Nutrition Research Reviews (2005), **18**, 241–248
DOI: 10.1079/NRR2005107

Can dietary patterns help us detect diet–disease associations?

Karin B. Michels[1,2]* and Matthias B. Schulze[3]

[1]*Obstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[2]*Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA*
[3]*Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany*

The role of diet in promoting health and preventing disease is difficult to elucidate due to its complex network of foods and nutrients. Besides total energy intake, dietary composition is probably the most important discriminator within and between populations. Dietary composition is reflected in dietary patterns, which have recently gained popularity. The present paper reviews the most commonly applied methods to identify dietary patterns, data-driven methods such as factor and cluster analysis, investigator-driven methods such as indices and score, and methods combining the two, namely reduced rank regression. We describe the techniques and their application, discuss strengths and limitations, and discuss the usefulness of dietary pattern analyses.

Dietary patterns: Epidemiology: Factor analysis: Principal components: Cluster analysis: Dietary composition

Introduction

The role of diet in promoting health and preventing disease is difficult to elucidate due to its complex network of foods and nutrients. Nutrition is essential to maintain life, but dietary composition can differ widely among individuals. While a limited range of total energy intake and a minimum amount of essential nutrients are required for survival, it is unclear whether certain dietary preferences or dietary patterns are most relevant for the prevention or promotion of certain diseases or whether we may be able to link intake of individual foods or nutrients to disease outcomes. The complexities of diet make it difficult to consider the role of individual foods or nutrients in isolation. Nutrients may interact with each other and influence their bioavailability and absorption. Because many nutrients are contained in the same foods, isolating their individual effects is almost impossible. Furthermore, given the limited variability of energy intake, a high consumption of one food must be associated with lower intake of other foods, which makes inferences about the relevance of individual foods even more difficult. An individual with a high daily consumption of red meat might consume fish rarely and eat few fruits. The same individual may also have a high intake of refined carbohydrates. Hence, confounding by foods not eaten or by foods highly correlated with the index food may result in a distorted picture of the role of individual food items or nutrients in disease causation.

Conversely, including information in an analytical model on the frequency of consumption of all food items assessed poses different problems such as collinearity. A summary variable characterising a diet pattern may be easier to handle.

Therefore the question arises whether dietary preferences reflected in dietary patterns may be more suitable for analyses in nutritional epidemiology when large populations and their dietary habits are studied in the search of disease causation. Particular combinations of foods as described by dietary patterns may be more strongly related to health and disease than individual foods or nutrients. Furthermore, dietary preferences may be more consistent over time than consumption of individual foods.

Recently, a number of methods have been introduced to use the wealth of dietary information collected in observational studies to define dietary patterns and relate them to disease outcomes (Trichopoulos & Lagiou, 2001; Hu, 2002). These dietary pattern methods are either data driven, such as principal component analysis, factor analysis and cluster analysis, or determined *a priori* by the investigator, such as dietary indices or dietary scores. All previously mentioned patterns are constructed independent of the disease endpoint of interest – hence the same patterns would be used for the analysis of dietary predictors of heart disease and in an analytical model to explore cancer. To refine patterns and target them to a specific disease outcome, reduced rank regression (RRR) – the latest addition in

Abbreviation: RRR, reduced rank regression.

* **Corresponding author:** Dr Karin B. Michels, fax +1 617 732 4899, email kmichels@rics.bwh.harvard.edu

dietary pattern research – combines exploratory methods and the use of prior knowledge to identify a dietary pattern that is associated with a specific disease. In the following, we discuss each of these methods, their concepts and applications, and their strengths and limitations in our search for diet–disease associations.

Data-driven methods

Data-driven methods to identify dietary patterns are also called *a posteriori* methods because the available data determine the patterns. Among data-driven methods, factor analysis has emerged as the most frequently used method, while cluster analysis is popular with some groups.

Principal component analysis and factor analysis

Concepts and methods. Principal component analysis (Manly, 2004) identifies foods that are frequently consumed together (Schwerin *et al*. 1981; Jacobson, 1986; Slattery *et al.* 1998; Hu *et al.* 1999; Schulze *et al.* 2001). It aggregates food items or food groups on the basis of the degree to which they are correlated with one another (i.e. found to be consumed together in diet assessments). The goal is to identify linear composites of optimally weighted food items or food groups (principal components) that account for the largest amount of variation in diet between individuals where each observed food or food group contributes one unit of variance to the total variance in the dataset. In contrast, factor analysis assumes that the observed variables (food items or groups) are linear combinations of unobservable (latent) factors. While in principal component analysis the component score represents a mathematical transformation (a linear combination) of the observed variables, factor scores computed in factor analysis are considered only estimates of the individual's actual underlying unobservable factor. Besides theoretical differences between principal component analysis and factor analysis, factor analysis based on the principal factor method gives generally similar results as principal component analysis. However, factor analysis may involve other methods than the principal factor method, for example, maximum-likelihood algorithms, that clearly distinguishes it in these cases from principal component analysis. Because the vast majority of studies on dietary patterns applied principal component analysis or factor analysis with the principal factor method and because both methods share the same mathematical concept, we discuss issues of their application together. Studies on dietary patterns, which applied principal component analysis or factor analysis were comprehensively reviewed by Newby & Tucker (2004).

It has become common practice to pregroup individual food items into food groups before applying principal component analysis or factor analysis, since the proportion of explained variance per factor decreases with the number of variables entered (Slattery *et al.* 1998). This approach reduces the number of food groups available for analysis substantially (mostly to approximately forty items).

The importance of a component or factor is reflected in its Eigenvalue, which represents the amount of variance that is accounted for by a given component or factor. Any component or factor that displays an Eigenvalue greater than one is accounting for a greater amount of variance than is contributed by one individual food item or food group and therefore accounts for a meaningful amount of variance. In determining the number of components or factors to retain, Eigenvalues greater than one are generally used as a decision criterion in conjunction with the scree test and interpretability of the components or factors. A scree plot of the Eigenvalues can be used to graphically determine the optimal number of components or factors to retain. A break between the components or factors allows us to distinguish between those with relatively large Eigenvalues and those with small Eigenvalues. The components or factors that appear before the break are assumed to be meaningful and are retained for rotation.

Initial components or factors identified are usually rotated by an orthogonal transformation to achieve simple structure with greater interpretability. Ideally, after-rotation factor loadings (correlation between components or factors and foods) are either high or near zero and single food items have high loadings at only one component or factor.

The percentage of variance explained by each factor depends largely on the total number of variables included and is therefore an inferior decision criterion of how many and which factors to maintain. The proportion of variance accounted for by a factor can be calculated as the Eigenvalue for that component divided by the total Eigenvalues of the correlation matrix. The total Eigenvalues of the correlation matrix are equal to the total number of variables being analysed. The factor-loading matrix is usually used to determine what foods are important contributors to the factor, although this is purely interpretational and involves the arbitrary decision about which cut-off to use. The factor score for each pattern can be computed by combining the standardised food variables with weights that are proportionate to their component loadings. However, a simpler approach is to combine with equal weight only those standardised food groups that showed high factor loadings (Schulze *et al.* 2003). Such a simplified score appears to correlate highly with the more complex score, but has the advantage of much easier calculation and interpretation.

The factor-loading matrix for dietary patterns often does not represent a clear and simple structure that easily allows the investigator to determine which food items in fact are contributors to a pattern and which are not. As a potential solution, the structure obtained with principal component analysis or exploratory factor analysis can be tested with confirmatory factor analysis – a step that has rarely been taken in dietary pattern analyses (Schulze *et al.* 2003). The goodness of fit is determined on the basis of the significance of factor loadings and goodness-of-fit test statistics. As this method allows us to objectively determine the foods for each pattern, the pattern score can then be calculated easily as the combination of these standardised food groups. Omitting weights in this step again appears to lead to useful and easily interpretable pattern scores (Glass *et al.* 1997).

In a US adult population, two distinct patterns, 'prudent pattern' and 'Western pattern,' have been identified (Slattery *et al.* 1998). While the prudent pattern is defined by frequent consumption of a variety of fruits, vegetables, whole grains, legumes, fish, and poultry, the Western pattern mainly

https://doi.org/10.1079/NRR2005107

comprises red and processed meats, high-fat dairy products including butter, eggs, and refined carbohydrates such as sweets, desserts, and refined grains. Other patterns have been described in European populations (Schulze *et al.* 2001; Costacou *et al.* 2003; Newby & Tucker, 2004).

Strengths. Principal component analysis and factor analysis have been validated, and results appear to be reproducible over time and across different dietary assessment methods (Hu *et al.* 1999).

Limitations. By definition, components or factors obtained with principal component analysis or factor analysis are statistically independent of each other. This is not always entirely intuitive, as, for example, we would expect that the US-based prudent and Western patterns would be inversely correlated. Furthermore, there are no tests to aid as decision criteria for the formulation of components or factors, only empirical guidelines. Hence, while data driven, principal component analysis and factor analysis include subjective criteria in defining dietary patterns. Attempts to statistically test the pattern structure with confirmatory factor analysis are useful, but this approach has rarely been employed.

Correlated measurement error in assessing foods may lead to an overestimation of correlation between foods and may distort definition of a pattern. Errors in the assessment of foods, especially foods that are grouped together on the food-frequency questionnaire (for example, vegetables) have been found to be correlated (Michels *et al.* 2005). Thus, if consumption of one vegetable is overreported, consumption of other vegetables is probably overreported as well. This error correlation increases the apparent correlation among the foods and increases the probability that they become constituents of the same pattern.

The components obtained with principal component analysis commonly account for only a modest proportion of the total variance in diet in the dataset. Thus, the results represent the optimal model with respect to the explained variance, but leave sufficient room for other patterns to prevail in the study population. Which dietary pattern is most predictive for any specific disease cannot easily be answered with principal component analysis alone. Other patterns may be as important, but were not identified with this approach because they explained a slightly smaller amount of total variance (Schulze *et al.* 2004).

Martinez *et al.* (1998) have voiced concern regarding the use of principal component analysis and factor analysis in nutritional epidemiology, pointing out their subjectiveness in preselecting food groups, determining the number of factors, deciding on when components or factors are relevant and when factor loadings are important to maintain foods in the pattern, all of which make them less data-driven methods than assumed.

Another issue with factor analysis (which also applies to other data-driven methods) is that dietary patterns strongly interact with other lifestyle characteristics or rather are part of specific lifestyles (Martinez *et al.* 1998). While this might strengthen the opinion that the observed patterns may be meaningful, it might consequently be impossible to separate pattern effects from the effects of other lifestyle characteristics. Slattery *et al.* (1999) have used dietary and lifestyle characteristics to define lifestyle habits using factor analysis, further supporting the above argument.

Cluster analysis

Concepts and methods. Cluster analysis (Everitt *et al.* 2001) is another data-driven procedure that aims to build clusters of individuals with similar diets (rather than clusters of foods consumed together, as factor analysis does). First, as in principal component analyses, foods are pregrouped. Foods need to be standardised since variables with large variances tend to have a greater effect on the resulting clusters than those with small variances. Thus, foods and food groups are commonly divided by total energy intake and the percentage of energy contributed by each food group is calculated and used in the cluster analysis. Cluster analysis is based on distance measures between individuals. Initial cluster seeds are followed by repeated comparisons between the means of initial clusters and subsequent updates of cluster groupings and means. Subjects are moved between mutually exclusive clusters and new means are computed until the distances between the observations within clusters are smaller than the distances between cluster means. Ideally, cluster analysis leads to complete convergence so that the final cluster seeds will equal the cluster means or cluster centres. Alternatively, investigators may specify the maximum number of iterations or a specific convergence criterion by which their procedures terminate when a complete iteration fails to move a cluster centre by more than a percentage of the smallest distance between any of the centres.

The number of clusters has to be prespecified by the investigator. The set of clusters is selected where the nearest higher number of clusters would not give a considerably better separation. A smaller number of clusters is usually preferable as they are easier to interpret. Although the real number of clusters in the data is unknown, a cross-validation method can be applied to a range of numbers of clusters to observe the resulting average distance of the observations in a cross-validation or testing sample from their cluster centres. A higher between-cluster variance:within-cluster variance ratio indicates a better separation of clusters. A scree plot of the overall ratios against the number of selected clusters is used as a decision criterion: stop when the plot reaches a plateau. After applying the resulting number of clusters to the study population, the dietary differences between clusters are descriptively evaluated and dietary patterns are characterised by their average energy contribution from each individual food group.

Studies on dietary patterns which applied cluster analysis were comprehensively reviewed by Newby & Tucker (2004). The clusters described have varied among research groups and populations. Akin and colleagues identified patterns of light eaters, heavy eaters, or consumers of large amounts of alcoholic beverages, salty snack products, animal fat products, legumes, or sweets and desserts in a population of older Americans (Akin *et al.* 1986). In another population of elderly Americans, four patterns characterised by (a) alcohol, (b) milk, cereal, and fruits, (c) bread and poultry, and (d) meat and potatoes were found (Tucker *et al.* 1992). In the UK Women's Cohort Study seven patterns emerged: (a)

monotonous low-quantity omnivores, (b) health conscious, (c) traditional meat, chips and pudding eaters, (d) higher diversity, traditional omnivores, (e) conservative omnivores, (f) low diversity vegetarians, (g) high diversity vegetarians (Greenwood *et al.* 2000). Wirfält *et al.* reported clusters of (a) many foods and drinks, (b) fibre bread, (c) low fat and high fibre, (d) white bread, (e) milk fat, and (f) sweets and cakes, and of (a) drinks and fries, (b) ice-cream and cake, (c) dieters, (d) healthy, (e) traditional, and (f) Mediterranean diet from the Swedish Malmö Diet and Cancer Cohort (Wirfält *et al.* 2000, 2001). In the Framingham Heart Study, clusters of (a) sugar, (b) fish and grain, (c) meat, eggs, and fat, (d) milk and fruit, and (e) alcohol were found (Haveman-Nies *et al.* 2001); in the Framingham Nutrition Studies, patterns of (a) heart healthy, (b) light eating, (c) wine and moderate eating, (d) high fat, and (e) 'empty calories' were identified (Millen *et al.* 2001).

Strengths. Cluster analysis groups individuals; thus a specific dietary pattern is assigned to each individual of the corresponding study population. In contrast, factor analysis produces a variety of pattern scores for each individual.

Limitations. Within and across populations and diet questionnaires, different cluster patterns have been described and a dietary pattern identified with cluster analysis has not been consistently reported across different studies. Newby and colleagues, however, compared factor and cluster analysis methods and found comparability in the dietary patterns derived with the two methods (Newby *et al.* 2004*b*). In randomly generated split samples, they found clusters to be more accurately reproducible than factors. While factor analysis results in a pattern that is directly interpretable, further analysis is necessary to define the particular dietary profiles of the extracted clusters.

Summary

Factor analysis and cluster analysis, while both driven by the observed data, are still partially influenced by investigator decisions. Grouping of foods before the search for factors and clusters, determining the critical value of the Eigenvalue, visual examination of a scree plot, determining the cut-off for foods to be important contributors to a factor based on the factor-loading matrix, and prespecifying the number of clusters all require decisions to be taken by the investigator and may vary by investigator. More objective attempts, for example, the use of confirmatory factor analysis to determine the important contributors to a factor-based pattern, are rarely applied.

For both factor analysis and cluster analysis, examining the relevance of the emerging factors or clusters is part of defining a dietary pattern. In addition, sensitivity analyses of the identified patterns can be performed by repeating the analyses on split halves of the sample.

Both methods build patterns solely based on dietary information, which is valuable to characterise the most common culturally determined dietary patterns but may be suboptimal if the goal is to define patterns most relevant to predicting specific diseases.

Factor and cluster analyses provide population-specific results. Although similarities in dietary patterns may be found across populations, there are generally important differences in the food composition of these patterns, they only explain a limited proportion of total variation in diet, and additional population-specific patterns exist (Balder *et al.* 2003). Associations between factor and cluster analysis-based patterns and disease risk may not be reproducible across populations. Thus, both methods are limited by a lack of generalisability of results.

Investigator-determined methods

Investigator-determined methods define dietary patterns *a priori* and compare the performance of individuals to these prespecified standards.

Dietary indices

Concepts and methods. Dietary indices are summary measures of the degree to which an individual's diet conforms to specific dietary recommendations. For example, the diet quality index is a summary score of the degree to which an individual conforms to dietary recommendations from the 1989 National Academy of Sciences publication *Diet and Health* (Patterson *et al.* 1994). The index ranking of overall dietary patterns based on this score was found to be reflective of total diet quality. The healthy eating index (Kennedy *et al.* 1995) is a summary measure of the degree to which an individual's diet conforms to the recommendations set out by the US Department of Agriculture's food guide pyramid and to specific recommendations in the 1990 US Dietary Guidelines for Americans. The healthy eating index includes ten scoring criteria which add to a summary score for each individual: (a) grains, (b) vegetables, (c) fruit, (d) milk, (e) meat, (f) total fat, (g) saturated fat, (h) cholesterol, (i) Na, (j) variety (number of different food items over a 3 d period); high consumption of the first five foods items and a higher variety and low intake of the nutrients resulted in a high score. McCullough and colleagues found the healthy eating index to be only weakly inversely associated with the risk of major chronic diseases and suggested that adherence to it would have limited benefit (McCullough *et al.* 2000*a*,*b*, 2002). These authors suggested an improved index, the alternative healthy eating index, which was a better predictor of chronic disease, especially CVD (McCullough *et al.* 2002). The alternative healthy eating index incorporates aspects of the original healthy eating index (i.e. (a) vegetables and (b) fruit), but adds (c) nuts and soya protein, (d) white meat:red meat ratio, (e) cereal fibre, (f) *trans* fatty acids, (g) polyunsaturated:saturated fats ratio, (h) multivitamin use, and (i) alcohol intake.

Strengths. Dietary indices use general dietary guidelines as guiding principles, which makes them objective. Dietary indices are easy to understand for the general public.

Limitations. Availability of dietary guidelines is required to define dietary indices. Dietary indices are only as good as the underlying dietary guidelines. Dietary guidelines are

generally not disease specific; hence adherence to them may reduce the risk of some diseases but not others.

Diet scores

Concepts and methods. Diet scores generally count the number or frequency of foods consumed that are considered by the investigator to promote health (or disease). One of the earliest diet scores was introduced by Manousos and colleagues, who quantified the frequency of consumption of different food items as the number of times per month the food was consumed and assigned a value between zero and 30 for each food item on the diet questionnaire (Manousos *et al.* 1983). The dietary diversity score simply counts the number of foods or food groups consumed regularly (Randall *et al.* 1989; Miller *et al.* 1992; Kant *et al.* 1993, 1995; McCann *et al.* 1994). Kant introduced the concept of a recommended food score, which simply adds up the number of 'recommended' foods consumed regularly (Kant *et al.* 2000). Michels & Wolk (2002) complemented the recommended food score with a 'not recommended food score', which adds up the number of 'not recommended' foods consumed regularly. In a Swedish population the recommended food score was found to be more predictive of longevity than the 'not recommended food score' was of mortality (Michels & Wolk, 2002). A composite dietary score based on nutrient intakes of cereal fibre, folate, marine *n*-3 fatty acids, the polyunsaturated:saturated fats ratio, *trans* fatty acids, and glycaemic load has been found to be a good predictor of CHD risk, superior to the individual nutrients and to patterns derived with factor analysis (Stampfer *et al.* 2000). Similarly, a composite score based on cereal fibre, PUFA, *trans* fatty acids, and glycaemic load strongly predicted risk of type 2 diabetes (Hu *et al.* 2001). Adherence to a Mediterranean diet has been found to be associated with a reduction in total mortality in a Greek population (Trichopoulou *et al.* 2003). The Mediterranean diet score assigned a value of 1 for each of the following foods considered beneficial consumed above the population median and zero below the median: vegetables, legumes, fruit and nuts, cereal, fish. For components presumed to be detrimental (meat, dairy products), individuals whose consumption was below the median were assigned a value of 1, above the median a value of zero (Trichopoulou *et al.* 2003).

Strengths. Dietary scores are intuitive, analytically simple, and easily translatable into public health messages. Dietary scores can be targeted to specific diseases if built on prior knowledge of dietary predictors of that disease.

Limitations. Determining which foods will be classified as 'recommended' and 'not recommended' is up to the investigator's perception of which foods promote health and prevent disease. Hence, dietary scores require prior knowledge and will perform only to the extent that this knowledge accurately reflects diet–disease associations. The frequency of consumption deemed important is also decided by the investigator (for example, ever, at least once per month, at least once per week). Hence, dietary scores are quite subjective and their definition may vary substantially among investigators.

Summary

In general, the indices and scores of overall diet quality have been found to relate to the risk of disease outcomes more consistently than individual nutrients or foods (Kant *et al.* 1993; Stampfer *et al.* 2000). Making use of prior knowledge or suspected diet–disease associations can be helpful in building more clearly defined and sensible dietary patterns and in deriving disease-specific dietary patterns.

Methods combining data-driven procedures with prior knowledge

Reduced rank regression

Concepts and methods. RRR does not focus on explaining variance between foods like principal component analysis but identifies linear functions of predictors (for example, food groups) that explain as much variation as possible in a set of intermediate response variables (for example, biomarker) (Hoffmann *et al.* 2004*a*). Since it uses both available data and prior knowledge about the response variables, it is an *a posteriori* method. RRR is similar to factor analysis in its mathematical foundation and technique of deriving factors. For both methods, the coefficient vectors of the extracted linear functions are eigenvectors of a covariance matrix. Factor analysis uses the covariance matrix of predictors, whereas RRR starts from the covariance matrix of responses. RRR can be interpreted as a principal component analysis applied to responses and a subsequent linear regression of principal components on predictors although it is somewhat more sophisticated and efficient than this two-step procedure.

In contrast to the large number of studies that have used factor or cluster analysis to derive dietary patterns, very few studies have evaluated associations between RRR-derived patterns and disease risk thus far. Hoffmann *et al.* (2004*b*) identified a dietary pattern characterised by high intakes of meat, margarine, poultry, and gravy and low intakes of vegetarian dishes, wine, vegetables, and wholegrain cereals that best explained variance among a set of risk markers for CVD (HDL-cholesterol, LDL-cholesterol, lipoprotein (a), C-peptide, and C-reactive protein). The pattern was a strong predictor of coronary artery disease in this case–control study (Hoffmann *et al.* 2004*b*). Heidemann *et al.* (2005) identified a dietary pattern protective for type 2 diabetes in the prospective EPIC-Potsdam Study. The pattern was characterised by high intake of fresh fruits and low intake of high-energy soft drinks, beer, meat, poultry, processed meat, legumes, and refined-grain bread and was negatively associated with HbA1c and C-reactive protein and positively associated with HDL-cholesterol and adiponectin levels. Other recent prospective studies link RRR patterns to weight change (Schulze *et al.* 2005), type 2 diabetes (Schulze *et al.* 2005), CHD (Weikert *et al.* 2005) and mortality (Hoffmann *et al.* 2005).

https://doi.org/10.1079/NRR2005107

Strengths. Associations between patterns defined by RRR and disease endpoints appear to be stronger than with patterns defined by factor analysis. The advantage of this approach is that the derived dietary pattern may incorporate information on biological pathways, and thus it is not purely data driven but involves prior knowledge of potential diet–disease associations mediated by biological parameters. If RRR is successfully applied, the effect of a dietary pattern on disease risk can be interpreted, explained, and described by changes of biologically important intermediate variables.

Limitations. The RRR approach requires the availability of response (biomarker) information. This information may not be available in many studies otherwise suitable to evaluate diet–disease associations. In the extreme case of missing knowledge concerning the development of the disease, no response variables can be justified and the RRR approach cannot be used. For many chronic diseases a complex interplay of metabolic pathways may link dietary intake to disease. So far, it is unclear whether RRR is more efficient using biomarkers for only one pathway than using all potential pathways; it is also not clear how to select the best set of responses. The selection of response biomarkers always depends on the current state of knowledge.

What can we learn from dietary patterns?

Whether the use of dietary pattern analyses may shed new light on the role of diet in the prevention and causation of cancer, CVD, diabetes, and other diseases remains to be explored. Thus far, the dietary patterns identified and their relationship with disease outcomes have confirmed current knowledge. Pattern analysis may be most useful for exploring diseases whose causative relationship to dietary exposures is not known. Moreover, if analytical models including individual foods or nutrients have not revealed important associations, dietary patterns may still emerge as predictors of disease. Dietary pattern analysis may also be useful if several dietary exposures are associated with disease risk; a pattern may capture the overall effect of diet, accounting for interactions and synergistic effects. Conversely, if only a single dietary exposure truly affects disease development, this effect is most probably diluted in dietary pattern analysis. Furthermore, a link between a dietary pattern and a disease does not allow mechanistic insights into disease causation unless it is followed by analyses of individual foods and nutrients. Some of the methodological problems inherent in the assessment and analysis of dietary data in observational research apply to dietary patterns just as much as to the analysis of individual foods or nutrients. Measurement error affects dietary pattern definitions as it does the assessment of food or nutrient intake. Correlated errors may lead to distortions of the definition of dietary pattern and its associations with disease outcome. While higher correlations for disease outcomes have been found with dietary patterns than with individual foods, this could be an artifact of correlated measurement errors, which may deattenuate the correlation with disease.

The question remains whether the analytically more complex, largely data-driven methods provide any advantage over the simpler investigator-determined methods. Since dietary patterns identified with factor analysis do not explain a large proportion of variability between individuals with respect to their diet, there still remain important dietary habits that account for a considerable proportion of between-individual variation. As factor and cluster analysis do not consider information relevant to the disease endpoint, the patterns they produce are probably not optimal to explain diet–disease associations. Investigator-determined indices and scores appear to be at least as useful if they are based on sound evidence. Newer strategies to define dietary patterns that incorporate both *a priori* knowledge and the data at hand may advance this discussion. Here, RRR allows us to take the step from data-driven procedures that ignore the endpoint of interest to a more targeted approach that considers the metabolic pathways of the diet–disease association. First results using this approach are promising and future research seems warranted. However, this method requires information on intermediate markers.

What we have concluded from dietary pattern analyses thus far is that a healthy diet reduces weight gain (Fogelholm *et al.* 2000; Newby *et al.* 2003, 2004*a*; Togo *et al.* 2004; Schulz *et al.* 2005) and the risk of premature mortality (Kant *et al.* 2000; Michels & Wolk, 2002), CVD (Huijbregts *et al.* 1995; Hu *et al.* 2000; Trichopoulou *et al.* 2003; Fung *et al.* 2004), diabetes (Van Dam *et al.* 2002; Fung *et al.* 2004; Heidemann *et al.* 2005; Montonen *et al.* 2005), and hypertension (Schulze *et al.* 2003) – not surprising or ground breaking insights. Still, these results support the notion that pattern analysis is a useful approach, complementing the more commonly used analysis of single nutrients and foods. If developed further, dietary pattern analysis might provide valuable new insights into the role of diet in disease prevention, which may be particularly useful for defining food-based dietary guidelines.

Acknowledgements

K. B. M. was supported in part, by research grant R01 DK 54 900 from the National Institute of Diabetes and Digestive and Kidney Diseases and by Senior International Fogarty Fellowship F06 TW05568 from the National Institutes of Health US Department of Health and Human Services.

References

Akin JS, Guilkey DK, Popkin BM & Fanelli MT (1986) Cluster analysis of food consumption patterns of older Americans. *Journal of the American Dietetic Association* **86**, 616–624.

Balder HF, Virtanen M, Brants HA, Krogh V, Dixon LB, Tan F, Mannisto S, Bellocco R, Pietinen P, Wolk A, Berrino F, van den Brandt PA, Hartman AM & Goldbohm RA (2003) Common and country-specific dietary patterns in four European cohort studies. *Journal of Nutrition* **133**, 4246–4251.

Costacou T, Bamia C, Ferrari P, Riboli E, Trichopoulos D & Trichopoulou A (2003) Tracing the Mediterranean diet through principal components and cluster analyses in the Greek population. *European Journal of Clinical Nutrition* **57**, 1378–1385.

Everitt BS, Landau S & Leese M (2001) *Cluster Analysis*, 4th ed. New York: Oxford University Press.

Fogelholm M, Kujala U, Kaprio J & Sarna S (2000) Predictors of weight change in middle-aged and old men. *Obesity Research* **8**, 367–373.

Fung TT, Schulze M, Manson JE, Willett WC & Hu FB (2004) Dietary patterns, meat intake, and the risk of type 2 diabetes in women. *Archives of Internal Medicine* **164**, 2235–2240.

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.

Greenwood DC, Cade JE, Draper A, Barrett JH, Calvert C & Greenhalgh A (2000) Seven unique food consumption patterns identified among women in the UK Women's Cohort Study. *European Journal of Clinical Nutrition* **54**, 314–320.

Haveman-Nies A, Tucker KL, de Groot LC, Wilson PW & van Staveren WA (2001) Evaluation of dietary quality in relationship to nutritional and lifestyle factors in elderly individuals of the US Framingham Heart Study and the European SENECA study. *European Journal of Clinical Nutrition* **55**, 870–880.

Heidemann C, Hoffmann K, Spranger J, Klipstein-Grobusch K, Mohlig M, Pfeiffer AF & Boeing H (2005) A dietary pattern protective against type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study cohort. *Diabetologia* **48**, 1126–1134.

Hoffmann K, Boeing H, Boffetta P, Nagel G, Orfanos P, Ferrari P & Bamia C (2005) Comparison of two statistical approaches to predict all-cause mortality by dietary patterns in German elderly subjects. *British Journal of Nutrition* **93**, 709–716.

Hoffmann K, Schulze MB, Schienkiewitz A, Nothlings U & Boeing H (2004*a*) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *American Journal of Epidemiology* **159**, 935–944.

Hoffmann K, Zyriax BC, Boeing H & Windler E (2004*b*) A dietary pattern derived to explain biomarker variation is strongly associated with the risk of coronary artery disease. *American Journal of Clinical Nutrition* **80**, 633–640.

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.

Huijbregts PP, Feskens EJ & Kromhout D (1995) Dietary patterns and cardiovascular risk factors in elderly men: the Zutphen Elderly Study. *International Journal of Epidemiology* **24**, 313–320.

Jacobson HN (1986) Pattern analysis in nutrition. *Clinical Nutrition* **5**, 249–253.

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.

Kant AK, Schatzkin A, Harris TB, Ziegler RG & Block G (1993) Dietary diversity and subsequent mortality in the First National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *American Journal of Clinical Nutrition* **57**, 434–440.

Kant AK, Schatzkin A & Ziegler RG (1995) Dietary diversity and subsequent cause-specific mortality in the NHANES I epidemiologic follow-up study. *Journal of the American College of Nutrition* **14**, 233–238.

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.

McCann SE, Randall E, Marshall JR, Graham S, Zielezny M & Freudenheim JL (1994) Diet diversity and risk of colon cancer in western New York. *Nutrition and Cancer* **21**, 133–141.

McCullough ML, Feskanich D, Rimm EB, Giovannucci EL, Ascherio A, Variyam JN, Spiegelman D, Stampfer MJ & Willett WC (2000*a*) Adherence to the dietary guidelines for Americans and risk of major chronic disease in men. *American Journal of Clinical Nutrition* **72**, 1223–1231.

McCullough ML, Feskanich D, Stampfer MJ, Giovannucci EL, Rimm EB, Hu FB, Spiegelman D, Hunter DJ, Colditz GA & Willett WC (2002) Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. *American Journal of Clinical Nutrition* **76**, 1261–1271.

McCullough ML, Feskanich D, Stampfer MJ, Rosner BA, Hu FB, Hunter DJ, Variyam JN, Colditz GA & Willett WC (2000*b*) *Adherence to the dietary guidelines for Americans and risk of major chronic disease in women.* American Journal of Clinical Nutrition ***72**, 1214–1222.*

Manly FJM (2004) *Multivariate Statistical Methods: a Primer*, 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.

Manousos O, Day NE, Trichopoulos D, Gerovassilis F, Tzonou A & Polychronopoulou A (1983) Diet and colorectal cancer: a case–control study in Greece. *International Journal of Cancer* **32**, 1–5.

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

Michels KB, Welch AA, Luben R, Bingham SA & Day NE (2005) Measurement of fruit and vegetable consumption with diet questionnaires and implications for analyses and interpretation. *American Journal of Epidemiology* **161**, 987–994.

Michels KB & Wolk A (2002) A prospective study of variety of healthy foods and mortality in women. *International Journal of Epidemiology* **31**, 847–854.

Millen BE, Quatromoni PA, Copenhafer DL, Demissie S, O'Horo CE & D'Agostino RB (2001) Validation of a dietary pattern approach for evaluating nutritional risk: the Framingham Nutrition Studies. *Journal of the American Dietetic Association* **101**, 187–194.

Miller WL, Crabtree BF & Evans DK (1992) Exploratory study of the relationship between hypertension and diet diversity among Saba Islanders. *Public Health Reports* **107**, 426–432.

Montonen J, Knekt P, Harkanen T, Jarvinen R, Heliovaara M, Aromaa A & Reunanen A (2005) Dietary patterns and the incidence of type 2 diabetes. *American Journal of Epidemiology* **161**, 219–227.

Newby PK, Muller D, Hallfrisch J, Andres R & Tucker KL (2004*a*) Food patterns measured by factor analysis and anthropometric changes in adults. *American Journal of Clinical Nutrition* **80**, 504–513.

Newby PK, Muller D, Hallfrisch J, Qiao N, Andres R & Tucker KL (2003) Dietary patterns and changes in body mass index and waist circumference in adults. *American Journal of Clinical Nutrition* **77**, 1417–1425.

Newby PK, Muller D & Tucker KL (2004*b*) *Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods.* American Journal of Clinical Nutrition ***80**, 759–767.*

Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutrition Reviews* **62**, 177–203.

Patterson RE, Haines PS & Popkin BM (1994) Diet quality index: capturing a multidimensional behavior. *Journal of the American Dietetic Association* **94**, 57–64.

Randall E, Marshall J, Graham S & Brasure J (1989) Frequency of food use data and the multidimensionality of diet. *Journal of the American Dietetic Association* **89**, 1070–1075.

Schulze M, Noethlings U, Hoffmann K, Bergmann MM & Boeing H (2005) Identification of a food pattern characterized by high-fiber and low-fat food choices associated with low prospective weight change in the EPIC-Potsdam cohort. *Journal of Nutrition* **135**, 1183–1189.

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.

Schulze MB, Hoffmann K, Kroke A & Boeing H (2003) Risk of hypertension among women in the EPIC-Potsdam Study: comparison of relative risk estimates for exploratory and hypothesis-oriented dietary patterns. *American Journal of Epidemiology* **158**, 365–373.

Schulze MB, Hoffmann K & Boeing H (2004) Risk of hypertension among women in the EPIC-Potsdam Study: comparison of relative risk estimates for exploratory and hypothesis oriented dietary patterns. *American Journal of Epidemiology* **159**, 913–914.

Schulze MB, Hoffman K, Manson JE, Willett WC, Meigs JB, Weikert C, Heidemann C, Colditz GA & Hu FB (2005) Dietary pattern, inflammation, and incidence of type 2 diabetes in women. *American Journal of Clinical Nutrition* **82**, 675–684.

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 HANES 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.

Slattery ML, Edwards SL, Boucher KM, Anderson K & Caan BJ (1999) Lifestyle and colon cancer: an assessment of factors associated with risk. *American Journal of Epidemiology* **150**, 869–877.

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.

Togo P, Osler M, Sorensen TI & Heitmann BL (2004) A longitudinal study of food intake patterns and obesity in adult Danish men and women. *International Journal of Obesity and Related Metabolic Disorders* **28**, 583–593.

Trichopoulou A, Costacou T, Bamia C & Trichopoulos D (2003) Adherence to a Mediterranean diet and survival in a Greek population. *New England Journal of Medicine* **348**, 2599–2608.

Trichopoulos D & Lagiou P (2001) Dietary patterns and mortality. *British Journal of Nutrition* **85**, 133–134.

Tucker KL, Dallal GE & Rush D (1992) Dietary patterns of elderly Boston-area residents defined by cluster analysis. *Journal of the American Dietetic Association* **92**, 1487–1491.

Van Dam RM, Rimm EB, Willett WC, Stampfer MJ & Hu FB (2002) Dietary patterns and risk for type 2 diabetes mellitus in US men. *Annals of Internal Medicine* **136**, 201–209.

Weikert C, Hoffmann K, Dierkes J, Zyriax BC, Klipstein-Grobusch K, Schulze MB, Jung R, Windler E & Boeing H (2005) A homocysteine metabolism-related dietary pattern and the risk of coronary heart disease in two independent German study populations. *Journal of Nutrition* **135**, 1981–1988.

Wirfält E, Hedblad B, Gullberg B, Mattisson I, Andren C, Rosander U, Janzon L & Berglund G (2001) Food patterns and components of the metabolic syndrome in men and women: a cross-sectional study within the Malmo Diet and Cancer cohort. *American Journal of Epidemiology* **154**, 1150–1159.

Wirfält E, Mattisson I, Gullberg B & Berglund G (2000) Food patterns defined by cluster analysis and their utility as dietary exposure variables: a report from the Malmo Diet and Cancer Study. *Public Health Nutrition* **3**, 159–173.