Examining Dietary Patterns in Relation to Chronic Disease:
Matching Measures and Methods to Questions of Interest

Running title: Krebs-Smith et al.; Examining dietary patterns

Susan M. Krebs-Smith, PhD, MPH; Amy F. Subar, PhD, MPH, RD; Jill Reedy, PhD

Epidemiology and Genomics Research Program, National Cancer Institute, Bethesda, MD

Address for Correspondence:
Amy F. Subar, PhD, MPH, RD
National Cancer Institute
Division of Cancer Control and Population Sciences
9609 Medical Center Drive, Room 4E140
Bethesda, MD 20814-9692
Tel: 240-276-6695
Fax: 240-276-7920
E-mail: subara@mail.nih.gov

Journal Subject Code: Etiology:[8] Epidemiology

Key words: Editorial, diet, epidemiology, health outcomes
In this issue of *Circulation*, Shikany and colleagues add to the growing body of literature on dietary patterns and their relation to health outcomes.\(^1\) The purpose of their paper was to test the hypothesis that dietary patterns extracted through factor analysis are associated with risk of acute coronary heart disease (CHD) in the Reasons for Geographic and Racial Differences in Stroke Study. The authors note that studies examining associations between dietary patterns and CHD in samples with sociodemographic and regional diversity are lacking. Understanding the drivers of health disparities is critically important, and the topic of dietary patterns is of increasing interest, as noted by the most recent Dietary Guidelines Advisory Committee Report, “…because the totality of diet…may have synergistic and cumulative effects on health and disease.”\(^2\)

Shikany and colleagues state that the use of “rigorous” factor analysis was a major strength of their study because it is a “data-driven” method through which patterns are “empirically derived.” They conclude that “a dietary pattern characteristic of the southern US” was associated with greater hazard of CHD whereas a “plant-based pattern” was not. We believe these assertions deserve further consideration.

Factor analysis is one of several ways of defining dietary patterns, each of which addresses a particular question (Table 1). Methods for studying patterns can be grouped into three broad categories: data-driven, outcome independent; data-driven, outcome dependent; and investigator-defined. Note that the labels are simply a shorthand way to refer to how the patterns are derived. “Data-driven” does not mean a method is more evidence-based, and “investigator-defined” does not mean a method includes more subjectivity. Each method is built on some type of evidence and includes some degree of subjectivity.\(^3\)

Factor analysis answers the question, “What elements of the diet (here, author-defined food groups) correlate (i.e., what are the major underlying factors) in accounting for the variation
in food intakes reported by study participants?” It is data-driven because the patterns emerge \textit{a posteriori} from an analysis of dietary data and is outcome-independent because the patterns are derived independent of their potential relationship to a health outcome. It works by examining correlations between food group variables and finding linear combinations of those variables that explain overall variance in the data; these linear combinations are called “factors.” Individuals are characterized by the degree to which their diets conform to each of the identified factors, using continuously scaled scores. Following a common convention, Shikany and colleagues refer to these factors as “patterns.” However, note that any one factor does not represent the entire pattern of eating for any individual or group because the factors are not mutually exclusive, although each person’s pattern of eating can be inferred by assessing his/her multiple factor scores.

Although factor analysis and cluster analysis (the other data-driven, outcome independent method) are frequently used to study patterns in relation to health outcomes, neither is expressly designed to derive dietary patterns that are predictors of disease. Therefore, resulting factors may be significantly associated with a health outcome, but the interpretation is not that such factors are the best or worst possible with regard to that outcome. In the paper by Shikany and colleagues, among all the factors which explain variability in dietary intakes in their sample, the “Southern” pattern (as labeled by the authors) was the only one significantly associated with increased CHD risk. However, that does not mean that this factor necessarily represents the dietary pattern most associated with disease. To answer that question, data-driven, outcome dependent methods have been developed. However, because these methods are relatively new, fewer studies have used them.

Reduced rank regression (RRR) is used expressly to examine the relationship between
dietary patterns and health outcomes. Specifically, it addresses the question, “What combination of foods explains the most variation in a set of response variables?” It involves the investigator specifying disease-specific response variables and then analyzing the data to determine the combinations of food intake that explain as much variation in those variables as possible. These response variables are based on a priori hypotheses and might be, for example, a biomarker with a known relationship to the disease of interest. The RRR finds the factors that explain variation in those response variables; the pattern(s) that emerge can then be validated with final outcomes.

Classification and regression tree analysis (CART) is also an emerging method that may prove useful to examine associations between dietary patterns and some other factor such as diet quality or a health outcome. It could be used to answer the question “What key indicators of intake explain the most variation in a health outcome?” CART results in a decision tree progression of indicators depending on the value at each branch. For example, one might find that in predicting a health outcome the most important factor is the amount of added sugars consumed. If intake of added sugars is high, the next most important factor in predicting outcome might be the amount of, say, solid fats, but if added sugars are low, the next most important factor might be fruit intake. There has been only preliminary work done in this area, but the method holds promise.

Investigator-defined methods include selective diets as well as scores and indices. They are called investigator-driven because the dietary characteristics are specified by the researcher a priori. Shikany and colleagues seem to assert that this is a limitation, but specifying the criteria for defining patterns in advance has the key advantage of allowing other researchers to define their criteria exactly the same way, paving the way for comparability between studies. In index analysis, the dimensions of the pattern, and how those dimensions are scaled, are specified (and
thus standardized) by the researcher based on external evidence regarding what constitutes a
healthy diet and not by “authors’ opinions on what defines a health or other dietary pattern.”
Common examples in the literature include the Diet Quality Index, 5 Healthy Eating Index, 6
Alternate Healthy Eating Index, 7 Recommended Food Score, 8 DASH Index, 9 and
Mediterranean Diet Score. 10 Using these indices answers the question “How close is the
population to meeting a certain set of dietary recommendations?” For this reason, they are
sometimes called measures of “diet quality.”

As with all methods, factor analysis has some limitations and challenges. As noted by
Shikany and colleagues, subjectivity is introduced when constructing food groups, determining
the number of factors, and naming the factors (often called patterns). Unless methods of
collecting the data are comparable and food group construction is standardized, results are not
comparable across studies. Factor names are determined by the researcher and, given that they
tend to be short, often fail to adequately convey what the underlying factor is. For example,
Shikany and colleagues labeled one factor as “Plant-based” which exhibited sizeable factor
loadings for low-fat dairy and milk, fish, poultry and water, in addition to various plant foods.

They conclude that, despite “the high loading of putative ‘heart healthy’ foods … a Plant-based
pattern was not associated with hazard of CHD in this sample.” Given the various types of
subjectivity just mentioned, and that factors are not mutually exclusive, this should not be
interpreted to mean that persons in this population who eat a high quality plant-based diet
(quality which could be evaluated systematically and in a standardized way using an index) have
no better risk profile than persons who do not. For example, when factor, cluster, and index
analyses were compared to investigate colorectal cancer incidence in the NIH-AARP Diet and
Health Study, some parallel findings were found for men between factor and index analyses, but
not for women. Specifically, for men, as might be expected, the Healthy Eating Index, which measures diet quality as defined by US dietary guidelines (based on a 0-100 score), found a 28% reduced risk for those in the highest versus lowest quintile, and the so-called “fruit and vegetable factor,” which had high factor loadings for foods rated highly on the HEI, found a 19% reduced risk for those with the highest versus lowest factor scores. In contrast, for women, the so-called “fruit and vegetable factor,” was not protective; however, the Healthy Eating Index found a 20% reduced risk for those in the highest versus lowest quintile.

Another challenge is that the population under study can dramatically affect results. In any study, we are interested in whether the results can be generalized to the population at large. But data-driven dietary pattern analyses have a special caveat related to this: if the patterns are derived by explaining the variability among diets of one population, it is unlikely the exact same patterns would be found in another population, so there is a lack of generalizability of the results and reproducibility is necessarily limited. With investigator-driven approaches, pattern identification (within any one index or score) is standardized a priori and so allows for truer comparisons across cohorts. However, many cohort studies, each with different populations, have used different indices or scores to examine the same question, making it difficult to sort out the findings. To address this concern, the Dietary Patterns Methods Project was initiated. It provided a systematic comparison of 4 key diet quality indices across multiple cohorts, using standardized measures, to examine their associations with mortality. Results indicated higher diet quality was significantly and consistently associated with reduced risk of death due to all causes, cardiovascular disease, and cancer compared with lower diet quality, independent of known confounders. Diet quality as measured on each of the indexes was negatively associated with marked reductions in mortality. Additional studies of this type should bring some
standardization to the field.

In summary, Shikany and colleagues did indeed identify distinct factors which explain the variability in food intakes in this sample, and one of those factors was related to CHD in this sample. Questions remain regarding how those factors relate to the overall pattern of eating, whether those factors could be identified in other samples, and whether those factors are the most important with regard to CHD. If this study was reexamined with different methods, we would have additional insights regarding the relationship between diet and CHD in this population.

In the future, like all of nutritional epidemiology, patterns research will be advanced by using methods of dietary capture and analysis that better estimate usual intakes and how they may change over time. In addition, continued clarification of the most useful categorization and treatment of input variables for factor and cluster analysis, continued development and refinement of indices, and progress in methods to correct for measurement error would advance the field. Without restricting innovation, there is a need to identify ways to standardize dietary patterns analyses which are necessarily complex. By definition, such analyses are trying to make order out of chaos, so working toward standardized definitions of food groups, and establishing best practices for procedures would help achieve that goal.

In considering approaches to dietary pattern analyses the first priority is to determine which questions to answer and the second is to ascertain whether any of the currently available methods, alone or in combination, will do so. If not, we may need to turn to other disciplines with theories to explain the chaos and complexity inherent in dietary exposure. As always in research, what is of key importance is the questions for which we need answers.

**Conflict of Interest Disclosures:** None.
References:


Table 1. Methods used to identify dietary patterns, the corresponding questions they address, and other considerations for selection.

<table>
<thead>
<tr>
<th>Method</th>
<th>Question addressed by method</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data-driven, outcome independent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Factor analysis</td>
<td>• What food groups correlate in explaining variation in diets?</td>
<td>• Identifies underlying factors that explain variability according to food groups defined by investigator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Factors are not mutually exclusive and therefore do not represent the entire pattern of eating for a group of individuals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Limited reproducibility across studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Resulting patterns can be examined subsequently in relation to health outcome</td>
</tr>
<tr>
<td>• Cluster analysis</td>
<td>• Are there groups of people with distinct dietary patterns?</td>
<td>• Identifies clusters of individuals with distinct dietary patterns according to food groups defined by investigator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Clusters are mutually exclusive, although some individuals’ patterns could be similar to patterns of persons in other clusters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Limited reproducibility across studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Resulting patterns can be examined subsequently in relation to health outcome</td>
</tr>
<tr>
<td><strong>Data-driven, outcome dependent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Reduced rank regression (RRR)</td>
<td>• What combination of foods explains the most variation in a set of intermediate health markers? And then, in confirmatory analysis, does that pattern explain the outcome of interest?</td>
<td>• Identifies patterns which explain variation in health outcomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Reproducible across studies to the extent that study populations exhibit same relationship of diet to outcome</td>
</tr>
<tr>
<td>• Classification and regression tree analysis (CART)</td>
<td>• Which dietary components explain the most variation in a health outcome?</td>
<td>• Makes determinations in stepwise fashion, first determining which dietary component explains the most variation in the health outcome and then, depending on value for</td>
</tr>
</tbody>
</table>
that initial component, which component explains the next most variation, and so on.
- Emerging method; not widely used
- Reproducible across studies to the extent that study populations exhibit same relationship of diet to outcome

<table>
<thead>
<tr>
<th>Investigator-defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective diet</td>
</tr>
</tbody>
</table>
| Do persons who identify as following a particular dietary regimen (vegetarian, gluten-free, etc) have a different risk with regard to a health outcome? | Status is self-assessed on basis of descriptor only (e.g., yes/no in response to “Are you a vegetarian?” or “Is your diet gluten-free?”)
- Only one component of the diet is controlled for (usually the excluded food(s)), whereas other components may vary considerably. Hence, “pattern” is not fully characterized.
- Reproducible across studies in terms of that one dietary regimen only; other aspects of diet can vary widely

<table>
<thead>
<tr>
<th>Indexes/scores</th>
</tr>
</thead>
</table>
| How close is the population to meeting a set of dietary recommendations? | Provides standardized assessment of a particular set of recommendations (e.g., Dietary Guidelines for Americans)
- The same recommendations can lead to different in indexes/interpretations
- Reproducible across studies, assuming index is operationalized using standardized methods
- Resulting patterns subsequently can be examined in relation to health outcome
Examining Dietary Patterns in Relation to Chronic Disease: Matching Measures and Methods to Questions of Interest
Susan M. Krebs-Smith, Amy F. Subar and Jill Reedy

Circulation, published online August 10, 2015;

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/early/2015/07/29/CIRCULATIONAHA.115.018010

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/