*Proceedings of the Nutrition Society* (2013), **72**, 191–199 doi:10.1017/S0029665113000013
© The Author 2013 First published online 30 January 2013

*The Summer meeting of the Nutrition Society hosted by the Irish Section was held at Queen's University, Belfast on 16–19 July 2012*

## Conference on 'Translating nutrition: integrating research, practice and policy'
## Symposium 1: Innovation in diet and lifestyle assessment

# Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis

Marga C. Ocké
*National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands*

This paper aims to describe different approaches for studying the overall diet with advantages and limitations. Studies of the overall diet have emerged because the relationship between dietary intake and health is very complex with all kinds of interactions. These cannot be captured well by studying single dietary components. Three main approaches to study the overall diet can be distinguished. The first method is researcher-defined scores or indices of diet quality. These are usually based on guidelines for a healthy diet or on diets known to be healthy. The second approach, using principal component or cluster analysis, is driven by the underlying dietary data. In principal component analysis, scales are derived based on the underlying relationships between food groups, whereas in cluster analysis, subgroups of the population are created with people that cluster together based on their dietary intake. A third approach includes methods that are driven by a combination of biological pathways and the underlying dietary data. Reduced rank regression defines linear combinations of food intakes that maximally explain nutrient intakes or intermediate markers of disease. Decision tree analysis identifies subgroups of a population whose members share dietary characteristics that influence (intermediate markers of) disease. It is concluded that all approaches have advantages and limitations and essentially answer different questions. The third approach is still more in an exploration phase, but seems to have great potential with complementary value. More insight into the utility of conducting studies on the overall diet can be gained if more attention is given to methodological issues.

**Overall diet: Dietary quality scores: Dietary pattern analysis**

Studies of the overall diet have emerged as an important research field complementary to 'reductionist' single component studies[1–4]. The rationale for this is threefold. First, dietary exposure consists of a multitude of different nutrients and other bio-active constituents. Some components act synergistically, where other components work in opposition. The complex interactions and cumulative effects cannot be captured well by studying the effects of single dietary components[1,3]. Second, people do not eat nutrients or constituents. They consume foods; and food consumption often occurs in patterns of meals and in-between meal consumption. Consumption patterns are shaped by income, prices, individual preferences and beliefs, cultural traditions, as well as geographical, environmental, social and economic factors[5]. Third, dietary change is usually not restricted to one dietary component because of substitution and compensatory effects of other dietary characteristics[2].

Various approaches to study the overall diet can be distinguished[6]. In hypothesis-driven or *a priori* approaches, researchers define scores or indices of the overall dietary quality. The scores are usually based on guidelines for a healthy diet or on diets known to be healthy[7]. In contrast, *a posteriori* approaches are driven by the underlying dietary data. Statistical methods such as principal component analysis, exploratory factor analysis and cluster analysis are applied to derive dietary patterns that are available in the data[3]. In principal component and exploratory factor

**Corresponding author:** Dr Marga Ocké, fax +31 30 274 4466, email marga.ocke@rivm.nl

**Table 1.** Key characteristics of dietary quality scores and commonly used implementations

| Characteristic | Implementation choices |
|---|---|
| Overall concept | Based on pre-existing general dietary guidelines, guidelines for the prevention of a specific disease, or on a specific dietary pattern that is known to be healthy<br>Reflects only healthy aspects, only unhealthy aspects, or both |
| Included components | Nutrients (including ratios), foods, food groups (including ratios), other aspects (like dietary variety) |
| Cut-off values | Median level (or other percentile), fixed cut-off |
| Scoring in relation to cut-off values | Dichotomous, proportional to the extent the guideline is met |
| Energy intake | Included yes/no; other components adjusted for energy intake (yes/no) |
| Age–sex specificity | No*/yes, e.g. for a specific subpopulation, the use of different cut-offs |
| Weighting of various components to the total score | Equal or different weights for each component |
| Overall scoring range | |

*Indirectly through energy adjustment or standardisation.

analyses, new dietary pattern variables are obtained based on the underlying interrelationships between the dietary components. Whereas in cluster analysis, subgroups of the study population are created that combine people with similar dietary intakes within a cluster and put people with different dietary intakes in different clusters[1]. Finally, hybrid approaches were introduced to study the overall diet. Hybrid approaches are driven by a combination of biological pathways and the underlying dietary data. For example, reduced rank regression can be applied to define linear combinations of food intakes that maximally explain predictors of disease[8]. Methods such as decision tree analysis can be used to identify subgroups of a population whose members share dietary characteristics that influence a disease or disease predictor[9].

This paper aims to present an overview of approaches for assessing the overall diet. Attention is particularly given to newer methodologies. The advantages and limitations of each approach are presented, as well as ways to evaluate the obtained overall diets and their main uses. The description is given from a public health perspective rather than from focus on individuals or on clinical settings.

## Dietary quality scores

### *Principles*

Scores or indices of dietary quality express the overall healthiness of the diet. They are usually developed based on a specific dietary pattern that is known to be healthy or based on pre-existing dietary guidelines for the general population or for the prevention of a specific dietary-related disease[7,10]. Table 1 shows key aspects to be considered in the development and interpretation of dietary quality scores, as well as commonly chosen implementations for those aspects. A more extensive description of the make-up of dietary quality scores is given in Waijers *et al.*[7].

In the application and interpretation of a dietary quality score, the chosen make-up should be considered. For example, population-specific median values as cut-off levels for scoring the individual components of a score will limit the usefulness of comparing the quality scores across populations. However, it will make sure that all components of the score contribute to the overall score. In contrast, in case of a score with fixed cut-off levels and a binary scoring system, it might occur that all persons score the same for a specific component. This component does not then contribute to the overall score. However, a score with this system is suitable to compare the dietary quality of various populations[7].

### *Well-known examples*

Recent reviews, in 2007[7] and 2009[11], identified twenty and twenty-five different scores of overall dietary quality, respectively. Dietary quality scores have been widely used in adult populations, whereas in children the use is limited[12]. Two well-known examples of dietary quality scores are the Healthy Eating Index[13,14] and the Mediterranean Diet Score[15].

The Healthy Eating Index is a measure of dietary quality according to the United States Department of Agriculture Food Guide Pyramid. It was developed in 1995[14], revised according to the revision of the guidelines in 2005[13], and is again being updated at the moment[16]. Index components in the 2005 Healthy Eating Index were: total fruit, whole fruit, total vegetables, dark green and orange vegetables and legumes, total grains, whole grains (each five points), milk, meat and beans, oils, saturated fat, Na (each ten points), and energy from solid fats, alcoholic beverages and added sugars (twenty points). The included food groups were expressed in servings per 4184 kJ (1000 kcal); the included nutrients and energy from solid fats, alcoholic beverages and added sugars are also expressed using an energy density approach. The scoring is proportional to the extent to which the dietary guideline is met. The overall score can range from zero (poor diet) to 100 (excellent diet)[13]. Adapted versions were developed for other countries (e.g. Canada[17]) and for specific population groups (e.g. children[18]).

Advantages of the Mediterranean diet were described already in the 1950s by Keys[19] and by many others afterwards[20,21]. The first scoring system to express adherence to the Mediterranean diet, i.e. the Mediterranean Diet Score, was developed by Trichopoulou *et al.*[22]. The original Mediterranean Diet Score was composed of eight components: i.e. the MUFA:SFA ratio, consumption of

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

legumes, cereals, fruits and nuts, vegetables, meat and meat products, milk and dairy products, and alcohol. The components were adjusted for energy intake by standardising intakes for men to 10460 kJ (2500 kcal) and for women to 8368 kJ (2000 kcal). Cut-off values were sex-specific median intakes of the studied population, and scoring was either zero (worst) or one (best) for each component. With equal weights for each component, this led to a total score range of zero (poorest adherence to the Mediterranean diet) to eight (excellent adherence to the Mediterranean diet)(22). The Mediterranean Diet Score has been applied to Mediterranean and non-Mediterranean populations(23). Over time, various alternative Mediterranean Diet Scores were developed and tested(20,21). A recent meta-analysis quantified the protective effect of adherence to the Mediterranean diet for overall mortality and major chronic diseases. A two-point increase in the Mediterranean Diet Score was related to a 8% reduction in mortality(23).

### *Evaluation strategies*

A careful development process of a dietary quality score should include an in-depth evaluation. A majority of the dietary quality scores have been evaluated with regard to their nutrition adequacy only(24). Various other aspects are relevant to include in the evaluation. A good example can be found in the paper on the 2005 Healthy Eating Index(25). First, content validity was evaluated by assessing if the index captured all aspects of the dietary guidelines. Then construct validity was judged. This was done by checking if the index gave maximum scores to menus developed by nutrition experts to illustrate high diet quality; and by checking if the index distinguished between groups with known differences in dietary quality. The next step was the evaluation of reliability. In this step, the relationships among the index components was assessed, and it was checked which components had most influence on the overall index score(25). Moreover, it is important to assess the validity of the underlying dietary data of descriptive or epidemiological studies in which a dietary quality score is calculated(12).

For scores that intend to quantify the healthiness of the overall diet, the longitudinal relationship with overall health or total mortality is the ultimate evaluation. The majority of studies that assessed this relationship demonstrated that higher dietary quality was consistently inversely related to all-cause mortality, with a protective effect of moderate magnitude. The associations were stronger for men and for all-cause and CVD mortality(11).

### *Considerations, advantages and limitations*

The strength of scores of dietary quality is that they rely on the body of scientific evidence from studies on health and disease prevention. However, this is partly a theoretical strength. In practice, there is insufficient knowledge and consensus on what actually is the healthiest diet. This is clearly shown by the large number of existing scores that attempt to express overall dietary quality(7,26). Also, interpretations of the dietary guidelines are often needed to construct a dietary quality score, and subjectivity is introduced(6). This can, for example, be about the definition of the score components (e.g. what is Eat a varied diet?); but particularly the scoring and weighting of the various components is often not defined in the dietary guidelines(27). A second advantage of dietary quality scores is that they are usually easy to compute; and thereby easily reproducible and comparable(6).

A limitation of a dietary quality summary score is that it does not describe the overall diet pattern. This is partly due to the fact that many dietary quality scores focus on selected aspects of the diet and the correlated structure of the components is not considered(8,10). However, especially, persons who have a midrange score can have very different contributing components, and thus different dietary patterns(6).

An important consideration during the application of dietary quality scores is that they should be tailored to their aim. Obviously, if the diet quality for persons with a high risk of a given disease is aimed at, dietary guidelines for the prevention of this specific disease should be the reference. If the dietary quality of children is to be assessed, the guidelines should be applicable to children(12). It should also be considered that although dietary quality scores are often named hypothesis-driven methods, their application usually relies also on the underlying dietary data. In the case of a score with median values as cut-off values this is obvious. A second example is that many dietary guidelines include a recommendation to limit the intake of salt, while total salt intake cannot be captured well with self-reporting dietary assessment methods and is therefore often missing in food consumption databases. More generally, the type of dietary data used for the scoring of dietary quality should be appropriate for its purpose. In many studies, data from FFQ were used, since this dietary assessment method provides information about usual intake and dietary recommendations are intended to be met over time. Data derived by one or a few 24-h dietary recalls should not be used as such to calculate usual dietary quality, since they include too much day-to-day variation. Recently, a statistical model became available to overcome this challenge. The model was applied to the National Health And Nutritional Examination Survey 24-h dietary recall data and provided estimates of the population distribution of the Healthy Eating Index for the US(28). Overall, the general problems of dietary assessment through self-reporting of food consumption(29) are also reflected in the calculated dietary quality scores.

### *Applications*

Dietary quality scores can be useful tools to monitor the overall adherence to dietary guidelines, and the dietary quality of a population. Comparisons within and between populations can be made to formulate or evaluate the need for dietary interventions. In addition, dietary quality scores are useful tools to test if current dietary recommendations have a measurable protective effect against diseases, and to get insight into the magnitude of the overall effect(11).

An efficient application of dietary quality scores is the combination of a dietary quality score with a short dietary

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

**Table 2.** Key aspects to be decided during the process of factor/cluster analysis and commonly used implementations

| Aspects | Implementation choices |
|---|---|
| Overall concept | Creation of new variables (factor analysis), creation of mutually exclusive groups of individuals (cluster analysis; factor analysis followed by quantile grouping). |
| Included components | Based on FFQ or diet records, less often on dietary recalls or diet histories.<br>Usually food groups, but also individual food items, nutrients (including ratios), or combinations. The number of components.<br>Frequency, weight in g or servings. |
| Adjustment of input variables before analyses | In factor analysis sometimes adjusted for total energy intake; in cluster analysis (sensitive to outliers) usually *Z*-scores or energy adjusted data are used. Other transformations less commonly applied. |
| Statistical analysis type and decisions to be taken | Factor analysis: principal component analysis; type of rotation (often varimax rotation (orthogonal)), cut-offs for food group loadings to consider; criteria for the number of factors to retain (eigenvalues, scree plot and interpretability).<br>Cluster analysis: K-means or Ward's method. The number of clusters to retain or report usually based on interpretability (after trying several options), in combination with the cluster variance ratio or scree plot, and cluster sample size(38). |
| Labelling of the dietary pattern | Can be quantitatively (based on the highest factor loadings; or nutrient composition), or qualitatively (specific combinations of foods and nutrient composition). |
| Subgroup specificity | Sometimes sex specific; usually not age specific. Seldom stratified by race/ethnicity, culture or geography. |

assessment or screening method that only enquires about the relevant dietary components. See, for example, a web-based questionnaire to score the Dietary Approaches to Stop Hypertension Diet(30). This allows a quick and efficient assessment of overall dietary quality as an alternative to an extensive dietary assessment covering all components of the total diet. In the setting of developing countries, a quick screening method about the number of food groups consumed in a given period is often converted into a dietary variety score. The dietary variety score can subsequently be used as proxy for monitoring the overall dietary quality and household economic access to food(26,31).

## Empirically derived dietary patterns

### *Principles*

In contrast to dietary quality scores, empirically derived dietary patterns are driven by the dietary data from which they are derived. Two main approaches can be distinguished(1). In the first approach, the dietary variables are combined into fewer variables based on their inter-relationships. Common methods in this approach are principal component analysis and exploratory factor analysis. In principal component analysis, patterns or components are direct linear relationships of the underlying dietary variables. The created dietary pattern variables explain as much as possible the total variation of the original dietary variables. In exploratory factor analysis dietary patterns are modelled as underlying factors; only the variance that is shared with other variables is accounted for, excluding variance unique to each variable and random error variance(3). In dietary pattern analysis, principal component analysis is more commonly used than exploratory factor analysis. The obtained component scores are continuous variables(6).

In the second approach, i.e. cluster analysis, mutually exclusive non-overlapping clusters of individuals are created(32). Individuals within clusters share a similar dietary pattern, whereas individuals in other clusters have food patterns that are far apart. The K-means method is the most often applied method to obtain clusters of people with similar dietary patterns. It is an optimisation method to derive a specified number of clusters, by minimising an error criterion. Alternatively, Ward's minimum variance method is an agglomerative hierarchical clustering method. Although requiring a large computation time, it is also found in the dietary pattern literature(6).

Other empirical approaches have been applied to obtain dietary patterns. A recent example is the use of the treelet transform(33). This approach combines the quantitative pattern extraction capabilities of principal component analysis with the interpretational advantages of cluster analysis. The end result is a small number of naturally and hierarchical grouped variables. A disadvantage of the treelet transform method for dietary pattern analysis(33) is its assumption that only selected dietary components can contribute to the patterns, allowing no contributions of other dietary factors to the patterns(34).

To obtain empirically derived dietary patterns, the researcher has to make many decisions. The empirically derived dietary patterns are therefore not entirely data driven. Table 2 shows important aspects that should be considered during the preparation phase, statistical analysis and reporting phase, with often chosen implementations. As dietary input variables, most often food groups are used. An advantage of this is that together they can represent the total dietary intake, accounting for interactions between nutrients and other components within the groups(32). Many of the decisions are important for the interpretation of the dietary patterns. For example, expressing the input variables as contributions to energy intake has the disadvantage that the analysis is less sensible to detect variations in food group consumption that might be important for health but contribute little to energy intake. This is particularly the case for fruit and vegetable consumption(35). On the other hand, several studies found little differences in the derived dietary patterns with input variables that were or were not adjusted for energy intake before the dietary pattern analysis(36,37).

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

### Often observed patterns

Although dietary patterns will never be exactly the same across studies, it is apparent from the published studies that certain dietary patterns are frequently found. A large number of studies using principal component, exploratory factor or cluster analyses have identified variations of a healthy and a traditional or less-healthful dietary pattern. Also a pattern high in desserts or sweets and patterns high in alcohol appeared repeatedly[1]. In principal component and exploratory factor analyses, a healthy dietary pattern is often labelled 'prudent' and a less-healthful pattern 'Western'[1,38]. Obviously, patterns with the same label can be defined by different food components or by different weights of the components[39]. The Western pattern is usually characterised by high loadings of red meat, processed meat, butter, potatoes, refined grains and high-fat dairy. The prudent pattern, in contrast, has high loadings of vegetables, fruit, legumes, fish and seafood, and whole grains[40]. In general, the 'healthy', compared with the 'Western' pattern has been associated with more favourable biological profiles, slower progression of atherosclerosis and reduced incidence of CVD[1,40].

### Evaluation strategies

Evaluation strategies for empirically derived dietary patterns can focus on different aspects. These include the goodness of the solutions (using criteria such as explained variance in principal component analysis and exploratory factor analysis, or internal cluster validity indices), comparison of using dietary data obtained with different dietary assessment methods[41], comparison of using different types of input variables[35] or different strategies to derive the dietary patterns[42], and the reproducibility of derived dietary patterns. Reproducibility can be assessed internally using split sample techniques[42], or externally over time for the same population[41], and in different but similar study populations[36].

With split samples, for example, splitting the dataset into two equal parts, dietary patterns obtained in one-half of the sample (the derivation sample) can be confirmed in the second half (the validation sample). This can be done either by repeating the exploratory analyses in both samples or by using a confirmatory approach in the validation sample. Dietary patterns derived with principal component analysis or exploratory factor analysis can thus be validated using confirmatory factor analysis[43] and those derived by cluster analysis using discriminant analysis[44].

Some researchers indicate that empirically derived dietary patterns should be validated by assessing whether the dietary patterns can reliably predict diseases or mortality[6]. However, an empirically derived dietary pattern might be perfectly valid, i.e. existing in a study population, but without a relationship with health and disease.

### Considerations, advantages and limitations

Empirically derived dietary patterns have the advantage that they are independent of definitions of what is a healthy pattern, and they are multidimensional in nature. However, principal component, exploratory factor and cluster analyses are no prediction techniques and are study-population- and data-specific. The derived patterns 'simply' explain the variation in intake. There is no guarantee that the identified patterns will be related to specific health outcomes[3]. Moreover, the application of principal component, exploratory factor and cluster analyses relies on various subjective decisions to be taken by the researcher[1,6]. See Table 2 for an overview.

Specific advantages of principal component and exploratory factor analysis are that they have good statistical power and the resulting dietary patterns show the interrelationships between the dietary components. In contrast, translation of the obtained dietary patterns to the individual is difficult since each individual scores on all the obtained dietary patterns[45]. In practice, the obtained dietary patterns usually explain a limited part of the variation in food intake[2].

For cluster analysis, the translation to individuals is very easy to make since the dietary patterns are mutually exclusive[45]. In most cluster analysis procedures, food components with high variance and outliers have large impacts on the results. For this reason, standardised input variables or the percentage of energy contributed by the food groups are often used as input variables[1]. However, using standardised input variables might give minor food groups undue influence and the differences in the dietary patterns might be diluted[6], whereas expressing foods as their contribution to energy intake might give too little weight to health-related food groups such as fruit and vegetables[35]. Clusters obtained with the K-means method produced cluster solutions that were more reproducible than those obtained with Ward's method[42].

It has been suggested that the combination of factor and cluster analyses is the ultimate way of empirical dietary pattern analysis, since they are complementary and give a better perspective and understanding of dietary habits.[46]

The type of dietary assessment used to collect the dietary data is important to be considered. Interest will mostly be on usual dietary patterns, and in this case day-to-day variation such as present in dietary data collected with 24-h dietary recalls or diet records will behave like random measurement error[6]. The general problems of measurement error associated with self-reported dietary data transfer to the obtained dietary patterns; and might even be more severe because correlations in measurement error might distort the definition of the patterns[3]. This would, for example, occur if unhealthy foods are underestimated systematically by study participants. The effects of misreporting of energy intake on the results of dietary pattern analysis need further study[38].

### Applications

Principal component, exploratory factor and cluster analyses are very useful in obtaining insight into existing dietary patterns within a specific population. Such insight is essential for nutrition education and for developing public health interventions[47]. Principal component and exploratory factor analysis is of particular importance for

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

insight into combinations of foods, and how people score on this[39]. Cluster analysis is more useful in getting insight into different subgroups in the population with different diets, i.e. for identifying groups of people who may be at nutritional risk[38,39]. The thus obtained dietary patterns can be used to explore the combined health effects of commonly existing dietary habits. However, these approaches are not powerful in generating new hypotheses[2]. For hypotheses testing, follow-up is needed through confirmatory type analysis or intervention studies.

## Hybrid methods

### *Principles*

The principle of hybrid approaches to study the overall diet is, not surprisingly, the combination of the two previous approaches. Hybrid approaches are partly theoretically driven, by using predictor variables that are relevant for the purpose of the researcher. In addition, the hybrid approaches identify multivariate dietary patterns based on the study data, specifically relevant for the study population[8]. The predictor variables can be biomarkers that are intermediate risk factors for a dietary-related disease[48], but also other risk factors can be used such as nutrients that are related to the outcome of interest[8], a disease itself[49] or an overall dietary quality score based on recommendations for a healthy diet[9].

The most commonly used hybrid approach in the field of dietary pattern analysis is reduced rank regression (e.g.[8,48,50]). With this approach, linear combinations of food intakes are defined that maximally explain a set of response variables[3,8]. The response variables need to be continuous variables[51], such as levels of a biomarkers or nutrient intakes. The resulting dietary patterns are new dietary variable scores similar to factor scores. Partial least-squares regression is a compromise between principal component analysis and reduced rank regression. With this approach, patterns are obtained that explain both variation in response variables and in the dietary components[8].

Also, cluster analysis has a parallel methodology defining distinct subgroups in a population while making use of an outcome variable. Classification and regression tree analysis is a non-parametric decision tree procedure that identifies mutually exclusive and exhaustive subgroups of a population whose members share common characteristics that are associated with the dependent variable of interest[52]. In contrast to reduced rank regression, decision tree analysis uses one response variable only, e.g. a disease risk factor or disease outcome[49]. The dependent, or response variable, can be either categorical (i.e. classification tree analysis) or continuous (i.e. regression tree analysis). In classification and regression tree analysis, independent variables can be any combination of categorical and continuous variables; no data assumptions are required[52]. Decision tree analysis produces a visual output that is a multilevel structure that resembles branches of a tree. The results can thus be interpreted as hierarchical dietary patterns. The structure of the classification tree model is a set of nodes from the top to the bottom, in which the terminal nodes show the specific pattern features of the subpopulations in percentage, including the number of people and the probability or mean values of the outcome variable[53]. Until now, decision tree analysis was seldom applied for dietary pattern analysis[9] or in a broader risk factor pattern analysis including dietary and other variables[49].

Other data mining techniques, such as neural network approaches might also be promising to obtain insight into the multiple dietary factors or a combination of diet and other risk factors that predict a disease outcome. Only a few applications including dietary variables have been published[9,54–57].

### *Evaluation strategies*

The evaluation strategies of the hybrid approaches to study the overall diet are similar to those described for the empirically based type of analysis. In addition, observed relationships between the obtained dietary patterns and outcome variables should be confirmed. It is important to perform this confirmation in independent populations[8]. In general, more experience is needed with evaluation of the hybrid approaches to study the overall diet[58].

DiBello *et al.*[59] compared dietary patterns derived with principal component analysis, reduced rank regression and partial least-squares regression. Response variables for reduced rank regression and partial least-squares regression were adipose tissue levels of α-linolenic and *trans*-fatty acids and dietary intakes of saturated fat, fibre and folate. All three methods derived a similar vegetable pattern that was associated with myocardial infarction status. However, principal component and partial least-squares regression analysis derived additional dietary patterns that were associated with the health outcome. They conclude that reduced rank regression would have been the most appropriate method if the goal was to test hypotheses limited to the present group of response nutrients. However, to test any dietary pattern relationships with myocardial infarction, partial least-squares regression offered more flexibility[59].

Other studies compared dietary patterns derived by reduced rank regression and principal component or exploratory factor analyses. In three studies, the first dietary pattern derived by reduced rank regression was related to the health outcome, whereas the first dietary pattern obtained by principal component analysis was not[60–62]. In a fourth study, the Mediterranean type dietary patterns derived using both approaches were similar and were both related to the health outcome[50].

### *Considerations, advantages and limitations*

Hybrid approaches have the advantage of building on *a priori* knowledge of biological relations. In this way the derived dietary patterns should be better able to examine the importance of overall dietary patterns in the aetiology of diseases[3,58]. The associated disadvantage is that hybrid approaches require a clear picture of the biological mechanism underlying the development of a given disease. They can only provide answers in the current theoretical framework[51]. There is especially incomplete knowledge

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

as to whether a dietary nutrient or biomarker is causal or merely a marker(63).

One of the criticisms of reduced rank regression is that the observed relations between the dietary pattern and outcome of interest may arise due to the dietary pattern acting as a proxy for the biomarker(58). This requires confirmation of the results in randomly split samples and in independent studies. Confirmation in other studies can be done using the same weight and dietary components as in the original study, hence without actually having the biomarker information available(64).

The disadvantages of decision tree methods are that one key factor can dominate the model, misclassification can be rather large, and the methods might overfit(9). Further considerations, advantages and limitations of decision trees and other data mining techniques need to be learned through more experience in the field of nutrition science.

### *Applications*

Hybrid approaches to study the overall diet may be particularly useful in identifying combinations of dietary components that are relevant for given health outcomes. The application of reduced rank regression is limited to those health outcomes for which sufficient knowledge about intermediate risk factors is available(3,8,61). In the case of partial knowledge about the biochemical pathways, partial least-squares regression might be more appropriate. This technique offers the possibility to obtain dietary patterns that are constrained by the response variables, as well as dietary patterns that are unconstrained by the response variable(59).

In the context of hybrid methods to identify dietary patterns, decision tree type methods seem particularly useful in identifying at-risk subgroups for a health outcome based on combinations of several known dietary and other risk factors (prediction application). In these approaches it is logical to include dietary as well as non-dietary information, because the methodology offers no other option to adjust for non-dietary confounders. It should be noted that decision tree analysis is also a useful technique to generate new hypotheses in the case of no prior hypotheses and many potential risk factors(65). This selection application would, however, not be called a hybrid approach for deriving dietary patterns.

## Discussion and conclusion

In the past three decades, studies of the overall diet have emerged as an important area of research complementary to single component studies(4). This paper presents an overview of different approaches used in studies of the overall diet. The described approaches included hypothesis-driven scores of overall dietary quality, data-driven approaches such as principal component, exploratory factor and cluster analysis, and hybrid approaches such as reduced rank regression, partial least-squares regression and decision tree analysis. Several reviews have been published that present comprehensive overviews of existing dietary quality scores, empirically derived dietary patterns, and their relationships with demographic characteristics, risk factors, biomarkers, health and disease(1–3,6,7,10–12,20,21,23,24,26,27,38,58,63). The present paper did not attempt to update these reviews, but focused particularly on methodological aspects.

The results of studies of the overall diet have great potential for use in nutrition policy, particularly as it demonstrates the importance of total diet in health promotion. Dietary quality scores are primarily important for monitoring the quality of the overall diet, to evaluate the overall effects of dietary interventions(26) and to test the validity of dietary recommendations(64). Data-driven approaches are particularly important for nutrition education and setting priorities in the planning of nutritional interventions(64). They show the interrelationships between dietary components and differences in dietary patterns within a population(61). However, the extent to which dietary quality scores and data-driven approaches help to generate new insights into the relationships between dietary intake and diet-related diseases remains debatable(66).

Reduced rank regression seems to have greater potential for testing new hypotheses on diet–disease relationships through specific biological pathways(3). The hybrid approach is potentially strong because the derived dietary patterns are relevant for the population and related to health outcomes; whereas the *a priori* diet pattern scores might have little contrast within a population and the *a posteriori* derived diet patterns might not be relevant for health. To our knowledge this is the first overview paper that presented and reflected upon alternative hybrid approaches to reduced rank regression. For reduced rank regression Kant concluded that these methods require further development and innovation(2). This is even more the case of the alternative hybrid approaches, which require more applications in the field of dietary patterns before conclusions on their use can be drawn.

The potential for several of the hybrid approaches to study the overall diet strongly depends on the availability of early risk factors for diseases(63). Many chronic diseases develop over a period of many years, and are multi-causal in nature. This makes the studying of diet in relation to disease extremely complicated. Valid (bio)markers that are causal predictors for the development of disease might be an important help in this complex task. They can serve as response variables in the hybrid approaches to study the relationship with the overall diet. In a second and preferably independent step, the thus derived dietary patterns might subsequently be related to the incidence of diseases in long-term prospective studies(3).

Few intermediate risk factors such as LDL- and HDL-cholesterol for CVD have long been used as clinical biomarkers. Since the early 1990s, research on the discovery and validation of biomarkers with prognostic values for CVD, cancer, obesity, diabetes and neurodegeneration has expanded considerably(67). Although for many diseases valid predictor biomarkers are currently still lacking, several developments work to the advantage of discovering new biomarkers for disease risk. The wish for substantiating health claims is one of the important driving forces for more research on the identification of further relevant markers to measure food functionality in the human

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

body[68]. Moreover, recent advances in genomics and systems biology enable researchers to measure and model biomarker profiles and to translate these into dynamic processes[67]. Especially markers for suboptimal health before clinical signs of disease are of increasing interest. Work on the development of markers for overarching processes such as oxidative, inflammatory, metabolic and psychological stress[67] is of great potential value for hybrid approaches for studying the overall diet.

From this overview, it is concluded that the various approaches for studying the overall diet are complementary, and no method can be considered superior to the other methods. Further insight into the utility of conducting studies on the overall diet can be gained if more attention is given to methodological issues. These include clarification of the aims and assumptions of the analyses and a precise description of the make-up of dietary quality scores or derived dietary patterns. Moreover, in-depth evaluations of the derived measures of the overall diet in terms of reproducibility, validity and comparisons of different methodologies are essential. This is particularly the case for the still less often applied hybrid approaches such as reduced rank regression, partial least-squares regression and decision tree analysis.

## Acknowledgements

The author declares no conflicts of interest. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## References

1. Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, 177–203.
2. Kant AK (2004) Dietary patterns and health outcomes. *J Am Diet Assoc* **104**, 615–635.
3. Schulze MB & Hoffmann K (2006) Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. *Br J Nutr* **95**, 860–869.
4. Jacobs DR Jr & Steffen LM (2003) Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr* **78**, Suppl. 3, 508S–513S.
5. World Health Organisation (2003) *Diet, Nutrition and the Prevention of Chronic Diseases. Joint WHO–FAO Expert Consulation. World Health Organisation Technical Report Series No. 916*. Geneva: WHO.
6. Moeller SM, Reedy J, Millen AE *et al.* (2007) Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. *J Am Diet Assoc* **107**, 1233–1239.
7. Waijers PM, Feskens EJ & Ocke MC (2007) A critical review of predefined diet quality scores. *Br J Nutr* **97**, 219–231.
8. Hoffmann K, Schulze MB, Schienkiewitz A *et al.* (2004) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* **159**, 935–944.
9. Hearty AP & Gibney MJ (2008) Analysis of meal patterns with the use of supervised data mining techniques–artificial neural networks and decision trees. *Am J Clin Nutr* **88**, 1632–1642.
10. Arvaniti F & Panagiotakos DB (2008) Healthy indexes in public health practice and research: a review. *Crit Rev Food Sci Nutr* **48**, 317–327.
11. Wirt A & Collins CE (2009) Diet quality – what is it and does it matter? *Public Health Nutr* **12**, 2473–2492.
12. Lazarou C & Newby PK (2011) Use of dietary indexes among children in developed countries. *Adv Nutr* **2**, 295–303.
13. Guenther PM, Reedy J & Krebs-Smith SM (2008) Development of the healthy eating index-2005. *J Am Diet Assoc* **108**, 1896–1901.
14. Kennedy ET, Ohls J, Carlson S *et al.* (1995) The healthy eating index: design and applications. *J Am Diet Assoc* **95**, 1103–1108.
15. Trichopoulou A, Kouris-Blazos A, Wahlqvist ML *et al.* (1995) Diet and overall survival in elderly people. *BMJ* **311**, 1457–1460.
16. Krebs-Smith S, Guenther P, O'Connell K *et al.* (2012) Development and evaluation of the Healthy Eating Index-2010. In Eight International Conference on Diet and Activity Methods, 14–17 May 2012, FAO, Rome. Abstract book OC 038, 34–35.
17. Glanville NT & McIntyre L (2006) Diet quality of Atlantic families headed by single mothers. *Can J Diet Pract Res* **67**, 28–35.
18. Feskanich D, Rockett HR & Colditz GA (2004) Modifying the Healthy Eating Index to assess diet quality in children and adolescents. *J Am Diet Assoc* **104**, 1375–1383.
19. Keys A & Grande F (1957) Role of dietary fat in human nutrition. III. Diet and the epidemiology of coronary heart disease. *Am J Public Health Nations Health* **47**, 1520–1530.
20. Mila-Villarroel R, Bach-Faig A, Puig J *et al.* (2011) Comparison and evaluation of the reliability of indexes of adherence to the Mediterranean diet. *Public Health Nutr* **14**, 2338–2345.
21. Bach A, Serra-Majem L, Carrasco JL *et al.* (2006) The use of indexes evaluating the adherence to the Mediterranean diet in epidemiological studies: a review. *Public Health Nutr* **9**, 132–146.
22. Trichopoulou A, Kouris-Blazos A, Vassilakou T *et al.* (1995) Diet and survival of elderly Greeks: a link to the past. *Am J Clin Nutr* **61**, 1346S–1350S.
23. Sofi F, Abbate R, Gensini GF *et al.* (2010) Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis. *Am J Clin Nutr* **92**, 1189–1196.
24. Kant AK (1996) Indexes of overall diet quality: a review. *J Am Diet Assoc* **96**, 785–791.
25. Guenther PM, Reedy J, Krebs-Smith SM *et al.* (2008) Evaluation of the Healthy Eating Index-2005. *J Am Diet Assoc* **108**, 1854–1864.
26. Fransen HP & Ocke MC (2008) Indices of diet quality. *Curr Opin Clin Nutr Metab Care* **11**, 559–565.
27. Kourlaba G & Panagiotakos DB (2009) Dietary quality indices and human health: a review. *Maturitas* **62**, 1–8.
28. Zhang S, Midthune D, Guenther PM *et al.* (2011) A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Ann Appl Stat* **5**, 1456–1487.
29. Kipnis V, Subar AF, Midthune D *et al.* (2003) Structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol* **158**, 14–21; discussion 22–16.
30. Apovian CM, Murphy MC, Cullum-Dugan D *et al.* (2010) Validation of a web-based dietary questionnaire designed for the DASH (dietary approaches to stop hypertension) diet: the DASH online questionnaire. *Public Health Nutr* **13**, 615–622.

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press

31. Kennedy G, Berardo A, Papavero C *et al.* (2010) Proxy measures of household food consumption for food security assessment and surveillance: comparison of the household dietary diversity and food consumption scores. *Public Health Nutr* **13**, 2010–2018.
32. Devlin UM, McNulty BA, Nugent AP *et al.* (2012) The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proc Nutr Soc* **71**, 599–609.
33. Gorst-Rasmussen A, Dahm CC, Dethlefsen C *et al.* (2011) Exploring dietary patterns by using the treelet transform. *Am J Epidemiol* **173**, 1097–1104.
34. Imamura F & Jacques PF (2011) Invited commentary: dietary pattern analysis. *Am J Epidemiol* **173**, 1105–1108.
35. Bailey RL, Gutschall MD, Mitchell DC *et al.* (2006) Comparative strategies for using cluster analysis to assess dietary patterns. *J Am Diet Assoc* **106**, 1194–1200.
36. Balder HF, Virtanen M, Brants HA *et al.* (2003) Common and country-specific dietary patterns in four European cohort studies. *J Nutr* **133**, 4246–4251.
37. Northstone K, Ness AR, Emmett PM *et al.* (2008) Adjusting for energy intake in dietary pattern investigations using principal components analysis. *Eur J Clin Nutr* **62**, 931–938.
38. Devlin UM, McNulty BA, Nugent AP *et al.* (2012) The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proc Nutr Soc*, 1–11.
39. Reedy J, Wirfalt E, Flood A *et al.* (2010) Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: the NIH-AARP Diet and Health Study. *Am J Epidemiol* **171**, 479–487.
40. Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**, 3–9.
41. Hu FB, Rimm E, Smith-Warner SA *et al.* (1999) Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am J Clin Nutr* **69**, 243–249.
42. Lo Siou G, Yasui Y, Csizmadi I *et al.* (2011) Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns: The Tomorrow Project. *Am J Epidemiol* **173**, 956–967.
43. Lau C, Glumer C, Toft U *et al.* (2008) Identification and reproducibility of dietary patterns in a Danish cohort: the Inter99 study. *Br J Nutr* **99**, 1089–1098.
44. Quatromoni PA, Copenhafer DL, Demissie S *et al.* (2002) The internal validity of a dietary pattern analysis. The Framingham Nutrition Studies. *J Epidemiol Community Health* **56**, 381–388.
45. Hearty AP & Gibney MJ (2009) Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. *Br J Nutr* **101**, 598–608.
46. Engeset D, Alsaker E, Ciampi A *et al.* (2005) Dietary patterns and lifestyle factors in the Norwegian EPIC cohort: the Norwegian Women and Cancer (NOWAC) study. *Eur J Clin Nutr* **59**, 675–684.
47. van Dam RM, Grievink L, Ocke MC *et al.* (2003) Patterns of food consumption and risk factors for cardiovascular disease in the general Dutch population. *Am J Clin Nutr* **77**, 1156–1163.
48. Hoffmann K, Zyriax BC, Boeing H *et al.* (2004) A dietary pattern derived to explain biomarker variation is strongly associated with the risk of coronary artery disease. *Am J Clin Nutr* **80**, 633–640.
49. Camp NJ & Slattery ML (2002) Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes Control* **13**, 813–823.
50. Vujkovic M, Steegers EA, Looman CW *et al.* (2009) The maternal Mediterranean dietary pattern is associated with a reduced risk of spina bifida in the offspring. *BJOG* **116**, 408–415.
51. Kroke A (2004) Re: 'Application of a new statistical method to derive dietary patterns in nutritional epidemiology'. *Am J Epidemiol* **160**, 1132.
52. Lemon SC, Roy J, Clark MA *et al.* (2003) Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med* **26**, 172–181.
53. Teng JH, Lin KC & Ho BS (2007) Application of classification tree and logistic regression for the management and health intervention plans in a community-based study. *J Eval Clin Pract* **13**, 741–748.
54. Cooper JG & Purcell GP (2006) Data mining for correlations between diet and Crohn's disease activity. *AMIA Annu Symp Proc* **2006**, 897.
55. Valavanis IK, Mougiakakou SG, Grimaldi KA *et al.* (2010) A multifactorial analysis of obesity as CVD risk factor: use of neural network based methods in a nutrigenetics context. *BMC Bioinformatics* **11**, 453.
56. Huang S, Xu Y, Yue L *et al.* (2010) Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a Chinese area. *Hypertens Res* **33**, 722–726.
57. Park J & Edington DW (2004) Application of a prediction model for identification of individuals at diabetic risk. *Methods Inf Med* **43**, 273–281.
58. Tucker KL (2010) Dietary patterns, approaches, and multicultural perspective. *Appl Physiol Nutr Metab* **35**, 211–218.
59. DiBello JR, Kraft P, McGarvey ST *et al.* (2008) Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *Am J Epidemiol* **168**, 1433–1443.
60. Manios Y, Kourlaba G, Grammatikaki E *et al.* (2010) Comparison of two methods for identifying dietary patterns associated with obesity in preschool children: the GENESIS study. *Eur J Clin Nutr* **64**, 1407–1414.
61. Nettleton JA, Steffen LM, Schulze MB *et al.* (2007) Associations between markers of subclinical atherosclerosis and dietary patterns derived by principal components analysis and reduced rank regression in the Multi-Ethnic Study of Atherosclerosis (MESA). *Am J Clin Nutr* **85**, 1615–1625.
62. Hoffmann K, Boeing H, Boffetta P *et al.* (2005) Comparison of two statistical approaches to predict all-cause mortality by dietary patterns in German elderly subjects. *Br J Nutr* **93**, 709–716.
63. Kant AK (2010) Dietary patterns: biomarkers and chronic disease risk. *Appl Physiol Nutr Metab* **35**, 199–206.
64. van Dam RM (2005) New approaches to the study of dietary patterns. *Br J Nutr* **93**, 573–574.
65. Dasgupta A, Sun YV, Konig IR *et al.* (2011) Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genet Epidemiol* **35**, Suppl. 1, S5–S11.
66. Jacques PF & Tucker KL (2001) Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr* **73**, 1–2.
67. van Ommen B, Keijer J, Heil SG *et al.* (2009) Challenging homeostasis to define biomarkers for nutrition related health. *Mol Nutr Food Res* **53**, 795–804.
68. Gallagher AM, Meijer GW, Richardson DP *et al.* (2011) A standardised approach towards PROving the efficacy of foods and food constituents for health CLAIMs (PROCLAIM): providing guidance. *Br J Nutr* **106**, Suppl. 2, S16–S28.

https://doi.org/10.1017/S0029665113000013 Published online by Cambridge University Press