2011

Invited Commentary: Dietary Pattern Analysis

Fumiaki Imamura
Harvard University, fimamura@hsph.harvard.edu

Paul F. Jaques

American Journal of Epidemiology

Vol. 173, No. 10
DOI: 10.1093/aje/kwr063
Advance Access publication:
April 7, 2011

Invited Commentary

Invited Commentary: Dietary Pattern Analysis

Fumiaki Imamura* and Paul F. Jacques

* Correspondence to Dr. Fumiaki Imamura, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: fimamura@hsph.harvard.edu).

Initially submitted November 2, 2010; accepted for publication January 11, 2011.

The analytic approaches used in nutritional epidemiology for dietary pattern analyses share common characteristics with those of genetic epidemiology. In this issue of the *Journal*, Gorst-Rasmussen et al. (*Am J Epidemiol.* 2011;173(10):1097–1104) discuss one such approach. Application of methods used in genetic pattern analyses to nutritional epidemiology could prove valuable but raises important issues that need to be considered because dietary and genetic studies often address different types of questions in analyzing interrelated variables. These different aims require statistical methods that assume different characteristics of the underlying patterns. The authors briefly describe such differences to facilitate interpretation and applications of previous and future pattern studies.

diet; myocardial infarction; statistics

Abbreviations: PCA, principal component analysis; TT, treelet transform.

Pattern analyses have been of great interest in diverse scientific fields. The main goal of pattern analysis is to create a meaningful overall representation of a large and complex set of interrelated factors. For these reasons, dietary patterns have received much attention in the field of nutritional epidemiology as a means of characterizing dietary intake and relating intake to health outcomes. This concept of the whole diet being more predictive of health outcomes than individual foods or nutrients, which are traditionally the focus of study (1, 2), is supported by results from clinical trials of dietary interventions, which have shown substantive impacts of changes in diet in its entirety on primary and secondary prevention of diseases (3–5). Advancements in dietary assessment methods and statistical software have also contributed to the current trend.

The number and scope of dietary pattern studies have grown rapidly as researchers have applied relatively parsimonious methods to different populations with different sociodemographic backgrounds for consideration of different health-related outcomes. Statistical approaches that use parsimonious methods are well documented; however, many relevant underlying statistical and epidemiologic issues have been recognized in the field of nutritional epidemiology but remain understudied.

Gorst-Rasmussen et al. (6) elegantly introduced a novel statistical method called treelet transform (TT) in this issue of the *Journal*. In addition to introducing this technique to the field of nutritional epidemiology, the article by Gorst-Rasmussen et al. provides the opportunity to identify and discuss a few epidemiologic issues related to dietary pattern methods and their application. Below, we describe some of the many important differences between the TT and pattern methods typically applied in diet-pattern studies. We further discuss 3 topics related to the application of these methods: the aim of the pattern analysis, the use of patterns to control confounding, and the validation of patterns.

APPLICATION OF PATTERN ANALYSES TO DIETARY DATA

The original article by Lee et al. (7) that described TT introduced the method for analyses that assumed that intrinsic patterns are present as distinct clusters of selected individual factors, assuming and allowing no contributions of other factors to the patterns. This sparsity assumption may hold in certain genetic studies, as Lee et al. illustrated. For example, acute myeloid leukemia can be caused by a handful of gene mutations that are often referred to as a

''gene signature,'' making millions of other genes irrelevant (7, 8). Bullinger et al. (8) analyzed microarray gene-expression profiles from leukemia patients and excluded roughly 20,000 genes (70%) because they were screened as less informative. Gorst-Rasmussen et al. (6) and Lee et al. referred to this as ''noise.'' The exclusion is reasonable because many genes may be biologically irrelevant to a specific pathologic pathway and the sparsity assumption is likely to hold. Furthermore, the genetic study aimed to identify a clinically important set of gene-expression profiles to predict leukemia prognosis and to show the pathological importance of major selected genes. Importantly, similar aims are very common in genetic studies (9, 10), but not in nutritional epidemiology studies, as discussed later.

The sparsity assumption should be justified when applying TT to any data set or epidemiologic question. Attributes of a data set, aims, and assumptions can affect the method, for example, when using pattern analyses of biologic markers of metabolic syndrome, including glucose homeostasis, lipids, and blood pressures (11). Generally, correlation between systolic and diastolic blood pressures is so high that a pattern analysis assuming sparsity might mistakenly consider factors with relatively low pairwise correlations as noise and may limit detection of potentially important patterns (11).

In dietary pattern studies, investigators typically assume no sparsity and aggregate dietary variables to capture overall diet, including foods that minimally influence patterns. Empiric motivations or underlying assumptions are that both foods consumed and not consumed are interrelated informatively, and that the cumulative role of all foods is important in the biologic influence of diet and for public health messages (1, 2) This idea is in agreement with a principal goal of most diet modifications, as dietary intervention trials typically examine the overall impact of a favorable diet in place of a prior average diet in its entirety. Finally, inference based on individual foods from dietary pattern analyses, particularly in an exploratory analysis, is typically a secondary goal.

In considering whether or not we should assume sparsity, we recognize one reasonable exception. When examining whether a certain pattern identified in one cohort is present in an independent cohort (confirmatory analysis), the analyses often entail the sparsity assumption because of analytic parsimony and different properties of diet (12).

The question regarding the necessity of the sparsity assumption raises the following question: Are foods identified as noise truly not relevant to dietary patterns? For example, the article by Gorst-Rasmussen et al. (6) suggested that high-fat and low-fat dairy products and many beverages, including soft drinks, did not contribute to any dietary patterns, just like the 20,000 excluded genes in the analysis of Bullinger et al. (8). This is inconsistent with dietary patterns previously identified in another Danish cohort (13), in which sugar-sweetened beverages were identified as contributing factors to a dietary pattern. The exclusion under the sparsity assumption could be valid if those items were consumed independently of the ones included in dietary patterns or were consumed without any variation in the population. The exclusion could be problematic if those items were correlated in a meaningful manner or if a few foods appeared too strongly correlated, which could cause false identification of the sparsity.

The article by Gorst-Rasmussen et al. (6) allows us the opportunity to consider the different aims and assumptions underlying genetic and dietary pattern analyses. We entirely agree on the utility of TT, but comparison among available methods and TT remains insufficient and requires clarification of the aims and assumptions of pattern analyses.

DIVERSE AIMS OF PATTERN ANALYSES

The original aim of dietary pattern analysis was to capture overall diet. However, several studies were conducted to identify a subset of dietary factors that could aid in disease prediction. In the INTERHEART Study, Iqbal et al. (14) selected dietary factors by using principal component analysis (PCA) to classify myocardial infarction cases and controls. Schulze et al. (15) also identified a simple set of dietary factors based on associations with disease biomarkers. Findings from those studies were externally validated for disease prediction in limited populations (12, 14, 15). Those studies identified key foods in diet-disease associations rather than dietary patterns that represented overall diet. Their aims and approaches were similar to those of genetic studies and were clearly distinct from the majority of dietary pattern studies.

Analogously, genetic studies not only were conducted to identify clinically important genes but also to characterize overall genetic structure. Menozzi et al. (16) reported the utility of PCA in capturing genetic variability in European populations in 1978. More recently, Novembre et al. (17) and others (18) similarly demonstrated that simple scales from PCA based on microarray polymorphism data could represent demographic variation in Europe. Notably, but not surprisingly, the studies did not expect any subset of genes to play roles in determining the genetic structures of populations. Therefore, dietary and genetic pattern analyses can share the same aims. The examples suggest that investigators must clarify the aims and underlying assumptions of the analyses, align the methods with these aims and assumptions, and carefully consider inference from results of pattern analyses.

PATTERNS TO ADDRESS CONFOUNDING

Pattern analyses can be applied to adjustment for potential confounding, which is often called population admixture or population stratification in the field of genetics (18, 19). Price et al. (20) demonstrated the utility of PCA in controlling for population admixture in genetic epidemiology. Our previous study (21) similarly demonstrated that dietary patterns can control for dietary confounding. This is important because the cumulative influence of minor or ''noisy'' variations or ''net confounding'' could be large (20–22). Therefore, adjustment for net confounding is another example of pattern analyses that do not need the sparsity assumption or clear interpretation of identified patterns.

TT and similar approaches might be inappropriate methods to adjust for net confounding, but they still could

be helpful for visualizing patterns to identify the structure of dietary correlates in a given population. For example, Gorst-Rasmussen et al. (6) showed that different subsets of food consumption were correlated with red meat and processed meat in their population; margarine and potatoes were correlated with processed meat consumption, whereas vegetables and beans/legumes were correlated with red meat consumption. The observations indicated potentially different dietary confounders for processed meat and red meat consumptions, even though net confounding may be the same.

VALIDITY OF PATTERN ANALYSES

Dietary pattern studies and genetic pattern studies confront typical epidemiologic issues, including validity and reproducibility. Gorst-Rasmussen et al. (6) used unique iterative approaches to confirm the reproducibility of TT-derived dietary patterns in their study population. Here, we briefly describe the validation of dietary patterns, which can further improve the quality of evidence.

In dietary pattern studies, a common scheme is to derive dietary patterns and to examine the association between the patterns and health-related outcomes. The 2 steps suggest 2 types of validity: validity of identified dietary patterns compared with true dietary patterns and validity of strength of associations between identified patterns and health-related outcomes. The distinction of these 2 steps is important when assessing the validity of dietary pattern studies. This idea is commonly incorporated in genetic studies (8, 10), but seldom in dietary studies.

Assessment of validity in relation to true dietary patterns is challenging. Because pattern analyses may capture only limited portions of overall diet, for example $<40\%$ of the total variation of Danish diet in the article by Gorst-Rasmussen et al. (6), we cannot clearly determine whether or not the analytic techniques can capture true dietary patterns. However, the validity of the observed dietary patterns can be indirectly inferred as follows. If the 2 distinct analytic or data-reduction methods captured a similar dietary pattern (23–25), we can expect higher likelihood that the identified dietary pattern is close to an intrinsic true pattern than we could if the pattern was found using only a single method. Therefore, researchers conducting dietary pattern studies should be encouraged to use multiple analytic methods to assess the validity of dietary patterns, with consideration of the methodological similarity or independence between multiple methods (25). Hence, TT is a promising option for inclusion in future studies. Notably, genetic studies often use multiple methods and rigorous approaches to demonstrate the capability of identifying a genetic signature independent of methodological options (8, 10).

Testing the external validity of dietary patterns in relation to true dietary patterns or diet-disease association is more challenging than testing internal validity. Across different populations, data collection and preparation techniques should be standardized (12). Discrepancies between dietary data may be partially reconciled by imputation, as genome-wide studies often undertake imputation based on underlying biology (18). In determining the external validity of dietary patterns, one must also consider the temporal nature of diet. For example, dietary patterns identified decades ago would require verification of current use to be useful in future public health studies (26). This is a much more critical issue in nutritional epidemiology than in genetic epidemiology, because lifestyle has dramatically changed over the decades, whereas genomic patterns of populations cannot change unless there is substantial migration over time. Nevertheless, because of the recent population transitions, temporal considerations of dietary and genetic patterns may prove interesting.

For internal and external validation of the diet-disease association in dietary pattern studies, alternative approaches include cross-validation analyses, as are often elaborated in genetic studies (10), and evaluation of physiologic mediators for disease outcomes to support causal inference (27). In validation, use of prespecified or simplified dietary patterns may work well, as demonstrated in the INTERHEART Study (14). Findings from the validation should be carefully interpreted as irrelevant to underlying dietary patterns in an independent cohort and as necessitating the sparsity assumption. The approach misses the potential cumulative roles of all consumed foods.

During the past decades, we have observed tremendous advances in the fields of nutritional and genetic epidemiology. Pattern analyses will continue to contribute to both fields. Conceptual consolidation of different pattern analyses helps us rethink the epidemiologic characteristics and the importance of the aims and assumptions of such analyses. Pattern analyses from other fields, such as social science, and other epidemiologic concepts, such as multiple testing, will also be interesting topics of exploration in relation to dietary and genetic pattern studies. With increasing use of pattern analyses in epidemiology and public health and the application of new methods for pattern analyses, we should all remain vigilant to possible misuse and overuse of these parsimonious methods.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Fumiaki Imamura); and Nutritional Epidemiology Program, Jean Mayer USDA Human Nutrition Research Center on Aging, Tufts University, Boston, Massachusetts (Paul F. Jacques).

Conflict of interest: none declared.

REFERENCES

1. Slattery ML. Defining dietary consumption: is the sum greater than its parts? *Am J Clin Nutr.* 2008;88(1):14–15.
2. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol.* 2002;13(1):3–9.
3. de Lorgeril M, Salen P, Martin JL, et al. Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction: final report of the Lyon Diet Heart Study. *Circulation.* 1999;99(6):779–785.

4. Appel LJ, Moore TJ, Obarzanek E, et al. A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *N Engl J Med.* 1997;336(16):1117–1124.
5. Zarraga IG, Schwarz ER. Impact of dietary patterns and interventions on cardiovascular health. *Circulation*. 2006; 114(9):961–973.
6. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, et al. Exploring dietary patterns by using the treelet transform. *Am J Epidemiol*. 2011;173(10):1097–1104.
7. Lee AB, Nadler B, Wasserman L. Treelets—an adaptive multi-scale basis for sparse unordered data. *Ann Appl Stat*. 2008; 2:435–471.
8. Bullinger L, Döhner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1605–1616.
9. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531–537.
10. Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*. 2003;95(1):14–18.
11. Lawlor DA, Ebrahim S, May M, et al. (Mis)use of factor analysis in the study of insulin resistance syndrome. *Am J Epidemiol*. 2004;159(11):1013–1018.
12. Imamura F, Lichtenstein AH, Dallal GE, et al. Generalizability of dietary patterns associated with incidence of type 2 diabetes mellitus. *Am J Clin Nutr.* 2009;90(4):1075–1083.
13. Togo P, Heitmann BL, Sørensen TI, et al. Consistency of food intake factors by different dietary assessment methods and population groups. *Br J Nutr.* 2003;90(3):667–678.
14. Iqbal R, Anand S, Ounpuu S, et al. Dietary patterns and the risk of acute myocardial infarction in 52 countries: results of the INTERHEART study. INTERHEART Study Investigators. *Circulation*. 2008;118(19):1929–1937.
15. Schulze MB, Hoffmann K, Manson JE, et al. Dietary pattern, inflammation, and incidence of type 2 diabetes in women. *Am J Clin Nutr.* 2005;82(3):675–684.
16. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science*. 1978; 201(4358):786–792.
17. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature*. 2008;456(7218):98–101.
18. Rosenberg NA, Huang L, Jewett EM, et al. Genome-wide association studies in diverse populations. *Nat Rev Genet.* 2010;11(5):356–366.
19. Khoury MJ, Beaty TH. Applications of the case-control method in genetic epidemiology. *Epidemiol Rev.* 1994;16(1): 134–150.
20. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904–909.
21. Imamura F, Lichtenstein AH, Dallal GE, et al. Confounding by dietary patterns of the inverse association between alcohol consumption and type 2 diabetes risk. *Am J Epidemiol*. 2009;170(1):37–45.
22. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167(5):523–529.
23. Newby PK, Muller D, Tucker KL. Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin Nutr.* 2004;80(3):759–767.
24. Hu FB, Rimm E, Smith-Warner SA, et al. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am J Clin Nutr.* 1999;69(2):243–249.
25. Kaaks R, Riboli E, Estève J, et al. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. *Stat Med*. 1994;13(2):127–142.
26. Newby PK, Weismayer C, Akesson A, et al. Long-term stability of food patterns identified by use of factor analysis among Swedish women. *J Nutr.* 2006;136(3):626–633.
27. Hafeman DM, Schwartz S. Opening the black box: a motivation for the assessment of mediation. *Int J Epidemiol*. 2009;38(3):838–845.