Invited Commentary: The Challenge of Multi-Center Cohort Studies in the Search for Diet and Cancer Links

Elio Riboli and Rudolf Kaaks

The investigation of the role of nutrition in cancer etiology and the perception of its importance among concerned scientists has gone through various phases over the past 20 years, ranging from the enthusiasm of Candide, Voltaire's hero (we are in the best possible world and we know everything) to Hume's kind of skepticism (the subject is appealing but causality can never be shown). These ups and downs have been marked by the excitement caused by reports which have emphasized the predominant role of diet in cancer causation (1-3) or the disappointment of large cohort studies that did not find what they were supposed to find, e.g., that fat causes breast cancer (4) or that (vegetable) fiber protects against colorectal cancer (5).

In the 1970s, descriptive epidemiology studies clearly showed that the incidence of cancers of the breast, colorectum, and prostate is substantially higher in populations that share economic development and a Western life-style, and that the incidence is positively correlated with the typical Western diet characterized by high consumption of meat, total or saturated fat, and highly refined carbohydrates, and negatively correlated with consumption of starchy foods and plant foods in general. This so-called “ecologic evidence” has been corroborated by the results of the vast majority of case-control studies conducted in different populations around the world.

In a recent and comprehensive report by the World Cancer Research Fund (WCRF) (6), almost 200 case-control studies on food and nutrient intake were cited (not counting those on alcohol, body mass index, and physical activity) and, from a rough calculation, about two-thirds of them show significant results in the anticipated direction (e.g., that fruit and/or vegetables protect against a variety of cancers, that red meat increases colorectal cancer risk, and that salt-preserved foods increase stomach cancer risk).

The same report (6) cites about 50 major articles from cohort studies, and their results are much less convincing. While on the one hand there is some consistency in finding an association between high intake of meat and fat and increased risk of colorectal and prostate cancers in cohorts conducted in Western populations (pp. 243-6 and 321 of the WCRF report (6)), the results on vegetables and fruit are quite disappointing compared with what one would have expected based on ecologic evidence and case-control studies.

So one wonders if the vegetable and cancer association is some kind of a mirage in the desert caused by a constant distortion of the “image” of the dietary habits reported by study subjects who, whether they are Italian or Chinese, Indian or North American, always overestimate vegetable intake if they are controls and always underestimate such intake if they are cases. It is, however, difficult to believe that there is such consistency in recall bias around the world and for different cancers. Empirical evidence supporting the existence of such universal recall bias is very scanty. To our knowledge, very few studies have compared dietary data collected retrospectively with those collected prospectively. The results were not concordant on fat intake (7, 8), and one prospective study in Sweden found that, if anything, cancer patients overestimated vegetable intake more than control subjects when interviewed after diagnosis compared with years before diagnosis (9).

The point is therefore whether the prospective studies published so far were sensitive enough to find associations between diet and cancer which may be relatively weak in nature for any single dietary component and even weaker in empirical relative risk estimates due to random measurement error (10). The sensitivity will depend not only on sample size and relative risk but also on the variability of the exposure in the study population relative to the background noise due to exposure measurement imprecision. The papers by Kolonel et al. (11) and Stram et al. (12) illustrate an interesting new approach to these issues which, in our
opinion, can substantially improve the sensitivity of prospective cohort studies.

The first important feature of the study is to select populations within which dietary habits vary as much as possible and to include within the same study different populations with very different dietary habits in order to maximize true exposure variances. This is effected by the inclusion of sub-cohorts from different ethnic groups in Los Angeles and Hawaii.

The second important feature of the Hawaii-Los Angeles project is the attempt to calibrate dietary measurements across cohorts. In theory, accurate calibration of dietary measurements can improve the comparability of estimated diet-disease relations, and hence the overall statistical power of pooled analyses. Calibration can be achieved by collecting additional well standardized dietary measurements in a representative subsample of each cohort, as described by Stram et al. (12).

In this respect, the design of the Hawaii-Los Angeles project shares many conceptual similarities with the European Prospective Investigation into Cancer and Nutrition (EPIC) (13), in particular the decisions to take advantage of differences in dietary habits and cancer risk in different populations, with the aim of combining the advantages of within-cohort and between-cohort risk factor analyses, and to implement an in-built calibration study.

There are, however, some critical issues which deserve further consideration regarding the difference between validation and calibration studies and regarding the conditions under which calibration can actually improve pooled relative risk estimates.

**Calibration** basically can be seen as a scale-adjustment to the questionnaire measurements ($Q'$) so that, after rescaling, the regression of disease outcome on $Q'$ allows an unbiased estimation of relative risk. This approach allows the correction of relative risk estimates for quantitative differences in intake level only if expressed on a continuous scale. In its simplest form, following the linear calibration model used by Stram et al. (12), rescaled questionnaire measurements ($Q'$) are calculated from the original measurements ($Q$) as $Q' = a + bQ$, where the parameters $a$ and $b$ are, respectively, the intercept and slope of the regression of true intake level $X$ on the measurements $Q$. In practice, the regression parameters $a$ and $b$ cannot be estimated by regressing true intake levels $X$ on $Q$, as the true intake levels will obviously be unknown. The parameters are therefore estimated by regressing some reference measurement $R$ on $Q$ in at least a representative subsample of the study population. Usually, the reference measurements are based on 24-hour diet recalls or on food consumption records. The validity of the validation approach then depends, however, on a number of important assumptions (see discussion below). In principle, calibration requires only a single day's food record or 24-hour diet recalls per person. As discussed by Stram et al. (12), the slope $b$ is usually estimated to be below 1.0, so that the calibrated questionnaire measurements $Q'$ will generally have a standard deviation ($SD(Q') = b \times SD(Q)$) that is smaller than that of the original questionnaire measurements (14). The magnitude of this contraction in the standard deviation is directly related to the loss of statistical power for tests of diet-disease association, and to a decrease in precision of relative risk estimates (15).

Compared with calibration studies, **validation** studies go one step further, in that not only can they provide the scale corrections of a calibration substudy, but also aim to estimate the proportion of variation in questionnaire measurements that reflects variation in true intake levels. The latter basically means estimation of the correlation $\rho_{QX}$ between questionnaire measurements and true intake level. The estimation of the correlation coefficient $\rho_{QX}$ also makes it possible to adjust relative risk estimates for quantile levels (e.g., quartiles, quintiles) of intake, as well as estimates of attributable risk, for attenuation bias due to random errors in the questionnaire measurements. Validation studies require at least two additional measurements (e.g., 24-hour recalls for $\geq 2$ days) in addition to the questionnaire measurements. Validation relies on similar model assumptions to calibration. These assumptions are needed partially to make up for the fact that error-free ("gold standard") reference measurements are not available in practice. The attenuation-corrected correlations in table 3 of the paper of Stram et al. correspond to the results of such a validation approach and, conditional on the model assumptions, can be interpreted as estimates of the correlation $\rho_{QX}$.

It is, however, important to remember that the validity of the calibration and validation approaches depends on three crucial requirements:

1) Reference measurements must be obtained in representative subsamples of population groups to whom the results of the calibration or validation studies will be applied;

2) The reference method must provide unbiased intake estimates ($R$) at a group level conditional on the individuals' true intakes; and

3) Random errors in the reference method must be uncorrelated with the random errors in the questionnaire measurements. In validation studies, an additional requirement is that random errors are also statistically independent between repeat reference measurements within the same individual.

The first requirement may be fulfilled by careful study design, and by ensuring high participation rates.
in the calibration/validation substudies. The second requirement, essential for both the within-group and between-group (ecologic) components of calibration and data analysis, may be more difficult to satisfy in practice. Intakes of energy and macronutrients may be underestimated by 24-hour recalls or by food consumption records, and this also seems apparent from the energy intakes reported by Stram et al. (12). However, if the degree of underreporting is approximately constant across the population subgroups, the 24-hour recalls can still be used as reference measurements for between-group calibration. The questionnaire measurements are then calibrated onto a reference scale without absolute validity, but which is common across groups. The third requirement is crucial especially for the within-group calibration by regression of R on Q measurements. Current evidence suggests that individuals may vary systematically in their tendencies to underreport intake irrespective of whether measurements are obtained by questionnaire, 24-hour recall, or food consumption records. This might lead to positive correlations between random errors in the reference and questionnaire measurements, overestimation of the calibration slope \( b \), and, hence, to an under-adjustment of relative risk estimates. Detailed validation studies with biomarkers may help to investigate the possible magnitude of this problem. Violations of requirements 2 and 3 due to systematic underreporting could be much diminished by expressing intake measurements in terms of relative nutrient composition (e.g., energy densities) instead of absolute intakes, as in the analysis by Stram et al. (12).

Calibration, as a means to re-express relative risk estimates so as to make these correspond to exposure differences expressed on a common reference scale, is of particular interest for inter-group comparisons in studies based on different subpopulations. This is true both when it is impossible to use the same dietary questionnaire in each subgroup—because of important differences in food consumption patterns, or language, and when it is difficult to simply assume that the validity of the questionnaire measurements is equal across subgroups, even when the questionnaire is identical. The Hawaii-Los Angeles study falls into the second case, while the first applies to EPIC, which is a multi-center study with cohorts in 22 research areas scattered all over western Europe, from northern Scandinavia to the Mediterranean (13).

In such multi-center or multi-ethnic cohort studies, estimated relations between diet and disease risk can be broken down into two components.

The first is the average relation of diet with risk within each of the subgroups separately, as would be estimated in a relative risk analysis of the full study population but with stratification by subgroup. Stram et al. (12) focus on this first component of analysis, and on the calibration adjustment of relative risk estimates within strata defined by ethnic subgroups. Here, the calibration may serve to adjust for heterogeneity in relative risk estimates (apparent effect modification) by subgroup, which may be due to differences in the magnitude of biases caused by dietary assessment errors. If accurate, such correction for error-induced heterogeneity in effect magnitude will be equivalent to a more optimal weighting of the subgroup-specific relative risks into an overall, stratified estimate, decreasing the weightings proportionally to the amounts of contraction in the standard deviations of questionnaire measurements after calibration (16). In theory, this more optimal weighting may improve the overall statistical power of a pooled analysis stratified by subgroup, but only if the calibration itself is sufficiently accurate.

The second component is the "ecologic" relation of the mean dietary intake levels with average disease incidences between the groups. This second component captures the increase in power for tests of diet-disease association that may be obtained by combining the dietary exposure and disease outcome data across diverse populations with different life-styles and disease incidence rates (16, 17). With data on potential confounders available at the individual level, the ecologic relation can in principle be estimated with adjustment for the confounding factors (e.g., by simple stratification), contrary to classic ecologic studies, where only population averages of confounding variables may be known. Calibration study data, obtained in representative samples of subjects within each subgroup, may provide the measurements, against a common scale, of mean dietary intakes at the subgroup level. After correction for potential biases due to confounding and measurement errors, the within- and between-group component estimates of a diet-disease relation should ideally be of similar magnitude, and thus corroborate each other.

Besides these critical issues related to variability of dietary exposure and measurement errors, there is, in our opinion, an additional fundamental component in the design of new generation prospective studies which relates to the collection and use of biologic samples. The researchers of the Hawaii-Los Angeles study, who have already played a pioneering role in investigating the role of genetic polymorphisms in different ethnic groups, correctly emphasize the potential importance of studying genetic susceptibility to dietary and environmental risk factors. The collection of baseline blood samples in a small representative subsample of the cohorts can provide the material needed for genetic analyses among controls. Subsequent collection of blood samples from
incident cancer cases occurring within the cohort is
needed to conduct case-cohort studies on genetic factors
and on gene-diet interactions, if statistical power allows
it. This design has the appeal of offering a low-cost solu-
tion to the issue of how to investigate gene-environ-
ment interactions in large cohort studies. However, it has
the major limitation of not being useful for studies on meta-
\[\text{Continued on next page}\]