Public Health Nutrition: 19(2), 191–194 doi:10.1017/S1368980015003730

Editorial

What can we learn from dietary pattern analysis?

Many authors have chosen to investigate diet–disease associations using overall diet rather than looking at individual foods and nutrients. This is usually justified by considering the correlations and synergies between foods and nutrients which are eaten in combination so that it is difficult to identify associations for single items. It may also be that a dietary pattern has a stronger association with the outcome than any of the individual items that contribute to it.

There are two main ways of approaching dietary pattern analysis, *a priori* and *a posteriori*, as has been well explained elsewhere[1]. *A priori* methods typically use scores or indices to assess how well the diet agrees with some predetermined 'ideal' diet and are often referred to as 'measures of diet quality'. *A posteriori* methods use statistical methods to look for patterns within the study population. Given the widespread use of dietary pattern methods it is timely to consider under what circumstances these are helpful.

Application of *a priori* methods across populations

Scores that assess adherence to a good-quality diet would be expected to be associated with health outcomes if the dietary data are valid, but they may not work so well if the dietary intakes in the population under study do not result in a reasonable distribution of dietary pattern scores. For example, *a priori* scores such as the Alternative Healthy Eating Index (AHEI) and versions of the Healthy Eating Index (HEI) from which the former was developed are widely used and are based on meeting dietary recommendations for different countries or age groups. In our Australian data (not published), however, the majority of participants received the top score for the *trans*-fatty acid intake component of the AHEI because intakes in Australia are much lower than in the USA where the AHEI was developed. Similarly Gerber[2] modified the Diet Quality Index developed for the US population for use in a French Mediterranean population as there was no gradient of consumption with increasing scores for 'total fat' and 'maintain protein intake at moderate levels'.

The original Mediterranean Diet Score (MDS) avoids this problem as scoring is based on median intakes in the study population. But if the intakes for MDS components are low overall, the intakes of even the highest-scoring individuals may not reach the levels in a traditional Mediterranean diet and thus may not show the expected associations. In a recent meta-analysis on Mediterranean diet and diabetes incidence[3], differences between European and US studies were attributed by the authors to differences in confounders adjusted for, but they may simply reflect that even the most Mediterranean diet in the USA is not very high in items considered characteristic of the Mediterranean diet.

Among the many different methods to score adherence to a Mediterranean diet[4], some score according to dietary guidelines independent of the population in which they are being applied rather than using population-specific values[2,5,6]. In 2014 Sofi *et al.*[7] published a meta-analysis of the Trichopoulou MDS in relation to morbidity and mortality, updating their previous such analyses in 2008 and 2010. In their 2014 analysis the cut-offs for determining adherence were shown to vary widely between studies. From the data collected the authors developed a short literature-based tool for scoring adherence to the Mediterranean diet using nine components – fruit, vegetables, legumes, cereals, fish, meat and meat products, dairy products, alcohol and olive oil – which are assessed according to three levels of intake, with cut-offs based on portions per day from the literature reviewed. While this would not overcome differences in the dietary instruments used to collect the data, it would standardize the scoring across populations.

The Dietary Inflammatory Index is also an *a priori* score but unlike most other diet quality scores it is not based on agreement with an 'ideal' diet, but rather on intake of nutrients and foods that have been reported to be associated with circulating levels of inflammatory biomarkers, in particular C-reactive protein[8]. However, because it can be estimated when fewer than the complete list of forty-five possible items on which it was based are available, two similar scores would not necessarily be comparable depending on which components were taken into account.

Despite differences in the way various 'diet quality scores' have been derived, they all tend to promote the intake of fruits, vegetables and whole grains. Reedy *et al.*[9] reported that four different diet quality scores examined were similarly associated with mortality, most likely due to common characteristics including plenty of fruit, vegetables, whole grains and plant-based proteins, while small differences in how other components were scored could contribute to differences in the strength of associations with different outcomes. This increases confidence in the findings.

Application of *a posteriori* dietary patterns in different populations

Data-driven patterns reflect the intakes reported by the study population but do not always identify patterns that

can be easily identified as 'healthy' and 'unhealthy' and may not be associated with outcomes. In this issue of *Public Health Nutrition* we have six papers that use a data-driven or an *a posteriori* approach to consider dietary patterns in a broad range of populations. Four used factor or principal components analysis (PCA), one of them comparing it with reduced rank regression (RRR); one used a new method, treelet transform, which was compared with PCA; and the sixth was a methodological study evaluating a new way of performing cluster analysis.

Using data from three 24 h recalls collected for the China Health and Nutrition Survey, dietary patterns were derived using both PCA and RRR, and their association with diabetes prevalence assessed[10]. The idea behind using two different methods to define dietary patterns was that PCA would identify patterns existing in the study population but possibly not associated with the outcome, while RRR would identify a pattern associated with relevant biological risk markers, namely glycated Hb (HbA1c), homeostatic model assessment insulin resistance index (HOMA-IR) and fasting glucose, but which may not exist in reality. The authors found that the PCA-derived 'modern high-wheat' pattern was positively associated with diabetes while the 'traditional southern' pattern showed an inverse association; both associations were significant in unadjusted models but not after adjustment. The RRR-identified pattern combined elements of the two PCA-derived patterns (higher intakes of the modern wheat-based items and lower intakes of traditional southern items such as rice and seafood) and was significantly associated with diabetes, even after adjustment. The authors provide an interesting discussion of how food items in the RRR pattern all tended to be associated with the outcomes in the same way, while food items in the PCA patterns were not all associated in the same way as their inclusion was based on behavioural patterns rather than nutritional attributes of the food.

PCA was also used to identify dietary patterns for Iranian adults using food frequency data[11]. The authors report on the associations of the four dietary patterns identified ('fast food', 'traditional', 'lacto-vegetarian' and 'western') with depression, anxiety and psychological distress. A useful aspect of the report is the presentation of actual amounts of different foods and nutrients consumed according to quintiles of dietary pattern scores. Further, the discussion of this paper highlights one of the challenges in naming dietary patterns and interpreting them based on their names; a 'traditional' Iranian pattern differs substantially from a similarly named 'traditional' Australian pattern, reflecting quite different foods and showing different associations with the outcomes.

In a study from rural Bangladesh, PCA-derived dietary patterns were investigated in relation to carotid intima-media thickness (cIMT)[12]. Three different patterns were identified: 'balanced', 'animal protein' and 'gourd/root vegetable'. The study found that the 'gourd/root vegetable' pattern was positively associated with cIMT while the 'balanced' diet showed an inverse association. In this study, associations of the major food groups (meat, poultry, fish, fruit and vegetables) with cIMT were also evaluated and no associations were found, suggesting that the use of dietary patterns including different combinations of food groups was relevant for identifying associations.

Also in this issue is a systematic review and meta-analysis of dietary patterns defined by factor or principal components analysis and incidence of type 2 diabetes mellitus (T2DM). In their summary of ten cohort studies, the authors concluded that 'healthy' patterns including vegetables, fruits, whole grains and seeds were inversely associated with T2DM while 'unhealthy' patterns including red and processed meat, processed foods, high-fat dairy and refined grains' were positively associated with T2DM[13]. Tests for heterogeneity indicated that the results were not consistent over all the patterns that had been classified as 'unhealthy'. Further examination of the data suggested that when the so-called 'unhealthy' patterns also included high loadings for plant foods that may be rich sources of phytochemicals, the association with T2DM was not evident. In general these findings are consistent with a recent review of the MDS and diabetes incidence[3].

Assi *et al.*[14], using data from over 300 000 women in the European Prospective Investigation into Cancer and Nutrition (EPIC), report on the use of a new method of identifying patterns in data, known as treelet transform (TT). TT is a dimension reduction method that combines features of both PCA and cluster analysis to produce a cluster tree that allows a visual examination of the way the different variables group. Interpretation of PCA results can be quite challenging as each factor derived from PCA involves all of the original variables; in contrast, each TT factor involves a smaller number of naturally grouped variables. In this study, rather than using food groups, nutrient densities were used. While there are advantages to looking at food groups so that results can be directly interpreted, in this situation where dietary data were combined across countries with different food resources and cuisines, nutrients common across countries were easier to use. TT identified two main patterns: the first was rich in nutrients from animal foods, loading on cholesterol, protein, retinol, vitamins B_{12} and D, while the second loaded on β-carotene, riboflavin, thiamin, vitamins C and B_6, fibre, Fe, Ca, K, Mg, P and folate from fruit, vegetables and cereals. The second pattern was found to be associated with reduced risk of breast cancer. A secondary analysis using PCA reached a similar conclusion, with both methods explaining a similar amount of variation. As with other clustering techniques, TT users need to subjectively select a suitable cut-level for the cluster tree. However, cross-validation techniques can be used to identify the optimal cut-level. It will be interesting to see whether TT is adopted as eagerly as has been PCA in nutritional

epidemiology and if so whether it provides any new knowledge or understanding of associations between diet and disease.

The last paper in this issue to be considered is a methodological study evaluating different methods of performing cluster analysis[15]. Cluster analysis groups people according to the degree of similarity in their diets. The two commonly used methods of performing cluster analysis, *k*-means and Ward's method, were compared with a new method based on Gaussian mixed models (GMM), first in a simulation study and second using data from children in the IDEFICS (Identification and Prevention of Dietary- and Lifestyle-Induced Health Effects in Children and Infants) Study. In simulated data, the GMM method performed better than the other two methods, and *k*-means performed better than Ward's method. Among the IDEFICS dietary data each method identified three reasonably consistent clusters, based on relative consumption. These were labelled 'non-processed' (higher-than-average intakes of fruit, vegetables and wholemeal bread, and lower-than-average intakes of refined cereals, sweet drinks and fast food); 'balanced' (no foods particularly standing out but slightly higher-than-average intakes of sauces, butter, sweet drinks, meat and refined cereals, and slightly lower intakes of breakfast cereals, dairy products and fruit); and 'junk food' (higher intakes of fast food, breakfast cereals, meat alternatives and dairy products, and lower-than-average intakes of wholemeal bread, fruit and vegetables). For each method, the prevalence of overweight/obesity was lower in the 'non-processed' cluster compared with the 'junk food' cluster, which is an indication of validity. Although the GMM outperformed the other two cluster methods in simulated data, the authors noted that if there was a habitual non-consumption of foods, only models with strong geometric restrictions on the clusters should be fitted, reducing the flexibility of this method. Thus, as well as recommending the geometrically restricted GMM, they also suggested that the *k*-means approach could be used as it often gave similar results and was more easily applicable.

Overall, while these papers confirm the popularity of dietary pattern analysis in nutritional epidemiology, they also highlight the difficulties in summarizing studies where the patterns are not actually the same, despite similar names. 'Traditional' as a pattern description is likely to reflect different things across countries depending on what foods were traditionally available. This underscores the need to show data on what is actually consumed by participants at extremes of the score ranges so that patterns can be properly understood. Such data also help with translation to public health messages. Methods such as PCA, TT, factor analysis and cluster analysis may work to identify patterns in the population but the patterns do not always fall neatly into 'healthy' or 'unhealthy' and may not be related to the outcome. On the other hand, PCA-derived patterns tended to show stronger associations with the outcomes than individual food groups.

In conclusion, the use of dietary patterns in nutritional epidemiology appears to be here to stay, but the limitations must be kept in mind and perhaps the use of more than one approach should be considered. While there is no right or wrong method, whatever is used should be appropriate for the research question being studied and the results interpreted appropriately. It is also important to keep in mind that creating dietary patterns does not overcome inherent weaknesses in dietary data.

Allison Hodge
Deputy Editor
Email: allison.hodge@cancervic.org.au

Julie Bassett
Cancer Epidemiology Centre, Cancer Council Victoria
Melbourne, Victoria, Australia

References

1. Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, 177–203.
2. Gerber M (2006) Qualitative methods to evaluate Mediterranean diet in adults. *Public Health Nutr* **9**, 147–151.
3. Schwingshackl L, Missbach B, Konig J *et al.* (2015) Adherence to a Mediterranean diet and risk of diabetes: a systematic review and meta-analysis. *Public Health Nutr* **18**, 1292–1299.
4. 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.
5. Lapointe A, Goulet J, Couillard C *et al.* (2005) A nutritional intervention promoting the Mediterranean food pattern is associated with a decrease in circulating oxidized LDL particles in healthy women from the Quebec City metropolitan area. *J Nutr* **135**, 410–415.
6. Rumawas ME, Dwyer JT, McKeown NM *et al.* (2009) The development of the Mediterranean-style dietary pattern score and its application to the American diet in the Framingham Offspring Cohort. *J Nutr* **139**, 1150–1156.
7. Sofi F, Macchi C, Abbate R *et al.* (2014) Mediterranean diet and health status: an updated meta-analysis and a proposal for a literature-based adherence score. *Public Health Nutr* **17**, 2769–2782.
8. Shivappa N, Steck SE, Hurley TG *et al.* (2014) Designing and developing a literature-derived, population-based dietary inflammatory index. *Public Health Nutr* **17**, 1689–1696.
9. Reedy J, Krebs-Smith SM, Miller PE *et al.* (2014) Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *J Nutr* **144**, 881–889.
10. Batis C, Mendez MA, Gordon-Larsen P *et al.* (2016) Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults. *Public Health Nutr* **19**, 195–203.
11. Hosseinzadeh M, Vafa M, Esmaillzadeh A *et al.* (2016) Empirically derived dietary patterns in relation to psychological disorders. *Public Health Nutr* **19**, 204–217.

12. McClintock TR, Parvez F, Wu F *et al*. (2016) Major dietary patterns and carotid intima-media thickness in Bangladesh. *Public Health Nutr* **19**, 218–229.
13. Maghsoudi Z, Ghiasvand R & Salehi-Abargouei A (2016) Empirically derived dietary patterns and incident type 2 diabetes mellitus: a systematic review and meta-analysis on prospective observational studies. *Public Health Nutr* **19**, 230–241.
14. Assi N, Moskal A, Slimani N *et al*. (2016) A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Public Health Nutr* **19**, 242–254.
15. Greve B, Pigeot I, Huybrechts I *et al*. (2016) A comparison of heuristic and model-based clustering methods for dietary pattern analysis. *Public Health Nutr* **19**, 255–264.