American Journal of Epidemiology
Copyright © 2001 by the Johns Hopkins University Bloomberg School of Public Health
All rights reserved

Vol. 154, No. 9
Printed in U.S.A.

Measurement Error Correction in Nutritional Epidemiology based on Individual Foods, with Application to the Relation of Diet to Breast Cancer

Bernard Rosner and Rebecca Gore

Nutrient intake is often measured with error by commonly used dietary instruments such as the food frequency questionnaire (FFQ) or 24-hour recall. More accurate assessments of true intake are obtained by using weighed diet records, in which subjects record what they eat on a real-time basis, but these records are expensive to administer. Validation studies are often performed to relate "gold standard" intake to intake according to surrogate instruments and to correct relative risk estimates obtained in the main study for measurement error. Most measurement error correction methods use validation study data at the nutrient level. However, subjects almost always report intake at the food rather than the nutrient level. In addition, the validity of measurement of different foods can vary considerably; it is relatively high for some foods (e.g., beverages) but relatively low for others (e.g., meats, vegetables). This differential validity could be incorporated into measurement error methods and potentially improve on nutrient-based measurement error methods. In this paper, the authors discuss correction methods for food-based measurement error and apply them to study the relation between FFQ intake in 1980 and incident breast cancer in 1980–1994 among approximately 89,000 women in the Nurses' Health Study, in whom approximately 3,000 incident breast cancers were observed. *Am J Epidemiol* 2001;154:827–35.

breast neoplasms; diet records; food; measurement error; nutrition; validity

The role of diet in disease processes is of great scientific interest. One limitation in assessing associations between diet and disease is the often-large measurement error in reported dietary intake, which arises from two major sources. First, there is random within-person variation in reported dietary intake based on commonly used instruments such as a food frequency questionnaire (FFQ) or a 24-hour recall. Second, even if a surrogate instrument (e.g., FFQ) were perfectly reproducible, it might not be a valid measure of true dietary intake as might be captured in a weighed diet record, in which subjects record what they eat on a real-time basis.

To address the validity issue, it is becoming common to conduct a validation study. A small subset of persons, ideally from the same population as the main study, are administered both the surrogate instrument (e.g., FFQ) and a "gold standard" instrument (e.g., diet record) and the relation between them is ascertained.

Suppose the goal of the analysis is to estimate the relation between a dichotomous disease variable D and true intake x, where x is continuous, based on the model

$$\log[p/(1-p)] = \alpha + \beta x \qquad (1)$$

One method for estimating β is to use the regression calibration approach, in which true intake ($\hat{x}$) is estimated as a function of surrogate intake (X) based on a regression function derived from validation study data and a logistic regression analysis of disease is run on $\hat{x}$ rather than actual true intake (x). It can be shown in the special case of univariate logistic regression with a single exposure measured with error, if x is linearly related to X, then the estimated β is given by

$$\hat{\beta} = \hat{\beta}^{*}/\hat{\lambda} \qquad (2)$$

where $\hat{\beta}^{*}$ is the estimated logistic regression coefficient of D on X from the main study, and $\hat{\lambda}$ is the estimated regression slope of x on X from the validation study (1). Typically, x and X are nutrients of interest, such as dietary fat or alcohol.

One issue is that subjects do not directly report nutrients but instead report intakes of individual foods from which nutrients are indirectly calculated by using food composition databases. In addition, the validity of individual food items can vary considerably, with some types of foods being reported with high validity (e.g., beverages) and other types with low validity (e.g., vegetables, meats) (2). This differential validity should be taken into account when correcting for measurement error, so that food items measured with high validity are given more weight than those measured

Received for publication March 21, 2000, and accepted for publication March 19, 2001.

Abbreviations: CI, confidence interval; FFQ, food frequency questionnaire; OR, odds ratio; Q, quintile.

From the Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

Reprint requests to Dr. Bernard Rosner, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115 (e-mail: bernard.rosner@channing.harvard.edu).

with low validity when true nutrient indices are estimated on the basis of validation study data.

There are two major goals of this paper. First, we develop a measurement error model by using the validation study population to predict diet record food intake as a function of FFQ food intake. We also obtain an estimate of diet record nutrient intake and compare the accuracy of these estimates when food-based versus nutrient-based validation study models are used. Second, we model disease incidence as a function of estimated diet record food intake and also (separately) as a function of estimated diet record nutrient intake. We then apply these methods to prospective breast cancer data from the Nurses' Health Study.

MATERIALS AND METHODS

Statistical methods

Suppose there are *N* main study subjects, all of whom are disease free at baseline. At baseline, each subject fills out an FFQ with *J* food items and is followed for disease outcome *D* over some time period *t*. A subset, N_1, of the main study population participates in a validation study at baseline and provides diet record information on weighed consumption of the same *J* food items each day during 4 weeks spaced approximately 3 months apart over a 1-year period. In addition, a repeat FFQ is administered at the end of the 1-year period so that the time frame (regarding consumption during the previous year) for the FFQ and the diet record is the same. Finally, assume there are *R* variables measured without error among validation study subjects.

Let 1) f_{ij} = intake of the *j*th food recorded on the diet record from the *i*th validation study subject, $j = 1, \ldots, J$; $i = 1, \ldots, N_1$; 2) F_{ij} = intake of the *j*th food noted on the repeat FFQ from the *i*th validation study subject, $j = 1, \ldots, J$; $i = 1, \ldots, N_1$; and 3) Z_{ir} = the *r*th variable measured without error from the *i*th validation study subject, $r = 1, \ldots, R$; $i = 1, \ldots, N_1$.

We consider the following prediction model for diet record intake:

$$\log f_{ij} = \alpha_j + \sum_{m=1}^{J} \lambda_{jm} \log F_{im} + \sum_{r=1}^{R} \theta_{jr} Z_{ir} + e_{ij},$$

$$j = 1, \ldots, J; \quad i = 1, \ldots, N_1, \tag{3}$$

where $e_{ij} \sim N(0, \sigma_j^2)$, $j = 1, \ldots, J$.

Note that equation 3 assumes normality of log diet record intake ($\log f_{ij}$) conditional on log FFQ intake ($\log F_i$) = ($\log F_{il}, \ldots, \log F_{iJ}$) and Z_i as opposed to marginal normality of $\log f_{ij}$. The latter is unlikely to hold for many foods for which the distributions are typically asymmetric, highly skewed, and nonnormal. Equation 3 also assumes that the association between $\log f_{ij}$ and $\{\log F_{im}, Z_{iR}\}$ is linear. We used log transformations because the linearity assumption in equation 3 was better satisfied on the basis of examination of scatter plots when this scale was used. In addition, use of logs ensures that the estimated diet record food intake values are positive.

Using equation 3, we can estimate "true" (diet record) intake of the *j*th food for the *i*th main study subject (f_{ij}) conditional on F_i and Z_i by

$$\hat{f}_{ij} = \exp\left(\hat{\alpha}_j + \sum_{m=1}^{J} \hat{\lambda}_{jm} \log F_{im} + \sum_{r=1}^{R} \hat{\theta}_{jr} Z_{ir}\right) \tag{4}$$

If we estimate f_{ij} by $\hat{f}_{\mathrm{ij}}$, we can relate *D* to "true" intake of food *j* by using the logistic regression model

$$\ln[p_i/(1 - p_i)] = \alpha_j + \beta_j \hat{f}_{ij} + \delta' U_i \tag{5}$$

where $p_i = \Pr(D_i = 1 | \hat{f}_{ij}, U_i)$, δ and U_i are (s × 1) vectors, and U_i are nondietary variables measured without error in the main study (which may or may not include *Z*), $i = 1, \ldots, N$.

In addition, suppose there are *Q* nutrients, $s_{i1}, \ldots, s_{iQ}$ and let v_{qj} = amount of the *q*th nutrient/serving of the *j*th food, $q = 1, \ldots, Q$; $j = 1, \ldots, J$, where the v_{qj} are obtained from food composition data and are assumed known without error. We can estimate "true" intake of the *q*th nutrient for the *i*th subject by

$$\hat{s}_{iq} = \sum_{j=1}^{J} v_{qj} \hat{f}_{ij}, \quad q = 1, \ldots, Q \tag{6}$$

and estimate the logistic regression model

$$\ln[p_i/(1 - p_i)] = \alpha_q^* + \beta_q^* \hat{s}_{iq} + \delta^{*\prime} U_i \tag{7}$$

We can also extend equations 5 and 7 to allow for multiple foods or nutrients, respectively, or a combination of foods and nutrients in the same model. In this case, the point estimates of regression coefficients from a single multivariate linear regression of $\log f_i$ = ($\log f_{i1}, \ldots, \log f_{iJ}$) on $\log F_i$ and Z_i are the same as from separate multiple linear regressions of each food ($\log f_{ij}$) on $\log F_i$ and Z_i based on equation 3. However, the standard errors of the regression coefficients will be different.

In general, obtaining standard errors of the regression coefficients in equations 5 and 7 is difficult analytically, in the case of both single or multiple foods/nutrients measured with error when diet record intakes are estimated by using equation 3. Instead, we use bootstrap methods, in which we perform the following:

1. First, obtain a random sample with replacement of size N_1 from the validation study population. Second, obtain a random sample with replacement of size *N* from the main study population. Third, repeat the analyses in equations 3–7.
2. Repeat step 1 *M* times and, in the case of nutrients, obtain *M* sample estimates of β_q^*, denoted by $\hat{\beta}_q^{*(m)}$, m = 1, . . . , M, and estimate $\mathrm{Var}(\hat{\beta}_q^*)$ from

$$\sum_{m=1}^{M} [\hat{\beta}_q^{*(m)} - \bar{\beta}_q^*]^2/(m - 1)$$

and the 95 percent confidence interval for β_q^* from $\bar{\beta}_q^* \pm 1.96\sqrt{\mathrm{Var}(\hat{\beta}_q^*)}$, where

$$\bar{\beta}_q^* = \sum_{m=1}^{M} \hat{\beta}_q^{*(m)}/M.$$

Similar analyses can also be run for foods based on equations 3–5.

Data analysis issues

Three data analysis issues arise in implementing the methods described in the Materials and Methods section of this paper. The first is to determine which among all available FFQ foods should be used in the measurement error model in equation 3. In the Nurses' Health Study, there were 54 foods available on both the FFQ and diet record in 1980 and 173 subjects in the validation study population. Hence, the number of variables was large relative to the number of subjects. To avoid "overfitting" the model, with resulting poor predictive power when applied to an external population (e.g., the main study), several more parsimonious approaches were considered.

One possible approach is to use only log F_{ij} as a predictor of log f_{ij}. This approach has the advantage of extreme parsimony but ignores the possible reduction in prediction error by using information on intake of related foods by the same person. For example, the best predictor of diet record intake of french fries was FFQ hamburger intake rather than FFQ french fries intake. Therefore, for each diet record food, the corresponding FFQ food and a set of nondietary covariates Z were forced into the model, and a stepwise-up regression was used to identify other foods that added significant ($p < 0.01$) predictive power to the starting model. Doing so typically resulted in prediction equations for individual diet record foods with one to five FFQ foods as predictors. One problem is that the stepwise approach may not always yield the "best" prediction model. Thus, to account for error in the variable selection process, we allowed possibly different variables to enter into the fit of the model in equation 3 for each bootstrap sample (refer to the third part of step 1 above).

A second issue is the strong assumption of linearity inherent in equations 5 and 7. Many foods and nutrients have skewed, nonnormal distributions, often with outlying values, and the assumption of linearity may be unrealistic. To address this issue, the predicted diet record foods ($\hat{x}_{ij}$) and nutrients ($\hat{s}_{iq}$) were grouped into quintiles, and the median value of each quintile was used instead of the raw values to fit the functions

$$\ln[p_i/(1 - p_i)] = \alpha_j^{(\mathrm{med})} + \beta_j^{(\mathrm{med})}\,\hat{x}_{ij}^{(\mathrm{med})} + \delta^{(\mathrm{med})\prime}U_i$$

$$\ln[p_i/(1 - p_i)] = \alpha_q^{*(\mathrm{med})} + \beta_q^{*(\mathrm{med})}\,\hat{s}_{iq}^{(\mathrm{med})} + \delta{*}^{(\mathrm{med})\prime}U_i \qquad (8)$$

where $\hat{x}_{ij}^{(\mathrm{med})}$ and $\hat{s}_{iq}^{(\mathrm{med})}$ are the quintile-specific median values corresponding to $\hat{x}_{ij}$ and $\hat{s}_{iq}$, respectively. In addition, categorical analyses were also performed by using dummy variables for quintiles based on predicted foods and nutrients.

A third issue concerns data quality. As a preliminary check of whether the diet record and FFQ were filled out appropriately, we computed total caloric intake based on the predicted intake of diet record foods by using equations 4 and 6 and included in the analysis only those subjects whose estimated daily total caloric intake was 600–4,200 calories. In addition, subjects who left more than 10 blanks on the FFQ were excluded. For the remaining subjects, a food for which there was a blank was assumed to represent essentially no intake and was coded as 0.001 servings per day. This value was chosen because, in residual analyses and for most foods, the linearity assumption in equation 3 was approximately satisfied by using this coding. Otherwise, the FFQ was coded in terms of servings/day as follows: ≥6 servings/day = 6.0; 4–6 servings/day = 5; 2–3 servings/day = 2.5; 1 serving/day = 1.00; 5–6 servings/week = 0.80; 2–4 servings/week = 0.43; 1 serving/week = 0.14; 1–3 servings/month = 0.07; and <1 serving/month = 0.001.

Sample

The population for the validation study consisted of 173 members of the Nurses' Health Study cohort. All women were aged 34–59 years in 1980, the same year in which they filled out an FFQ that included a list of 61 foods. Between June 1980 and June 1981, participants completed a 7-day dietary record four times at approximately 3-month intervals. A repeat FFQ was administered at the end of either the third or fourth week of dietary recording. Two of the 61 FFQ foods were omitted from the analysis because data from the diet record were not comparable (e.g., other fruits, home-fried food). Data on two additional foods (low-calorie carbonated drinks and artificial sweetener) were not available at the time of analysis. Furthermore, two separate FFQ questions on meat and on pie were each combined into single items for comparability with the diet record, as detailed by Salvini et al. (2). Finally, mean reported diet record intake of sweet potatoes was very low (0.003 servings/day) and was eliminated from the analysis. Thus, 54 foods were available for analysis.

VALIDATION STUDY ANALYSES

In a previous analysis of these data (2), the Pearson's correlation coefficient between diet record intake and FFQ intake for specific foods (based on the $\ln(x + 1)$ transformation) ranged from 0.08 (spinach) to 0.90 (tea) (mean, 0.52), implying a wide variation in the validity of reporting of individual food items. For illustration, table 1 presents the relation between diet record and FFQ intake for two foods for which the measured validity was high (coffee, wine) and two foods for which it was low (carrots, hamburger). Although there was an apparent relation between diet record and FFQ intake for all food items, the correspondence between median FFQ intake and actual mean diet record intake was much stronger for coffee and wine than for carrots and hamburger.

For each of the 54 foods included, we ran stepwise-up regressions with a p value for inclusion of <0.01, with

TABLE 1. Actual mean intake (servings/day) according to dietary records, by categories of food consumption reported in the second food frequency questionnaire, Nurses' Health Study, 1980–1981

FFQ* category	FFQ servings/day	Coffee		Wine		Carrots		Hamburger	
		Mean	No.	Mean	No.	Mean	No.	Mean	No.
Missing			0	0.41	1		0	0.09	1
<1/month	0.001	0.24	22	0.04	48	0.06	6	0.07	4
1–3/month	0.07	0.80	3	0.17	40	0.10	49	0.09	46
1/week	0.14	1.68	4	0.28	18	0.17	66	0.15	92
2–4/week	0.43	1.09	11	0.40	40	0.23	42	0.18	29
5–6/week	0.80	1.10	5	0.67	9	0.22	7	0.07	1
1/day	1.0	0.97	17	1.24	9	0.16	3		0
2–3/day	2.5	1.85	69	2.07	7		0		0
4–6/day	4.5	2.39	26	2.60	1		0		0
≥6/day	6.0	4.09	16		0		0		0

* FFQ, food frequency questionnaire.

ln(diet record intake) as the dependent variable, in which a starting model of ln(FFQ intake for the same food), age in 1980, body mass index in 1980, and smoking status in 1980 (ever/never) was used. These nondietary covariates were chosen because, in the preliminary analyses for some foods, they were significant predictors of ln(diet record intake) after adjustment for ln(FFQ intake). The results for the four foods listed in table 1 are shown in table 2.

The two FFQ foods for which validity was high (coffee and wine) had R^2 values of 0.63–0.64, while the two FFQ foods for which validity was low (carrots and hamburger) had R^2 values of 0.06–0.14—a major difference. For some food items, other foods on the FFQ were predictive of diet record intake. For example, fruit punch consumption on the FFQ was negatively associated with diet record coffee consumption, even after we controlled for FFQ coffee consumption; also, beef and pork consumption reported on the FFQ was negatively associated with diet record wine consumption, even after controlling for FFQ wine consumption. The latter association may reflect the possibility that both wine consumption and avoidance of beef are indicators of a "healthy lifestyle." The nondietary covariates considered were each significantly associated with 6–9 of the 54 diet record food items. Body mass index was negatively associated with wine consumption ($p = 0.017$) and showed a trend toward an inverse association with carrot consumption ($p = 0.088$), again possibly because of healthy-lifestyle associations.

Overall mean R^2 values for each of the 54 diet record food regressions are shown in figure 1. The overall results of these regressions indicated that the validity of measuring beverages and dairy products was generally high, while that for meats and vegetables was generally low.

The 54 food regression equations were then used to calculate predicted diet record food intake for each subject in the main Nurses' Health Study based on equation 4.

TABLE 2. Regression analyses of ln(diet record intake) on ln(food frequency questionnaire intake, same food), age, body mass index, smoking status, and ln(food frequency questionnaire intake of selected other foods), by stepwise regression analysis (p < 0.01), Nurses' Health Study, 1980–1981

	Diet record food							
	Coffee (servings/day)*		Wine (servings/day)*		Carrots (servings/day)*		Hamburger (servings/day)*	
	β	p value	β	p value	β	p value	β	p value
Constant	−1.421		−1.490		−2.616		−2.935	
Age (years)	0.015	0.20	0.017	0.28	0.022	0.17	0.027	0.14
Body mass index−23 kg/m^2	−0.008	0.71	−0.072	0.017	−0.050	0.088	−0.018	0.61
Smoking status (ever/never)	0.130	0.46	−0.213	0.37	0.000	1.0	−0.336	0.21
Same food FFQ*,†	0.516	<0.001	0.683	<0.001	0.426	<0.001	0.369	0.005
Fruit punch FFQ*	−0.127	0.003						
Beef, pork FFQ*			−0.391	0.001				
R^2	0.640		0.630		0.135		0.063	

* Natural logarithm scale, with reported 0 intake recoded as 0.001 servings/day.
† FFQ, food frequency questionnaire.

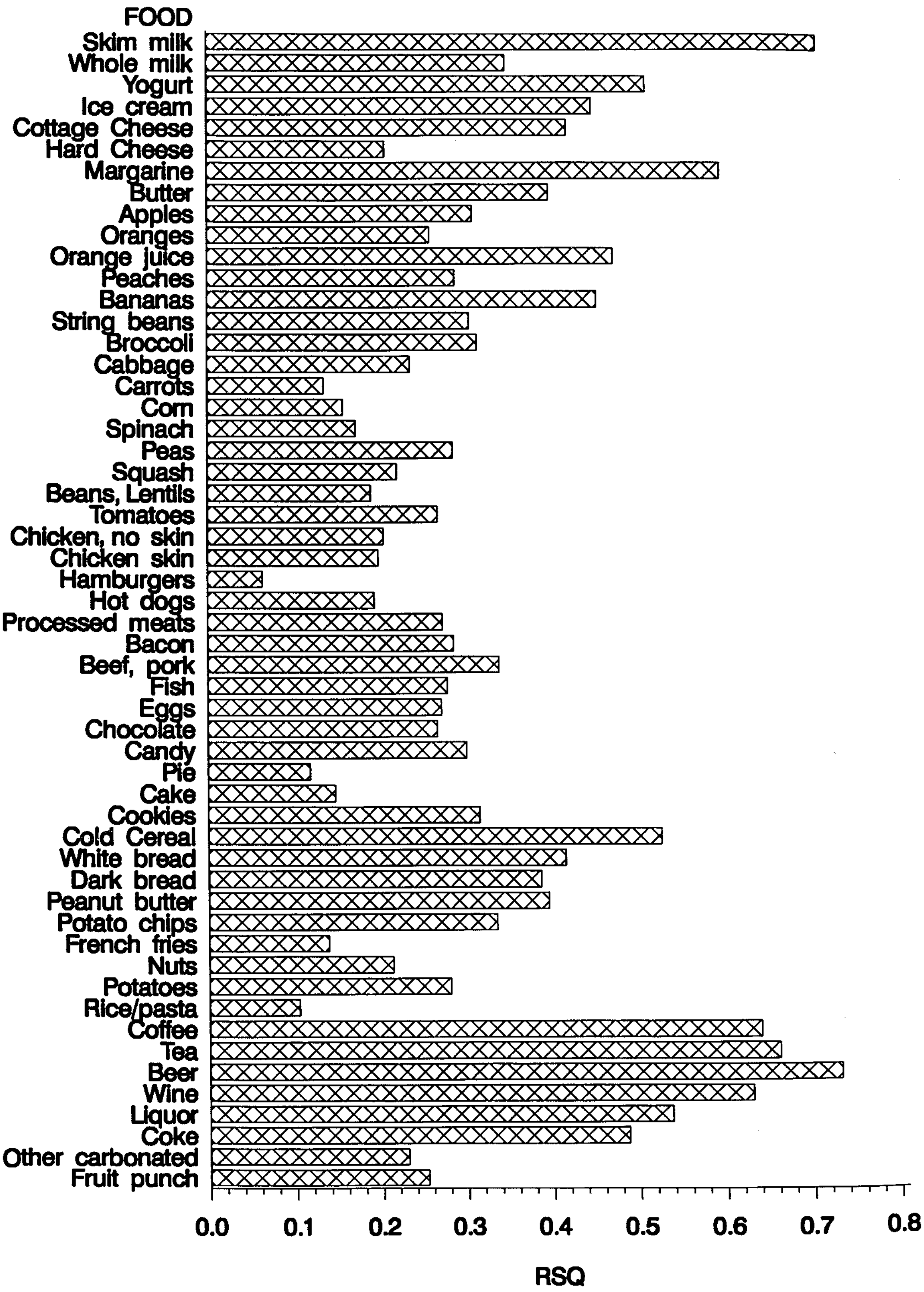

FIGURE 1. Plot of R^2 (RSQ) values for stepwise regression analyses for each of 54 diet record foods in a validation study (n = 173), Nurses' Health Study, 1980–1994.

Furthermore, from equation 6, the predicted foods were used to calculate predicted diet record nutrient intake as derived from these 54 foods for selected nutrients.

An assumption made in the Materials and Methods section of this paper was that for different foods reported on the FFQ, the validity may be different. Therefore, a more precise method of estimating diet record nutrient intake would be to weight foods that contribute to this nutrient differently according to their validity as indicated by using equation 6. A more standard approach would be to estimate diet record nutrient intake directly based on FFQ nutrient intake, irrespective of which foods the FFQ nutrient intake comes from. To implement this approach, we computed nutrient intake for the qth nutrient and ith subject based on the FFQ by

$$S_{iq} = \sum_{j=1}^{J} v_{qj} F_{ij} \qquad (9)$$

and then ran the regression

$$\log s_{iq} = \alpha + \gamma \log S_{iq} + \theta_1^* U_1 + \theta_2^* U_2 + \theta_3^* U_3 + e_{iq}^* \qquad (10)$$

where U_1 = age (years), U_2 = (body mass index − 23 kg/m^2), U_3 = smoking status (ever/never), and $e_{iq}^* \sim N(0, \sigma_q^{2*})$.

An alternative estimator of diet record nutrient intake is then

$$\hat{s}_{iq}^* = \exp(\hat{\alpha} + \hat{\gamma} \log S_{iq} + \hat{\theta}_1^* U_1 + \hat{\theta}_2^* U_2 + \hat{\theta}_3^* U_3). \qquad (11)$$

To compare the food-based ($\hat{s}_{iq}$, equation 6) with the nutrient-based ($\hat{s}_{iq}^*$, equation 11) estimator of diet record nutrient intake, we computed Spearman correlation coefficients between the actual diet record intake (s_{iq}) and each estimator for each of 37 nutrients available in the 1980 Nurses' Health Study database.

To assess statistical significance, we used the method of Wolfe (3) to compare dependent correlation coefficients. The results are given in table 3.

For 27 of the 37 nutrients, the food-based estimates had a higher correlation with actual diet record nutrient intake than the nutrient-based estimates. However, differences were usually, but not always, small. Significant differences were found for protein (food-based r = 0.467, nutrient-based r = 0.336; p = 0.027) and carotene intake (food-based r = 0.446, nutrient-based r = 0.368; p = 0.012). Nonsignificant trends in the same direction were also found for carbohydrates (r = 0.684 vs. r = 0.595, p = 0.081), sucrose (r = 0.627 vs. r = 0.545, p = 0.094), dietary fiber (r = 0.658 vs. r = 0.558, p = 0.065), and folate (r = 0.688 vs. r = 0.587, p = 0.065). Further inspection of protein intake indicated that major sources of protein were dairy foods, meat, poultry, and fish. On the FFQ, dairy foods tended to be reported more accurately than meat, poultry, and fish. This differential validity is taken into account when the food-based, but not the nutrient-based, estimation methods are used. Similar issues may hold for the other five nutrients noted above. Conversely, some nutrients were measured equally well by using either method. For example, the correlation coefficient for total fat was 0.436 for the food-based versus 0.463 for the nutrient-based regression method. The dominant components of total fat were derived from beef and poultry intake, which had similar R^2 values (figure 1). Thus, food-based and nutrient-based methods are likely to provide similar rankings for total fat.

TABLE 3. Rank correlation between actual diet record intake and predicted diet record intake for regression models based on foods and nutrients, respectively (*n* = 173), Nurses' Health Study, 1980–1981

Diet record nutrient	Food-based estimate*	Nutrient-based estimate†	*p* value
Calories	0.578	0.493	0.13
Protein	0.467	0.336	0.027
Total fat	0.436	0.463	0.65
Animal fat	0.568	0.510	0.32
Vegetable fat	0.651	0.602	0.42
Saturated fat	0.532	0.515	0.77
Monounsaturated fat	0.428	0.456	0.64
Polyunsaturated fat	0.565	0.548	0.77
Carbohydrates	0.684	0.595	0.081
Sucrose	0.627	0.545	0.094
Fructose	0.649	0.647	0.97
Crude fiber	0.636	0.563	0.11
Dietary fiber	0.658	0.558	0.065
Calcium	0.528	0.545	0.74
Iron	0.473	0.397	0.21
Magnesium	0.604	0.516	0.14
Phosphorus	0.516	0.455	0.28
Potassium	0.585	0.543	0.43
Zinc	0.440	0.349	0.13
Vitamin C	0.669	0.696	0.59
Vitamin B1	0.592	0.545	0.34
Vitamin B2	0.584	0.518	0.23
Niacin	0.402	0.362	0.52
Pantothenic acid	0.609	0.563	0.36
Vitamin B6	0.561	0.503	0.27
Folate	0.688	0.587	0.065
Vitamin B12	0.451	0.416	0.51
Retinol	0.515	0.599	0.11
Carotene	0.446	0.368	0.012
Vitamin D	0.544	0.615	0.16
Vitamin E	0.567	0.509	0.32
Oleic acid	0.426	0.459	0.59
Linoleic acid	0.581	0.570	0.86
Cholesterol	0.496	0.510	0.79
Methionine	0.414	0.306	0.075
Alcohol	0.851	0.895	0.10
Caffeine	0.612	0.647	0.41

* Based on equation 6; refer to the Materials and Methods section of the text.

† Based on equation 11; refer to the Validation Study Analyses section of the text.

EXAMPLE

As an example of the methodology used, we consider several logistic regression models relating breast cancer incidence from 1980 to 1994 to estimated sucrose and alcohol intake in 1980. The rationale for this analysis is that sucrose intake is inversely associated with alcohol intake (4), and alcohol intake is an established positive risk factor for breast cancer. Therefore, we categorized estimated sucrose intake into quintiles and related breast cancer incidence to quintile of sucrose intake while controlling for age (5-year age groups), total alcohol intake (quintiles), and total caloric intake (quintiles). The results are given in table 4.

For the FFQ, after controlling for age and total caloric intake, we found significant effects of alcohol (quintile (Q)-5 vs. Q1: odds ratio (OR) = 1.29, 95 percent confidence interval (CI): 1.16, 1.43; $p < 0.001$) and borderline significant effects of sucrose (Q5 vs. Q1: OR = 0.87, 95 percent CI: 0.76, 1.01; $p = 0.068$). After controlling for measurement error at the food level, we used equation 7 with 100 bootstrap replications to estimate the standard errors of the regression coefficients. The effect of alcohol was slightly diminished, although still statistically significant (Q5 vs. Q1: OR = 1.21, 95 percent CI: 1.04, 1.42; $p = 0.013$), while the effect of sucrose was slightly enhanced (Q5 vs. Q1: OR = 0.84, 95 percent CI: 0.70, 1.01; $p = 0.069$). There was no significant effect of total caloric intake based on either the FFQ or estimated diet record intake.

In addition to the models shown in table 4, the dummy variables for caloric intake, alcohol quintile, and sucrose quintile were replaced by single continuous variables with quintile medians (refer to equation 8), with p values interpreted as tests for trend. In addition, we also ran nutrient-based error corrections based on the quintile medians (5, 6). The results in table 5 are based on comparisons of women whose intake was at the approximate 10th versus 90th percentile.

Based on the FFQ, after adjustment for age and total caloric intake, there were significant effects for alcohol intake (OR = 1.22, 95 percent CI: 1.11, 1.35; $p < 0.001$) and borderline significant effects for sucrose intake (OR = 0.89, 95 percent CI: 0.78, 1.02; $p = 0.086$). When the food-based error correction methods were used, the effect of alcohol intake decreased modestly, with slightly wider confidence limits (OR = 1.18, 95 percent CI: 1.05, 1.33; $p = 0.005$), while the effect of sucrose increased slightly and became statistically significant, albeit with slightly wider confidence limits (OR = 0.86, 95 percent CI: 0.74, 0.99; $p = 0.033$). With nutrient-based error correction, the estimated odds ratios for both alcohol intake and sucrose intake became stronger, although the confidence intervals became much wider than either the odds ratios based on the FFQ or the odds ratios based on food-based error correction (alcohol: OR = 1.42, 95 percent CI: 1.17, 1.73; $p < 0.001$ and sucrose: OR = 0.73, 95 percent CI: 0.37, 1.48; $p = 0.39$). Total caloric intake was not statistically significant for either the FFQ or either method of error correction for the diet record.

DISCUSSION

Most published dietary analyses have been based on surrogate instruments with large measurement error, such as

TABLE 4. Association of breast cancer incidence from 1980 to 1994 with total caloric intake, sucrose intake, and alcohol intake (in quintiles) in 1980, Nurses' Health Study*

Variable	Food frequency questionnaire				Estimated diet record intake			
	β (SE†)	*p* value	OR†	95% CI†	β (SE)	*p* value	OR	95% CI
Age 40–44 years	0.298 (0.067)	<0.001	1.35	1.18, 1.54	0.294 (0.074)	<0.001	1.34	1.16, 1.55
Age 45–49 years	0.569 (0.063)	<0.001	1.77	1.56, 2.00	0.566 (0.068)	<0.001	1.76	1.54, 2.01
Age 50–54 years	0.676 (0.062)	<0.001	1.97	1.74, 2.22	0.676 (0.064)	<0.001	1.97	1.74, 2.23
Age 55–60 years	0.875 (0.062)	<0.001	2.40	2.12, 2.71	0.879 (0.069)	<0.001	2.41	2.11, 2.76
Caloric intake, Q2†	0.019 (0.059)	0.75	1.02	0.91, 1.14	−0.001 (0.076)	0.99	1.00	0.86, 1.16
Caloric intake, Q3	0.016 (0.062)	0.80	1.02	0.90, 1.15	0.060 (0.084)	0.48	1.06	0.90, 1.25
Caloric intake, Q4	−0.006 (0.065)	0.93	0.99	0.87, 1.13	0.117 (0.085)	0.17	1.12	0.95, 1.33
Caloric intake, Q5	0.018 (0.072)	0.80	1.02	0.88, 1.17	0.031 (0.090)	0.73	1.03	0.86, 1.23
Alcohol intake, Q2	0.066 (0.073)	0.37	1.07	0.93, 1.23	0.029 (0.067)	0.66	1.03	0.90, 1.17
Alcohol intake, Q3	0.162 (0.053)	0.003	1.18	1.06, 1.31	0.142 (0.071)	0.045	1.15	1.00, 1.32
Alcohol intake, Q4	0.075 (0.054)	0.17	1.08	0.97, 1.20	0.126 (0.072)	0.081	1.13	0.98, 1.31
Alcohol intake, Q5	0.253 (0.053)	<0.001	1.29	1.16, 1.43	0.194 (0.078)	0.013	1.21	1.04, 1.42
Sucrose intake, Q2	−0.034 (0.058)	0.56	0.97	0.86, 1.08	−0.088 (0.077)	0.26	0.92	0.79, 1.07
Sucrose intake, Q3	−0.047 (0.062)	0.44	0.95	0.85, 1.08	−0.056 (0.073)	0.45	0.95	0.82, 1.09
Sucrose intake, Q4	−0.022 (0.065)	0.74	0.98	0.86, 1.11	−0.147 (0.083)	0.076	0.86	0.73, 1.02
Sucrose intake, Q5	−0.134 (0.073)	0.068	0.87	0.76, 1.01	−0.170 (0.094)	0.069	0.84	0.70, 1.01

* 3,084 cases; 89,755 subjects.

† SE, standard error; OR, odds ratio; CI, confidence interval; Q, quintile.

TABLE 5. Association of breast cancer incidence from 1980 to 1994 with total caloric intake, sucrose intake, and alcohol intake (based on ordinal scores*) in 1980, Nurses' Health Study†

Variable	β (SE‡)	p value	OR‡	95% CI‡
	Food frequency questionnaire (FFQ) intake			
Age 40–44 years	0.297 (0.067)	<0.001	1.35	1.18, 1.53
Age 45–49 years	0.567 (0.063)	<0.001	1.76	1.56, 1.99
Age 50–54 years	0.673 (0.062)	<0.001	1.96	1.73, 2.21
Age 55–60 years	0.871 (0.062)	<0.001	2.39	2.11, 2.70
Total caloric intake§	0.009 (0.071)	0.90	1.01	0.88, 1.16
Alcohol intake¶	0.202 (0.048)	<0.001	1.22	1.11, 1.35
Sucrose intake#	−0.118 (0.069)	0.086	0.89	0.78, 1.02
	Estimated diet record intake			
Age 40–44 years				
Foods	0.294 (0.067)	<0.001	1.34	1.18, 1.53
Nutrients	0.251 (0.072)	<0.001	1.28	1.12, 1.48
Age 45–49 years				
Foods	0.563 (0.060)	<0.001	1.76	1.56, 1.97
Nutrients	0.522 (0.069)	<0.001	1.69	1.47, 1.93
Age 50–54 years				
Foods	0.672 (0.062)	<0.001	1.96	1.73, 2.21
Nutrients	0.650 (0.068)	<0.001	1.91	1.68, 2.19
Age 55–60 years				
Foods	0.873 (0.068)	<0.001	2.39	2.10, 2.73
Nutrients	0.870 (0.068)	<0.001	2.39	2.09, 2.73
Total caloric intake				
Foods**	0.046 (0.066)	0.49	1.05	0.92, 1.19
Nutrients§	−0.071 (0.377)	0.85	0.93	0.44, 1.95
Alcohol intake				
Foods††	0.169 (0.051)	0.005	1.18	1.05, 1.33
Nutrients¶	0.353 (0.098)	<0.001	1.42	1.17, 1.73
Sucrose intake				
Foods‡‡	−0.154 (0.072)	0.033	0.86	0.74, 0.99
Nutrients#	−0.309 (0.356)	0.39	0.73	0.37, 1.48

* Using median scores for specific quintiles.

† 3,084 cases; 89,755 subjects.

‡ SE, standard error; OR, odds ratio; CI, confidence interval.

§ Comparing women with total caloric intake of 989 calories vs. 2,177 calories (the approximate 10th and 90th percentiles of total caloric intake on the FFQ).

¶ Comparing women with total alcohol intake of 0.76 g vs. 17.3 g (the approximate 10th and 90th percentiles of alcohol intake on the FFQ).

# Comparing women with sucrose intake of 14 g vs. 52 g (the approximate 10th and 90th percentiles of sucrose intake on the FFQ).

** Comparing women with total caloric intake of 698 calories vs. 1,074 calories (the approximate 10th and 90th percentiles of estimated diet record total caloric intake).

†† Comparing women with total alcohol intake of 1.30 g vs. 13.56 g (the approximate 10th and 90th percentiles of estimated diet record alcohol intake).

‡‡ Comparing women with sucrose intake of 6.97 g vs. 16.16 g (the approximate 10th and 90th percentile of estimated diet record sucrose intake).

the FFQ or the 24-hour recall. Correction methods attempt to account for this measurement error but are usually performed at the nutrient level. However, subjects report intake based on individual foods with differential validity, which should be taken into account in the analysis. To the best of our knowledge, this is the first paper to consider measurement error correction at the food level. If a large number of days of diet record are available for each subject, then this approach may reduce bias relative to FFQ-based inference. Another approach for reducing bias in FFQ-based inference is to use cumulative averages of the FFQ over multiple years (7); however, this approach assumes that the repeated measures of the FFQ have approximately equal predictive power for subsequent incident disease.

Note that diet record estimates of food intake are based on average intake over 28 days per subject and thus are subject to error. However, the regression parameter estimates in the validation study regression shown in equation 3 (λ, θ) will still be unbiased as long as average log(diet record intake) over 28 days for specific foods provides an unbiased estimate of long-term average log(diet record intake). The impact of using a small number of days of diet record is that standard errors of validation study parameters will increase, and measurement-error-corrected confidence limits will widen.

The analyses described in this paper indicate that sucrose intake is inversely associated and alcohol intake is positively associated with breast cancer risk. A more complete analysis of these associations would consider other known breast cancer risk factors and/or other nutrients but is beyond the scope of this paper.

An assumption of the regression calibration approach based on either foods or nutrients is that measurement error from the diet record and FFQ are independent in relation to a true gold standard. To test the validity of this assumption, Spiegelman et al. (8) extended the regression calibration approach to assess the case in which the diet record was considered an alloyed gold standard relative to a biomarker considered the true gold standard. The correlation between errors in the alloyed gold standard and the surrogate with respect to the biomarker was taken into account. For all examples considered, the results from the ordinary regression calibration approach (e.g., equation 2) and the extended regression calibration approach gave virtually the same results. Finally, we included body mass index as an additional covariate in each of the food-based validation study regressions in equation 3. Hence, even if there is differential reporting bias by subjects with high versus low body mass index for specific foods (e.g., for wine consumption in table 2), this bias is taken into account when predicted diet record intake for each specific food is estimated.

In conclusion, we have presented a method of measurement error correction at the food level. Subjects report data at the food level, with differential reporting accuracy for different foods. Therefore, compared with nutrient-based error correction, our approach has the potential for more accurate prediction of disease as a function of either true food or nutrient intake, particularly when foods that make important contributions to nutrient show large differences in validity (e.g., dairy foods vs. meat or poultry as constituents of protein). In addition, our example demonstrates some notable differences between inference based on these two approaches. However, to enable this methodology to be used, validation study data should be collected and reported at the food level. This approach warrants further investigation in other study settings.

ACKNOWLEDGMENTS

This work was supported by grant CA-50597 from the National Cancer Institute.

REFERENCES

1. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8: 1051–69.
2. Salvini S, Hunter DJ, Sampson L, et al. Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. Int J Epidemiol 1989;18:858–67.
3. Wolfe DA. On testing equality of related correlation coefficients. Biometrika 1976;63:214–15.
4. Colditz GA, Giovanucci E, Rimm EB, et al. Alcohol intake in relation to diet and obesity in women and men. Am J Clin Nutr 1991;54:49–55.
5. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990;132:734–45.
6. Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London, United Kingdom: Chapman and Hall, 1995.
7. Willett WC. Nutritional epidemiology. New York, NY: Oxford University Press, 1998.
8. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an “alloyed gold standard.” Am J Epidemiol 1997;145:184–96.