Irish Section Postgraduate Symposium

The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting

Una M. Devlin*, Breige A. McNulty, Anne P. Nugent and Micheal J. Gibney
UCD Institute of Food and Health, University College Dublin, Belfield, Dublin 4, Republic of Ireland

Over the last three decades, dietary pattern analysis has come to the forefront of nutritional epidemiology, where the combined effects of total diet on health can be examined. Two analytical approaches are commonly used: *a priori* and *a posteriori*. Cluster analysis is a commonly used *a posteriori* approach, where dietary patterns are derived based on differences in mean dietary intake separating individuals into mutually exclusive, non-overlapping groups. This review examines the literature on dietary patterns derived by cluster analysis in adult population groups, focusing, in particular, on methodological considerations, reproducibility, validity and the effect of energy mis-reporting. There is a wealth of research suggesting that the human diet can be described in terms of a limited number of eating patterns in healthy population groups using cluster analysis, where studies have accounted for differences in sex, age, socio-economic status, geographical area and weight status. Furthermore, patterns have been used to explore relationships with health and chronic diseases and more recently with nutritional biomarkers, suggesting that these patterns are biologically meaningful. Overall, it is apparent that consistent trends emerge when using cluster analysis to derive dietary patterns; however, future studies should focus on the inconsistencies in methodology and the effect of energy mis-reporting.

Dietary patterns: Cluster analysis: Adults: Energy mis-reporting

With the global prevalence of chronic diseases increasing, it is now widely accepted that diet has an important role to play, as many of these diseases may have a nutritional base or may be promoted by inappropriate dietary habits(1,2). Traditionally, nutritional epidemiology focused on a detailed examination of single nutrient intake; however, over the last three decades research has moved towards examining the combined effect of total food intake. This significant shift reflects a need to explore the complexity of individual total dietary intake and it is hoped that this alternative approach will help to increase our understanding of the role of diet in chronic diseases and improve the effectiveness of public health recommendations(3). Furthermore, it has been recognised that individuals consume diverse diets consisting of many foods containing complex combinations of nutrients and it is likely that these nutrients will interact with each other, an effect that may be confounded within the single nutrient approach(4).

One way to examine the combined effect of total food intake on health is to derive dietary patterns. Dietary patterns are typically characterised on the basis of habitual food intake and can be described as a measure of usual intake of food combination in individuals and groups where nutritional variables are grouped according to some criterion of nutritional status(5). Two analytical approaches are commonly used: *a priori* and *a posteriori*. The *a priori* approach is a theoretically driven method that focuses on constructing dietary scores using a predefined combination of diet quality based on published dietary guidelines(6). The *a posteriori* approach is an exploratory method that

**Abbreviations:** hcy, homocysteine; %TE food, percentage total energy contribution from food.

*Corresponding author:* Una M. Devlin, email una.devlin@ucd.ie
uses multivariate statistical techniques to derive dietary patterns where large datasets representing total food intake are aggregated and reduced to smaller datasets to summarise total dietary exposure\(^{(7)}\). Factor analysis and cluster analysis are two *a posteriori* methods commonly used to derive dietary patterns in nutritional epidemiology. In factor analysis, linear combinations (factors) are created based on correlations between dietary intakes where each individual receives a score for the derived factors; however, these scores are difficult to interpret as an individual can belong to more than one factor\(^{(8)}\). Cluster analysis, on the other hand, offers the advantage of deriving dietary patterns which represent homogenous groups that can be related to other variables\(^{(14)}\).

In studies where factor and cluster analysis were used simultaneously to derive dietary patterns, results have shown good evidence of comparability. Two studies have indicated that there is a high resemblance between some of the clusters and factors identified due to similarities in food types\(^{(8,10)}\). In addition, one study reported that three patterns dominated irrespective of which method was used\(^{(8)}\). Dietary patterns derived using both methods have also been compared with plasma lipid markers. Newby *et al.* reported that a cluster and a factor dominated by healthy foods were both inversely associated with plasma TAG, whereas a cluster and a factor dominated by alcohol were both directly associated with HDL and cholesterol\(^{(11)}\). Although both methods are directly comparable, it has been suggested that the choice of the dietary pattern analysis technique should depend on the type of outcome that is needed from the dataset as each method approaches the data from different angles and thus answers different questions\(^{(8)}\). Other authors have suggested that the ultimate way to approach dietary pattern analysis is to use a combination of factor and cluster analysis as complementary approaches\(^{(12)}\) in order to give a better perspective and understanding of dietary habits\(^{(13)}\).

Clustering methods separate individuals into mutually exclusive, non-overlapping clusters, where an individual can belong to one cluster only, therefore representing a unique cluster or dietary pattern\(^{(8)}\). Differences between clusters are based on mean dietary intake of each individual, where the dietary patterns derived are specific to individuals within each cluster and each cluster has a specific food and nutrient composition\(^{(14)}\). Clusters are then labelled based on shared characteristics of dietary intake, where individuals with similar dietary intake will cluster together, away from others in dissimilar clusters. Dietary input variables can include nutrients, foods or food groups or a combination of all three\(^{(15)}\). However, within the literature, food groups are most commonly used\(^{(8,16–19)}\). One reason for using food groups as the preferred dietary input variable is that these groups can represent total dietary intake, accounting for any interaction between nutrients within the groups. Furthermore, various algorithms can also be used in the clustering procedure. The principle of all clustering algorithms is to calculate the Euclidean distance, which measures the distance between each dietary variable consumed together by similar individuals. Individuals are then grouped into clusters where the distance is maximised between the defined centre of each cluster from others, while the distance is minimised between any single individual and the centre of their closest cluster\(^{(5)}\). Of these algorithms, the *k*-means approach is most frequently used\(^{(8,19–21)}\), although this algorithm has limitations which will be discussed later. This review examines the literature on dietary patterns derived by cluster analysis in adult population groups only, focusing in particular on methodological considerations, reproducibility, validity and the effect of energy mis-reporting.

### Methodological considerations

Many dietary assessment tools are available to researchers to estimate dietary intake of an individual or a population group. These methods can be split into two categories: one is the prospective method, i.e. those that record data at the time of eating (dietary records) and the other is the retrospective method, i.e. those that collect data about the diet eaten in the past (diet histories, FFQ and dietary recalls)\(^{(22)}\). Within dietary pattern analysis, consideration should be given to the most appropriate method, as some may provide more ‘favourable’ results than others as several may not accurately identify the usual food pattern\(^{(23)}\). The impact of the dietary assessment methods used in cluster analysis will be discussed later in the review.

In recent years, scrutiny of the statistical methodology concerning cluster analysis has been undertaken by many researchers, due to its highly exploratory nature. One issue of concern is researcher bias, which can ultimately influence the grouping of the dietary variables and the number of clusters in the final solution\(^{(8)}\). The frequently used *k*-means approach has a subjective element as the number of clusters needs to be predefined prior to analysis. To overcome this problem varying cluster solutions are usually run and then the clusters are examined for the best fit using cross-validation methods. Two approaches that can be used to examine the final cluster solution are to calculate the within cluster variance ratio\(^{(20,24,25)}\) or to generate scree plots\(^{(26,27)}\), where higher ratios indicate a better separation of clusters. It has been suggested, however, that there is no gold standard for determining the number of clusters\(^{(15)}\). In many cases, the appropriate number of clusters is determined by the author, taking into consideration those which are clearly distinct and nutritionally meaningful, while also maintaining a reasonable sample size\(^{(25)}\). In a similar way, there is no gold standard concerning the format of the dietary variable for the clustering procedure. Preferably, the dietary variables should be grouped to suitably represent the dataset to increase the likelihood of identifying sensible dietary patterns. When using food groups as the dietary variable, it has been suggested that food items consumed need to be aggregated into a limited number of groups avoiding the exclusion of subjects due to missing data\(^{(28)}\). Previous studies have joined food groups together based on similarities in food group types\(^{(18,16,18)}\) or on nutrient content and culinary preference\(^{(19,29,30)}\). In most cases authors have also differentiated between food groups, e.g. low- or high-energy and low- or high-fat\(^{(8,16,19,29,30)}\). Food groups are usually presented using three different methods (1) the frequency of
the food consumed (servings)\(^{(17,19)}\), (2) the portion size of the food consumed (grams)\(^{(8,21)}\) or (3) the percentage total energy contribution from food (%TE food)\(^{(8,30,33)}\). Few studies have examined the impact of the methodological differences between these different methods. One author has proposed that when using the %TE food method, differences in energy needs due to sex, age, body weight and level of physical activity can be accounted for\(^{(25)}\). One study that compared two methods (servings and %TE food) reported similar clusters for food groups high in energy. However, clusters arising from %TE food were less likely to differentiate between low-energy foods such as fruit and vegetables. The authors therefore concluded that the servings approach best represented the patterns\(^{(32)}\). In contrast, a second study that clustered using the grams and %TE food methods showed that the %TE food method best characterised the patterns, which were fully interpretable based on their contributing food group\(^{(8)}\). To the best of our knowledge no studies have examined the results obtained comparing all three methods in one dataset, therefore, it is difficult to make firm conclusions on the best method to use. One way to overcome the issue of high- or low-energy food groups affecting the patterns is to standardise the variables prior to analysis ensuring that variables with large variances which may have greater effects on resulting patterns than those with small variances can be accounted for\(^{(24)}\). Ideally, by standardising the input variables, all food groups will have equal influence on the clustering procedure. Research carried out by Wirfalt et al.\(^{26}\) examining the effect of standardising variables found that the distribution of individuals was more evenly spread and differences in nutrient intake across patterns were improved when using the un-standardised approach\(^{(33)}\). Furthermore, in a follow-up study, Wirfalt reported that the transformation of variables by standardisation may have an effect on the dietary patterns identified as low-energy foods may be given equal weights to high-energy foods, which may represent poor dietary patterns\(^{(34)}\). Overall, there is insufficient evidence regarding the standardisation procedure and more research is needed.

### Dietary patterns in healthy population groups

Throughout the last three decades many studies have identified meaningful dietary patterns in healthy population groups using cluster analysis as the patterning method. Initial studies focused on identifying patterns where nutrient intakes were inadequate v. published dietary recommendations, thus acknowledging that cluster analysis is a useful tool for identifying groups of people who may be at nutritional risk\(^{(35,36)}\). Later studies have accounted for the influence of sex, age, socio-economic status, geographical area and weight status. A range of dietary assessment methods were used including FFQ, dietary recalls and diet records. Only one study used nutrients as the clustering variable\(^{(35)}\), whereas another used meal type\(^{(37)}\); therefore, food groups were predominantly used and were presented using servings\(^{(9,19,36,38-44)}\), grams\(^{(13,16,21,45-47)}\) and %TE food\(^{(8,18,31)}\). It is noteworthy that no matter which dietary assessment method or clustering variable was used, similar dietary patterns have been found across a collection of studies in healthy population groups.

In all studies, labels or names are normally assigned to characterise each pattern, based on the dietary intake that contributes relatively greater proportions\(^{(11,31,48)}\). Two commonly used terms are ‘healthy’ patterns characterised by the consumption of fruits and vegetables and ‘unhealthy’ patterns characterised by the consumption of foods high in fat and salt\(^{(8,30,33,39)}\). ‘Healthy’ patterns can also be referred to as ‘prudent’, while ‘unhealthy’ patterns can also be referred to as ‘western’ or ‘traditional’\(^{(8,21,45)}\). A strength of these studies is large sample size \((n>1379)^{(8,21,35,38,39)}\) (only one study of sample size \(n \in 213^{(45)}\)) though many were carried out in females\(^{(9,36)}\) or older adults\(^{(31)}\) only. In one study of London adults aged 39–63 years, differences were reported in the type of ‘healthy’ patterns identified by using terms such as ‘very healthy’ or ‘moderately healthy’, similarly for ‘unhealthy’ patterns\(^{(39)}\). Other descriptive labels used to characterise dietary patterns relate to ‘high- or low-nutrient density’\(^{(40,43)}\) or ‘glycaemic level’\(^{(42)}\); however, these findings are limited to three US studies in either females or older adults. Furthermore, many studies have examined differences in socio-economic status according to dietary patterns, reporting that typically ‘healthy’ patterns are associated with increased socio-economic status in males and females\(^{(13,21,36,39,46)}\).

Significant differences among dietary patterns by sex have also been reported, highlighting the need to examine males and females separately in healthy population groups\(^{(26,49)}\). In a study carried out in a representative sample of UK adults aged 16–64 years, it was reported that dietary patterns differ by sex\(^{(16)}\) but these differences were lost in an older cohort aged 65+ years of the same study\(^{(46)}\). Confirmation that dietary patterns differ by sex was reported in a cohort of older Italian adults aged 65+ years\(^{(41)}\). Swedish adults aged 30–60 years\(^{(19)}\), African–American adults aged 18+ years\(^{(44)}\) and American adults aged 20–70 years\(^{(17)}\). These studies suggest that dietary patterns differ by sex and this should therefore be accounted for in public health recommendations. Few studies have reported differences among age across dietary patterns\(^{(16,35,40,45)}\) and to the best of our knowledge no studies have examined the effect of age groups on dietary patterns in a large representative sample.

Dietary pattern analysis is also influenced by geography. Within large cohorts of older European adults, specific dietary patterns have been found to represent those living in Northern and Southern regions where one of these patterns is usually considered as more healthy\(^{(18,41,47,50)}\). Differences have also been found at a national level; in a large study of Norwegian females aged 41–56 years, one dietary pattern was dominated by those living in a certain region of Norway\(^{(13)}\). These results could therefore indicate that dietary patterns are influenced by geography and are associated with cultural perceptions, beliefs and attitudes about foods which can ultimately affect food choice. Although these studies are of large sample sizes, a limitation is that they are limited to groups of older adults and female populations only.
Three studies have also examined differences in weight status according to dietary patterns in healthy population groups. These studies have reported that BMI of individuals is significantly different across all patterns after controlling for age, sex, exercise and total energy intake in US adults (mean age 37 years)\(^{(20)}\) and UK adults aged 16–64 years\(^{(16)}\). In the US study, the dietary pattern with the highest mean BMI was found to be predominantly male and had high intake of soft drinks. In contrast, in a large sample of Swedish adults aged 47–68 years, Holmback reported that the ‘fruit’ pattern had the greatest proportion of overweight individuals\(^{(37)}\). These differences may perhaps be explained by the different types of clustering variables used (servings, %TE and meal type); however, further research is required.

The earlier studies in general show consistent findings across dietary patterns in healthy population groups. One issue of concern is that few have accounted for energy misreporters, with only two studies excluding such reporters from their analysis. This issue will be discussed later in the review. It is evident that from these studies, literature is accumulating in relation to using cluster analysis to derive dietary patterns taking into account sex, age, socioeconomic status, geographical area and weight status; however, the lack of consensus of some studies warrants further research in this area.

### Dietary patterns and associations with chronic diseases

The effect of diet on chronic diseases is a key consideration in nutritional epidemiology. By considering the effect of total diet using dietary pattern analysis, it is believed that various patterns may influence the development and possibly increase the risk of many diet related chronic diseases over time. An overview of the literature examining the association of dietary patterns and chronic diseases is outlined in Table 1 and reviewed briefly later.

As previously discussed, evidence has suggested that weight status can differ according to dietary patterns in cross-sectional cohorts\(^{(16,26,37)}\). In studies, specifically examining the risk of obesity, it has been reported that in comparison with ‘healthy’ patterns and after adjustments for confounders, patterns that are considered ‘less healthy’ have a significantly larger BMI and waist circumference\(^{(29,51)}\), higher total percentage body fat (males only)\(^{(25)}\) and are associated with an increased risk of overweight (14–17\%)\(^{(52,53)}\) and obesity (20\%)\(^{(53)}\). Interestingly, Carrera et al. found that no one pattern was associated with increased risks of obesity as it was reported that BMI and waist circumference were high among all patterns identified\(^{(54)}\). Overall, arising from these large studies involving a wide variety of age groups, the consensus appears that subjects in ‘healthy’ patterns following current dietary recommendations are at lesser risk of becoming overweight or obese. Furthermore, it has been suggested that due to the complexity of total diet, future studies should consider the influence of total food volume on energy balance\(^{(29)}\).

Dietary patterns have also been associated with CVD risk mainly in prospective studies. As before, ‘healthy’ patterns have been shown to be protective, lowering the risk of subclinical heart disease\(^{(55)}\) and carotid atherosclerosis\(^{(56)}\) by 4\% and are favourably associated with anthropometric, blood pressure and blood lipid values\(^{(28)}\) and with markers of inflammation\(^{(57)}\) in comparison with the other patterns identified. However, one study relied on the analysis of non-fasting blood samples\(^{(28)}\). In one case-control study, food groups associated with increased risk of acute myocardial infarction after adjustments for confounders were a ‘red meat and alcohol’ pattern in males and females and a ‘low fruit and vegetables’ pattern in females only, where the ‘red meat and alcohol’ pattern had significantly higher risks of CVD risk markers than those in a ‘healthy’ pattern\(^{(58)}\). Interestingly, in one study no one pattern was associated with increased CVD risk although a ‘sweets’ pattern, showed a protective effect against CVD risk factors as significant associations were reported among HDL and elevated systolic blood pressure\(^{(59)}\). These results provide support for the protective effects of ‘healthy’ dietary patterns against CVD.

Dietary patterns have also been linked to risk factors for diabetes. In one study, where 67 and 33\% of subjects had normal and impaired glucose tolerance, respectively, it was reported that the ‘white bread’ pattern was associated with poorest insulin sensitivity and adiposity levels, whereas a ‘wine’ and ‘dark bread’ pattern was associated with improving these markers\(^{(60)}\). In non-diabetic cohorts, it has been reported that a pattern that is high in dairy products and low in staple foods is associated with a lower prevalence of type-2 diabetes\(^{(61)}\), and a ‘healthy’ pattern improves insulin concentration and anthropometric profiles\(^{(62)}\). One study also reported that a pattern with high intake of animal and soyabean products had a higher prevalence of glucose tolerance abnormalities, after adjustment for confounders\(^{(63)}\). The cross-sectional study design of most of these studies is a limitation as information on diet (mainly collected using FFQ) and indicators of diabetes were collected at one specific point in time. This highlights the need for more prospective studies to be carried out in order to determine how the dietary patterns affect diabetes over a certain time frame.

Specific dietary patterns have also been associated with cancer risk, mainly in case–control studies. As before, ‘healthy’ dietary patterns were shown to have protective effects, and to reduce the risk of oesophageal cancer\(^{(64)}\), gastric cancer\(^{(65)}\), ovarian cancer\(^{(66)}\) and lung cancer in subjects who smoke\(^{(67)}\). ‘Unhealthy’ patterns increased the risk of oesophageal and colorectal cancer\(^{(64,68)}\) and one pattern with high intake of bread and pasta was unfavourable for breast and ovarian cancer risk\(^{(66)}\). Although these results have shown patterns that may increase cancer risk and others that are protective, a difficulty in epidemiological studies of diet and cancer is lack of specific biomarkers for the disease. Further research needs to be carried out to establish environmental factors that may increase cancer risk.

The effect of dietary patterns on a combination of chronic diseases has also been evaluated. In one study, it was reported that after 16 years of follow-up, levels of overweight and obesity increased from 67 to 76\% and 81 to 91\%, respectively, whereas the rates of diabetes
<table>
<thead>
<tr>
<th>Reference</th>
<th>Study type</th>
<th>n</th>
<th>Disease</th>
<th>Cohort</th>
<th>Patterns associated with chronic disease</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(29) L</td>
<td>459</td>
<td>Obesity</td>
<td>Adults (30–80 years)</td>
<td>Meat and potatoes</td>
<td>Red and processed meat, potatoes and fast food</td>
<td>Increase in mean annual change: BMI 0.30 v. 0.05 in healthy pattern (P&lt;0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>White bread</td>
<td>White bread</td>
<td></td>
</tr>
<tr>
<td>(30) CS</td>
<td>825</td>
<td>Obesity</td>
<td>Adults (60–92 years)</td>
<td>Rice</td>
<td>Rice, added fats (mainly cooking oil), beans and poultry</td>
<td>BMI was greater than all other patterns (P&lt;0.05) and WC was greater than a fruit and cereal cluster (P&lt;0.05)</td>
</tr>
<tr>
<td>(25) P</td>
<td>3075</td>
<td>Obesity</td>
<td>Older adults (70–79 years)</td>
<td>Meat, snacks, fat and alcohol</td>
<td>Processed meat, meat, fried poultry, alcohol, high energy drinks, snacks, nuts, salad dressings and miscellaneous fats</td>
<td>Higher total percentage body fat (P&lt;0.05) in comparison with healthy pattern – males only</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Breakfast cereal</td>
<td>Breakfast cereals</td>
<td>Higher total percentage body fat (P&lt;0.05) and higher abdominal visceral fat (P&lt;0.05) in comparison with healthy pattern – males only</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(51) L</td>
<td>737</td>
<td>Obesity</td>
<td>Females (30–89 years)</td>
<td>Empty calorie pattern</td>
<td>Animal and vegetable fats, sweets and desserts, meat and sweetened beverages</td>
<td>17% absolute increased risk in comparison with healthy pattern</td>
</tr>
<tr>
<td>(53) CS</td>
<td>15 890</td>
<td>Obesity</td>
<td>Adults (20–59 years)</td>
<td>Refined foods and sweets</td>
<td>Alcohol, soft drinks, white bread, fast food, sweets and snacks</td>
<td>Both patterns were associated with a 14–17% increased risk of being overweight (P&lt;0.01) and 20% increased risk of being obese (P&lt;0.001) in comparison with a traditional pattern</td>
</tr>
<tr>
<td>(54) CS</td>
<td>659</td>
<td>Obesity</td>
<td>Adults (18+ years)</td>
<td>N/A</td>
<td>N/A</td>
<td>No one pattern was associated with lowering of the risk of obesity, as the levels of BMI and WC were high across all patterns</td>
</tr>
<tr>
<td>(55) P</td>
<td>1423</td>
<td>CVD</td>
<td>Females (18–76 years)</td>
<td>Less heart healthy</td>
<td>High-fat foods</td>
<td>Higher total, LDL cholesterol and total to HDL cholesterol ratio (P&lt;0.05) than heart healthy pattern. In all, 11% of the sample had subclinical heart disease at follow-up, in comparison with 7% of the heart healthy pattern</td>
</tr>
<tr>
<td>(56) L</td>
<td>1423</td>
<td>CVD</td>
<td>Females (18–76 years)</td>
<td>Light eating</td>
<td>Lowest energy content†</td>
<td>Follow up: 11% had carotid atherosclerosis in comparison with 7% of the heart healthy pattern (P&lt;0.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Empty energy</td>
<td>Sweetened beverages, red meat and desserts</td>
<td>Follow up: 18% had carotid atherosclerosis in comparison with 7% of the heart healthy pattern (P&lt;0.05)</td>
</tr>
<tr>
<td>(28) CS</td>
<td>3452</td>
<td>CVD</td>
<td>Adults (25–74 years)</td>
<td>Traditional</td>
<td>Medium fat milk, offal, boiled coffee and potatoes</td>
<td>Significantly higher BMI, WHR and lower serum HDL (P&lt;0.05) in comparison with the healthy pattern</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fast energy</td>
<td>Soft drinks, white bread, fast food, full fat milk, cheese, alcohol, sweets and snacks</td>
<td>Significantly higher BMI, WHR, blood pressure, serum TAG and lower serum HDL (P&lt;0.05) in comparison with the healthy pattern</td>
</tr>
<tr>
<td>Reference</td>
<td>Study type</td>
<td>n*</td>
<td>Disease</td>
<td>Cohort</td>
<td>Labels</td>
<td>Food groups (highest contribution)</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>----</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>(57)</td>
<td>P</td>
<td>4999</td>
<td>CVD</td>
<td>Adults (46–73 years)</td>
<td>Milk fat, Sweets and cakes</td>
<td>Cheese, whole milk, white bread and sweets, sugar, sweets, jam, cakes, biscuits and soft drinks</td>
</tr>
<tr>
<td>(58)</td>
<td>CC</td>
<td>820 v. 2196</td>
<td>CVD</td>
<td>Adults (18–73 years)</td>
<td>Red meat and alcohol</td>
<td>Red meat, fast food and alcohol</td>
</tr>
<tr>
<td>(59)</td>
<td>CS</td>
<td>1313</td>
<td>CVD</td>
<td>Females (50+ years)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>(60)</td>
<td>CS</td>
<td>980</td>
<td>Diabetes</td>
<td>Adults (40–69 years)</td>
<td>White bread</td>
<td>White bread, tomatoes, cheese, dried beans, eggs, meat, fats and oils and beer</td>
</tr>
<tr>
<td>(61)</td>
<td>P</td>
<td>64191</td>
<td>Diabetes</td>
<td>Females (40–70 years)</td>
<td>Cluster 1</td>
<td>Staple foods</td>
</tr>
<tr>
<td>(62)</td>
<td>CS</td>
<td>2875</td>
<td>Diabetes</td>
<td>Adults (age NR)</td>
<td>Soda</td>
<td>Meat, chocolate and miscellaneous sweets</td>
</tr>
<tr>
<td>(63)</td>
<td>CS</td>
<td>20, 210</td>
<td>Diabetes</td>
<td>Adults (45–69 years)</td>
<td>Refined grains and sweets</td>
<td>Refined grains, sweets, beer and soda</td>
</tr>
<tr>
<td>(64)</td>
<td>CC</td>
<td>124 v. 449</td>
<td>Oesophageal and stomach cancer</td>
<td>Gastric cancer</td>
<td>Adults (21+ years)</td>
<td>High meat</td>
</tr>
<tr>
<td>(65)</td>
<td>CC</td>
<td>591 v. 1463</td>
<td>Gastric cancer</td>
<td>Adults (18–93 years)</td>
<td>Pattern II</td>
<td>Low intake of fruits, salads, vegetables, meat, fish and dairy products</td>
</tr>
<tr>
<td>(66)</td>
<td>CC</td>
<td>2569 v. 1041 v. 3413</td>
<td>Breast and ovarian cancer</td>
<td>Females (17–79 years)</td>
<td>G5</td>
<td>Bread, pasta</td>
</tr>
<tr>
<td>(67)</td>
<td>CC</td>
<td>254 v. 184</td>
<td>Lung cancer</td>
<td>Adults (age NR)</td>
<td>Unhealthy</td>
<td>Total fat, saturated fat, animal fat, cholesterol and alcohol</td>
</tr>
<tr>
<td>(68)</td>
<td>CC</td>
<td>465 v. 426 v. 171 v. 309</td>
<td>Colorectal cancer</td>
<td>Adults (30–79 years)</td>
<td>Cluster 2</td>
<td>White bread, pork, processed meat, potatoes, rice and pasta</td>
</tr>
<tr>
<td>(69)</td>
<td>L</td>
<td>1666</td>
<td>Chronic diseases</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>(70)</td>
<td>P</td>
<td>7731</td>
<td>Chronic diseases</td>
<td>Adults (age NR)</td>
<td>Unhealthy</td>
<td>White bread, processed meat, fries and full cream milk</td>
</tr>
</tbody>
</table>
Dietary patterns derived by cluster analysis

605

nearly doubled from 10 to 18% in the total population\(^{69}\).

No significant difference in risk was found according to dietary patterns, as it was reported that chronic disease risk factors were high in all patterns; however, the sample consisted of only males living in one suburban community of the US. In another study, a pattern characterised by the consumption of wholemeal bread, fruits, vegetables, pasta and rice lowered cancer mortality rate and myocardial infarction rates and a pattern characterised by wholemeal bread, fruits, vegetables and polyunsaturated margarine lowered the incidence of obesity\(^{70}\). This provides extra support for the health promoting effects of healthy diets.

Dietary patterns have also been explored in relation to the metabolic syndrome. In one study of Italian non-diabetic adults, the highest prevalence of the metabolic syndrome was found in the ‘starch’ and ‘animal products’ patterns and the lowest prevalence found in a ‘vegetable oil and fat spread’ pattern and a ‘vegetable and fruit’ pattern\(^{71}\). Furthermore, in a Swedish study, it was reported that in males the ‘many foods and drinks’ and the ‘white bread’ pattern and in females the ‘white bread’ pattern only had increased risks of metabolic risk factors\(^{34}\). Song et al. also found increased risks of metabolic risk factors, although this time with a ‘meat and alcohol’ pattern, where it was also reported that a ‘traditional’ pattern that was characterised by high intake of white rice and vegetables had a 23% lower likelihood of having low HDL-cholesterol\(^{72}\). One limitation of these studies is that divergent definitions were used to define the metabolic syndrome prior to analysis.

Few studies have examined the association of dietary patterns with a risk of osteoporosis. In one study an association with bone mineral density was reported, as it was demonstrated that a diet consisting of high intake of fruits, vegetables and breakfast cereals and limited in less nutrient dense foods may contribute to better bone mineral density in both males and females, though this association was not as strong in females, as levels of bone mineral density in both males and females, though this association was not as strong in females, as levels of bone mineral density were fairly equal among all patterns identified\(^{73}\).

Overall strength of these studies includes large sample sizes where a wide variety of clustering variables were also used; nevertheless as with healthy population groups the issue of energy mis-reporting is overlooked, as few authors have excluded these mis-reporters from their analysis. Findings mostly from cross-sectional studies have linked dietary patterns and numerous foods associated with these patterns to chronic diseases; however, further research including targeted nutrition interventions is warranted to fully assess the relationship taking into account all other environmental factors that may influence the disease. As it is well known that the progression of these chronic diseases gradually worsens over time, future studies should also consider the importance of prospective and case–control studies, to help advancements in the area.

### Dietary patterns and associations with nutritional biomarkers

More recently, cluster analysis has been used firstly to derive dietary patterns, and thereafter differences in
nutritional biomarkers explored in an attempt to examine the relationship between the two. It is hoped that this will enhance the knowledge base as to whether these dietary patterns are biologically meaningful.

In addition to the earlier studies on markers of lipid metabolism and inflammation, dietary patterns have been associated with markers of homocysteine (hcy) and vitamin B status. Hcy is an important and well-recognised biomarker in nutritional epidemiology as high levels have been linked to increasing the risk of CVD(79). In a sample of 119 Chinese adults aged 35–49 years, it was found that relative to the ‘fruit and milk’ pattern, those subjects consuming a ‘refined cereals’ pattern were 4 and 5.2 times more likely to have high hcy and low vitamin B12 concentration, respectively(75). Another study investigated the levels of folate and hcy in a sample of 354 American males aged 21–88 years, following the folic acid fortification programme in the US. Within this study it was reported that plasma folate increased in all three dietary patterns identified, although plasma hcy decreased in the low fruit and vegetable pattern only(79). Limitations of these studies include small sample sizes where one study was limited to males only.

A study has also linked dietary patterns to metabolic profiles in a small sample of Irish adults aged 18–63 years. Three dietary patterns were identified, and when compared with metabolic profiles (using metabolomics(77)), it was reported that food groups within patterns could be associated with concentration of metabolites(30). A pattern that had high intake of fruits and vegetables and a pattern that had high intake of red meat were associated with phenylacetylglutamine and O-acetylcarnitine, respectively. Although one major limitation of this study is its small sample size, the findings of this study underline the ability of metabolomics to identify novel biomarkers of dietary intake. Future studies should consider advancing these results in larger studies, in order to strengthen findings.

Reproducibility and validity

Although dietary pattern analysis has become of major interest in the field of nutritional epidemiology, the reproducibility and validity of the patterns derived are not clear, and few studies have fully evaluated this issue. As part of the Framingham Nutrition Studies, dietary patterns were identified for adult males and females aged 18–76 years separately. Five patterns were found to best represent each sex, with some patterns being associated with healthier nutrient profiles, while others were associated with disease risk(17). The internal validity of the five dietary patterns identified for women was assessed and it was found that 80% of the sample was correctly classified when using a discriminant analysis technique to measure the stability of the patterns(48). Furthermore, the authors used the results of this study to derive a statistical scoring system or algorithm that would classify a subject from a newer Framingham Nutrition Study into one of the previously identified patterns for males and females. Using the scoring system it was reported that 80% of new males and females under study were correctly classified into one of the previous patterns already identified(78). The results from this large population based study show that dietary patterns are reproducible across similar population groups, although it should be noted that reproducibility does not guarantee validity. As mentioned previously, cluster analysis can be carried out using different algorithms; however, to date just one study has investigated the differences between these. Lo Siou et al. reported that when the clustering variable was presented as the %TE food method, the k-means approach (in comparison with Ward’s and flexible beta methods) had the highest reproducibility of cluster solutions for Canadian adults aged 35–69 years(20). When the sample was split by sex, a strong relationship was only seen for males; similar results were not found in females, therefore, highlighting the need for further research in the area. One study has also evaluated the influence of the dietary assessment method used (FFQ and 3-d diary), by comparing the classification rate of subjects into the same dietary patterns using either method, where it was found that four out of ten subjects were misclassified(79). Furthermore, the question is raised as to what is the appropriate threshold for acceptable correct classification. As few studies have assessed both reproducibility and validity, it is clear that there is insufficient evidence to make firm conclusions; therefore highlighting the need for further research.

Energy mis-reporting

Energy mis-reporting is a major issue in dietary surveys(22). Research has indicated consistent errors in self-reported dietary intake, using the available dietary assessment methods(80). Dietary intake is commonly over- or under-reported leading to implausible energy intake in population groups, where the latter may be considered the most detrimental to research studies. Under-reporting of dietary intake can happen in three ways, where subjects can (1) deny ever eating the food at all; (2) fail to report the correct portion size consumed or (3) fail to report how many times the food is actually consumed. Approaches to identify under-reporters are to calculate the ratio of energy intake to BMR where cut-off values are applied described by Goldberg et al. or by using the gold standard doubly labelled water technique. In studies of under-reporting, it has been found that females, overweight and obese subjects are more likely to under-report their dietary intake. This is no exception in dietary pattern analysis studies as significant differences have been reported among males and females(37,46) and healthy dietary patterns have been found to contain the greatest proportion of females and overweight subjects(19,37). In contrast, Pryer et al. found that there was no difference in the proportion of under-reporters across the patterns(16), although Martikainen et al. demonstrated that differences in the numbers of under-reporters exist across all patterns; however, these differences are not systematically associated with good or bad diets. Other studies have found that under-reporting of energy intake is not uniformly distributed among dietary patterns(87,88). In one study the highest prevalence of under-reporting fell among those in the healthy
pattern, where although this study measured under-reporting using the doubly labelled water method, the sample consisted of females only aged 18–57 years.*

To the best of our knowledge, no studies have examined the effects of energy mis-reporting by identifying patterns for adequate and under-reporters separately. Two studies have although demonstrated that patterns generated following the removal of under-reporters are relatively similar in comparison with patterns of the total sample (including adequate and under-reporters)**. In one study 70% of participants fall into the same pattern regardless of their reporting status***. The limitations of both these studies are that the authors have only briefly acknowledged under-reporting and there is a lack of published statistical analysis. Similarly, in another study patterns were identified in the total population and adequate reporters, where it was shown that the correlation between energy intake and weight status was improved for females only after removal of under-reporters****. Although it is not clear the effect energy mis-reporting may have on dietary pattern analysis, only two studies have removed such reporters from their analysis in healthy population groups***** and eight studies in chronic disease groups******.

**Conclusion and future work**

From the numerous studies mentioned in this review, some consistent trends emerge when using cluster analysis to derive dietary patterns. It can be argued that there is homogeneity of dietary patterns across populations, where the consistency of patterns identified suggests that they are reproducible. Despite this, given the data driven nature of this statistical technique, the extent to which the identified patterns are reproducible and the extent to which they can be used to develop the understanding of nutritional epidemiology remains debatable. Several important issues have been highlighted, specifically regarding the methodological aspect of cluster analysis and these should be considered in future studies. However, in the earlier studies, different clustering techniques and procedures have been used, making it difficult to draw firm conclusions. Few studies have examined the effect of energy mis-reporting and it is clear that this effect is not fully understood. This review demonstrates the need for large representative cross-sectional and longitudinal studies to assess the effects of energy mis-reporting by carrying dietary pattern analysis on (1) the total population, (2) adequate reporters and (3) under-reporters.

**Acknowledgements**

U.M.D. wrote the review, B.A.McN., A.P.N. and M.J.G. provided expert advice in the drafting of the paper and commented on drafts of the paper. The authors declare no conflict of interest. The work was supported by joint funding from the Irish Department of Agriculture, Fisheries and Food and the Health Research Board under the Food for Health Research Initiative (2007–2012).

**References**


