Empirical evidence of correlated biases in dietary assessment instruments and its implications

Victor Kipnis¹, Douglas Midthune¹, Laurence S. Freedman², Sheila Bingham³, Arthur Schatzkin⁴, and Raymond J. Carroll⁵
1 Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.

2 Department of Mathematics, Statistics and Computer Science, Bar Ilan University, Ramat Gan, Israel.

3 Medical Research Council, Dunn Human Nutrition Unit, Cambridge, United Kingdom.

4 Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

5 Department of Statistics, Texas A&M University, College Station, TX and Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA.

Reprint Requests: Dr. Victor Kipnis, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Executive Plaza North, Room 344, 6130 Executive Blvd., MSC 7354, Bethesda, MD 20892-7354.

Running Head: Nutrient Biomarkers and Dietary Assessment Instruments

Word Count: 4,992
Abstract

Multiple-day food records or 24-hour recalls are currently used as “reference” instruments to calibrate food frequency questionnaires (FFQs) and to adjust findings from nutritional epidemiologic studies for measurement error. The standard adjustment is based on the critical assumptions that the reference instrument is unbiased at the individual level, with errors independent of those in the FFQ. Using data on urinary nitrogen as a biomarker for nitrogen intake, together with conventional dietary assessment measurements, it is demonstrated that a self-report reference instrument does not meet the critical requirements. A new model is introduced that includes, for both the FFQ and the reference instrument, group-specific biases related to true intake and correlated person-specific biases. Using the biomarker measurements, the new model is compared with alternative measurement error models proposed in the literature and is demonstrated to provide the best fit to the data. The new model suggests that, for these data, measurement error in the FFQ leads to a 51% greater attenuation of true nutrient effect and the need for a 2.3 times larger study than would be estimated by the standard approach. The implications of the results for the ability of FFQ-based epidemiologic studies to detect important diet-disease associations are discussed.

Key Words: biomarkers; dietary assessment methods; epidemiologic methods; measurement error; models statistical; model selection; nutrient intake; regression analysis.

Word Count: 200
Scientists have long sought a connection between dietary factors and cancer. A number of large prospective studies have now challenged conventional wisdom, derived in large part from international correlation studies and animal experiments, in reporting no association between dietary fat and breast cancer (1) and, most recently, dietary fiber and colorectal cancer (2). These null epidemiologic findings may ultimately be shown to reflect the truth about these diet and cancer hypotheses. Alternatively, however, the studies themselves may have serious methodologic deficiencies.

Usually, in large studies, a relatively inexpensive and rapid method of measurement, such as a food frequency questionnaire (FFQ), is employed. Over the years, investigators have recognized that the reported values on FFQ’s are subject to substantial error that can profoundly affect the results and interpretation of studies in nutritional epidemiology (3-5). Usually, measurement error in an exposure variable attenuates (biases toward zero) the estimated disease risk for that exposure and reduces statistical power to detect an effect. An important direct relation between diet and disease, therefore, may be obscured.

Realization of this problem has prompted the integration of calibration sub-studies in large epidemiologic investigations that involve a more intensive, but presumably more accurate, dietary assessment method, called the “reference” instrument. A critical assumption underlying the use of such sub-studies has been that the reference instrument is unbiased at the individual level and contains only within-person random error. In other words, the average of multiple repeat reference measurements converges to the true long-term intake of each individual. Typically, the instruments chosen as the reference measurements have been multiple-day food records, sometimes with weighed quantities reported, or multiple 24-hour recalls. Firstly, FFQ’s have been “validated” against such instruments, and correlations between FFQ’s and reference
instruments have been quoted as evidence of FFQ validity (6), although more sophisticated approaches adjusting for within-person random errors in the reference instrument are available (7). Secondly, statistical methods have been used for adjusting FFQ-based relative risks for measurement error (8), using the regression calibration approach.

The standard application of the regression calibration approach relies on the requirement that errors in the reference instrument be uncorrelated with true intake and also with errors in the FFQ (9). This fits with the standard assumptions about a reference instrument, that it contains only within-person random errors that are indeed uncorrelated with true intake and with errors in the FFQ.

Recent evidence suggests that these assumptions may be unwarranted. Studies involving biomarkers, such as doubly-labeled water and urinary nitrogen (10-16), suggest that reports using food records or recalls are biased (on average towards under-reporting), and that individuals may systematically differ in their reporting accuracy. This may mean that all dietary self-report instruments involve systematic bias at the individual level. Part of the bias may depend on true intake (which manifests itself in what we call group-specific bias or what might be thought of as the flattened slope phenomenon), therefore violating the first crucial condition for the standard measurement error adjustment. Part of the bias may also be person-specific, a concept we define later in detail, and correlate with its counterpart in the FFQ, thereby violating the second crucial assumption underlying the standard approach.

For this reason, Kipnis et al. (9) proposed a new measurement error model that allows for person-specific bias in the reference instrument as well as in the FFQ. Using sensitivity analysis, they showed that if the correlation between person-specific biases in the FFQ and reference instrument was 0.3, or greater, the adjustment for measurement error in the FFQ calculated under
the standard assumptions would be seriously wrong. However, the paper presented no empirical evidence that such correlations exist in practice.

In this paper, we present a re-analysis of the calibration study conducted in Cambridge, UK, by the Medical Research Council’s (MRC) Dunn Nutrition Unit (17-19). The study represents a new type of a calibration study that employs biomarkers for assessing dietary intake as well as the conventional dietary instruments. In this study the biomarker is the level of urinary nitrogen excretion in a 24-hour period which has been shown to provide a truly unbiased measure of protein intake (20). The biomarker measurements allow us to generalize the model by Kipnis et al. (9) and further explore the structure of measurement error in dietary assessment instruments and its implications for nutritional epidemiology.

MODELS AND METHODS

Effect of measurement error

Consider the disease model

\[ R(D|T) = \alpha_0 + \alpha_1 T, \]  

where \( R(D|T) \) denotes the risk of disease \( D \) on an appropriate scale (e.g., logit or logarithmic) and \( T \) is true long-term usual intake of a given nutrient. The slope \( \alpha_1 \) represents an association between the nutrient intake and disease. Let \( Q = T + e_Q \) denote the nutrient intake obtained from a FFQ, where the difference between the reported and true intakes, \( e_Q \), defines measurement error. Note that short-term variation in diet is included in \( e_Q \), as well as systematic and/or random error components resulting from the instrument itself. We assume throughout that error \( e_Q \) is nondifferential with respect to disease \( D \); i.e., reported intake contributes no additional information about disease beyond that provided by true intake.
Fitting model 1 to observed intake $Q$, instead of true intake $T$, yields a biased estimate $\tilde{\alpha}_1$ of the exposure effect. To an excellent approximation (21), the expected observed effect is expressed as

$$E(\tilde{\alpha}_1) = \lambda_1 \alpha_1,$$  \hspace{1cm} (2)

where the bias factor $\lambda_1$ is the slope in the linear regression calibration model

$$T = \lambda_0 + \lambda_1 Q + \xi.$$  \hspace{1cm} (3)

Although, in principle, when measurement error $e_Q$ is correlated with true exposure $T$, $\lambda_1$ could be negative or greater than one in magnitude, in nutritional studies $\lambda_1$ usually lies between 0 and 1 (22) and can be thought of as an attenuation of the true effect $\alpha_1$.

Measurement error also leads to loss of statistical power for testing the significance of the disease-exposure association. Assuming that the exposure is approximately normally distributed, the sample size required to reach the requested statistical power for a given exposure effect is proportional to (22)

$$N \propto 1/\{\rho^2(Q,T)\sigma_T^2\} = 1/\{\lambda_1^2\sigma_Q^2\},$$  \hspace{1cm} (4)

where $\rho(Q,T)$ is the correlation between the reported and true intakes and $\sigma_T^2$ is the variance of true intake. Thus, the asymptotic relative efficiency of the “naïve” significance test, compared to one based on true intake, is equal to the squared correlation coefficient $\rho^2(Q,T)$.

**Standard measurement error adjustment**

Following formulas 2-3, the unbiased (adjusted) effect can be calculated as $\lambda_1^{-1}\tilde{\alpha}_1$, where $\hat{\lambda}_1$ is the estimated attenuation factor. Estimation of $\hat{\lambda}_1$ usually requires simultaneous evaluation of additional dietary intake measurements made by the reference instrument in a calibration
substudy. Ideally, such a reference measurement would be the “gold standard” representing true nutrient intake, but it does not exist in dietary studies. The standard approach in nutritional epidemiology, introduced and made popular by Rosner et al. (8), uses food records/recalls as reference measurements, assuming that they are unbiased instruments for true long-term nutrient intake at the personal level. More precisely, for person $i$ and repeat measurement $j$ the standard model can be expressed as

$$Q_i = T_i + e_{Q_i},$$

where it is assumed that

$$E(e_{Fij} | i) = 0,$$ \hspace{1cm} (7)

$$\text{cov}(e_{Fij}, e_{Fj'}) = 0, j \neq j',$$ \hspace{1cm} (8)

$$\text{cov}(e_{Fij}, e_{Q_i}) = 0.$$ \hspace{1cm} (9)

Based on recent evidence, we introduce below a new measurement error model and compare it to the standard model and its modifications using the MRC data.

**The MRC data**

The data come from a dietary assessment methods validation study carried out at the MRC Dunn Clinical Nutrition Center, Cambridge, UK (17). One hundred and sixty women aged 50-65 years were recruited through two general medical practices in Cambridge, UK. Subjects from practice 1 (group 1) were studied from October 1988- September 1989, and those from practice 2
(group 2) from October 1989- September 1990. The principal measures for this study were four-day weighed food records and two 24-hour urine collections obtained on each of four occasions (seasons) over the course of one year. Season 1 was October – January; season 2, February – March; season 3, April – June; and season 4, July – September.

The four-day weighed food record (WFR) was the primary dietary assessment instrument of interest. The weighed records were obtained using portable electronic tape recorded automatic scales (PETRA) that automatically record verbal descriptions and weights of food without revealing the weight to the subject. Each four-day period included different days chosen to ensure that all days of the week were studied over the year, with an appropriate ratio of weekend to weekdays.

The urine collections were checked for completeness by p-amino benzoic acid (PABA) and used to calculate urinary nitrogen (UN) excretion. When the body is in nitrogen balance, UN was demonstrated to provide an unbiased and relatively accurate biomarker for nitrogen or, which is essentially equivalent, protein intake (20). Subjects were asked to make the first 24-hour urine collection on the third or fourth day of their food record, and the second collection was usually made 3-4 days later.

Other dietary assessment methods investigated in the study included the Oxford FFQ, the Cambridge FFQ, two types of 24-hour recall, and three types of seven-day food record. Some of those additional methods were repeated on two or more occasions, although we only had access to data on their first administration.

In this paper, we study nitrogen intake, measured in g/day. We chose to analyze the Oxford FFQ as it was based on the widely used Willett FFQ (23), modified to accommodate the characteristics of a British diet. It was administered one day before the start of the four-day
weighed food record in season 3. We used the WFR, which was considered the best dietary assessment method in the study, as the reference instrument and the UN measurements as the biomarker. In all our analyses, we applied logarithmic transformation to data to achieve normality.

Check of standard reference instrument assumptions

As mentioned above, it is typically assumed that the reference instrument in a calibration study contains only within-person random error that is unrelated to true nutrient intake and is independent of error in the FFQ. In this section, we provide an indirect check of these assumptions for the WFR in the MRC data. A critical assumption in our analysis is that the biomarker is a truly unbiased measure of intake, which seems reasonable for the UN method.

Suppose that the standard assumptions for a reference instrument hold for the WFR. Then we would expect that using the standard approach (8) with the WFR as the reference instrument should lead to nearly the same estimated attenuation as using the UN as the reference instrument. We checked this assumption, finding that the former method yielded an estimated attenuation factor of 0.282, while the latter estimated it as 0.186; using the bootstrap, the difference between these two estimates is statistically significant (p = 0.022). This important finding means that the attenuation caused by measurement error in the FFQ is in fact more severe than it would appear when using the WFR as the reference instrument. One explanation of this result is that the WFR does not satisfy at least one of the two major requirements for a reference instrument, namely that its error is unrelated to true intake and is independent of error in the FFQ.

A new dietary measurement error model

Model for the FFQ. The error in a FFQ is thought likely to include a systematic within-person bias $b$ that may depend on the individuals’ true intake $T$, as well as within-person variation $\varepsilon$ (19,
21, 24), so that

\[ Q = T + e_Q = T + b + \varepsilon. \]

We model the relation between bias \( b \) and true intake \( T \) as the linear regression

\[ b = \beta_{Q0} + \beta_{Qi}^* T + r, \]

where \( r \) has zero mean and is independent of \( T \). The component \( E(b \mid T) = \beta_{Q0} + \beta_{Qi}^* T \) is common to all persons with the same true intake and may be called group-specific bias. It can be thought of as arising from correlation between error and true intake. For example, given the social/cultural pressure to follow the ‘correct’ dietary pattern, persons with a low intake of supposedly healthy food may be tempted to overreport their intake and those with a high intake of supposedly unhealthy food to underreport. In this case, as in many other instances, \( \beta_{Qi}^* \) is negative, giving rise to the flattened slope phenomenon in the regression of the reported on true intake \( E(Q \mid T) = \beta_{Q0} + (\beta_{Qi}^* + 1)T \).

The difference \( r \) between within-person bias and its group-specific component varies from person to person, and may be caused by personality characteristics such as susceptibility to social/cultural influences. We will call it person-specific bias. Note that this error component is part of within-person systematic error and will be reproduced in repeated measurements on the same individual.

Gathering all the error components together, we model the FFQ intake \( Q_{ij} \) for individual \( i \) and repeat measurement (season) \( j \) as

\[ Q_{ij} = \mu_Q + \beta_{Q0} + \beta_{Qi} T_i + r_i + e_{ij}, \quad (10) \]
where $\beta_{Qi} = \beta_{Qj}^* + 1$. Term $\mu_{Qi}$ represents a possible seasonal effect at the population level, a factor that usually improves model fit (25). Throughout, any term $\mu_j$ refers to such a seasonal effect. If there is only one FFQ measurement per individual it will be impossible to disentangle person-specific bias $r_i$ and within-person error $\epsilon_{ij}$. This happens to be the case in the MRC study. We will therefore have to content ourselves with measuring the joint effect of these two types of error.

**Model for the reference instrument.** As argued previously, we need to allow for systematic group-specific and person-specific biases in self-report reference instruments. Thus, we now make the same assumptions regarding the error structure for the reference instrument as for the FFQ and use a model that is analogous to model 10.

In the MRC study each individual $i$ was requested to provide the WFR in each $j$ of the 4 seasons. We model these data as

$$F_{ij} = \mu_{Fj} + \beta_{F0} + \beta_{F1j}T_i + s_i + u_{ij}, \quad i=1,2,\ldots,n; \quad j=1,2,3,4; \quad (11)$$

where $\beta_{F0} + \beta_{F1j}T_i$ indicates the possibility of group-specific bias, $s_i$ and $u_{ij}$ denote person-specific bias and within-person random error, respectively, and are assumed to be independent of each other and of true intake $T_i$. As before, $\mu_{Fj}$ represents a seasonal effect at the population level.

Note that term $s_i$ in expression 11 is parallel to term $r_i$ in model 10 for the FFQ. Since the same personality characteristics underlie both person-specific biases, one may anticipate that the two will be correlated.
Model for the biomarker. As was mentioned above, it is reasonable to assume that the UN provides an unbiased biomarker for nitrogen intake at the individual level. The MRC study included two repeat UN measurements in each of the 4 seasons. Letting \( j \) denote season \((j=1,2,3,4)\), as before, and \( k \) denote the repeat measurement within the season \((k=1,2)\), we write this model as

\[
M_{ijk} = \mu_{Mj} + T_i + \nu_{ijk},
\]

where \( \nu_{ijk} \) denotes within-person random error independent of true intake \( T_i \), and \( \mu_{Mj} \) represents a seasonal effect at the population level.

Unlike model 10-11 for dietary assessment methods, which is not identifiable without biomarker data (9), model 12 is fully identifiable on its own. Fitting it to the MRC data has revealed that the within-person random errors \( \nu_{ijk} \) are mutually independent and have constant variances within, but not between seasons. In particular, season 2 has a different error variance than the other three seasons, which have similar variances, so that, denoting the variance of \( \nu_{ijk} \) by \( \sigma_{\nu_1}^2, \sigma_{\nu_2}^2, \sigma_{\nu_3}^2, \sigma_{\nu_4}^2 \neq \sigma_{\nu_2}^2 \).

Details of the Full Model. To complete the precise description of the full model, we state the following assumptions behind each of the equations 10-12.

- Usual intake \( T \) has mean \( \mu_T \) and variance \( \sigma_T^2 \), and is independent of person-specific biases and within-person random errors.

- The person-specific biases \( s_i \) and \( r_i \) are independent of all other random variables, have mean zero, correlation \( \rho(r,s) \), and variances \( \sigma_s^2 \) and \( \sigma_r^2 \), respectively.
• The within-person random errors $\varepsilon_{ij}$, $u_{ij}$, and $\nu_{ijk}$ are mutually independent except when the instruments are administered in the same season, in which case seasonal fluctuations in diet could lead to non-zero correlations $\rho(\varepsilon, u)$, $\rho(\varepsilon, \nu)$, and $\rho(u, \nu)$ among $\varepsilon_{ij}$, $u_{ij}$, and $\nu_{ijk}$.

• The variances of $\varepsilon_{ij}$ are constant and equal $\sigma_{\varepsilon}^2$.

• The variances of $u_{ij}$ are constant and equal $\sigma_u^2$.

Since there was only one application of the FFQ in the MRC study, we cannot estimate $\sigma_{\varepsilon}^2$ and $\sigma_r^2$ separately, but only their sum. Similarly, we can estimate the covariance between $r$ and $s$, and the correlation between $r + \varepsilon$ and $s$, but not the correlation between $r$ and $s$. The correlation between $r + \varepsilon$ and $s$ will be smaller than the correlation between $r$ and $s$, because $\varepsilon$ is independent of $s$.

**Alternative measurement error models**

Several alternatives to measurement model 10-12 have been proposed in the literature. The standard model 5-9 allows for any error structure in a FFQ, such as systematic bias or within-person random variation, but assumes no group- or person-specific biases in the reference instrument. In our notations, it means that $\beta_{F1} = 1; \sigma_s^2 = \rho(r, s) = 0$. In addition, within-person random errors in the FFQ and reference instrument are assumed uncorrelated ($\rho(\varepsilon, u) = 0$).

Freedman et al. (26) provided a generalization of the standard model. First, they introduced structure in the measurement error model for the FFQ by specifying (under different terminology) both group- and person-specific bias in addition to within-person random variation, so that their model is equivalent to model 10. Second, although their model retains the standard assumptions for the reference instruments ($\beta_{F1} = 1; \sigma_s^2 = 0$), it allows within-person random
errors in the FFQ and reference instruments to be correlated if both instruments are applied contemporaneously.

Kaaks et al. (24) were the first to introduce, at least conceptually, person-specific bias in the reference measurement. However, they assume no group-specific bias in it \((\beta_{r1} = 1)\) and, more importantly, no correlation between person-specific biases in the FFQ and reference instrument \((\rho(r, s) = 0)\).

Spiegelman et al. (27) suggested a different modification of the standard model. They assume no group-specific bias in the reference measurement \((\beta_{r1} = 1)\) and no person-specific biases in either the FFQ or reference instrument \((\sigma_r^2 = \sigma_s^2 = 0)\), but allow for correlated within-person random errors in both instruments, even when the two measures are taken far apart \((\sigma(r + \epsilon, s + u) = \rho(r, s)\sigma_\epsilon\sigma_u)\).

Kipnis et al. (9) introduced a model that includes group-specific bias in the FFQ related to true intake and correlated person-specific biases in both the FFQ and reference instruments. Their model is described by equations 10-11 with the restriction that \(\beta_{r1} = 1\).

For comparison purposes, we slightly modified the above models by adding term \(\mu_{Fj}\) to represent a possible seasonal effect in the WFR. Also, we included the UN measurements that were modeled by equation 12. Note that all the above models are nested within model 10-12, i.e., included as special cases of this model. To test the significance of the correlation between person-specific biases in the FFQ and the WFR, we also included in the comparison a version of model 10-12 with \(\rho(r, s) = 0\).

Plummer and Clayton suggested a quite general model (19, model II(c)) that includes our model as a special case. They do not consider person-specific biases but allow group-specific
biases to vary in repeat administrations of the same instrument. In addition, within-person random errors are assumed freely correlated, both in repeat administrations of the same instrument and across instruments, with the exception of errors in the biomarker. The latter are assumed to be freely correlated with errors in repeat biomarker measurements, as well as errors in dietary assessment instruments, that are administered in the same season, but to be independent of measurements taken in different seasons.

Prentice (28) suggested a model that is close to ours in spirit. It allows for group-specific bias in the FFQ, although not the reference instrument, which is related to true intake and correlated person-specific biases in both the FFQ and the reference instrument. In addition, all model parameters are allowed to depend on body mass index (BMI) tertiles, making the model more general at the expense of tripling the number of parameters. However, this model involves an implicit assumption that $\rho(r, s) = \sigma_r / \sigma_r$ (9). This condition imposes a restriction on the joint distribution of person-specific biases in the FFQ and the reference instrument that may not accord with empirical fact.

MODEL COMPARISON USING MRC DATA

Model comparison criteria

All models mentioned above were fitted to the MRC data by the method of maximum likelihood under multivariate normality, a reasonable assumption after the log transformation, and compared using three methods. First, we tested the models’ goodness-of-fit by comparing each model with the unstructured model using the likelihood ratio chi-square test. A model that fits the data should produce a non-significant p-value, thereby indicating that it does not explain the data statistically significantly worse than the most general model possible. We also applied the likelihood ratio test to evaluate statistically significant differences in model fit for nested
models. In addition, all models were compared using two standard model selection criteria, namely the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) (29), defined as

\[
AIC = \log(likelihood) - d ; \\
BIC = \log(likelihood) - \log(n) \times d / 2 ,
\]

where \(d\) is the number of parameters and \(n\) is the sample size. Larger values of AIC and BIC are desirable. Both AIC and BIC penalize more complex models: the “best” models chosen by BIC tend to be simpler than those chosen by AIC.

Model comparison results

The results of model comparison are given in table 1. Ideally, one aims to find a model that passes the goodness-of-fit test, is not statistically significantly different from any more complex model, provides statistically significantly better fit than all models nested within it, and has the highest AIC and BIC scores among all models. For the MRC data, model 10-12 emerges as best by these criteria. First, it is one of only four models, together with its two simplified versions and the model of Plummer and Clayton, to pass the goodness-of-fit test. Second, it does not fit the data statistically significantly different than the model of Plummer and Clayton in which it is nested. The likelihood ratio chi-squared statistic comparing the two models is \(38.8 = 1173.2 - 1134.4\), based on \(37 = 56 - 19\) degrees of freedom, and hence produces a non-significant p-value of 0.38. Third, model 10-12 provides a statistically significantly better fit \((p \leq 0.0011)\) than any model nested in it. For example, comparing it with its version with uncorrelated person-specific biases, the likelihood ratio chi-square statistic is \(10.7 = 1134.4 - 1123.7\), based on \(1 = 19 - 18\) degree of freedom, with a p-value of 0.0011. This indicates that there is a non-zero correlation.
between person-specific biases in the FFQ and WFR. Lastly, model 10-12 has the highest AIC and BIC values among all models.

These results suggest that group and person-specific biases exist in both the FFQ and WFR, and that these person-specific biases are indeed correlated. As was mentioned above, in the absence of repeat measurements on the FFQ, we cannot estimate this correlation directly. However, it is estimated as at least 0.35 (the low bound for $\rho(r,s)$ corresponding to $\sigma^2 = 0$) and may be considerably higher. For example, if the variance of the person-specific bias is the same for the FFQ and WFR, then this correlation is estimated as 0.81.

**Effect of body mass index**

To explore a possible effect of body composition on biases in dietary assessment instruments, we added BMI as a covariate to equations 10-11 in our model, as well as to other models. Typically the BMI term was statistically significant for the WFR, although not for the FFQ, but the estimated parameters of interest remained essentially unchanged.

The Prentice idea refits the models separately within each BMI tertile, thus allowing for intake-BMI-error interactions. Since the Prentice model neither contains other models nor is it nested within any of them (including even the unstructured one), to provide a better comparison we also fit model 10-12 to each tertile. Stratification by BMI tertiles triples the number of parameters in the models. As is demonstrated in the last two rows of table 1, the result is to decrease AIC and BIC to values below that of our original model. This does not invalidate the idea that BMI might have an impact on the parameter estimation: AIC and BIC penalize such complex models severely, and the power to detect differences among models caused by tertile effects is not great. Our model fit to each tertile includes the Prentice model as a special case and can be compared to it by the likelihood ratio test. As follows from table 1, our model provides a
statistically significantly better fit than the Prentice model (p=0.003), has higher AIC but lower BIC scores.

**Attenuation of estimated effect and statistical power**

Table 2 displays the estimates of the most interesting parameters for model 10-12 and the standard model. They include the attenuation factor $\lambda_1$, the variance of true intake $\sigma_T^2$, the correlation $\rho(Q,T)$ between the FFQ and true usual intake, and the slopes $\beta_Q$ and $\beta_F$ that represent group-specific biases in the FFQ and WFR, respectively. For all parameters, except $\sigma_T^2$, there are major differences between model 10-12 and the standard approach. First, the slope of the regression of the WFR on true intake, $\beta_F$, assumed to be 1 in the standard approach, is estimated as 0.766 in our model demonstrating the flattened slope phenomenon in the reference instrument. Also, the standard approach suggests that the slope in the regression of the FFQ on true intake, $\beta_Q$, is 0.661 and the correlation $\rho(Q,T)$ between the FFQ and true usual intake is 0.432, while our model estimates them as 0.430 and 0.284, respectively, indicating much less accuracy.

The major parameter controlling the ability to detect disease-nutrient relationships using a FFQ is the attenuation factor $\lambda_1$. The standard approach yields the attenuation factor of 0.282, while our model estimates it as 0.187. Since the true effect of an exposure is calculated as the observed effect divided by the attenuation factor, our model suggests that the true effect would be 51% greater than the one estimated by the standard approach. There is also a much greater impact on the design of epidemiologic studies. As follows from formula 4, for any two models, the ratio of the sample sizes for the same required statistical power is the same as the squared ratio of their attenuation factors. Thus, our model suggests that the study size based on the
standard model should be increased by the factor $(0.282 / 0.187)^2 = 2.3$, i.e., studies would have to be more than twice as large as suggested by the standard model in order to maintain a nominal power.

DISCUSSION

Our purpose has been to propose a statistical framework for evaluating the common dietary-assessment reference instruments (multiple-day food records, 24-hour recalls), and to use this framework to evaluate the WFR as a reference instrument for nitrogen intake using the MRC data. We have demonstrated that our model produces the best fit to these data when compared to several other models proposed in the literature:

- It is not statistically significantly different from any other more complex model.
- It provides a statistically significantly better fit than the simpler models, which are special cases of it.
- It has the highest values of AIC and BIC, two numerical measures of model fit.

Our statistical framework is particularly important because it allows evaluation of three major common assumptions about a dietary assessment reference instrument: (a) no bias at the personal level; (b) no correlation of measurement error with that of the FFQ; and (c) no correlation between measurement errors in repeat administrations. In the MRC data, our results demonstrate that all three assumptions are violated due to presence of both group- and person-specific biases in the WFR and the correlation of the person-specific bias with that in the FFQ.

We have thus demonstrated that, at least for these data, the WFR is a flawed reference instrument. There still remains the question: do these flaws translate into anything of importance? We believe that they do. As was shown above, using the standard approach yields the estimated attenuation factor of 0.282, but it is estimated as 0.187 when using the new model.
Also, the correlation between the FFQ-based nitrogen intake and true intake is estimated as 0.432 by the standard approach and as only 0.284 by the new model. Besides the fact that the correlation between the reported and true intakes is being used as a measure of the FFQ validity, its squared value represents the loss in statistical power to test the significance of disease-exposure association. Thus, for these data, the real effect of measurement error in the FFQ is a far greater (51%) attenuation and a far greater loss of power (52%) for testing the true effect than would be estimated by the standard approach.

Whenever an attenuation factor is evaluated in a calibration study, one can estimate the sample size necessary to achieve a required level of statistical power to detect a given nutrient-disease relationship. Our results indicate that using the standard approach may lead to unexpectedly under-powered studies. For the MRC data, our model suggests the need for a study 2.3 times larger than would have been designed had the standard approach been used.

In summary, our results suggest that the impact of measurement error in dietary assessment instruments on the design, analysis, and interpretation of nutritional studies is much greater than has been previously suspected, at least regarding protein intake. Both the attenuation of nutrient effect and the loss of statistical power in FFQ-based epidemiologic studies may be far more profound than has been estimated due to flaws in the standard use of dietary assessment reference instruments. This means that current and past studies may be under-powered, and that the null results in nutritional epidemiology are inevitable. There is quite obviously a need to confirm these results by further studies, if for no other reason than to confirm that they are not peculiar to the MRC data.

Equally important is the need for simultaneous consideration of energy intake, using a biomarker such as doubly labeled water (10). Our results are specific to protein intake, and need
not necessarily apply to either protein density (percent of energy due to protein intake) or to energy adjusted protein intake (6). As was reported before (30), the effect of measurement error in energy-adjusted models may be much more complex than in the univariate analysis. The real effects may be hidden, observed data may show effects that do not really exist, and even the directions of effects may be wrong. Thus, it is entirely conceivable that adjustment for energy intake could lead to different statistical properties. Clearly, studies are needed in which questionnaires, various reference instruments, and biomarkers for protein and energy intakes are all collected and analyzed simultaneously.

Acknowledgments

Carroll’s research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-E509106).
REFERENCES


Table 1: The results of model comparison using the MRC* data.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mathcal{L}_2 \times LL$</th>
<th>df</th>
<th>$p$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td>-1222.3</td>
<td>104</td>
<td>-</td>
<td>507.2</td>
<td>224.7</td>
</tr>
<tr>
<td>Plummer and Clayton II(c)</td>
<td>-1173.2</td>
<td>56</td>
<td>.426</td>
<td>530.6</td>
<td>378.5</td>
</tr>
<tr>
<td>New model (equations 10-12)</td>
<td>-1134.4</td>
<td>19</td>
<td>.393</td>
<td>548.2</td>
<td>496.6</td>
</tr>
<tr>
<td>New model with $\rho(r,s) = 0$</td>
<td>-1123.7</td>
<td>18</td>
<td>.167</td>
<td>543.9</td>
<td>495.0</td>
</tr>
<tr>
<td>Kipnis et al., 1999</td>
<td>-1122.2</td>
<td>18</td>
<td>.142</td>
<td>543.1</td>
<td>494.2</td>
</tr>
<tr>
<td>Kaaks et al.</td>
<td>-1112.4</td>
<td>17</td>
<td>.049</td>
<td>539.2</td>
<td>493.0</td>
</tr>
<tr>
<td>Spiegelman et al.</td>
<td>-1058.0</td>
<td>17</td>
<td>&lt;0.001</td>
<td>512.0</td>
<td>465.8</td>
</tr>
<tr>
<td>Freedman et al.</td>
<td>-1050.1</td>
<td>16</td>
<td>&lt;0.001</td>
<td>509.1</td>
<td>465.6</td>
</tr>
<tr>
<td>Standard (Rosner et al.)</td>
<td>-1050.1</td>
<td>15</td>
<td>&lt;0.001</td>
<td>510.1</td>
<td>469.3</td>
</tr>
<tr>
<td>New model stratified by BMI tertiles</td>
<td>-1183.9</td>
<td>57</td>
<td>-</td>
<td>535.0</td>
<td>380.2</td>
</tr>
<tr>
<td>Prentice</td>
<td>-1164.0</td>
<td>51</td>
<td>-</td>
<td>531.0</td>
<td>392.6</td>
</tr>
</tbody>
</table>

* MRC, Medical Research Council, Dunn Human Nutrition Unit, Cambridge, United Kingdom, dietary assessment methods validation study (1988-1990)
† $\mathcal{L}_2$ times the log likelihood
‡ Number of parameters
§ $p$-value for goodness-of-fit test relative to the unstructured model
¶ Akaike Information Criterion for model selection
# Bayes Information Criterion for model selection
Table 2. Estimated parameters for the new and standard models using the MRC* data

<table>
<thead>
<tr>
<th>Model</th>
<th>Attenuation factor $\lambda_i$</th>
<th>$\sigma_r^2$</th>
<th>$\rho(Q,T)$</th>
<th>$\beta_{Q1}$</th>
<th>$\beta_{F1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>.187 (.056)†</td>
<td>.031 (.004)</td>
<td>.284 (.082)</td>
<td>.430 (.129)</td>
<td>.766 (.066)</td>
</tr>
<tr>
<td>Standard</td>
<td>.282 (.054)</td>
<td>.030 (.004)</td>
<td>.432 (.076)</td>
<td>.661 (.131)</td>
<td>1</td>
</tr>
</tbody>
</table>

* MRC, Medical Research Council, Dunn Human Nutrition Unit, Cambridge, United Kingdom, dietary assessment methods validation study (1988-1990)
† Standard deviation in parentheses.