fries	High-fat dairy products	High-energy drinks	Eggs	Simplified pattern variable§
Meat	1·00									0·48
Processed meat	0·20	1·00								0·52
Refined grains	0·08	0·30	1·00							0·43
Chips, popcorn	0·04	0·06	0·02	1·00						0·37
Sweets and desserts	0·10	0·12	0·10	0·12	1·00					0·49
French fries	0·23	0·07	0·01	0·08	0·09	1·00				0·44
High-fat dairy products	0·01	0·01	0·01	0·01	0·19	0·02	1·00			0·34
High-energy drinks	0·07	0·10	0·08	0·06	0·10	0·05	0·01	1·00		0·39
Eggs	0·17	0·17	0·08	0·05	0·11	0·18	0·07	0·07	1·00	0·48

* Data were from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study (Riboli & Kaaks 1997; Boeing *et al.* 1999); for details of calculations and procedures, see p. 410 and Appendix.
† 'Western' dietary pattern reported by Hu *et al.* (2000).
‡ Mean correlation between original variables 0·09, mean correlation between original variables and simplified pattern variable 0·44.
§ Simplified pattern variable = sum of unweighted standardised food variables.

intakes of carbohydrates and polyunsaturated fat showed only weak correlations with the simplified pattern variable (Pearson correlation coefficients −0·06 and 0·05 respectively).

Discussion

The present study demonstrated that the approach of 'simplified pattern variables' might be successfully used to construct less data-dependent pattern variables from exploratory analysis of the same or of other populations. These simplified pattern variables represent the sum of the unweighted standardised food variables which loaded high at the pattern of interest. This approach is conceptually meaningful. It conforms to the interpretability criteria of factor analysis (Hatcher, 1994), where observed variables should either show high factor loadings with the pattern if they are measures of the latent construct or should show near zero loadings if not. Since the weights, (β_i) in the original pattern variable (equation 1) are equivalently either high or near zero too, food variables with near zero loadings will contribute only a minor part to the factor score. In the case that all high-loading food variables have approximately similar loadings, they will almost have similar contributions to the pattern variable. The similarity of the original factor score and the simplified pattern variable to reflect the same pattern therefore depends on the extent of how much the factor solution resembles the so-called 'simple structure' and on the variation of high factor loadings. The more the original factor loadings of those food items that will be retained varies, the more will the simplified pattern variable depart from the weights originally assigned. In our present example, the six highest original factor loadings ranged from 0·74 to 0·52, with the simplified pattern having a correlation >0·95 with the original factor score. While this simplified approach has been found to yield acceptable estimates of the true underlying score and is considered robust and appropriate when factor loadings are reasonably high (Comrey, 1988), it remains unclear whether the loss of information might be unacceptably high in the case that the variation of factor loadings is more pronounced than in our present example.

We demonstrated that the construction of a simplified pattern variable resulted only in a minor loss of information compared with the more elaborate method of factor score determination in factor analysis. A simplified pattern variable based on the six highest loading food variables showed a correlation >0·95 with the original factor score, which consisted of forty-seven food variables. In other words, $(1 - 0{\cdot}95^2) \times 100\,\%$, i.e. <10 %, of information was lost by using the simplified instead of the original pattern variable. In our present example, the advantage lays clearly with the simplified pattern variable, which is much more easily interpretable. Only few food variables were retained, all data-dependent weights have been neglected and all retained variables have equal importance. However, whether the simplified pattern approach always leads to a minor loss of information compared with the more precise factor score determination remains unclear and should be addressed in future research.

Moreover, only a few attempts have been undertaken to insure the internal validity of identified pattern structures. In particular, confirmatory factor analysis (Hatcher, 1994) can be sued to test the goodness of fit of the extracted factor structures. So far, this method has hardly been used in dietary pattern analysis (Maskarinec *et al.* 2000). To ensure the independence of the exploratory and confirmatory analysis, both analyses can be performed in split samples. This procedure has, however, the drawback that factor scores cannot be directly saved from the confirmatory analysis. Here, the simplified pattern approach might represent an option to calculate pattern variables that reflect the confirmed structure, since the simplified variables correspond with the tested measurement model that consists of only those food variables with high loadings which are seen to be indicators of the latent construct (Hatcher, 1994). Glass *et al.* (1997) reported that the simplified method of score determination in confirmatory factor analysis yields essentially identical results compared with a pattern score incorporating weights from the measurement model.

Furthermore, it seems possible to construct pattern variables that approximate factor-analysis-based patterns even if the original food variables do not demonstrate strong intercorrelations, but rather weak intercorrelations. This was theoretically proven, as well as practically, by constructing a simplified pattern variable corresponding to a pattern reported from a different population. Prior exploratory factor analysis of the German EPIC-Potsdam study population did not yield a pattern similar to the 'western' pattern (Schulze *et al.* 2001), indicating that this pattern does not explain a major portion of total variance of food intake in this specific study population. However, the simplified pattern variable showed similar associations with food as well as nutrient intake as the originally reported 'western' pattern (Hu *et al.* 2000). One might argue that some associations between the simplified pattern and food items are too low to be acceptable as measuring this pattern (e.g. correlation for high-fat dairy products r 0·34). However, cut-off points as low as 0·2 for factor loadings have frequently been used in the context of determining those food items being significantly associated with factor-analysis-based dietary patterns (Slattery *et al.* 1998; Hu *et al.* 1999). Clearly, relatively low loadings limit the interpretability of the pattern structure (Martinez *et al.* 1998), but a minimum limit has not been agreed on so far.

One might question whether it is reasonable to transplant factor-analysis-based patterns from one population to another, especially if the food variables aggregated correlate poorly in the latter. Clearly, it is not reasonable to transplant all patterns derived by exploratory methods, but it might be reasonable to transplant those that have been proven to be associated with disease. Simplified pattern variables aim to reflect dietary exposure patterns across study populations rather than to maintain high intercorrelation between variables within a pattern. This disease-specific focus should not be confused with exploratory attempts to explain a maximum of variance. High intercorrelation between original variables are therefore not a prerequisite for constructing meaningful *a priori* pattern variables (Kant, 1996; Kant *et al.* 2000; Osler *et al.* 2001). For example, Kouris-Blazos *et al.* (1999) have demonstrated that adherence to principles of the Mediterranean diet, as determined by an *a priori* pattern score, was associated with overall mortality in elderly Anglo-Celtic Australians. While this pattern variable was based on evidence from prior observational studies on Mediterranean diet and mortality in Mediterranean populations (Trichopoulou *et al.* 1995), this pattern was unlikely to explain a high portion of variance within the study population in which it was tested. We have demonstrated that the simplified pattern approach might be applied to reflect a dietary pattern, which has been reported to be associated with cardiovascular disease. This view is also in contrast to that of Randall *et al.* (1990), who suggested that a link between patterns and disease risk is most likely to be identifiable among those patterns contributing most to the variance in dietary intake. This statement implies that *a posteriori* approaches are superior to *a priori* approaches. While this has never been proven, we have demonstrated in a previous study (Schulze *et al.* 2001) that factor-analysis-based patterns might explain the intake of single food items and nutrients quite differently. In cases where food items that are likely to be related to the outcome are not well explained, the exploratory factor solution might not be very useful to explain risk.

Factor solutions are generally not reproducible and risk estimates of single studies are therefore not comparable (Kant, 1996; Martinez *et al.* 1998; Jacques & Tucker, 2001). Hypothesis-driven patterns have the advantage of being created with respect to the specific outcome (and thus might be more easily interpretable with respect to their biological plausibility) and of being constructs that are applicable across different populations. The latter makes them favourable in multi-centre studies, like EPIC, and assures that risk estimates of single studies are comparable and might be aggregated to an overall measure of effect. Thus, if the simplified pattern variables reflect the specific pattern of interest and can be applied uniformly to different populations, this approach might be useful to standardise and compare factor-analysis-based risk estimates across different populations. This has not been possible so far.

It is noteworthy that different populations might vary largely with regard to their average quantitative intake of specific food items or groups and its variation. The standardisation of food items in the process of constructing simplified pattern variables, however, sets any population average value to 0 and the standard deviation to 1. The simplified pattern approach is therefore unable to account for quantitative differences in food intake across different populations. This is an important point, as the application of a simplified pattern to a population with only minor variation in food intake will not be very useful, and as effects of pattern measures might vary widely between populations with largely varying average intakes. However, this drawback applies to the comparison of factor-analysis-based patterns retained in different populations as well, since both methods rely on standardised original variables. Other methods of defining *a priori* patterns, particularly based on quantitative cut-off points (Kant, 1996), have their advantage here, although differences in dietary assessment or differences in the response of different

populations to the same questionnaire might limit the usefulness of absolute intake levels to define pattern scores (Willett, 1998). Food-frequency questionnaires, as the one applied in the EPIC-Potsdam study, have limited usefulness for defining absolute quantitative intake values and are seen to be more useful to rank individuals with regard to their intake (Willett, 1998). A transformation of food variables without defining quantitative cut-off points, as is realised with the standardisation in factor analysis and the simplified pattern approach, seems therefore to better account for the semiquantitative character of food-frequency questionnaire data. However, other semiquantitative methods of defining the pattern score might be applicable. For example, Stampfer *et al.* (2000) and Hu *et al.* (2001) used quintiles for several dietary variables to calculate pattern scores. This approach might be more appropriate to avoid overweighting of extreme intakes that cannot be ruled out by standardising the food variables. Still, the simplified pattern approach focuses on constructing patterns similar to those from factor analysis, and consequently the same disadvantage of standardised food variables applies to it as is present for original factor scores.

A major point of concern applying the simplified pattern approach across different populations is the comparability of food variables across studies. Only studies with comparable data will benefit optimally from this approach. While it might be possible to construct a pattern variable in a northern European country reflecting an US-based 'western' pattern, as was shown in the present study, this approach might not be applicable in study populations where food intake is different. For example, the interpretation of a 'western' pattern variable in study populations not consuming meat and eggs will largely depart from the original 'western' pattern that is characterised by high intakes of both food items. Furthermore, while studies might use the same food groups, not every study assesses the same kind of food items and those food items aggregated to the groups might therefore differ. For example, 'vegetables' might reflect different items in a southern than in a northern European population. The latter is not a specific concern for pattern analysis, but rather a problem of any comparison of food-group-based dietary data across different populations. The applied food grouping has been a matter of debate for factor analysis as well (Martinez *et al.* 1998), with only a little information yet on whether the applied grouping influences retained patterns and subsequent risk estimates (McCann *et al.* 2001). Few attempts have been made so far to develop food-grouping schemes that are applicable across various population, such as the EUROCODE (Kohlmeier, 1992), and no 'standard' has been agreed upon in the context of dietary pattern analysis, where the number of analysed food groups ranged from fifteen (Kumagai *et al.* 1999; Schwerin *et al.* 1981) to ninety-five (Randall *et al.* 1992). Clearly, further research is needed to define optimal food groups for factor analysis, including studies comparing different populations, as has been done recently within the EPIC study (Slimani *et al.* 2002). Furthermore, evidence from observational and interventional epidemiological studies might be used to affirm the importance of the selected food groups.

A further criticism of applying simplified pattern variables to reflect factor-analysis-based patterns across different populations is the possibly arbitrary decision on which food items correspond to the originally observed pattern. In our present study, other simplified pattern variables were observed to correlate almost as highly with the original factor score as the variable from six food items, indicating that a variety of solutions might almost similar adequately reflect the original pattern score. Variables omitted from the simplified pattern variable are not necessarily unimportant for the originally observed pattern; rather the ones remaining might be well enough correlated with some of those omitted to capture the flavour of the original score. The decision on which set of food items characterises the original pattern and therefore the exact nature of the pattern might vary from study to study, especially if the original factor solution did not demonstrate a clear simple structure with either high or near zero factor loadings.

A further disadvantage of simplified pattern variables might be that patterns can be markedly correlated with other food variables not included in the pattern. In our present example, the simplified pattern variable for the 'western' pattern showed only small correlation coefficients with most food items, among them fruits, vegetables, potatoes, nuts, low-fat dairy products, juices, coffee, alcoholic beverages, butter, margarine, fish, poultry and soups, but was markedly correlated with sauce intake. This issue of possible high intercorrelation is not unique to the proposed simplified pattern variables, but rather is a characteristic of any dietary pattern approach that is not data-dependent. Clearly, factor analysis has its advantage here, since it assures that no original variables are highly correlated with the pattern other than those with high factor loadings.

Our present study suggests that simplified pattern variables allow to reduce a number of food variables into a pattern variable that can be predefined from exploratory analysis even of other populations. These simplified pattern variables can subsequently be used to study the association of the pattern with disease risk. It might therefore be an approach that could overcome one of the most frequently claimed limitations of factor analyses applied in epidemiological research: their non-reproducible risk estimates (Kant, 1996; Martinez *et al.* 1998, Jacques & Tucker, 2001). However, the usefulness of this approach needs to be determined in future studies that address its applicability in different settings and test whether the simplification generally leads to a minor loss of information across different factor solutions.

Acknowledgements

We wish to thank all study participants for their cooperation and all interviewers, programmers and document

coordinators who devoted their energy to collect and process all the data. The recruitment phase of the EPIC-Potsdam study was mainly supported by the Federal Ministry of Science, Germany; grant no. 01 EA 9401. Further financial support was given by the 'Europe against Cancer' programme of the European Community (grant no. SOC 95 201408 05F02). The EPIC-Potsdam study is now supported by the Deutsche Krebshilfe (grant no. 70-2488-Ha I) and the European Community (grant no. SOC 98 200769 05F02). This present study was furthermore financially supported by the German Research Foundation (grant no. BO 807/6-1).

References

Boeing H, Wahrendorf J & Becker N (1999) EPIC-Germany – A source for studies into diet and risk of chronic diseases. *Annals of Nutrition and Metabolism* **43**, 195–204.

Comrey AL (1988) Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology* **56**, 754–761.

Dehne LI, Klemm C, Henseler G & Hermann-Kunz E (1999) The German Food Code and Nutrient Data Base (BLS II.2). *European Journal of Epidemiology* **15**, 355–359.

Fung TT, Willett WC, Stampfer MJ, Manson JE & Hu FB (2001) Dietary patterns and the risk of coronary heart disease in women. *Archives of Internal Medicine* **161**, 1857–1862.

Gex-Fabry M, Raymond L & Jeanneret O (1988) Multivariate analysis of dietary patterns in 939 Swiss adults: sociodemographic parameters and alcohol consumption profiles. *International Journal of Epidemiology* **17**, 548–555.

Glass TA, Mendes de Leon CF, Seeman TE & Berkman LF (1997) Beyond single indicators of social networks: a LISREL analysis of social ties among the elderly. *Social Science and Medicine* **44**, 1503–1517.

Hatcher L (1994) *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute Inc.

Hoffmann K, Schulze MB, Boeing H & Altenburg HP (2002) Dietary patterns: report of an international workshop. *Public Health Nutrition* **5**, 89–90.

Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Current Opinion in Lipidology* **13**, 3–9.

Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG & Willett WC (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine* **345**, 790–797.

Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, Sampson L & Willett WC (1999) Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *American Journal of Clinical Nutrition* **69**, 243–249.

Hu FB, Rimm EB, Stampfer MJ, Ascherio A, Spiegelman D & Willett WC (2000) Prospective study of major dietary patterns and risk of coronary heart disease in men. *American Journal of Clinical Nutrition* **72**, 912–921.

Jacques PF & Tucker KL (2001) Are dietary patterns useful for understanding the role of diet in chronic disease? *American Journal of Clinical Nutrition* **73**, 1–2.

Kant AK (1996) Indexes of overall diet quality: a review. *Journal of the American Dietetic Association* **96**, 785–791.

Kant AK, Schatzkin A, Graubard BI & Schairer C (2000) A prospective study of diet quality and mortality in women. *Journal of the American Medical Association* **283**, 2109–2115.

Kennedy ET, Ohls J, Carlson S & Fleming K (1995) The Healthy Eating Index: design and applications. *Journal of the American Dietetic Association* **95**, 1103–1108.

Kohlmeier L (1992) The Eurocode 2 food coding system. *European Journal of Clinical Nutrition* **46**, Suppl. 5, S25–S34.

Kouris-Blazos A, Gnardellis C, Wahlqvist ML, Trichopoulos D, Lukito W & Trichopoulou A (1999) Are the advantages of the Mediterranean diet transferable to other populations? A cohort study in Melbourne, Australia. *British Journal of Nutrition* **82**, 57–61.

Kumagai S, Shibata H, Watanabe S, Suzuki T & Haga H (1999) Effect of food intake pattern on all-cause mortality in the community elderly: a 7-year longitudinal study. *Journal of Nutrition, Health and Aging* **3**, 29–33.

McCann SE, Marshall JR, Brasure JR, Graham S & Freudenheim JL (2001) Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. *Public Health Nutrition* **4**, 989–997.

Martinez ME, Marshall JR & Sechrest L (1998) Invited commentary: Factor analysis and the search for objectivity. *American Journal of Epidemiology* **148**, 17–19.

Maskarinec G, Novotny R & Tasaki K (2000) Dietary patterns are associated with body mass index in multiethnic women. *Journal of Nutrition* **130**, 3068–3072.

Osler M, Heitmann BL, Gerdes LU, Jørgensen LM & Schroll M (2001) Dietary patterns and mortality in Danish men and women: a prospective observational study. *British Journal of Nutrition* **85**, 219–225.

Randall E, Marshall JR, Brasure J & Graham S (1992) Dietary patterns and colon cancer in western New York. *Nutrition and Cancer* **18**, 265–276.

Randall E, Marshall JR, Graham S & Brasure J (1990) Patterns in food use and their association with nutrient intakes. *American Journal of Clinical Nutrition* **52**, 739–745.

Riboli E & Kaaks R (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *International Journal of Epidemiology* **26**, Suppl. 1, S6–S14.

Schulze MB, Hoffmann K, Kroke A & Boeing H (2001) Dietary patterns and their association with food and nutrient intake in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study. *British Journal of Nutrition* **85**, 363–373.

Schwerin HS, Stanton JL, Riley AM Jr, Schaefer AE, Leveille GA, Elliott JG, Warwick KM & Brett BE (1981) Food eating patterns and health: a reexamination of the Ten-State and NHANES I surveys. *American Journal of Clinical Nutrition* **34**, 568–580.

Slattery ML, Boucher KM, Caan BJ, Potter JD & Ma KN (1998) Eating patterns and risk of colon cancer. *American Journal of Epidemiology* **148**, 4–16.

Slimani N, Fahey M, Welch AA, Wirfält E, Stripp C, Bergström E, Linseisen J, Schulze MB, Bamia C, Chloptsios Y, Veglia F, Panico S, Bueno-de-Mesquita HB, Ocké MC, Brustad M, *et al.* (2002) Diversity of dietary patterns observed in the EPIC project. *Public Health Nutrition* (In the Press).

Stampfer MJ, Hu FB, Manson JE, Rimm EB & Willett WC (2000) Primary prevention of coronary heart disease in women through diet and lifestyle. *New England Journal of Medicine* **343**, 16–22.

Trichopoulou A, Kouris-Blazos A, Wahlqvist ML, Gnardellis C,

Lagiou P, Polychronopoulos E, Vassilakou T, Lipworth L & Trichopoulos D (1995) Diet and overall survival in elderly people. *British Medical Journal* **311**, 1457–1460.
Trichopoulos D & Lagiou P (2001) Dietary patterns and mortality. *British Journal of Nutrition* **85**, 133–134.
Tseng M, DeVellis RF, Maurer KR, Khare M, Kohlmeier L, Everhart JE & Sandler RS (2000) Food intake patterns and gallbladder disease in Mexican Americans. *Public Health Nutrition* **3**, 233–243.
US Department of Agriculture (1995) *Nutrition and Your Health: Dietary Guidelines for Americans*. Washington, DC: US Government Printing Office.
Whichelow MJ & Prevost AT (1996) Dietary patterns and their associations with demographic, lifestyle and health variables in a random sample of British adults. *British Journal of Nutrition* **76**, 17–30.
Willett W & Stampfer MJ (1986) Total energy intake: implications for epidemiologic analyses. *American Journal of Epidemiology* **124**, 17–27.
Willett WC (1998) *Nutritional Epidemiology*. New York: Oxford University Press.
Williams DEM, Prevost AT, Whichelow MJ, Cox BD, Day NE & Wareham NJ (2000) A cross-sectional study of dietary patterns with glucose intolerance and other features of the metabolic syndrome. *British Journal of Nutrition* **83**, 257–266.

Appendix

Correlation between standardised variables and its sum

Consider m standardised variables $X_1, \ldots, X_m$. Denote the covariance between X_i and X_j, $i \neq j$, by ρ_{ij} and the mean covariance by $\bar{\rho}$. Then, the mean Pearson correlation coefficient between standardised variables and their sum can be written as:

$$\frac{1}{m}\sum_{i=1}^{m} Corr(X_1 + \ldots + X_m, X_i)$$

$$= \frac{1}{m}\sum_{i=1}^{m} \frac{Cov(X_1 + \ldots + X_m, X_i)}{\sqrt{Var(X_1 + \ldots + X_m)}\sqrt{Var(X_i)}}$$

$$= \frac{1}{m}\frac{Cov(X_1 + \ldots + X_m, X_1 + \ldots + X_m)}{\sqrt{Var(X_1 + \ldots + X_m)}}$$

$$= \frac{1}{m}\sqrt{Var(X_1 + \ldots + X_m)} \quad = \frac{1}{m}\sqrt{m + m(m-1)\rho}$$

$$= \frac{1}{\sqrt{m}}\sqrt{1 + (m-1)\rho}.$$

In the special case of identical covariance $\rho_{ij} = \rho$, each standardised variable X_i has the same covariance with its sum, namely:

$$Corr(X_1 + \ldots + X_m, X_i) = \frac{1}{\sqrt{m}}\sqrt{1 + (m-1)\rho}.$$

Table 2 presents estimates of mean correlation coefficients between simplified pattern variables and original variables. The correlation coefficients were calculated for varying numbers of original variables combined in the simplified pattern variable as well as for varying intercorrelations between the original variables.