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 \textit{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, \textit{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 analysis, scores or indices are derived from the underlying dietary data. These scores or indices are usually based on one or 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.

Corresponding author: Dr Marga Ocké, fax +31 30 274 4466, email marga.ocke@rivm.nl
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...
legumes, cereals, fruits and nuts, vegetables, meat and meat products, milk and dairy products, and alcohol. The components were adjusted for energy intake by standardizing 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 Score) to eight (excellent adherence to the Mediterranean diet). The Mediterranean Diet Score has been applied to Mediterranean and non-Mediterranean populations. Over time, various alternative Mediterranean Diet Scores were developed and tested. 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 an 8% reduction in mortality.

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 nutritional adequacy only. 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. 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 were assessed, and it was checked which components had most influence on the overall index score. 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.

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.

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. Also, interpretations of the dietary guidelines are often needed to construct a dietary quality score, and subjectivity is introduced. 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. A second advantage of dietary quality scores is that they are usually easy to compute; and thereby easily reproducible and comparable.

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. However, especially, persons who have a midrange score can have very different contributing components, and thus different dietary patterns.

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. 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. Overall, the general problems of dietary assessment through self-reporting of food consumption 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.

An efficient application of dietary quality scores is the combination of a dietary quality score with a short dietary
### 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\(^2\). In dietary pattern analysis, principal component analysis is more commonly used than exploratory factor analysis. The obtained component scores are continuous variables\(^3\).

In the second approach, i.e. cluster analysis, mutually exclusive non-overlapping clusters of individuals are created\(^4\). 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\(^5\).

Other empirical approaches have been applied to obtain dietary patterns. A recent example is the use of the treelet transform\(^6\). 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\(^7\) is its assumption that only selected dietary components can contribute to the patterns, allowing no contributions of other dietary factors to the patterns\(^8\).

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\(^9\). 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\(^10\). 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\(^11\).

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

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Implementation choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall concept</td>
<td>Creation of new variables (factor analysis), creation of mutually exclusive groups of individuals (cluster analysis; factor analysis followed by quantile grouping).</td>
</tr>
<tr>
<td>Included components</td>
<td>Based on FFQ or diet records, less often on dietary recalls or diet histories. Most often food groups, but also individual food items, nutrients (including ratios), or combinations. The number of components.</td>
</tr>
<tr>
<td>Adjustment of input variables</td>
<td>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.</td>
</tr>
<tr>
<td>before analyses</td>
<td></td>
</tr>
<tr>
<td>Statistical analysis type</td>
<td>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).</td>
</tr>
<tr>
<td>and decisions to be taken</td>
<td>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(^12).</td>
</tr>
<tr>
<td>Labelling of the dietary pattern</td>
<td>Can be quantitatively (based on the highest factor loadings; or nutrient composition), or qualitatively (specific combinations of foods and nutrient composition).</td>
</tr>
<tr>
<td>Subgroup specificity</td>
<td>Sometimes sex specific; usually not age specific. Seldom stratified by race/ethnicity, culture or geography.</td>
</tr>
</tbody>
</table>
**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-healthy dietary pattern. Also a pattern high in desserts or sweets and patterns high in alcohol appeared repeatedly. In principal component and exploratory factor analyses, a healthy dietary pattern is often labelled ‘prudent’ and a less-healthy pattern ‘Western’. Obviously, patterns with the same label can be defined by different food components or by different weights of the components. 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. 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.

**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, comparison of using different types of input variables or different strategies to derive the dietary patterns, and the reproducibility of derived dietary patterns. Reproducibility can be assessed internally using split sample techniques, or externally over time for the same population, and in different but similar study populations.

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 and those derived by cluster analysis using discriminant analysis.

Some researchers indicate that empirically derived dietary patterns should be validated by assessing whether the dietary patterns can reliably predict diseases or mortality. 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. Moreover, the application of principal component, exploratory factor and cluster analyses relies on various subjective decisions to be taken by the researcher. 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. In practice, the obtained dietary patterns usually explain a limited part of the variation in food intake.

For cluster analysis, the translation to individuals is very easy to make since the dietary patterns are mutually exclusive. 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. However, using standardised input variables might give minor food groups undue influence and the differences in the dietary patterns might be diluted, whereas expressing foods as their contribution to energy intake might give too little weight to health-related food groups such as fruit and vegetables. Clusters obtained with the K-means method produced cluster solutions that were more reproducible than those obtained with Ward’s method.

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.

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. 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. 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.

**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. Principal component and exploratory factor analysis is of particular importance for...
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 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...
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 (64). 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
body\textsuperscript{(68)}. Moreover, recent advances in genomics and systems biology enable researchers to measure and model biomarker profiles and to translate these into dynamic processes\textsuperscript{(67)}. Especially markers for suboptimal health before clinical signs of disease are of increasing interest. Work on the development of markers for overarching processes\textsuperscript{(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


