Invited Commentary: A Further Look at Dietary Questionnaire Validation

Walter Willett

Knowledge about the long-term consequences of diet is centrally dependent on the methods used to measure the intake of foods by individual participants in epidemiologic studies. To avoid recall and selection biases, which occur in many case-control investigations of diet, large-scale prospective studies are highly desirable. In this context, self-administered questionnaires are usually a practical necessity, and several different self-administered food frequency questionnaires have been developed. Because of the complex nature of individual diets and the substantial variability of food intake over time, measurement errors are inevitable. Thus, the evaluation and refinement of dietary questionnaires are appropriately an important component of nutritional epidemiology. The paper by Subar et al. (1) in this issue is a valuable contribution, as it provides further insights into progress in the development of dietary assessment methods as well as limitations in the methods to assess validity.

In the analysis by Subar et al. (1), three food frequency questionnaires are each compared with a “gold standard” method—up to four 24-hour dietary recalls administered by telephone over a 1-year period. In principle, this design can be informative about the relative validity of different food frequency questionnaires, even when there is some uncertainty regarding the validity of the telephone-administered recalls. The core finding of the Subar analysis is that, after adjusting for total energy intake, the correlations between all three of the food frequency questionnaires and the recalls were remarkably similar (corrected for variation in the recalls, the correlations averaged across nutrients were all approximately 0.60–0.70). The similarity in results for energy-adjusted intakes is remarkable given the quite different nature of the dietary questionnaires: The Eating at America’s Table Study (EATS) questionnaire was a 34-page form based on very recent food pattern data, and the Willett form was an older four-page version that did not reflect the form based on very recent food pattern data, and the Willett questionnaire is associated with fasting triglyceride levels, as predicted by metabolic studies (7). In addition, we and others have documented the ability of this questionnaire to predict reproducibly important diet and disease relations in prospective studies (8). More recently, we have shown that both baseline dietary patterns and change in these patterns predict...
coronary heart disease (9), which adds important evidence for causality. Thus, the field of nutritional epidemiology has matured sufficiently so that we can be confident that important associations with diet can be detected in large, well-conducted studies. Just as importantly, the lack of an observed association can be informative, of course taking into consideration the confidence intervals and the potential effects of measurement error and the temporal relation between dietary assessment and disease occurrence.

The similariy of correlation coefficients for energy-adjusted nutrients assessed by the three dietary questionnaires in the Subar analysis (1) is reassuring, because this suggests that epidemiologic findings will not be highly sensitive to the specific dietary questionnaire that was used. However, the findings are also somewhat troublesome because they suggest a ceiling of validity. Even with a major expansion in length of the questionnaire and careful attention to recent changes in the food supply, the results for the EATS were not appreciably better than those obtained using a considerably shorter and out-of-date questionnaire. This is consistent with the general literature on dietary validation studies; in general, correlation coefficients greater than 0.7 are rare. This ceiling is probably related to the inherent complexity of diet that cannot be fully captured by a structured questionnaire, although some of the error is inevitably attributable to the comparison methods as well. One solution to this limitation appears to be the use of repeated questionnaires in prospective studies. We have found that adjusted correlations can be increased to approximately 0.8 by the use of three questionnaires over a 6-year period (F. B. Hu et al., unpublished data). Moreover, the use of cumulatively averaged questionnaire assessments over time has quite consistently provided improved prediction of cardiovascular disease and diabetes in our large prospective studies compared with the use of a single baseline assessment (10). These repeated measurements presumably provide a better assessment of long-term intake because using the average of repeated measures reduces the impact of errors in completing the questionnaire and also takes into account true changes in diet. The use of repeated questionnaires increases demands on participants. Thus, the modestly lower response rate and the higher percentage of exclusions due to implausible intakes with the long EATS questionnaire could be a disadvantage in this context.

Although absolute intakes have limited use in epidemiologic studies for the reasons described above, several reasons might explain the lower correlations before energy adjustment seen with our questionnaire in the Subar analysis. First, and as noted by Subar et al. (1), the EATS and Block nutrient analysis calculations include portion sizes that are specific to age and gender, whereas this was not done with the Willett questionnaire. Because serving sizes and total energy intakes tend to be greater among men and younger persons, this will tend to increase correlations with absolute intakes when the population is heterogeneous for age and gender, as in the Subar study. We have not routinely used age- and gender-specific serving size values because in epidemiologic studies we would always adjust for these variables and thus remove the variation that was provided by these specific values. Although Subar et al. argue that the diverse population they studied (age range, 20–70 years) was an advantage, the resulting correlations for absolute intakes are misleadingly high because in epidemiologic analysis age would have been controlled. In general, a more realistic validation study would adjust the correlations for the same demographic variables that would ordinarily be controlled in an actual application or would restrict the sample to a more homogeneous group.

A second likely reason for the findings on absolute nutrient intakes is greater correlated error between both the Block and EATS questionnaires and the 24-hour recalls, as compared with the Willett questionnaire. The Block and EATS questionnaires both used portion size questions, as did the 24-hour recalls. From the design of our first validation study (11), we specifically attempted to avoid overstatement of validity due to correlated error between methods and thus selected weighed-diet records as the comparison method least similar to the food frequency questionnaire. Specifically, weighed-diet records do not depend on memory because they are recorded in real time. In addition, the use of dietary scales avoids judgments in assessing portion sizes, which are notoriously poor (12, 13). The use of perceived serving sizes with 24-hour recalls will thus create error, which will tend to reduce correlations with the Willett questionnaire. However, these same errors will tend to be correlated with perceived serving sizes in the EATS and Block questionnaires, which ironically will tend to increase observed correlations. Correlated errors in estimating serving sizes will primarily affect the absolute intakes, rather than energy-adjusted intakes because the latter are primarily determined by the mix of foods eaten rather than the amounts. Subar et al. briefly discuss the issue of correlated errors and dismiss it as an explanation for their finding, arguing that recalls involve episodic memory and that food frequency questionnaires use generic memory. However, these forms of memory are not entirely independent because social desirability and serving size perception can affect both. The degree to which correlated errors account for the findings of Subar et al. is difficult to establish. However, the increase in correlation coefficients with adjustment for total energy intake that they observed with our questionnaire (average deattenuated $r$ increasing from 0.34 to 0.60 in women and from 0.40 to 0.66 in men) was appreciably greater than we have seen when weighed-diet records were used as the comparison in male health professionals (average $r$ without deattenuation increased from 0.46 to 0.54) (14), in female nurses (average $r$ without deattenuation increased from 0.45 to 0.53) (11), or in residents of South Dakota (average $r$ without deattenuation increased from 0.41 to 0.42) (15). These findings add to the suspicion that the data of Subar et al. for absolute intakes are in part artifact due to the use of 24-hour recalls. However, only the use of a comparison method with truly uncorrelated error, such as biochemical indicators of diet, could fully resolve the issue of relative validity for absolute intakes assessed by different food frequency questionnaires.

In summary, Subar et al. (1) provide further evidence that self-administered food frequency questionnaires can pro-
vide informative data on nutrient intakes in epidemiologic applications. For dietary composition, the similarity in correlations is notable despite the substantial differences in the length and design of the questionnaires, which suggests that further advances in questionnaire validity are likely to be modest. Although direct comparisons of different dietary questionnaires can be valuable, their findings suggest that combinations of biochemical measurements and weighed-diet records will be more useful as gold standards than the 24-hour recalls because of lower correlated errors.

ACKNOWLEDGMENTS

The author acknowledges the helpful suggestions of Drs. Meir Stampfer, Eric Rimm, and Liz Lenart.

REFERENCES