A key concern in nutritional epidemiology is the complex nature of the exposure assessed by means of dietary questionnaires. Unlike other exposures, diet is so complex that special knowledge is required to ensure that the correct exposure measure is used. A clear definition of the food or nutrient of interest and knowledge of how it may affect the disease or outcome of interest is imperative.

Food consists of nutrients as well as other substances, such as additives, naturally occurring compounds, and unknown components, that may affect disease risk. Therefore, it is often not adequate to represent food intake by nutrient intake. The approach of Slattery et al. (1) is innovative: Using data from a detailed diet history questionnaire administered to participants in a large case-control study, the authors employed exploratory factor analysis to typify food patterns that might be associated with colon cancer risk. They then evaluated the factors they identified as risk predictors in conventional logistic regression analysis.

Factor analysis is one of the most widely used quantitative techniques in the social sciences. It was developed primarily for psychometric measurement; its origins and early development were largely in the field of abilities testing. Factor analysis has been employed virtually since its inception for two somewhat distinct purposes: data reduction and theory building (2, 3). It does this by characterizing the covariance among many variables in terms of a few underlying but unobservable quantities called “factors.” The factor model is driven by the idea that correlated variables can be grouped or aggregated—i.e., that all correlated variables belong together, and they should be recognized as distinct from groups of variables with which they are not correlated. Factor analysis has great intuitive appeal in nutritional epidemiology: It offers a means of factoring intakes of a variety of foods as manifestations of underlying eating patterns.

The use of factor analysis raises some concerns. Several arise from the degree of subjectivity involved throughout the analytical process. In spite of the elaborate mathematics involved, the principal components approach to factor analysis is atheoretical and pragmatic, typified by a posteriori analysis. Since factor analysis is not a common methodological tool among epidemiologists, it may be useful to outline here the main steps involved (figure 1) and note the kind of subjectivity encountered at each step.

**THE ARBITRARY NATURE OF THE FACTOR ANALYTICAL PROCESS**

In factor analysis, crucial decisions are made when an investigator first selects the variables to examine. That selection is critical in relation to dietary intake, as the number of foods available for analysis can be enormous. Slattery et al. noted that there were over 800 individual food items on their dietary questionnaire (1), making it impractical to include all of these items. Thus, from this list, Slattery et al. identified 35 food groups for analysis. Although the authors probably approached the selection of these foods in an eminently sensible manner, it is clear that it was subjective, not driven by the structure of the data. Thus, no investigator can avoid making judgmental decisions, and it will be difficult to justify the assumption that a given set of foods constitutes all potentially relevant variables. Bias in either inclusion or deletion can be problematic: Inclusion of unrelated variables can have the effect of redefining factors because of shared extraneous variance, whereas deletion of variables in order to simplify the factorial structure can lead to erroneous conclusions (4).

Probably the most important decision in factor analysis is the choice of the number of factors to be extracted. In this analytical step, the identification of dietary patterns from factor loadings is conducted, where the minimum number of factors that can account for the observed correlations is identified. The principal components method, by setting the communality elements at 1.0, permits as many factors as there are variables in the analysis, so unless two variables
are completely redundant (i.e., perfectly correlated), mathematically there can be as many factors as there are variables. Usually, however, only a few of these factors are considered important, and often few are clearly interpretable. In practice, a factor is usually taken to be of consequence if its eigenvalue is greater than 1.0—i.e., if the factor explains more of the variance in the correlations than is explained by a single variable. However, this number can be arbitrarily set at another value; Slattery et al. set it at >1.25. There are no established objective criteria for deciding whether a factor is noteworthy.

Principal components solutions to factoring are often rotated (the next step) to achieve a simple structure of the factor loadings. However, methods of rotation do not improve the degree of fit of the factor structure to the data: The covariation explained is only redistributed among factors. Rotation is aimed at generating an easily interpretable solution. The choice of approach to rotation is similarly arbitrary, and what constitutes a "simpler" structure is a matter of judgment, even if the solutions are completely mathematical. An important but arbitrary decision is whether the method of rotation should result in orthogonal (uncorrelated) or oblique (correlated) factors. Varimax rotation, used by Slattery et al., is one of several options that identifies orthogonal factors.

Once the rotational solution has been chosen, labeling of the factors leads to interpretation of the results. Theory exerts a critical effect as the investigator elects to "name" the factors. This process is essentially an attempt to guess at—or construct—the latent variable that might have produced the observed factor loadings. In the analysis by Slattery et al., the largest factor, labeled the "Western diet," accounted for only 9.7 percent and 7.8 percent of the variance for males and females, respectively (1). A second factor, labeled the "prudent diet," emerged, explaining an additional 8.4 percent of the total variance for men and an additional 7.4 percent for women. Although these two factors appear logical, some others seem less so. For example, the loading for coffee on the "coffee and roll" factor is 0.21, but coffee loads higher (0.33) on the "drinker" factor (see Slattery et al.'s table 3). It could be argued that the label applied to this factor is inappropriate or that this factor does not identify a meaningful dietary pattern.

**FIGURE 1.** Analytical steps in factor analysis methodology in dietary studies.

---

**FACTOR ANALYSIS IN NUTRITIONAL EPIDEMIOLOGY**

A number of issues must be addressed when considering factor analysis in nutritional epidemiologic studies. The first issue pertains to heterogeneity in the data. Although dietary variables are intercorrelated, their intercorrelations are generally modest. Slattery et al. used a criterion of 0.20 as the lower limit for a meaningful factor loading. This level of factor loading would mean that two variables with a predicted correlation of only 0.04 could be considered to load on the same factor. With low loadings, one variable can flip from one factor to another, obscuring its interpretability. The identification of six factors in the analysis by Slattery et al. accounted only for 36.9 percent of the cumulative proportion of the total variance for men and 34.3 percent of that for women (1). Thus, the overall proportion of variance accounted for by these factors is not large. If there are "factors" in people's diets, they stem from very limited relations among dietary elements, and they account for only a very limited portion of the between-person variance in diet.

A major concern is whether the results can be reproduced. If another investigator were to repeat the analytical process, even using the same data set, would he or she be likely to obtain the same results? If factor analysis were applied to a different data set in a similar population, would the results be similar to those of Slattery et al.? If the results were different, how would we judge which model was most correct? Is it likely that there would be consistency across studies? Inconsistency among findings of nutritional epidemiologic studies is already common. In the case of colon cancer, one of the inconsistencies deals with its relation to total dietary fiber (5). While this association is supported by the results of some but not all case-control studies, prospective cohort studies have shown weak or nonexistent associations. Will factor analysis assist in clearing up the inconsistencies? Inconsistent among findings of nutritional epidemiologic studies is already common. In the case of colon cancer, one of the inconsistencies deals with its relation to total dietary fiber (5). While this association is supported by the results of some but not all case-control studies, prospective cohort studies have shown weak or nonexistent associations. Will factor analysis assist in clearing up the inconsistencies?
searchers to support what is already known—that a diet rich in fruits and vegetables reduces risk of colon cancer. Clearly, we will need to gain some experience with this approach before we know whether it is likely to produce consistent results.

MAKING BETTER USE OF FACTOR ANALYSIS

Slattery et al. used what is referred to as “exploratory factor analysis” to interpret the covariance in a large number of variables by a much smaller number of hypothetical factors. Exploratory factor analysis is susceptible to the vagaries of sampling and to instabilities in the data. Slight variations in correlation, which are reflected in much larger (square root) variations in factor loadings, can result in variables’ shifting from one factor to another, depending on exactly how the rotation proceeds (3). Furthermore, the appropriateness of the factor analytical interpretation cannot be tested.

Over the past decade or so, the advantages of confirmatory factor analysis over exploratory factor analysis have become increasingly evident; the rapid development of sophisticated software is threatening to make exploratory factor analysis irrelevant. A major advantage of confirmatory factor analysis is that the results can be formally tested (6). Its disadvantage is that it requires considerable thought ahead of time. Slattery et al. might have derived a theoretically defensible factor structure for the dietary data which they had available and then tested a specific model against the observed data. The factor structure they derived might have been different but possibly more interpretable than that of their purely exploratory approach.

The factors Slattery et al. identified may not be dietary factors at all but may be more generally indicative of broader lifestyle patterns. The fact that the dietary factors correlate with such variables as sex and socioeconomic status indicates that there is more to them than simply diet. It might have been useful to test dietary factors in a hierarchical regression model in which variables such as sex and socioeconomic status were entered as the initial predictors. The question would then be whether diet per se adds anything to predictions based on more general lifestyle and demographic variables.

It is clear that we need better means of using the dietary data we collect. Examining dietary patterns may be useful in many situations. We poorly understand most of the diseases we study, and we have few a priori hypotheses. In situations where reasonable explanations in terms of a few interpretable factors are possible, factor analysis may be very useful. Each application of this technique must be examined on its own merits for determination of its success. As noted above, criteria for judging the quality of any factor analysis have not been well quantified. Thus, it may be difficult to judge whether the inferences generated are misleading or uninformative.

In epidemiologic studies, statistical analyses are used to promote objectivity in the conclusions drawn. Statistical analyses play a central role in nutritional epidemiology. The adoption of today’s statistical methods has led to enormous improvements in the understanding of disease etiology, particularly that of colon cancer. Clearly, we need some way of dealing with the huge numbers of variables on which our dietary instruments collect data. All of these variables are measured with substantial amounts of error which may be correlated or poorly behaved (7). Factor analysis induces numerous uncertainties that allow considerable judgmental discretion on the part of the investigator. In addition, whether it will advance nutritional epidemiology will depend on the effectiveness of its future applications. Nevertheless, Slattery et al. (1) are to be commended for their imaginative adoption of a powerful technique that has, to date, been seldom considered by nutritional epidemiologists.

REFERENCES