Use of Semiquantitative Food Frequency Questionnaires to Estimate the Distribution of Usual Intake

R. J. Carroll,1 L. S. Freedman,2 and A. M. Hartman3

The authors consider whether semiquantitative food frequency questionnaires can be used to survey a population to estimate the distribution of usual intake. They take as an assumption that, if they were possible to obtain, the mean of many food records or recalls would be an accurate representation of an individual's usual diet. They then assume that nutrient intake as measured by a questionnaire follows a linear regression model when regressed against the usual intake of that nutrient. If the coefficients in this regression relation were known, then the distribution of usual intake could be constructed from the responses to the questionnaire. Since one generally does not know the values of the coefficients, they need to be estimated from a calibration study in which respondents complete the questionnaire together with multiple food records or recalls. This can be done either through an internal subset of the data or through an independent external study. With an internal substudy, the authors find that food frequency questionnaires typically provide little information about the distribution of usual intake in addition to that obtained from the multiple records or recalls in the substudy. When the substudy is external, if it is small then having very large numbers of subjects completing food frequency questionnaires in the survey is no more efficient than having a few subjects completing food records or recalls. However, if the external substudy is large and accurately characterizes the relation between the questionnaire response and usual intake, food frequency questionnaires can provide a cost-efficient way of estimating the distribution of usual intake. These results do not apply to the different problem of correcting relative risks for the effects of measurement error. Am J Epidemiol 1996;143:392-404.

As is apparent in these debates, dietary measurement is an activity that is fraught with difficulties. There is no one method of measurement that can be regarded as a gold standard. Nevertheless, certain assumptions or "working hypotheses" about the proper-
ties of the various methods have been quite widely, although not universally, adopted and form the basis for the work we present in this paper.

First, we assume that the usual dietary intake can be recorded without bias using a method such as a 24-hour recall or a self-reported food record. The phrase “without bias” is crucial to our argument. We mean that were we able to repeat such an assessment of an individual many times over a period of time, then the average over such assessments would indeed represent the usual intake of the individual. This assumption is not without challenge. There is evidence (8) that both multiple-day food records and multiple 24-hour recalls are biased assessments, but there are at least some grounds for optimism about an interviewer-conducted 24-hour recall that is administered only once or twice on the same individual. While there remain doubts regarding the unbiasedness of recalls or food records, there is graver concern that the food frequency questionnaire method carries the potential for serious bias. To reflect such concern, we assume in our work that nutrient or food intake as measured by a food frequency questionnaire has, on average, a linear relation with the true usual intake and is therefore correlated with usual intake but that, on average, the food frequency questionnaire value may underestimate or overestimate the true usual intake.

Under these assumptions, we pose the following question in this paper. By calibrating the food frequency questionnaire against an unbiased method (such as several 24-hour recalls and/or food records) in a relatively small calibration study, are we then able to use the food frequency questionnaire in the population survey to estimate nutrient or food usual intake distributions? The similar problem of developing a calibration regression method for adjusting the nutrient intakes obtained from the less accurate reporting method has been considered (9) but did not discuss the estimation of usual intake distributions.

STATISTICAL MODEL FOR CALIBRATION

We propose that the typical survey that uses a food frequency questionnaire will have two components. In the main survey, for a large number (n) of individuals, nutrient intake is measured by a food frequency questionnaire. In the substudy, on a smaller number (n_R) of individuals, the food frequency questionnaire nutrient intake is calibrated against usual intake by measuring nutrient intake using the questionnaire as well as two or more food records/recalls. We will call this small number of individuals the calibration study.

The statistical calibration model is that introduced previously (10). It models the relation between questionnaires, denoted by Q, and food records/recalls, denoted by F. The aim is to understand true usual intake, namely, T. In a typical substudy, a food frequency questionnaire is obtained as are m food records or recalls. The model relating Q and F is a standard linear errors-in-variables model, namely

$$Q = \beta_0 + \beta_1 T + \epsilon;$$  \hspace{1cm} (1)

$$F_j = T + U_j; \quad j = 1, \ldots, m.$$  \hspace{1cm} (2)

In model 1, \(\beta_0\) is the intercept and \(\beta_1\) is the slope. The term \(\epsilon\) represents the usual error about the line in linear regression relating Q to T. In model 2, we are assuming that F is an unbiased measure of usual intake, and the \(U_j\) are the intraindividual measurement errors, perhaps after a transformation. As discussed previously, one may think then of T as the best measure of an individual’s intake, if one could obtain many records/recalls.

Among these random variables, T has mean \(\mu_T\) and variance \(\sigma_T^2\), \(U_j\) has mean zero and variance \(\sigma_U^2\), and \(\epsilon\) has mean zero and variance \(\sigma^2\). All random variables are uncorrelated.

The distribution of usual intake is the probability distribution of the random variable T. Our goal is to understand the role that food frequency questionnaires play in estimating this distribution.

The simplest case occurs when T has a normal distribution with mean \(\mu_T\) and variance \(\sigma_T^2\). For such populations, the \((1 - \alpha)\) × 100th percentile of the distribution is \(\mu_T + z_{1-\alpha} \sigma_T\), where \(z_{1-\alpha}\) is the \((1 - \alpha)\) × 100th percentile of the standard normal distribution. Thus, \(z_{1-\alpha} = 0.84, 1.645, \) and 1.96 for the 80th, 95th, and 97.5th percentiles. All the percentiles of usual intake thus depend only on the mean and variance of the usual intake.

The assumption of normality is plausible in our experience to a first approximation for the percentage of energy from fat, as well as for simple transformations (using the logarithm or square root) of other macronutrients such as fat, protein, and carbohydrate, and for some micronutrients.

NUMERIC RESULTS FOR INTERNAL CALIBRATION

As described above, in a typical survey n individuals are given a food frequency questionnaire, and of these \(n_R\) participate in the calibration substudy. This section explores the efficiency of food frequency questionnaires numerically, using population values for the percentage of calories from fat as determined by the Women’s Health Trial Vanguard Study (11) and by the Helsinki Diet Pilot Study (12). There are many
variables that go into the design of a calibration study, including the number of food frequency questionnaires and the number of food records or recalls completed by each individual. We have found that different plans have qualitatively the same behavior. So, for purposes of illustration, we consider the simplest possible calibration study, where two food records or recalls are taken, along with the food frequency questionnaire. To avoid possible correlation between the food frequency questionnaire and food record or recall errors within an individual, we will assume that the records/recalls are completed sufficiently far apart in time from one another and from the questionnaire that the errors in these random variables can be considered as independent. When this assumption cannot be made, other, more complex models can account for the correlations (see the Appendix).

In the Appendix, we discuss a method, based on the Fisher information matrix, for computing the standard errors of estimates of the distribution of $T$, under the assumption of normal distributions. The parameters in the model 3–4 were estimated using the techniques of Freedman et al. (10) from the Women’s Health Trial Vanguard Study data (which used food records), namely, $\mu = 38.25$, $\sigma^2 = 24.45$, $\sigma^2_u = 40.92$, $\sigma^2_u = 30.36$, $\beta_0 = 5.95$, and $\beta_1 = 0.83$. While the estimate of $\sigma^2_u$ is 40.92, as part of a sensitivity analysis we will consider the cases $\sigma^2_u = 24.71$, 40.92, and 57.12. The choices around the estimated value reflect either no or a doubling of the so-called “equation error” in the expanded model 3 in the Appendix.

In what follows we contrast the standard errors of the estimates of certain parameters of interest (such as the true mean usual intake in the population) obtained under two designs. The first design is where $n$ individuals complete a food frequency questionnaire, and a subset $n_R$ also complete two food records. The second design is where $n_R$ subjects complete two food records and the food frequency questionnaire is not used in the survey. We are interested to know what decrease in variance is caused by the extra information provided by the food frequency questionnaires. We will examine the ratios of the standard errors in these two designs for three estimates: the population mean usual intake, the standard deviation of the usual intake, and the 80th percentile of the usual intake.

In table 1, we show these ratios for the case $\sigma^2_e = 40.92$ over a range of sample sizes for the main survey and substudy. The main survey provides only a small reduction in the standard deviation for estimating the mean usual intake. For example, a survey with $n = 4,000$ participants completing a food frequency questionnaire and a calibration substudy with $n_R = 100$ participants gives a standard error only 9.2 percent smaller than the standard error obtained directly from 100 participants in the substudy. To give some perspective, the same decrease in standard error could be obtained by increasing the calibration study sample size by 21 percent ($(1 - (1 - 0.092)^2)$, i.e., adding 21 additional participants to the calibration study).

The picture is even less optimistic when one considers estimating the distribution of usual intake instead of its mean. For example, in estimating the 80th percentile of the distribution of true usual intake, having 4,000 participants completing a food frequency questionnaire results in only a 6.03 percent decrease in the standard error; increasing the calibration study by 14 percent has the same net effect on the standard error. Thus, once one has performed a calibration study of size $n_R = 100$, 4,000 participants completing a food frequency questionnaire contribute the same degrees of information as do 14 participants completing two food records.

The value of food frequency questionnaires decreases as one attempts to estimate extremes of the distribution of the usual intake. When estimating the 95th percentile, with $n_R = 100$ individuals in the calibration study, one finds that the use of food frequency questionnaires can decrease the standard error by no more than 3.64 percent versus 6.19 percent when estimating the 80th percentile of usual intake.

Figure 1 gives a graphical representation of the ratio of the standard errors for estimating the 80th percentile of the distribution of usual intake when questionnaires are used to the standard error when they are not used, for the three values of $\sigma^2_e$. This ratio depends slightly but not heavily on this value. It is important to note that empirically, in this example, it is impossible to decrease standard errors by more than 10 percent, no matter how many food frequency questionnaires are given.

We also considered what might happen if the calibration study is based on 24-hour recalls instead of food records. Using data from the 1985 and 1986 US Department of Agriculture Continuing Survey of Food Intake by Individuals, we found that an estimate of the within-individual variance for a single 24-hour recall is 83.35. While we do not display the results here, food frequency questionnaires are even less valuable in this case. Indeed, food frequency questionnaires become progressively less valuable as the within-individual variance in the 24-hour recall increases, because the calibration study estimates the relation between food frequency questionnaires and the usual intake with progressively poorer precision. To see this, consider figure 2, where we have plotted the relative standard errors for using versus not using food frequency questionnaires as $\sigma^2_u$ ranges from 0 to 200. There we see
TABLE 1. Efficiency of food frequency questionnaires with internal validation using food records when $\mu = \mu_3 = 38.25$, $\sigma^2 = 24.45$, $\sigma^2_\mu = 40.92$, $\sigma^2_\sigma = 30.36$, $\beta_0 = 5.95$, and $\beta_1 = 0.83$

<table>
<thead>
<tr>
<th>No of food records</th>
<th>No of food frequency questionnaires</th>
<th>$\mu_\sigma$ Standard error$^*$</th>
<th>Efficiency (%)$^t$</th>
<th>$\sigma^2$ Standard error</th>
<th>Efficiency (%)</th>
<th>$\mu + 0.84 \sigma_\mu$ Standard error</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>0.62</td>
<td>100.00</td>
<td>6.00</td>
<td>0.81</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>95.39</td>
<td>96.28</td>
<td>99.93</td>
<td>95.75</td>
<td>94.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>93.00</td>
<td>96.50</td>
<td>98.81</td>
<td>94.60</td>
<td>94.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>91.79</td>
<td>98.75</td>
<td>98.71</td>
<td>94.45</td>
<td>93.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>91.54</td>
<td>98.66</td>
<td>98.64</td>
<td>94.23</td>
<td>94.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>91.21</td>
<td>98.62</td>
<td>98.62</td>
<td>94.02</td>
<td>93.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>91.05</td>
<td>98.60</td>
<td>98.60</td>
<td>93.94</td>
<td>94.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,000</td>
<td>90.88</td>
<td>98.58</td>
<td>98.58</td>
<td>93.94</td>
<td>94.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5,000</td>
<td>90.75</td>
<td>98.57</td>
<td>98.57</td>
<td>93.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>90.65</td>
<td>98.57</td>
<td>98.57</td>
<td>93.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000,000</td>
<td>90.56</td>
<td>98.57</td>
<td>98.57</td>
<td>93.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>0.36</td>
<td>100.00</td>
<td>3.46</td>
<td>0.46</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>97.72</td>
<td>96.44</td>
<td>99.00</td>
<td>95.71</td>
<td>95.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>95.39</td>
<td>96.28</td>
<td>98.81</td>
<td>94.78</td>
<td>94.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>94.21</td>
<td>98.71</td>
<td>98.71</td>
<td>94.45</td>
<td>94.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>93.49</td>
<td>98.68</td>
<td>98.68</td>
<td>94.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>92.52</td>
<td>98.65</td>
<td>98.65</td>
<td>94.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>92.03</td>
<td>98.61</td>
<td>98.61</td>
<td>94.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,000</td>
<td>91.54</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>91.30</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5,000</td>
<td>91.15</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>90.85</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000,000</td>
<td>90.56</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>0.28</td>
<td>100.00</td>
<td>2.68</td>
<td>0.36</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>98.48</td>
<td>99.75</td>
<td>99.46</td>
<td>97.72</td>
<td>97.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>96.56</td>
<td>99.46</td>
<td>99.46</td>
<td>97.72</td>
<td>97.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>95.39</td>
<td>99.28</td>
<td>99.28</td>
<td>96.95</td>
<td>95.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>93.81</td>
<td>99.04</td>
<td>99.04</td>
<td>95.92</td>
<td>95.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>93.00</td>
<td>98.93</td>
<td>98.93</td>
<td>95.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,000</td>
<td>92.19</td>
<td>98.81</td>
<td>98.81</td>
<td>94.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>91.79</td>
<td>98.71</td>
<td>98.71</td>
<td>94.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5,000</td>
<td>91.54</td>
<td>98.71</td>
<td>98.71</td>
<td>94.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>91.05</td>
<td>98.64</td>
<td>98.64</td>
<td>94.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,000,000</td>
<td>90.56</td>
<td>98.57</td>
<td>98.57</td>
<td>93.81</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Standard errors are the large-sample standard error of the parameter when only the food records from the calibration study are used.

† Percentages are the ratio of the standard errors when not using versus using the food frequency questionnaires in the main survey (and the smaller calibration study).

that, as the within-individual variance increases, the benefit for using questionnaires decreases.

In figure 3, we reproduce figure 2 but for different amounts $\sigma^2_\sigma$ of variance of food frequency questionnaires for a given usual intake. The correlation $corr(Q,T)$ between a food frequency questionnaire and usual intake is listed. These correlations are, in practice, rarely much more than 60 percent (approximately 55 percent in the Women’s Health Trial Vanguard Study). It is only when food frequency questionnaires are rather highly correlated (perhaps unrealistically so) that the food frequency questionnaires in the main study add much information to the food records or recalls obtained in the calibration substudy.

We have repeated this analysis using parameters estimated from the Helsinki Study, using baseline and final food-use questionnaires, namely, $\mu_\sigma = 38.90$, $\sigma^2_\sigma = 19.30$, $\sigma^2_\mu = 20.15$, $\sigma^2_\sigma = 15.20$, $\beta_0 = 14.60$, and $\beta_1 = 0.59$ (figure 4). Qualitatively, the results are much the same as for the parameters from the Women’s Health Trial Vanguard Study (compare figure 4 with figure 1).
FIGURE 1. Efficiency of food frequency questionnaires with internal calibration for estimating the upper quintile of the distribution of usual intake, with measurement error for food records and distribution of usual intake based on the Women's Health Trial Vanguard Study. var(e), within-individual variance of a food frequency questionnaire, is the same quantity as \( \sigma_f^2 \). For \( \sigma_f^2 = 24.71, 40.92, \) and 57.12, a plot is shown of the ratio of the standard error when using 100 replicate food records and the number of food frequency questionnaires as indicated on the abscissa to that using 100 replicate food records only. Here, \( \mu_e = 38.25, \sigma_f^2 = 24.45, \sigma_u^2 = 30.36, \beta_0 = 5.95, \) and \( \beta_1 = 0.83. \)

NUMERIC RESULTS FOR EXTERNAL CALIBRATION

In some instances, one is unable to perform a calibration study within the population of interest, and one only has the choice of conducting the survey either with food frequency questionnaires or with food records or recalls on \( n_F \) participants. We will call this population the primary population. Alone, neither a single food frequency questionnaire nor a single food record or recall suffices to characterize the usual intake distribution in the primary population, and for that one must rely on an external calibration study of size \( n_e \), performed on a different population, the external population. The issue is whether it is more efficient to have a small number of participants completing food records or recalls or a large number of participants completing food frequency questionnaires in the main study.

To use a calibration study on an external population, we must assume that the regression relation in model 3-4 is the same in both populations. However, the mean and variance of usual intake may differ in the two populations. The intake reported from food records or recalls in the primary population has mean \( \mu_u \) but variance \( \sigma_u^2 + \sigma_e^2 \). It is only possible to estimate \( \sigma_u^2 \) if the external study provides information on \( \sigma_e^2 \). Similarly, the intake reported from food frequency questionnaires in the primary population has mean \( \beta_0 + \beta_1 \mu_e \) and variance \( \beta_1^2 \sigma_f^2 + \sigma_u^2 \), so that estimation of \( \mu_e \) requires that the intercept and the slope not differ between the primary and external studies, while estimation of \( \sigma_f^2 \) requires that the errors of food frequency questionnaires about individual usual intake have the same variance \( \sigma_f^2 \) in the two populations.
Distribution of Usual Nutrient Intake

1.0 - 
0.9 - 
0.8 - 
0.7 - 
0.6 - 

corr(Q,T)=0.79 
corr(Q,T)=0.68 
corr(Q,T)=0.54 
corr(Q,T)=0.47 

0 20 40 60 80 
Food Record Error Variance 

1.00 - 
0.95 - 
0.90 - 
0.85 - 
0.80 - 

var(e)=12.59 
var(e)=20.15 
var(e)=27.71 

0 1,000 2,000 3,000 4,000 5,000 
Number of Questionnaires 

FIGURE 3. Efficiency of food frequency questionnaires with internal calibration for estimating the upper quintile of the distribution of usual intake for differing correlations between a food frequency questionnaire and usual intake. For \( \sigma_e^2 = 0-100 \), and \( \sigma_f^2 = 10, 20, 40, \) and 60, a plot is shown of the ratio of the standard error when using 100 replicate food records only to that using the food records and 4,000 food frequency questionnaires. \( \text{corr}(Q,T) \), correlation between the food frequency questionnaire and usual intake. Here, \( \mu_f = 38.25, \sigma_f^2 = 24.45, \beta_0 = 5.95, \) and \( \beta_1 = 0.83 \)

FIGURE 4. Efficiency of food frequency questionnaires with internal calibration for estimating the upper quintile of the distribution of usual intake, with measurement error for food records and distribution of usual intake based on the Helsinki Diet Pilot Study. \( \text{var}(e) \), within-individual variance of a food frequency questionnaire, is the same quantity as \( \sigma_e^2 \). For \( \sigma_e^2 = 12.59, 20.15, \) and 27.71, a plot is shown of the ratio of the standard error when using 100 replicate food records only to that using the food records and the number of food frequency questionnaires as indicated on the abscissa. Here, \( \mu_e = 38.90, \sigma_f^2 = 19.30, \beta_0 = 14.60, \beta_1 = 0.59, \) and \( \sigma_e^2 = 15.20 \)

In the Appendix, we indicate how one can perform the calculations leading to a comparison of the two designs, i.e., using food frequency questionnaires or food records or recalls. We have performed these calculations for the same hypothetical population based upon the parameters from the Women's Health Trial Vanguard Study considered previously. To make the problem simple, we have taken the mean and variance of usual intake in the primary and external populations to be the same, namely, 38.25 and 24.45, respectively.

In figure 5, we display the number of participants, \( n_F \), completing food records that gives the same standard error for estimating the 80th percentile of usual intake to that given by 100-4,000 participants completing food frequency questionnaires. It is striking that the qualitative conclusions depend on the size of the external study. For example, if there are 100 participants in the external study, then approximately 50 participants completing food records are equivalent to many thousands of participants completing food frequency questionnaires. However, if the external study is large (\( n_e = 500 \)), then as few as 1,000 participants completing food frequency questionnaires are equivalent to 175 participants completing food records.

The conclusion is that the larger the external study, the more value there is to using food frequency questionnaires in surveys.

The reason for this phenomenon can be understood intuitively. If the external study is very large, then the slope and the intercept are essentially known, and the food frequency questionnaire value can be corrected to be unbiased for usual intake, by

\[
Q(\text{adjusted}) = \frac{Q - \beta_0}{\beta_1} = T + \frac{\epsilon}{\beta_1}
\]
With a very large external study, this adjusted food frequency questionnaire has measurement error variance \( \frac{\sigma_e^2}{\sigma_f^2} \), as opposed to food records or recalls that have measurement error variance \( \frac{\sigma_e^2}{\sigma_r^2} \). Thus, with a very large external study, one internal food record or recall is approximately equivalent to the following number of food frequency questionnaires:

\[
\frac{\sigma_e^2}{\sigma_f^2} \beta_f^2
\]

Using the estimates from the Women’s Health Trial Vanguard Study, we find that this number is approximately equal to two for food records.

The value of increasingly large external studies can also be seen in figure 6, where we display the actual large sample standard error for estimating the 80th percentile of the distribution of usual intake, as a function of the number of participants completing food frequency questionnaires in the main survey (\( n_x \)). For example, with an external study of size \( n_x = 100 \), the standard error for 2,000 food frequency questionnaires is approximately 1.4, but this standard error decreases by a factor of approximately 2 when the external study is of size \( n_x = 500 \).

We repeated these analyses with parameters from the Helsinki Diet Pilot Study (see figures 7 and 8). Note that figure 8 is virtually identical to figure 6. Qualitatively, the conclusions remain the same.

It is clear from the the results given in this section that relative costs will play a large consideration in whether food frequency questionnaires should be
used, especially when a satisfactory calibration study is done on an external population. For example, from figure 5 when an external study with \( n_e = 400 \) is available, the same standard error of the 80th percentile of usual intake is obtained using approximately 3,000 food frequency questionnaires or 200 food records, a 15:1 ratio. If the per person cost ratio of food records to food frequency questionnaires is greater than 15 to 1, then 3,000 food frequency questionnaires would be preferable to the 200 food records, because the same standard error could be obtained at less cost.

We summarize these considerations in figure 9, for external studies of size \( n_e = 200 \) or 400. The parameters used in the model 3-4 were the same as those reported earlier for the Women's Health Trial Vanguard Study. This figure gives the cost ratio line at which food frequency questionnaires and food records are equivalent in terms of the standard error for estimating the 80th percentile of usual intake, as a function of the number of food frequency questionnaires. Any cost ratio above the line favors food frequency questionnaires. For example, in an external study with \( n_e = 200 \) participants, if the primary study requires that 1,000 individuals complete a food frequency questionnaire, then food frequency questionnaires are preferred only if the cost ratio is above 10 to 1.

Some typical food record (one 4-day record per person) to food frequency questionnaire cost ratios range from 14 to 67, considering a range of current food frequency questionnaire costs in the United States between $3.00 and $8.00 per person and food record costs between $115.00 and $200.00 per person.
DISCUSSION

This paper is concerned with the use of food frequency questionnaires in population surveys for estimating usual intake distributions. We want to emphasize that our results do not have any direct bearing on the problem of relating nutrition to disease, which is an entirely different topic. It is noteworthy that food frequency questionnaires were developed mainly by nutritional epidemiologists for use in nutritional epidemiologic studies (13).

It is important to note that in this paper we have been assuming that internal calibration substudies are representative of the population to be surveyed. Thus, with an internal calibration substudy, the food records or recalls completed by the subjects give direct information concerning the distribution of intake among the population. We have shown that any number of extra subjects completing only food frequency questionnaires does little to add to the information provided by the food records or recalls of calibration substudy subjects. However, when there is an external calibration study, we have been assuming that the subjects in that study are not representative of the population to be surveyed. Consequently, the food records or recalls from the external study do not contribute direct information about the population intake but do still enable us to estimate the relation between food frequency questionnaire reported intake and the usual intake. In this case, when there is a large external calibration study already available, using the food frequency questionnaire in the new population survey for estimating usual intake distributions may sometimes be justified. However, it is important to emphasize that the use of external data must be considered carefully. One must assume that the relation between food frequency questionnaires, food records or recalls, and usual intake is exactly the same in the primary and external populations, and that only the mean and variance of usual intake may differ. This is an assumption that will often be difficult to justify on a priori grounds. Comparison of the estimates from the Women's Health Trial Vanguard Study with those from the Helsinki Dietary Pilot Study presented in this paper clearly indicates that these can differ, at least for the percentage of energy from fat. The difference may have been due to the different instruments used and/or the nature of the populations surveyed, which differed in sex and nationality.

In practice, the assumption that the subjects in an internal calibration substudy are representative of the population may not always be tenable. For example, those who would consent to complete food records...
or recalls may form a special subgroup of the population, perhaps with more education or affluence and, hence, possibly with different dietary intake. Such internal calibration substudies have to be thought of, in our terminology, as external calibration studies, since the completed food records or recalls do not provide direct information on the population distribution of intake.

In view of the above comment, one may wonder whether it is better to have an external calibration study (or a nonrepresentative internal calibration study) together with a food frequency questionnaire survey of the population, or simply an internal calibration study representative of the population and no food frequency questionnaire survey. Figure 6 shows that an external calibration study with 500 subjects together with a food frequency questionnaire survey of 4,000 subjects gives a standard error for the 80th percentile of approximately 0.65; this compares with a standard error of 0.36 for a representative internal calibration study with 500 subjects (table 1). Thus, a representative internal calibration study provides a considerably smaller standard error and is therefore to be preferred.

Our general conclusion is that to use food frequency questionnaires in surveys that aim to estimate the population distribution of usual intake, a large external calibration study must have already been carried out using the same instrument in a similar population. The most reasonable scenario might be where a series of surveys are to be conducted over time to monitor the dietary intake of a population. It may be reasonable to conduct a single large calibration study of a food frequency questionnaire at the outset and then repeated surveys using that food frequency questionnaire. Even in this case, one would need to remain alert to secular changes in the food supply that may change the regression relation between the food frequency questionnaire and usual intake.

In summary, we concur with the remarks of Briefel et al. (2) regarding the use of food frequency questionnaires in population surveys for measuring distributions of usual intake, with the modification that such use may be justified in some situations, especially after considering cost, as outlined above. We reiterate that our comments neither address nor challenge the use of food frequency questionnaires for epidemiologic studies of the association between nutrition and disease.

ACKNOWLEDGMENTS

R. J. C.'s research was supported by a grant from the National Cancer Institute (CA-57030).

REFERENCES

APPENDIX

Computation of Asymptotic Standard Errors

The model

Here we discuss the statistical calibration model (10) in full generality. In a typical calibration study, an individual is surveyed at time points denoted by \( j \), and either \( Q \), or \( F \), or both are observed. The model relating \( Q \) and \( F \) is a standard linear errors-in-variables model, namely

\[
Q_j = \beta_0 + \beta_1 T + r + \epsilon_j; \quad j = 1, \ldots, m_1 + m_2; \tag{3}
\]

\[
F_j = T + U_j; \quad j = m_1 + 1, \ldots, m = m_1 + m_2 + m_3. \tag{4}
\]

This model is explained as follows. There are \( m_1 \) time points at which only \( Q \) is measured, these denoted by \( j = 1, \ldots, m_1 \). There are \( m_2 \) time points \( (j = m_1 + 1, \ldots, m_1 + m_2) \) at which \( Q \) and \( F \) are measured. Finally, there are \( m_3 \) time points \( (j = m_1 + m_2 + 1, \ldots, m = m_1 + m_2 + m_3 = m) \) at which only \( F \) is measured.

For example, the text describes a sampling situation where a food frequency questionnaire is obtained initially, and some months later a food record or recall is obtained, followed by a second food record or recall obtained not too soon thereafter. Then, \( m_1 = 1, m_2 = 0, \) and \( m_3 = 2. \)

In model 3, the term \( r \) is called the equation error in the errors-in-variables literature (14), and its implication is that, even if we could have many questionnaires on each individual, the sample means of these questionnaires would not all fall on a straight line when regressed against usual intake. Another way to think of \( r \) is as the residual error from the regression line when there is no within-individual variation. The terms \( \epsilon_j \) represent the within-individual variation in food frequency questionnaires.

An important feature to keep in mind is the possible correlation in the measurement errors \( \epsilon_j \) and \( U_j \) when a questionnaire is given nearly coincidentally in time to a record or recall. The main text describes a sampling situation where these errors are uncorrelated.

Among these random variables, \( T \) has mean \( \mu_T \) and variance \( \sigma_T^2 \); \( U_j \) has mean zero and variance \( \sigma_U^2 \); \( \epsilon_j \) has mean zero and variance \( \sigma^2_\epsilon \); \( r \) has mean zero and variance \( \sigma^2_r \), and \( \sigma_{ue} \) is the intraindividual covariance \( \text{cov}(\epsilon_j, U_j) \) between food record and questionnaire intakes measured at the same point in time. All other random variables are uncorrelated.

The model 3-4 has been extended (15, 16) to account for possible systematic trends in measurements over time and to account for possible covariation in errors among questionnaires and/or food records/recalls.

Internal calibration

When all random variables are normally distributed, it is possible to compute asymptotic standard errors for estimates of the mean, variance, and percentiles of the distribution of usual intake \( T \), by using classical ideas of Fisher information theory. First, consider an internal calibration study. The joint distribution of any possible combination of \( Q \)s and \( F \)s can be described by the following parameters.

\[
\begin{align*}
\theta_1 & = \beta_0 + \beta_1 \mu_T; \quad \theta_2 = \mu_U; \\
\theta_3 & = \sigma_T^2; \quad \theta_4 = \beta_1^2 \sigma_T^2 + \sigma_r^2 + \sigma_\epsilon^2; \\
\theta_5 & = \beta_1 \sigma_T^2 + \sigma_\epsilon^2; \quad \theta_6 = \beta_1 \sigma_T^2 + \sigma_{ue}; \\
\theta_7 & = \beta_1 \sigma_T^2; \quad \theta_8 = \sigma_r^2 + \sigma_u^2.
\end{align*}
\]

The meaning of these parameters is that \( \theta_1 \) is the mean of any \( Q \), \( \theta_2 \) is the mean of any \( F \), \( \theta_3 \) is the covariance between any two \( F \)s, \( \theta_4 \) is the variance of \( Q \), \( \theta_5 \) is the covariance between any two \( Q \)s, \( \theta_6 \) is the variance of \( F \), \( \theta_7 \) is the covariance between any \( Q \) and \( F \) measured contemporaneously, \( \theta_8 \) is the covariance between any \( Q \) and any \( F \) not measured contemporaneously, and \( \theta_9 \) is the variance of any \( F \). Depending on the sampling situation, not all of these parameters will be estimable. However, \( \mu_T \) and \( \sigma_T^2 \) can be estimated as long as some of the data have two \( F \)s measured on an individual.

Write \( \Theta_m = (\theta_1, \theta_2) \), the parameters depending on the means, and \( \Theta_\nu = (\theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) \), the parameters involving variances. If all the random variables are normally distributed, then \( Q_{ij} \) is normally distributed with mean \( \theta_1 \) and variance \( \theta_4 \). If we write \( Z \) to be the vector of data observed for the calibration study, \( Z \) is normally
distributed with a mean that we denote by \( \mu(\Theta_m) \) and a covariance matrix denoted by \( \Sigma(\Theta) \). Both \( \mu(\Theta_m) \) and \( \Sigma(\Theta) \) depend on the sampling scheme being used.

The total sample size is \( n \), all of which receive a food frequency questionnaire. The calibration data are of size \( n_R \). For notational convenience, we arrange the calibration data to be the first \( n_R \) observations, the remaining \( n - n_R \) observations consisting only of food frequency questionnaires. Because each \( Z_i \) is a multivariate normal random variable, it follows that, except for a constant, the loglikelihood for the observed data is

\[
l(\Theta_m, \Theta_s) = -(1/2) \sum_{i=1}^{n_R} \left[ \log|\Sigma(\Theta)| + (Z_i - \mu(\Theta_m))^\prime \Sigma^{-1}(\Theta_s)(Z_i - \mu(\Theta_m)) \right]
- (1/2) \sum_{i=n_R+1}^{n} \{\log(\theta_i) + (Q_{ii} - \theta_i)^2/\theta_i\}. \tag{6}
\]

In equation 6, \( |\Sigma(\Theta_s)| \) is the determinant of \( \Sigma(\Theta_s) \).

For example, if \( (m_1, m_2, m_3) = (0, 2, 0) \), then

\[
Z_i = (Q_{11}, Q_{21}, F_{11}, F_{12})';
\]

\[
\mu(\Theta_m) = (\theta_1, \theta_2, \theta_3)';
\]

\[
\Sigma(\Theta_s) = \begin{pmatrix}
\theta_1 & \theta_2 & \theta_3 \\
\theta_2 & \theta_4 & \theta_5 \\
\theta_3 & \theta_5 & \theta_6 \\
\theta_1 & \theta_4 & \theta_6 \\
\theta_2 & \theta_5 & \theta_6
\end{pmatrix}. \tag{7}
\]

As an alternative, in the case that the food records are taken well after the administration of a single questionnaire, consider \( (m_1, m_2, m_3) = (1, 0, 2) \), then

\[
Z_i = (Q_{11}, Q_{21}, F_{11})';
\]

\[
\mu(\Theta_m) = (\theta_1, \theta_2, \theta_3)';
\]

\[
\Sigma(\Theta_s) = \begin{pmatrix}
\theta_1 & \theta_2 \\
\theta_2 & \theta_3 \\
\theta_1 & \theta_3
\end{pmatrix}. \tag{8}
\]

Note that \( (\theta_5, \theta_6) \) do not appear in these expressions, so that this design does not allow them to be estimated. When translated back to the original model, this means that, with this design, neither \( \sigma^2 \) nor \( \sigma_{ae} \) can be estimated.

The information matrix \( \Im(\Theta_m, \Theta_s) \) for the parameters (17) is readily calculated analytically by using numeric differentiation to compute derivatives. For large \( n \) and \( n_R \), the covariance matrix of the maximum likelihood estimators is approximately \( (nR\Im(\Theta_m, \Theta_s))^{-1} \). To compute this information matrix, it is helpful to note that it necessarily has the form

\[
\Im(\Theta_m, \Theta_s) = \begin{bmatrix}
\Im_m(\Theta_m) & 0 \\
0 & \Im_s(\Theta_s)
\end{bmatrix}.
\]

Terms in \( \Im_s(\Theta_s) \) are calculated as follows. Fix any matrix \( A \), and for any \( (\theta_j, \theta_k) \) that are elements of \( \Theta_s \) (and these will coincide for the diagonal elements), let \( G_{jk}(\Theta_s, A) \) denote the second partial derivative of the trace of \( \Sigma^{-1}(\Theta_s) A \). Then, the corresponding entry in \( \Im_s(\Theta_s) \) is just

\[
(1/2)(n - n_R)\theta_k \frac{\partial^2}{\partial \theta_j \partial \theta_k} \left(1/\theta_j \right) + (1/2)(n - n_R) \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log(\theta_j) + (nR/2) \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log(\Sigma(\Theta_s))
+ (nR/2)G_{jk}(\Theta_s, A)|_{A=\Sigma(\Theta_s)}.
\]
Terms in $\mathbb{S}_m(\Theta_m)$ are calculated as follows. For $j, k = 1, 2$, the corresponding $(j, k)$th entry in $\mathbb{S}_e(\Theta_e)$ is

$$(1/2)(n - n_R)\theta_k^{-1} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \theta_k^2 + (n_R/2) \frac{\partial^2}{\partial\theta_j \partial\theta_k} [(Z - \mu(\Theta_m))^t \Sigma^{-1}(\Theta_e)(Z - \mu(\Theta_m))]_{Z=\mu(\Theta_e)}.$$

**External calibration**

With an external calibration study, the likelihood in the external study is the same as before, with the following exceptions. The parameters $\mu_e$ and $\sigma_e^2$ in equation 5 should be changed to $\mu_{e, e}$ and $\sigma_{e, e}^2$, reflecting that they are the mean and variance of usual intake in the external study. The sample size is no longer $n$, but it is instead $n_e$. The number of observations in the external calibration study that measure food records is now $n_{R, e}$.

In the primary study, there are two additional parameters, namely $\mu_t$ and $\sigma_t^2$, the mean and variance of usual intake. If the primary study consists of $n_F$ food records, define $\theta_9 = \mu_t$, $\theta_{10} = \sigma_t^2 + \sigma_u^2$. The added contribution to the likelihood in equation 6 is

$$-(1/2) \sum_{i=1}^{n_F} \{(log(\theta_{10}) + (F_i - \theta_9)^2)/\theta_{10}\}. \quad (7)$$

We can obtain $\mu_t = \theta_9$ and $\sigma_t^2 = \theta_{10} - (\theta_9 - \theta_3)$.

If the primary study consists of $n_Q$ questionnaires, define $\theta_9 = \beta_0 + \beta_1 \mu_t$, and $\theta_{10} = \beta_0^2 \sigma_t^2 + \sigma_u^2$, and add

$$-(1/2) \sum_{i=1}^{n_Q} \{log(\theta_{10}) + (Q_i - \theta_9)^2/\theta_{10}\} \quad (8)$$

to the likelihood in equation 6. We find that $\mu_t = -\theta_2 + (\theta_9 - \theta_1)\theta_3/\theta_2$ and $\sigma_t^2 = \theta_3 + (\theta_{10} - \theta_4)(\theta_3/\theta_2)^2$.

The Fisher information matrix is of the form

$$\mathbb{S}(\Theta_m, \Theta_e) = \begin{bmatrix} \mathbb{S}_m(\Theta_m) & 0 & 0 \\ 0 & \mathbb{S}_e(\Theta_e) & 0 \\ 0 & 0 & \mathbb{S}_e(\theta_9, \theta_{10}) \end{bmatrix}, \quad (9)$$

where $\mathbb{S}_m(\Theta_m)$ and $\mathbb{S}_e(\Theta_e)$ are calculated from the external study. For records/recalls,

$$\mathbb{S}_e(\theta_9, \theta_{10}) = \begin{bmatrix} n_F/\theta_{10} & 0 \\ 0 & n_F/(2\sigma_{10}^2) \end{bmatrix},$$

while for questionnaires

$$\mathbb{S}_e(\theta_9, \theta_{10}) = \begin{bmatrix} n_Q/\theta_{10} & 0 \\ 0 & n_Q/(2\sigma_{10}^2) \end{bmatrix}.$$