This supplement contains the results from studies in the eight countries participating in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Each collaborating centre has assessed the relative validity and reproducibility of the dietary assessment method being used in their centre. One study has also looked at the relative validity of a measure of physical activity, and another study presents the results on an inter-observer study of anthropometric measures. The paper by Kaaks\textsuperscript{1} discusses some fundamental assumptions (particularly the statistical independence of random measurement errors between different measurements) that govern the design and analysis of the preliminary validity studies of the pilot phase of the EPIC project. He presents the rationale for conducting additional ‘calibration’ studies on representative sub-samples of participants in the main cohort studies, and the way the results from the calibration study in each centre may be used to allow the data to be pooled and used to explore the relationship between a wider range of intake and cancer risk. The paper by Kaaks, Slimani and Riboli\textsuperscript{2} gives an overview and summary of the dietary results from each centre.

HISTORICAL PERSPECTIVE

Research in this area has progressed rapidly over the last 10 years. When one of us published a validation study in 1988 there were few comparable studies and little understanding of the need for such studies. It is now widely accepted by centres that do nutritional epidemiological research that assessing the relative validity of the instrument to be used in the study is essential before the study begins, or at the very least during the study. Ideally what is required is a comparison of the test method with a reference method that measures the ‘truth’. In reality there is no such reference method. This is why we use the term relative validity, rather than validity. More recently the term calibration study has been used to describe studies that assess the relative validity of measures. The differences between calibration and validation studies need to be emphasized to avoid confusion. Validation can be defined, as suggested by Armstrong \textit{et al.}\textsuperscript{3} as the estimation of the total relevant measurement error. A validation study estimates a set of unknown statistical parameters in a model that specifies the assumed relations between measurements and true intake.\textsuperscript{4} Given that it is not possible to measure an individual’s true intake absolutely accurately, the measurement model used to assess the relationship between the test measure and the ‘true’ intake must be based on a number of \textit{a priori} assumptions. Validation implies the estimation of a full set of unknown parameters in the measurement model and, under the assumptions of the model, permits the separation of different sources of measurement errors. Validation therefore requires at least two reference measurements.
Calibration studies do not estimate the sources of measurement errors separately and require only a single reference measurement. They are designed to assess those parameters needed to correct measures of association for given increments of intake for measurement error. Kaaks and others have clarified a measurement error model where the measured intake is related to the true intake times a proportional scaling factor PLUS systematic constant measurement error PLUS random individual measurement error. In calibration studies the systematic measurement error and the variance of the calibrated measurements are estimated, but not the separate variances of the random measurement error and the true intake. Thus, a calibration study is only a part of what would be required for a proper validation study (even assuming a measure of the ‘true’ intake). Most validation studies have not considered all the sources of errors mentioned above.

It is implicit in these validation studies that the comparison is between two measures derived from methods that seek to describe dietary habits, and that one of these measures is assumed to be more accurate (the reference measure derived from the reference method). It may not always be clear which method to use as the reference method, especially when the research aims to find out about long term habitual intakes. Up to now, a weighed record of several days’ duration has most often been taken as the reference method. Once the construct validity of the test method has been tested in the study population to ensure that the general approach ‘works’ in practical terms, the next step is to assess whether there is a statistically significant association between the measurements obtained by the test and reference method. It is not clear whether statistical significance alone should be used to judge the level of association or agreement between the measures obtained by the test and reference method. Most studies have used a correlation coefficient and considered that the measure obtained by the test method was ‘good enough’ if a statistically significant association was observed between the measures obtained from the test and reference method. Generally, the correlation coefficients have been between 0.4 and 0.7. This suggests that for most studies, less than half the variation in the measure obtained from the test method can be explained by variation in the measure obtained from the reference method. This shows that there is poor agreement between the measures and that the use of the test method will lead to substantial misclassification and attenuation of risk estimates.

Once a statistically significant correlation has been observed, most researchers have generally said that the test method is good enough. The test method is then used in the main study. Rarely has the measure of association between the test and reference measure been used in the interpretation of the results, except perhaps as a one line statement that the methods used were valid and so the results obtained are not due to biased ascertainment of exposure. The paper by Kaaks, and work by Plummer and Clayton and others suggest that more should be done.

SUMMARY OF MAIN RESULTS AND MAIN CONCLUSIONS FROM STUDIES PRESENTED

These pilot studies have tested the reproducibility and validity of the main method of dietary assessment to be used in the EPIC study, a semi-quantitative food frequency questionnaire (FFQ) or a dietary history (Spain). The reference method has been 12 monthly 24-hour recalls covering one calendar year. Two centres that started their studies earlier have used multiple food records as the reference method (Sweden and UK). Most centres also used plasma levels of vitamin C, vitamin E and β-carotene and 24-hour urinary excretion of nitrogen as reference methods. The dietary questionnaires have been either self-administered or administered by an interviewer. All have used photographs of different portion sizes of food items and commonly known units of portions. The list of food items has been based on the analyses of previous national dietary surveys in many centres and many, but not all, have used a meal-based structure in the questionnaire (especially those using an interviewer). Thus, although the main idea, asking the usual food intake over the previous 12 months, is the same, the methods vary between the centres.

The repeatability of the questionnaire is generally good, but the validity is from modest to good and varies between the centres. Most of the questionnaires overestimated the consumption of vegetables, fruit, added fats, and fish more than other food items and, thus, the intakes of β-carotene and vitamin C were especially overestimated by the questionnaires. However, there were differences between the countries; for example, coffee consumption seemed to be very easy to report in some countries while in some others it was problematic. The relative validity of reporting alcohol intakes was generally good, but in some centres the level of intakes was underestimated much more than in others. The reference method (multiple 24-hour recalls) seemed to underestimate food intakes seriously in some countries as judged by the 24-hour urinary nitrogen and the ratio of energy intake to basal metabolic rate (BMR); in some other centres there was good agreement between the measure. The UK centre, which tested several methods in their pilot study in addition to the FFQ, decided to
include a 7-day food record as well as one 24-hour recall in their EPIC study; they did this because of the poor performance of the FFQ compared to the biomarkers in their validation study. Two of the centres (UK and France) have only women in their study.

Physical activity and anthropometry are also being measured in the EPIC study. The repeatability of the physical activity questionnaire, and its comparability with a more extensive questionnaire, was tested in the Netherlands centre. Intra- and inter-interviewer variability in anthropometric measurements, and estimates of body composition, were investigated in the German centre.

SUMMARY OF STRENGTHS AND WEAKNESSES OF EPIC
The requirements for measuring dietary habits differ for different epidemiological study designs, the food or nutrient of interest, and the variability of that measure in the study population. A major strength of the EPIC study is that dietary habits differ substantially between the centres involved in the study, enhancing the capacity of the study to assess the effects of these dietary differences on cancer risk. The total EPIC sample is more than 400,000 subjects and will represent the largest cohort study undertaken to date. Thus, a large study with large dietary differences makes it more likely that biologically important differences in diet, and their effects on cancer, will be able to be assessed with reasonable statistical confidence. No study published to date either is as large or includes a sample with such a heterogeneous dietary pattern. A number of cohort studies conducted within populations with similar diets, and smaller total sample size, had lower statistical power and may therefore miss important differences. Studies based in one country may also miss the range of exposure that is relevant (that which is related to the outcome) in terms of understanding causation. For the researchers running the study, keeping closer control on aspects of quality control for studies conducted within one country may be easier. Between-centre differences may occur in the way questions have been asked, and this may increase the errors in between-country comparisons. These factors are less likely to affect the results of the study than surveying samples with similar intakes.

All studies can increase their statistical power by reducing the noise around the estimate of dietary habits by using more accurate methods to assess diet. Although it may be feasible, it is unlikely that using a more detailed interviewer-supported method will be widely adopted for reasons of time and money. The methods used by the EPIC study centres cover the range of approaches that are feasible within the study populations, and are similar to methods used in most other large cohort studies. The Spanish data presented in this supplement were collected by interview using a dietary history, and the agreement between the dietary history and the reference method were generally better than in other centres that used an FFQ. The costs of doing the study in Spain, using an interviewer-based approach, were not that much higher than in other centres using an FFQ, primarily because of lower wages in the field workers.

While doing these validation studies the researchers have developed a standardized 24-hour recall method against which all study centre data can be calibrated. This has not been developed for other cohort studies yet.

GENERAL ISSUES
Diet can be measured, despite the difficulties. If understanding the way diet influences disease risk is important then diet must be measured. In the whole spectrum of issues affecting the generalizability of nutritional epidemiological research, however, the technical difficulty of measuring diet is not the weakest link in the chain of doing good relevant research in this area. Of greater importance is concern about understanding the effects of what people eat (or tell the researcher that they eat) in a relevant biological model. Also, to date, there have been few links made between the work undertaken in molecular biology and population-based epidemiological studies. Considerable effort has been invested in animal experimental work describing the consequences of large doses of chemicals on the development of cancer, without always solving the problem of the relevance of these animal models to humans, or whether the chemical or dose is likely to be relevant to humans in terms of what people eat, or whether one substance changed in isolation will have different effects when whole patterns of eating change. The results from three recent β-carotene intervention studies raise important questions about the potential harmful effects of giving people high doses of nutrients, and about how we decide what constituents of foods are the key aetiological factors.9–11

Over the last 20 years there has been limited development of biological measures of the functional state of nutrients in the body that can be used in epidemiological studies. Relatively few data exist on the causal pathways whereby diet is meant to affect cancer risk.

There is virtually no research which has attempted to link interests in nutrient-gene-nutrient interactions to
the design and conduct of epidemiological studies; what gene markers are relevant and how many subjects are needed to measure these markers?

Nevertheless, there are still issues that need to be sorted out with respect to assessing dietary intakes in nutritional epidemiological research. Riboli et al. have recently highlighted four important issues: 1) current measures of nutrient intake lack precision and specificity; 2) nutrients are highly correlated and attribution of causation to one nutrient in isolation may be misleading; 3) biological measures of dietary intake, measured as a concentration in blood at one point in time, may not reliably reflect dietary intake because the biological regulation of these measures is complex and may be influenced by levels of other nutrients; 4) most studies undertaken to date have not considered the effects of the physical characteristics of food on the metabolic activity of the constituents of the food. Data presented in this supplement help address the first of these issues, but more work is required on finding solutions to the remaining three issues.

Another area of concern is defining the relevant time frame over which to measure diet. A perfectly accurate measure of intake, but which covers an irrelevant time frame in terms of either disease induction or promotion, will be of no value. There have been a number of workers who have recently suggested that, for example, the relevant time frame for exploring the effect of fat intake on risk of breast cancer is during adolescence, not in late adulthood.

Differentiating between the technical error of measurement (between observers etc.) and true within-person variability is difficult, but not impossible. The assumption that biological measures will replace dietary assessment as the preferred method for assessing nutrient intakes needs to be made with caution. If what is required is a measure of what people eat, for a biological measure to be an appropriate marker of that measure it must be demonstrated that there is a clear relationship (not necessarily only linear) between the level of intake and the level of the biological measure, within the study population.

From a public health perspective what is required is evidence about the effects of what people eat, in order to frame advice about change (or maintenance) of habits to reduce risk, while ensuring optimal functional levels. In epidemiological research exploiting mechanisms, the interest may be more on the effect of the constituents of foods on the supply of nutrients at the site of biological activity—the functional intake, or on the effect of diet on the processes affecting disease state. This requires an understanding, or at least a plausible hypothesis, of the causal pathway between diet and the outcome of interest. For cancer, the concern is likely to focus on how diet promotes, rather than induces the cancer. A key question is; what is the relevant measure of exposure (dose, intensity and time), and can this exposure be measured, either by a dietary assessment or by a biological marker?

CONCLUSION

We should not underestimate the progress that has been made in nutritional epidemiology and acknowledge that assessing the relative validity of measures has been an important step forward in helping us to understand the relationship between diet and disease, and the role that chance, bias and confounding has on that relationship. The challenge over the next 10 years is to use the results of validation studies to help us to improve the design and interpretation of diet-disease studies. We need to explore ways of developing valid and reliable indicators of food patterns, as well as exploring statistical techniques to handle the multicollinearity which exists between the consumption of groups of foods. We need to develop more relevant functional markers of intake that can be measured in biological samples. We also need to forge links between those working at the molecular level and those working in epidemiology, and ensure that those who are charged with guiding public health have the best possible information upon which to make difficult decisions.

REFERENCES


