Current Research

A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution

JANET A. TOOZE, PhD, MPH; DOUGLAS MIDTHUNE, MS; KEVIN W. DODD, PhD; LAURENCE S. FREEDMAN, PhD; SUSAN M. KREBS-SMITH, PhD, MPH, RD; AMY F. SUBAR, PhD, MPH, RD; PATRICIA M. GUENTHER, PhD, RD; RAYMOND J. CARROLL, PhD; VICTOR KIPNIS, PhD

ABSTRACT

Objective We propose a new statistical method that uses information from two 24-hour recalls to estimate usual intake of episodically consumed foods.

Statistical analyses performed The method developed at the National Cancer Institute (NCI) accommodates the large number of nonconsumption days that occur with foods by separating the probability of consumption from the consumption-day amount, using a two-part model. Covariates, such as sex, age, race, or information from a food frequency questionnaire, may supplement the information from two or more 24-hour recalls using correlated mixed model regression. The model allows for correlation between the probability of consuming a food on a single day and the consumption-day amount. Percentiles of the distribution of usual intake are computed from the estimated model parameters.

Results The Eating at America's Table Study data are used to illustrate the method to estimate the distribution of usual intake for whole grains and dark-green vegetables for men and women and the distribution of usual intakes of whole grains by educational level among men. A simulation study indicates that the NCI method leads to substantial improvement over existing methods for estimating the distribution of usual intake of foods.

Conclusions The NCI method provides distinct advantages over previously proposed methods by accounting for the correlation between probability of consumption and amount consumed and by incorporating covariate information. Researchers interested in estimating the distribution of usual intakes of foods for a population or subpopulation are advised to work with a statistician and incorporate the NCI method in analyses.

J Am Diet Assoc. 2006;106:1575-1587.

J. A. Tooze is an assistant professor, Department of Biostatistical Sciences, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC. D. Midthune, K. W. Dodd, and V. Kipnis are mathematical statisticians and S. M. Krebs-Smith and A. F. Subar are nutritionists, National Cancer Institute, Bethesda, MD. L. S. Freedman is director, Biostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel. P. M. Guenther is a nutritionist, US Department of Agriculture, Center for Nutrition Policy and Promotion, Alexandria, VA. R. J. Carroll is distinguished professor, professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX.

Address correspondence to: Janet A. Tooze, PhD, MPH, Assistant Professor, Department of Biostatistical Sciences, Wake Forest University School of Medicine, Medical Center Blvd, Winston-Salem, NC 27157. E-mail: jtooze@wfubmc.edu

Copyright © 2006 by the American Dietetic Association.

0002-8223/06/10610-0016$32.00/0
doi: 10.1016/j.jada.2006.07.003

When assessing dietary intake among populations or individuals, investigators are often interested in capturing usual intakes—ie, long-term averages. The 24-hour dietary recall provides rich details about dietary intake for a given day, but collecting more than two 24-hour recalls per individual is impractical in large surveys such as the National Health and Nutrition Examination Survey (NHANES). Therefore, it is necessary to use statistical methods to estimate usual dietary intake.

Researchers are interested in estimating the usual intake of foods to assess compliance with food-based dietary recommendations and to relate food intake to health parameters. Unlike most nutrients, which are consumed daily, estimating usual intake of episodically consumed foods presents unique challenges for statistical modeling. (See Figure 1 for a definition of *statistical modeling* and related terms.) These challenges are: (a) accounting for days without consumption of a particular food or food group; (b) allowing for consumption-day amount data that are generally positively skewed and have extreme values in the upper tail of the intake distribution; (c) distinguishing within-person variability, which consists of day-to-day variation in intake and random reporting errors, from between-person variation; (d) allowing for the correlation between the probability of consuming a food and the consumption-day amount; and (e) relating covariate information (eg, sex, age, race, ethnicity, or education level) to usual intake (1).

As discussed by Dodd and colleagues (1), two other methods have been used to estimate the distribution of

© 2006 by the American Dietetic Association

Statistical term	Definition	Use in usual food intake model
Statistical Model	A model is a mathematical formula used to quantify the relationship between two or more variables. The model is "statistical" when it also incorporates uncertainty in the relationship between the variables.	A statistical model is used to estimate usual food intake and to relate it to other variables of interest.
Two-part model	Sometimes there is a need for a more complex statistical model that includes two component parts.	In this article, the model for food consumption models the probability of consuming a food as well as the usual amount consumed.
Outcome Variable	The variable of interest in the analysis. Sometimes referred to as the dependent variable.	Usual food intake is the outcome variable for the two-part model. It is derived as the probability of consuming the food multiplied by the amount consumed on a consumption day.
Covariates	Variables that are related to the outcome variable. The term "covariate" is used to describe a general class of variables that may be of most interest, define a subpopulation, need to be adjusted for in the statistical analysis. Sometimes referred to as independent variables.	Responses to line items from a food frequency questionnaire may be used as a covariate to improve estimation from the 24-hour recall alone; variables such as age and race may be used to define subpopulations for which usual intake is estimated.
Person-specific random effect	A term that is specific to an individual that refers to how an individual's value deviates from the average. It is considered "random" because the individuals in the study are considered as a random sample from a larger population.	Both parts of the statistical model include a person-specific random effect that describes the individual's frequency of consuming a particular food and the amount consumed.
Normality	Refers to a statistical distribution of a variable, specifically a bell-shaped curve. The tails of the normal distribution refer to the extreme values. If the data do not follow a normal distribution, then many commonly used statistical methods cannot appropriately be used. By applying a function, such as the logarithm, data can be transformed to a more normal distribution.	The amount part of the model is transformed to normality (using a Box-Cox [power] transformation). In the model, the normality assumption must hold for the random effects after including the covariates of interest in the model. Including the covariates may help to make the distribution of these random effects more normally distributed.
Correlation	If two variables are associated with each other, they are said to be correlated. The opposite of correlation is independence, in which a change in one variable does not impact the value of another variable.	In this model, there are two types of correlation. First, the two person-specific effects are correlated. This means that we allow the individual's tendency to consume a food to be related to the amount that he or she consumes. Second, the covariates in the model are correlated with the outcome (food intake). For example, persons who report a higher frequency of intake on the food frequency questionnaire generally have a higher probability of consuming a food on the 24-hour recall.
Simulation Study	A simulation study is a method that statisticians use to validate their models. Many hypothetical random samples are generated (ie, simulated), and statistical estimates are computed for each sample. The results are then averaged and compared to the "truth" that was used to generate the model.	In this paper, simulations were used to generate 365 days of pseudo-data for a series of individuals. Then different statistical methods to obtain estimates of the distribution of usual intake were run using the same generated data sets. Finally, these estimated distributions were compared to truth.

Figure 1. Definitions of common statistical terms and their use in the usual food intake model.

usual intake of episodically consumed foods with a few days of 24-hour recalls: (a) the distribution of within-person means, and (b) the method developed at Iowa State University for estimating the distribution of foods (ISUF) (2). The within-person means method usually leads to biased estimates of the prevalence of either inadequate or excess food intake because it does not meet any of the challenges listed earlier. In particular, because the within-person means method does not meet the challenge of distinguishing within-person variability from between-person variation (and thereby includes within-person variability), the variance of usual intake is inflated. The ISUF method meets the first three challenges (accounting for days without consumption of a particular food or food group, allowing for consumption-day amount data that are generally positively skewed and have extreme values in the upper tail of the intake distribution; and distinguishing within-person variability from between-person variation), but it does not allow for correlation between probability and amount and cannot incorporate covariate information regarding usual intake.

This article describes a new statistical method that was developed at the National Cancer Institute (NCI) to meet all five of the challenges noted, using two 24-hour recalls, and evaluates the new statistical method's application to estimating the distribution of usual intake of episodically consumed foods.

METHODS

Assumptions of the NCI Method

In the NCI method, we assume that the 24-hour recall is an unbiased instrument for usual intake of episodically consumed foods. This assumption has two components. First, we assume that the 24-hour recall does not misclassify the respondent's food consumption (ie, if a food actually is consumed on a surveyed day, the food will be reported on the 24-hour recall; and if a food is not consumed, it will not be reported on the 24-hour recall). Second, we assume that the 24-hour recall is an unbiased measure for the amount of food consumed on the consumption day. This does not mean that the 24-hour recall captures the amount of food consumed by an individual exactly on each recall—at a given time an individual may report more or less than was actually consumed—but over many days it produces the correct average intake. (This assumption will be discussed in detail later.) In addition to the assumptions made about the 24-hour recall, we make the usual assumptions for parametric regression analysis in our models. In particular, we assume that, after an appropriate transformation, the amount of food consumed on a consumption day is approximately normally distributed.

Overview of the NCI Method

The NCI method for estimating the usual intake of foods has two steps. The first step consists of fitting a two-part statistical model that describes the relationship between usual intake and covariates and estimates the variability of intake both within and between individuals. We adapted a two-part model with correlated person-specific random effects, developed by Tooze and colleagues (3), for this purpose. Similar to the ISUF method, the statistical model represents usual intake as the product of the probability to consume a food on a given day and the usual consumption-day amount. The amount data are transformed to approximate normality, using the Box-Cox transformation (4), as part of the model-fitting process. To account for the correlation between the probability and amount that exists for most foods, as described later, the two parts of the model are linked.

The second step of the method involves additional statistical procedures that, depending on the application of interest, are used to obtain the final "product" of the analysis. Examples of those products include estimates of the distribution of usual intake in a population or subpopulation of interest, or estimates of individual intakes. The latter may be used to assess diet–health relationships. Because these are clearly varying endpoints, different statistical procedures are required. However, because the data share a common structure in each case, the same statistical model is used to obtain parameter estimates, which are the inputs for the final step. This article focuses on the statistical model and its application for estimating the distribution of usual intake for populations and subpopulations. Predicting individual usual intake and relating it to health outcomes are beyond the scope of this article.

Details of the NCI Method

Statistical Model. The first part of the model estimates the probability of consuming a food (positive intake reported on the 24-hour recall) using logistic regression with a person-specific random effect (mixed model). The logistic regression model incorporates covariates to represent the effect of personal characteristics, such as age, sex, or body mass index, on the probability of food consumption. The person-specific effect is a factor that allows an individual's consumption probability to differ from the population level. It may be thought of as the individual's personal tendency to consume a food. The probability of consumption is estimated from two or more 24-hour recalls, accounting for covariates. Symbolically, Part I may be represented as:

$$Logit(24\text{-}Hour\ Recall\ Probability) = Intercept_I + Slope_I \times Covariate + Person\text{-}Specific\ Effect_I \quad [A]$$

where, for probability p, $\text{logit}(p) = \log(p/1 - p)$. The intercept, slope, and variance of the person-specific effect are the model parameters, and subscript I indicates their association with Part I. Although one covariate is shown in equation [A], the model allows for multiple covariates or no covariates. The second part of the model specifies the consumption-day amount of a food using the 24-hour recall data on a transformed scale. Similar to Part I, Part II may incorporate covariate information to estimate amount. As before, the covariates, which need not be the same covariates as in Part I, represent the effect of personal characteristics on the consumption-day amount. This part of the model also includes a person-specific effect as well as within-person variability due to day-to-day variation in an individual's intake and other sources of random error. The model for Part II is:

$$\textit{Transformed 24-Hour Recall Amount} = \textit{Intercept}_{II} + \textit{Slope}_{II} \times \textit{Covariate} + \textit{Person-specific Effect}_{II} + \textit{Within-person Variability}_{II} \quad [B]$$

where subscript II indicates that these parameters are associated with Part II, and differ from those in Part I. Two or more 24-hour recalls on a number of individuals with reports of the food of interest are required to distinguish between- and within-person variation. The model is specified on the transformed scale where the person-specific effect and within-person random variability are normally distributed.

Links Between Parts I and II

Unlike the ISUF method, in which the two parts of the model are assumed to be independent and are estimated separately, the NCI method fits both parts simultaneously, which associates probability to amount in two ways. First, the two person-specific effects are modeled as correlated random variables. Second, some covariates may be the same in both parts of the model, inducing correlation between them. By linking the two parts of the statistical model, the relationship between probability and amount is accounted for, meeting the challenge (stated previously) of allowing for the correlation between the probability of consuming a food and the consumption-day amount.

Fitting the Model

The model is fit by the maximum likelihood method, using an SAS software (version 8.2, 1999-2001, SAS Institute, Cary, NC) macro. To account for the correlation, all of the model parameters are estimated at the same time using a nonlinear mixed-effects model. In addition, the Box-Cox transformation parameter is also estimated as part of the likelihood maximization procedure. The advantage of using the normality transformation within the modeling step, not before modeling as in other methods, is that the amount reported on the 24-hour recall is transformed to normality conditionally on the covariates in the model. Estimates are obtained for the model parameters presented in equations [A] and [B], and for the correlation between the person-specific effects.

Adding Information from Food Frequency Questionnaires as Covariates

As described by Subar and colleagues (5), frequencies from a food frequency questionnaire (FFQ), such as NCI's Diet History Questionnaire (DHQ), are generally positively related to the proportion of 24-hour recalls with reported consumption of those foods. This demonstrates that food frequency information could be a useful covariate in estimating the probability of consumption. In addition, because of the correlation between the probability of consumption and the amount consumed, frequency responses can contribute to estimating not only probability of consumption, but also amount consumed (5). The sum of the frequencies of several individual FFQ items may be used to represent the frequency of consumption of a food group containing them. The relationship between the FFQ and the 24-hour recall is often nonlinear. Consequently, a polynomial model may be used to model the relationship between the FFQ and 24-hour recall.

Estimating the Distribution of Usual Intake for the Population

To estimate the distribution of usual intake, the estimated model parameters are used to simulate a population that has the same characteristics (as described by the values of the covariates) and between-person variability as the sample on which the model was fit. The within-person variation in model [B] is not included because, by definition, it does not contribute to long-term intake.

First, the estimated intercept and slope(s) for the covariate(s) are used to obtain for each individual in the sample an estimate based on the covariate values used to fit the model. To these, it is necessary to add an estimate of the person-specific random effects. Because these effects are unobservable, an estimate from their statistical distribution (bivariate normal with mean zero and variance parameters estimated from fitting the statistical model) is generated. To improve the precision of the estimated usual intake distribution, 100 pseudo-persons for each individual in the sample are generated, each with the same covariate values but with different simulated person-specific effects.

Because the consumption-day amount data are transformed using the Box-Cox transformation during the model fitting process, it is necessary to back-transform the amount data to the original scale before estimates of the distribution of usual intake may be obtained. The back-transformation is similar to the approach used in the method developed at Iowa State University for estimating usual nutrient intake distributions (6). It adds an adjustment term to make the mean of the back-transformed variable match the mean on the original scale as described by Dodd and colleagues (1). Lastly, the mean, standard deviation, and percentiles are estimated empirically from this simulated population.

Estimating the Distribution of Usual Intake for a Subpopulation

Estimates of the distribution of usual intake for a subpopulation are made in the same way as estimates for the total population, except that covariates that define the subgroup are included in the NCI model. When making the estimates for a subpopulation, only the covariate values differ; all other variance components remain the same. This specification leads to smaller standard errors of estimated parameters than stratifying by subpopulation.

Correlation of Probability of Consumption and Amount Consumed on 24-hour Recalls in the Eating at America's Table Study

One of the challenges in developing this statistical model was to account for the phenomenon that occurs for some food groups: that those individuals who eat a food most frequently tend to eat more of it (the challenge, mentioned earlier, of allowing for the correlation between the probability of consuming a food and the consumption-day amount). To determine how often this happens, we used the Eating at America's Table Study (EATS) data to assess the proportion of food groups that exhibit a posi-

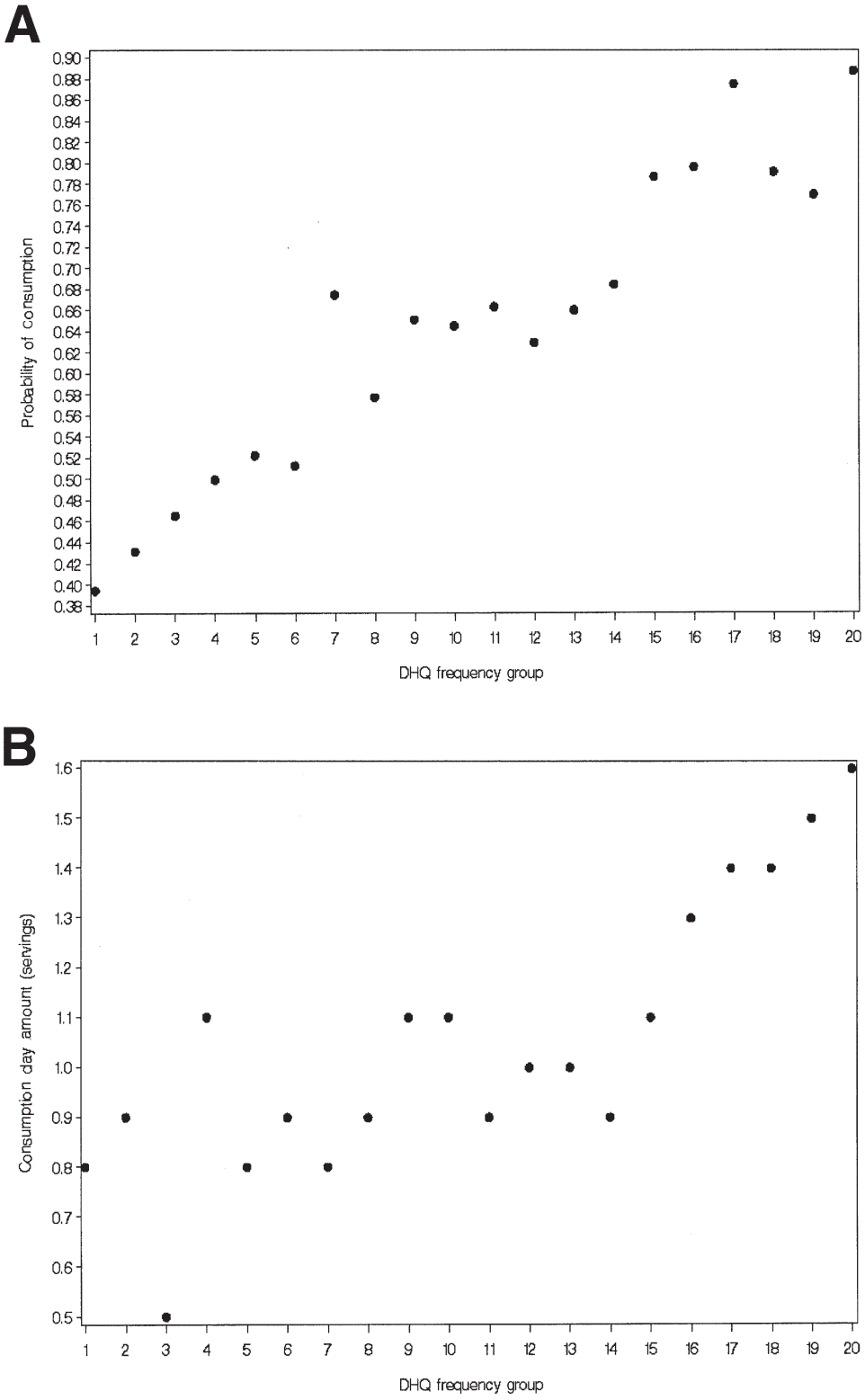

Figure 2. (A) Mean probability of whole-grains consumption by Diet History Questionnaire (DHQ) whole-grains frequency group for women in the Eating at America's Table Study. (B) Mean whole-grains consumption-day amount (servings) by DHQ whole-grains frequency group for women in the Eating at America's Table Study.

Table 1. Percentage of individuals in the Eating at America's Table Study consuming from food group and median amount[ab] of food group consumed per day on 24-hour recall by number of days food group was reported[c]

	Men (n=446)					Women (n=519)				
	Number of 24-Hour Recalls with Reported Intake of Food Group					Number of 24-Hour Recalls with Reported Intake of Food Group				
	0	1	2	3	4	0	1	2	3	4
Alcohol[d]										
% consuming	46.4	18.4	13.5	9.2	12.6	58.4	21.0	10.0	5.8	4.8
Amount (drinks)	0.0	1.0	2.1	2.3	3.0	0.0	1.0	1.4	1.9	2.6
Cheese[d]										
% consuming	6.3	14.1	22.6	33.6	23.3	6.9	18.7	26.6	29.7	18.1
Amount (servings)	0.0	0.6	0.7	0.9	1.0	0.0	0.4	0.6	0.7	0.7
Milk[d]										
% consuming	1.1	7.0	12.6	27.1	52.2	2.5	7.7	16.4	25.4	48.0
Amount (servings)	0.0	0.3	0.6	0.7	1.3	0.0	0.2	0.4	0.6	1.0
Total dairy[d]										
% consuming	0.2	2.0	4.7	14.8	78.3	0.2	1.7	6.2	18.1	73.8
Amount (servings)	0.0	0.9	0.9	1.2	1.7	0.0	0.2	0.5	0.8	1.3
Citrus, melon, berries[d]										
% consuming	5.2	12.3	22.6	30.0	29.8	6.2	14.5	25.2	27.4	26.8
Amount (servings)	0.0	0.1	0.8	0.9	1.6	0.0	0.1	0.5	0.9	1.3
Other fruit[d]										
% consuming	12.3	21.7	21.1	22.2	22.6	11.6	20.0	24.7	22.5	21.2
Amount (servings)	0.0	1.0	1.1	1.4	1.8	0.0	0.7	0.9	1.1	1.4
Total fruit[d]										
% consuming	1.1	6.3	14.6	22.0	56.1	2.5	5.4	15.2	25.0	51.8
Amount (servings)	0.0	0.3	1.1	1.2	2.2	0.0	0.2	0.6	1.3	1.9
Non–whole grains[d]										
% consuming	0.0	0.2	0.4	2.0	97.3	0.0	0.0	0.0	6.0	94.0
Amount (servings)	0.0	0.6	5.5	4.5	6.7	0.0	0.0	0.0	3.2	4.8
Whole grains[d]										
% consuming	8.3	13.2	23.3	26.5	28.7	7.5	16.0	22.5	27.0	27.0
Amount (servings)	0.0	1.1	1.5	2.1	2.4	0.0	0.8	1.1	1.3	1.5
Total grains[d]										
% consuming	0.0	0.2	0.2	1.6	98.0	0.0	0.0	0.0	2.7	97.3
Amount (servings)	0.0	0.6	6.4	5.0	8.1	0.0	0.0	0.0	4.2	5.8
Eggs[d]										
% consuming	5.2	18.8	29.6	24.2	22.2	8.3	24.7	30.4	24.3	12.3
Amount[e]	0.0	0.1	0.4	0.5	0.7	0.0	0.1	0.3	0.5	0.4
Fish[d]										
% consuming	45.7	33.2	16.1	3.6	1.3	51.8	32.9	11.6	2.9	0.8
Amount[e]	0.0	2.0	3.0	4.0	4.5	0.0	1.6	1.8	2.3	4.9
Frankfurters and sausages										
% consuming	20.4	32.1	29.6	13.5	4.5	34.5	32.8	22.4	9.2	1.2
Amount[e]	0.0	2.0	2.1	2.1	2.3	0.0	1.5	1.3	1.4	1.7
Meat[f]										
% consuming	6.5	14.1	22.0	30.5	26.9	7.1	17.0	28.1	31.0	16.8
Amount[e]	0.0	3.4	3.3	3.6	3.7	0.0	1.8	2.1	2.4	2.4
Nuts and seeds[d]										
% consuming	28.0	30.0	22.9	14.1	4.9	30.6	33.1	19.8	11.0	5.4
Amount[e]	0.0	0.3	0.4	0.5	0.4	0.0	0.2	0.3	0.4	0.5
Organ meats										
% consuming	97.3	2.5	0.2	0.0	0.0	96.3	3.5	0.2	0.0	0.0
Amount[e]	0.0	1.8	1.2	0.0	0.0	0.0	2.4	0.9	0.0	0.0
Poultry[f]										
% consuming	19.1	32.3	28.7	15.5	4.5	18.9	32.8	28.7	15.8	3.9
Amount[e]	0.0	3.2	3.7	3.3	4.0	0.0	2.4	2.4	2.9	3.0
Soy[d]										
% consuming	82.7	13.2	3.6	0.4	0.0	82.1	13.3	2.7	1.2	0.8
Amount[e]	0.0	0.1	0.2	6.0	0.0	0.0	0.0	0.9	0.6	1.4

(continued)

Table 1. Percentage of individuals in the Eating at America's Table Study consuming from food group and median amount[ab] of food group consumed per day on 24-hour recall by number of days food group was reported[c] (continued)

	Men (n=446)					Women (n=519)				
	Number of 24-Hour Recalls with Reported Intake of Food Group					Number of 24-Hour Recalls with Reported Intake of Food Group				
	0	1	2	3	4	0	1	2	3	4
Meat, fish, poultry[d]										
% consuming	1.6	0.9	3.4	11.4	82.7	0.8	1.2	4.6	22.0	71.5
Amount[e]	0.0	1.8	5.0	5.3	6.1	0.0	1.7	2.8	3.2	3.7
Deep-yellow vegetables[d]										
% consuming	24.0	33.6	24.4	13.2	4.7	19.3	35.8	26.6	13.7	4.6
Amount (servings)	0.0	0.3	0.5	0.6	0.8	0.0	0.3	0.4	0.4	0.7
Dark-green vegetables[f]										
% consuming	52.0	28.5	14.6	4.3	0.7	43.4	29.5	19.3	6.4	1.5
Amount (servings)	0.0	0.8	1.1	1.1	1.1	0.0	0.7	0.8	1.1	1.2
Legumes[f]										
% consuming	56.7	31.2	9.0	2.7	0.4	61.3	28.1	8.3	2.3	0.0
Amount (servings)	0.0	1.1	1.1	1.6	1.1	0.0	0.7	0.8	0.7	0.0
Other vegetables[d]										
% consuming	0.0	0.9	6.3	23.5	69.3	0.0	1.3	6.9	26.2	65.5
Amount (servings)	0.0	0.3	0.9	1.0	1.5	0.0	0.5	0.5	0.8	1.1
Potatoes[d]										
% consuming	12.1	24.2	29.6	22.4	11.7	14.5	26.2	28.3	21.6	9.4
Amount (servings)	0.0	2.0	2.3	2.7	2.7	0.0	1.5	1.6	1.9	1.8
Starchy vegetables[g]										
% consuming	45.1	37.9	12.6	4.5	0.0	45.5	35.6	15.0	3.5	0.4
Amount (servings)	0.0	1.0	0.9	1.4	0.0	0.0	0.6	0.7	0.8	1.3
Tomatoes[d]										
% consuming	1.8	7.8	22.2	37.4	30.7	1.9	11.8	26.2	34.5	25.6
Amount (servings)	0.0	0.7	0.6	0.8	0.9	0.0	0.5	0.5	0.6	0.6
Total vegetables[d]										
% consuming	0.0	0.2	1.1	11.9	86.8	0.0	0.0	1.2	12.7	86.1
Amount (servings)	0.0	0.7	3.3	3.4	4.3	0.0	0.0	1.6	2.7	3.2

[a]Median amount consumed by category.
[b]Servings in this table refer to Food Guide Pyramid servings (see www.ba.ars.usda.gov/cnrg/services/foodlink.html).
[c]Data from reference 7.
[d]Correlation between number of consumption days and amount consumed (Spearman correlation): $P<0.05$ for men and women.
[e]Ounces of cooked lean meat equivalents.
[f]Correlation between number of consumption days and amount consumed (Spearman correlation): $P<0.05$ for women; $P>0.05$ for men.
[g]Correlation between number of consumption days and amount consumed (Spearman correlation): $P<0.05$ for men; $P>0.05$ for women.

tive correlation between the probability of consumption and consumption-day amount on 24-hour recalls. Details of the EATS, conducted in 1997-1998, are published elsewhere (7). We used data from a national sample from 965 men and women, 20 to 70 years of age, who completed four 24-hour recalls 3 months apart, followed by NCI's DHQ. The NCI Special Studies Institutional Review Board approved the study.

To determine what proportion of 27 Food Guide Pyramid food groups (described in reference 5) exhibited a significant correlation between probability of consumption and consumption-day amount, we calculated the proportion of respondents who reported consumption on zero, one, two, three, or four of the four 24-hour recalls and the median portion size consumed for each category by sex. Spearman correlation coefficients between the number of recalls with consumption and the consumption-day amount were then computed.

Applying the Method: Example of Estimating the Distribution of Usual Intake

We also used the EATS data to (a) illustrate the application of the NCI method, in comparison with the within-person means and ISUF methods, for estimating the distribution of usual intake for two food groups (whole grains and dark-green vegetables) for men and women, and (b) estimate the distribution of usual intake by education level for whole-grain consumption by men. In these analyses, only DHQ reported frequencies were used; the average portion-size information (small, medium, large) was not used. This variable of reported frequencies is similar to the Food Propensity Questionnaire (FPQ) used in the 2003-2006 NHANES (5).

Daily frequencies were determined for each respondent for each food group by summing the line items belonging to each group. The covariate for whole grains included the

frequency data for the following DHQ items: breads, crackers, hot cereals, popcorn, ready-to-eat cereal, rice or other cooked grains, and potato/tortilla/corn chips. The covariate for dark-green vegetables included frequencies of broccoli, raw greens, cooked greens, and lettuce. The proportions of the line items in each category (eg, the proportion of bread that was whole grain) were considered when deriving the frequency variables (see reference 5 for further details). To accommodate the nonlinear relationship between the 24-hour recall and the food-frequency variables, a polynomial model was used—that is, square root, linear, and quadratic functions of the food frequency variable were included as covariates in the statistical model.

We also estimated the distribution of usual intake of whole grains for three subpopulations of men: those with a high school education or less, those with some college, and college graduates. This was achieved by including the indicator variables for education level as covariates in both parts of the statistical model.

Applying the Method: Simulation Study for Estimating the Distribution of Usual Intake

Simulation studies provide a means of comparing statistical methods to each other, as well as to a measure of true usual intake. We conducted a simulation study to evaluate the NCI method and to compare it with other currently available methods for estimating usual food intake distributions. Because we were interested in the performance of the methods for the minimum number of 24-hour–recall days, two days were chosen to compare six different methods of estimating usual intake: (a) the 2-day within-person means; (b) the ISUF method; (c) the standard NCI method with correlated person-specific effects and the FPQ as a covariate (as described earlier); (d) the NCI method with correlated person-specific effects but without the FPQ (ie, the covariate[s] generated by the FFQ were removed from the model fittings of equations [A] and [B]); (e) the NCI method with uncorrelated person-specific effects (ie, the person-specific effects in equations [A] and [B] were assumed to be uncorrelated in the model fittings) with the FPQ; and (f) the NCI method with uncorrelated person-specific effects without the FPQ. By comparing the NCI method both with and without the correlated person-specific effects and with and without the FPQ, we were able to isolate the effects of these components of the model in this simulation study.

We simulated 200 datasets, each with 2,000 pseudo-persons, based on whole-grain consumption by women in the EATS. For each pseudo-person in a simulated data set, an FFQ frequency value was selected from the actual values in the EATS. Next, a probability was generated for that individual, using the mean probability from the EATS data, stratified by FFQ frequency group (Figure 2A). For example, a pseudo-person who fell into FFQ group 5 (corresponding to a value between the 20th and 25th percentile, or approximately four times per week), would have approximately a 52% chance of consuming whole grains, based on the recall data for everyone in that group. Using this probability, 365 pseudo-days were generated for this person, with each one having the underlying chance of 52% of consuming a whole grain on that day.

Next, for the days that were simulated to be consumption days, a mean amount was generated for each consumption day using the mean consumption-day amounts from the EATS, represented in Figure 2B. For example, for a woman in group 5 (Figure 2B), the mean value of approximately 0.8 servings, plus or minus a randomly generated value reflecting the variability about the mean, was used. Finally, two correlated person-specific effects (from a bivariate normal distribution) corresponding to probability and amount, were generated for each pseudo-person with a correlation equivalent to that found in the EATS data.

After combining the simulated probability, consumption-day amount, and person-specific effects, each person in the dataset has 365 days of pseudo-data, the mean of which was used to estimate true intake. The mean percentiles estimated from all six methods described earlier were compared with the percentiles of this true intake.

RESULTS

Correlation of Probability of Consumption and Amount Consumed on 24-hour recalls in EATS

Table 1 presents the proportion of respondents in the EATS who consumed a food group on 0, 1, 2, 3, or 4 of the recall days, and the corresponding median amount, expressed as the number of Food Guide Pyramid servings (8) consumed per day. Forty-five of the 54 (83%) food group by sex combinations (27 food groups for each sex) exhibited a positive correlation between the probability of consuming the food and the mean consumption-day amount.

Example of Estimating the Distribution of Usual Intake in the EATS

Estimating the Distribution of Usual Intake. The smoothed distribution curves for whole-grain consumption by women from the EATS data is represented in Figure 3. This figure illustrates the differences between the within-person means, ISUF, and NCI methods using 4 days of 24-hour recalls. First, even with 4 days of recalls, it is still difficult to estimate empirically the lower tail of the distribution using within-person means, resulting in a large spike at zero. Second, because the within-person means method does not distinguish within-person from between-person variability, its distribution has a longer tail. Although the ISUF software produced a warning that it should not be used for whole grains because significant correlation existed between the probability of consuming and the amount consumed, its results are presented here for illustration purposes.

The curve for the NCI method is to the left of the ISUF curve below approximately 1.7 servings, at which point it shifts to the right of the ISUF curve for larger amounts. The difference between these curves is due to the positive correlation between probability and amount. Because those women who are less likely to eat whole grains eat smaller amounts when they do eat them, the area under the NCI method curve is larger in the lower part of the distribution than the area below the curve produced using the ISUF method, which assumes no relationship between probability and amount. For the same reason, this relationship reverses in the upper part of the area

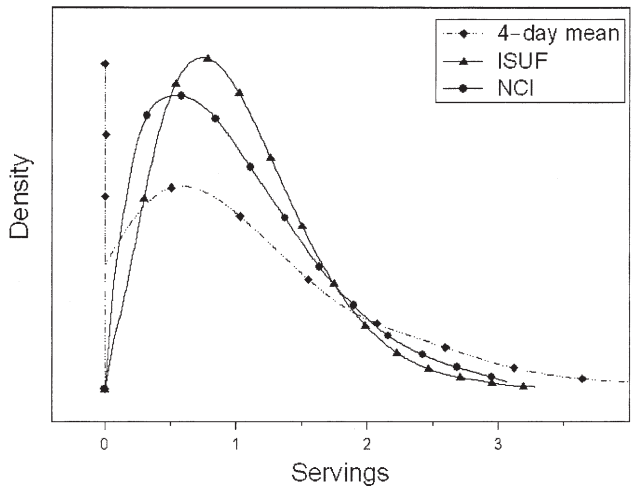

Figure 3. Estimated distributions of usual intake of whole grains for women in the Eating at America's Table Study using different methods. The spike at zero for the 4-day mean (within-person mean of four 24-hour recalls) represents 7.5% of the distribution. ISUF=Iowa State University Foods method. NCI=National Cancer Institute method with correlated random effects and a food frequency questionnaire as a covariate. (This figure is available online at www.adajournal.org as part of a PowerPoint presentation featuring additional online-only content.)

under the NCI method curve, reflecting that women who are more likely to eat whole grains eat larger amounts when they do eat them.

Estimating the Percent of Population More or Less than a Cutoff. Table 2 illustrates the differences among the three methods. For both dark-green vegetables and whole grains, the 4-day within-person means method produces estimates that are higher than the ISUF and NCI methods in both of the tails due to the large spike at zero and the longer tail in the upper end of the distribution.

Other differences between the NCI method and the ISUF method are determined by the strength of the correlation between probability of consumption and the amount consumed. This correlation coefficient is not significantly different from zero for dark-green vegetable consumption by men in the EATS, and, as shown in Table 2, the estimates from the NCI and ISUF methods are similar for men. In contrast, due to substantially correlated probability and amount for whole grains ($r=0.29$ for women and $r=0.34$ for men), the estimates of the proportion of people consuming less than 0.5 serving and more than three servings of this food group are considerably lower in the ISUF method than the NCI method for both women and men.

Estimating the Distribution of Usual Intake for a Subpopulation. The percentages of men who consume more or less than specified cutoff values for servings of whole grains by education levels are given in Table 3. It is clear from this table that whole-grain consumption by men in the EATS differs by level of education attainment.

Simulation Study for Estimating the Distribution of Usual Intake. Figure 4 illustrates the mean bias, defined as the difference between the estimate and simulated truth (the 365-day mean), of each of the methods. Except at the mean of the distribution, the 2-day within-person means estimate has a much greater bias than any of the other methods. The NCI method with correlated person-specific effects either with or without the FPQ produced estimates that are close to the estimate from the 365-day mean, with essentially no bias for all percentiles.

Figure 5 illustrates that the NCI method with uncorrelated person-specific effects and without the FPQ produced a similar curve to the ISUF method, and that both curves are shifted from the 365-day mean curve. By ignoring correlation between probability and amount, these two methods tend to overestimate the amounts consumed by those with a low probability of consumption, and underestimate the amounts consumed by those with a high probability of consumption, leading to biased estimates. The NCI method with uncorrelated person-specific effects but with the FPQ led to the best results when compared with the same method without the FPQ. Including the shared covariate (the FPQ information) in both parts of the model captured some of the correlation between probability and consumption-day amount, although not as much as incorporating the correlated person-specific effects. Figures 4 and 5 illustrate that modeling the correlation between probability of consumption and consumption-day amount leads to the largest gains over the other methods.

DISCUSSION

A statistical model used for estimating the usual intake of episodically consumed foods using two or more 24-hour recalls per subject needs to appropriately account for the characteristics of such data. First, it must account for the spike at zero due to nonconsumption of a food on the recalled days. This is achieved in the NCI method by representing usual intake as the product of the probability of consuming a food and the amount consumed on a consumption day. In addition, the model must transform skewed distributions of the consumption-day amounts to approximate normality. It also must have the ability to distinguish within-person variability that results from day-to-day differences in intake and random reporting error from the variability among individuals. Like the ISUF method, the NCI method addresses all of these challenges. The NCI method overcomes the limitations of the ISUF method, however, by incorporating covariate information and by accounting for the correlation between the probability of consuming a food and the consumption-day amount. Because probability and amount are substantially correlated in a majority of food groups, as our findings from EATS illustrate, the ISUF method should not be used on most foods according to its own criterion.

The importance of using covariates in the model is highly dependent on the application of interest. When interest is in estimating the distribution of usual intake in a subpopulation or population, it is not important how the between-person variation is partitioned between the part that is explained by covariates and the unexplained component captured by the person-specific effect. Rather, the focus should be on how well the variability of person-specific effects and within-person random error can be transformed to normality. Including covariates in the model may make normality more realistic, with conse-

Table 2. Percentage of individuals in the Eating at America's Table Study who consume less or more than a cutoff number of servings using different methods to estimate the distribution of usual intake[a]

	Women (n=519)			Men (n=446)		
	4-day WPM[b]	NCI[c]	ISUF[d]	4-day WPM[b]	NCI[c]	ISUF[d]
Dark green vegetables (servings)						
<0.10	54.3	30.4	26.4	62.3	31.7	30.9
<0.25	67.6	63.7	51.9	70.6	65.5	59.3
<0.5	82.0	86.6	89.0	82.3	91.6	92.8
>1	6.7	2.0	0.5	5.2	0.1	0.0
Whole grains (servings)						
<0.5	37.2	26.0	17.9	27.6	17.2	9.4
<1	60.5	57.4	55.5	43.3	37.9	29.7
>2	14.8	8.6	5.4	30.0	29.0	26.4
>3	3.9	1.1	0.3	13.5	11.1	7.9

[a]Data from reference 7.
[b]WPM=Within-person Mean Method (mean of 4 days).
[c]NCI=National Cancer Institute Method with correlated person-specific effects and food frequency information as a covariate.
[d]ISUF=Iowa State University Foods Method.

Table 3. Percentage of men in the Eating at America's Table Study who consumed less or more than a cutoff number of servings of whole-grains by education levels[a]

No. servings of whole-grains	High school education or less (n=83)	Some college (n=157)	College graduate (n=206)
	←	%	→
<0.5	32.3	20.6	9.3
<1	57.9	44.2	26.4
>2	14.5	23.3	38.7
>3	4.2	8.1	16.3

[a]Data from reference 7.

quent improvement in estimating the distribution, especially its tails. This was not the case with our example. After allowing for correlated person-specific effects, our simulation study results indicated little difference between the NCI method with and without the FPQ. This may be because the 24-hour recall reported amounts for whole grains (the basis for the simulated data) can be transformed to normality almost equally well unconditionally or conditionally on the FPQ. For foods without this characteristic, including the FPQ in the model may lead to some improvement in estimating the tails of the distribution. Because the FPQ may not be necessary to estimate the usual intake distribution of foods, it should be possible to estimate the usual intake distribution of foods from previous survey data with at least two 24-hour recalls per participant.

When the distribution of usual intake is estimated in subpopulations, incorporating covariates that characterize the subpopulation, such as age, sex, race, income, or education, may provide a substantial improvement by leading to more efficient estimation than does stratification. This advantage in efficiency is expected to increase as size of the subpopulation decreases. In addition, covariates may be used to adjust for temporal effects, such as seasonality and day-of-week effects, and the reduction in mean levels of intake that can occur with repeat 24-hour recalls, the time-in-sample effect. This is done simply by including an indicator variable in the statistical model as a covariate to indicate that the second recall is being modeled, allowing its mean to be adjusted for repeat application of the recall, if necessary.

For predicting individual usual intake and relating it to health outcomes, the goal is to reduce unexplained between-person variation of intake, which may be achieved by including appropriate covariates in the model. Different types of covariates may be incorporated in the estimation of usual intake for this purpose, including personal characteristics associated with intake as well as FPQ information. Because the FPQ frequencies are associated both with the probability of consuming a food and the amount consumed on the consumption day (5), incorporating food frequency information as a covariate in both parts of the model can explain at least part of between-person variation in the 24-hour recall, therefore providing a better estimate of diet–health relations. Although the NCI statistical model is the first step in estimating individual usual intake, the development of additional statistical methodology for the second step is necessary; this work is currently underway by the authors of this article.

Another benefit of including covariates in the statistical model is the improved ability to make inferences

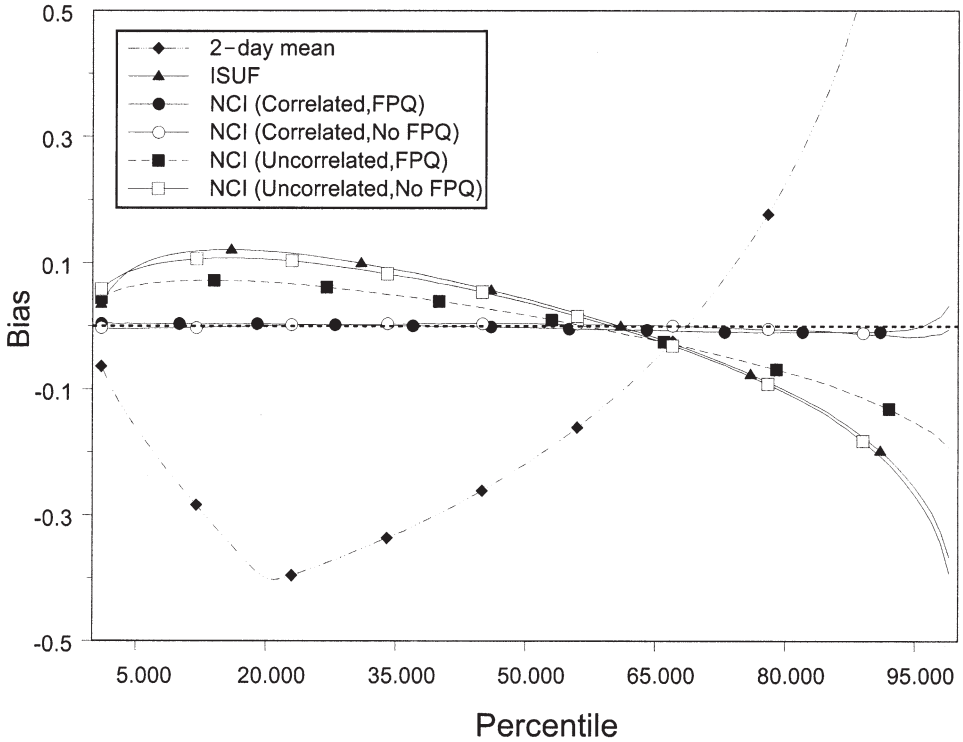

Figure 4. Bias of percentile estimates from simulations based on whole grains for women (from the Eating at America's Table Study). The dashed line at zero represents no bias. 2-day mean=within-person mean of 2 days of simulated 24-hour recalls. ISUF=Iowa State University Foods method. NCI=National Cancer Institute method, specifying whether the person-specific effects are correlated or uncorrelated and whether the simulated Food Propensity Questionnaire (FPQ) is used as a covariate in the model. (This figure is available online at www.adajournal.org as part of a PowerPoint presentation featuring additional online-only content.)

regarding the effects of these covariates on food consumption, similar to the purpose of the statistical models described by Haines and colleagues (9) and Guenther and colleagues (10). Like these approaches, the NCI model allows the analyst to separate the effect of covariates on the decision to consume a food from the effect of covariates on the amount that is consumed when the food is eaten. The NCI model may be used to approximate the effects of individual covariates to determine which variables are associated with the probability of consumption and the consumption-day amount by making inferences about the strength of the relationships of the covariates with individual intake.

The NCI method was developed to meet the special challenges for estimating intake of episodically consumed foods. Although some of those challenges are unique to episodically-consumed foods (accounting for nonconsumption days, allowing for the correlation between the probability of consuming a food and the consumption-day amount), others apply as well to nutrients or foods that are consumed on a daily basis. Part II of the NCI model (equation [B]) alone could be used to estimate usual intake for foods or nutrients that are consumed on a daily basis by nearly everyone. The ability to easily incorporate covariates in the estimation of usual intake makes the NCI method an attractive alternative to the ISU nutrient method.

Although the NCI method seems to present a substantial improvement over existing methods for estimating the distribution of usual intake for foods, it does have some limitations. First, the model never produces a true zero intake because the logistic regression that is used to model the probability of consumption does not predict a zero value. Furthermore, the model requires that a sufficient number of people consume a given food on at least 2 recalled days. For foods that are consumed episodically in the population, such as organ meats, this condition may not be satisfied.

Most importantly, the model is based on the assumption that the 24-hour recall is an unbiased instrument for measuring usual food intake. Many recent studies with doubly labeled water have found misreporting of energy intake on both the 24-hour recall and FFQ, almost always in the direction of underreporting (11-14). This suggests that at least some foods are underreported as well. A few studies have investigated the extent and type of underreporting by food, and it seems that underreporting may be differential by food (15). Unfortunately, however, it is impossible to know which foods, and by how much, are misreported on the 24-hour recall. Due to this uncertainty, we follow the only available practical convention and assume that the 24-hour recall is unbiased. For those foods that are reported with bias on the 24-hour recall, the estimated intake will be biased as well. When the FFQ is used as a covariate in the model, it is allowed to involve systematic bias as well as random measurement

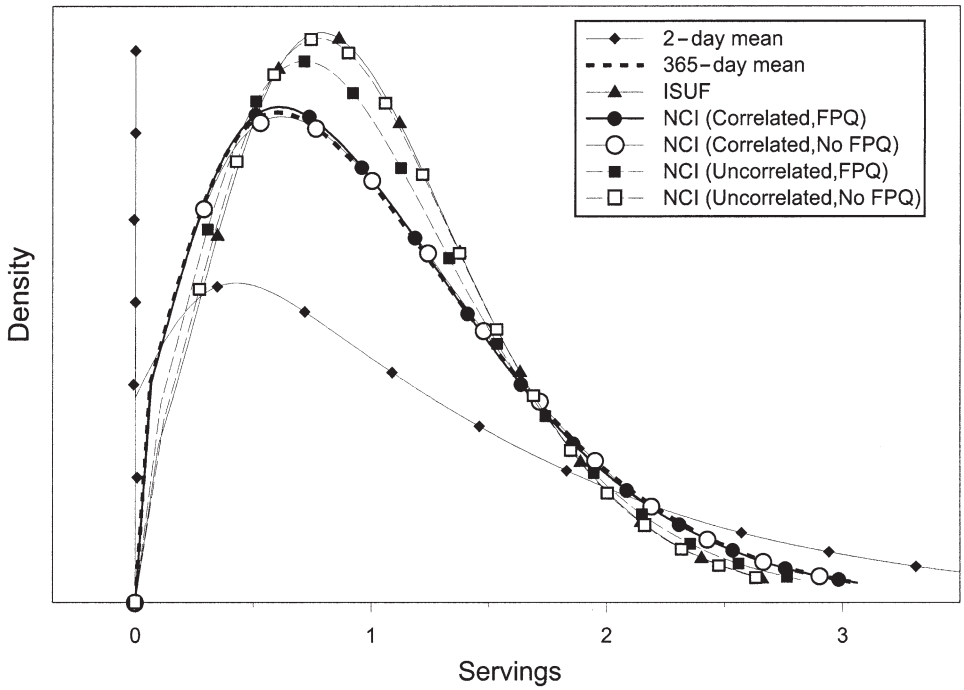

Figure 5. Smoothed distribution curves from simulations based on whole grains for women (from the Eating at America's Table Study). The spike at zero for the 2-day mean represents 18% of the distribution. The 365-day mean represents true usual intake. 2-day mean=within-person mean of 2 days of simulated 24-hour recalls. ISUF=Iowa State University Foods method. NCI=National Cancer Institute method, specifying whether the person-specific effects are correlated or uncorrelated and whether the simulated Food Propensity Questionnaire (FPQ) is used as a covariate in the model.

error. It is important to note that in the NCI method the FFQ does not replace information from the 24-hour recall but is being calibrated using the 24-hour recall as a reference instrument. The method, therefore, includes useful information from the FFQ without subjecting the final estimates to its measurement error.

The NCI method has been developed and illustrated here using data from what we implicitly assumed to be a random sample of independent individuals, not a complex survey sample. We are presently working on its extension for complex surveys, such as NHANES, so that estimates of the distribution of usual intake in the US population may be made using the NCI method.

CONCLUSIONS

The NCI method to estimate usual intake of even episodically consumed foods using two 24-hour recalls represents an advance in dietary assessment. It provides distinct advantages over previously proposed methods by accounting for the correlation between probability of consumption and amount consumed and by incorporating covariate information. A macro that automates the procedure and addresses other special aspects of dietary recall data will be available as part of a Web-based tutorial on the NHANES survey in the future (see http://www.cdc.gov/nchs/tutorials). Researchers interested in estimating the distribution of usual intakes of food groups for a population or subpopulation or the percentage of people that consume more or less than a given standard, are advised to work with a statistician to incorporate this method in their analyses. A further application of the NCI method for estimating individual usual intake and relating it to health outcomes is forthcoming.

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-ES09106).

The authors thank Phillip S. Kott, Joseph D. Goldman, Richard P. Troiano, and Amy Millen for their thoughtful reviews and Anne Brown Rodgers for her expert editing assistance.

References

1. Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, Tooze JA, Krebs-Smith SM. Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. *J Am Diet Assoc*. 2006;106:1640-1650.
2. Nusser SM, Fuller WA, Guenther PM. Estimation of usual dietary intake distributions: Adjusting for measurement error and nonnormality in 24-hour food intake data. In: Trewin D, ed. *Survey Measurement and Process Quality*. New York, NY: Wiley; 1996:689-709.
3. Tooze JA, Grunwald GK, Jones RH. Analysis of re-

peated measures data with clumping at zero. *Stat Methods Med Res*. 2002;11:341-355.
4. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Ser B*. 1964;26:211-252.
5. Subar AF, Dodd KW, Guenther PM, Kipnis V, Midthune D, McDowell M, Tooze JA, Freedman LS, Krebs-Smith SM. The Food Propensity Questionnaire: Concept, development, and validation for use as a covariate in a model to estimate usual food intake. *J Am Diet Assoc*. 2006;106:1556-1563.
6. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semi-parametric transformation approach to estimating usual nutrient intake distributions. *J Am Stat Assoc*. 1996;91:1440-1449.
7. Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, McIntosh A, Rosenfeld S. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *Am J Epidemiol*. 2001;154:1089-1099.
8. USDA Agricultural Research Service. FoodLink. Available at: http://www.ba.ars.usda.gov/cnrg/services/foodlink.html. Accessed April 14, 2006.
9. Haines PS, Guilkey DK, Popkin BM. Modeling food consumption decisions as a two-step process. *Am J Agric Econ*. 1988;70:543-552.
10. Guenther PM, Jensen HH, Batres-Marquez, SP, Chen C-F. Sociodemographic, knowledge, and attitudinal factors related to meat consumption in the United States. *J Am Diet Assoc*. 2005;105:1266-1274.
11. Hill RJ, Davies PS. The validity of self-reported energy intake as determined using the doubly labelled water technique. *Br J Nutr*. 2001;85:415-430.
12. Trabulsi J, Schoeller DA. Evaluation of dietary assessment instruments against doubly labeled water, a biomarker of habitual energy intake. *Am J Physiol Endocrinol Metab*. 2001;281:E891-E899.
13. Macdiarmid J, Blundell J. Assessing dietary intake: Who, what and why of under-reporting. *Nutr Res Rev*. 1998;11:231-253.
14. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, Sunshine J, Schatzkin A. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN Study. *Am J Epidemiol*. 2003;158:1-13.
15. Krebs-Smith SM, Graubard BI, Kahle LL, Subar AF, Cleveland LE, Ballard-Barbash R. Low energy reporters vs others: A comparison of reported food intakes. *Eur J Clin Nutr*. 2000;54:281-287.