

Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships between Episodically Consumed Foods and Health Outcomes

Victor Kipnis,^{1,*} Douglas Midthune,¹ Dennis W. Buckman,² Kevin W. Dodd,¹
Patricia M. Guenther,³ Susan M. Krebs-Smith,⁴ Amy F. Subar,⁴ Janet A. Tooze,⁵
Raymond J. Carroll,⁶ and Laurence S. Freedman⁷

¹Biometry, Division of Cancer Prevention, National Cancer Institute, 6130 Executive Boulevard, EPN-3131, Bethesda, Maryland 20892-7354, U.S.A.

²Information Management Services, Inc., 12501 Prosperity Drive, Silver Spring, Maryland 20904, U.S.A.

³Center for Nutrition Policy and Promotion, U.S. Department of Agriculture, 3101 Park Center Drive, Ste 1034, Alexandria, Virginia 22302, U.S.A.

⁴Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 6130 Executive Boulevard, EPN-4005, Bethesda, Maryland 20892, U.S.A.

⁵Department of Biostatistical Sciences, Wake Forest University, School of Medicine, Medical Center Boulevard, Winston-Salem, North Carolina 27157, U.S.A.

⁶Department of Statistics, Texas A&M University, 3143 TAMU, College Station, Texas 77843-3143, U.S.A.

⁷Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer 52161, Israel

**email*: kipnisv@mail.nih.gov

SUMMARY. Dietary assessment of episodically consumed foods gives rise to nonnegative data that have excess zeros and measurement error. Tooze et al. (2006, *Journal of the American Dietetic Association* **106**, 1575–1587) describe a general statistical approach (National Cancer Institute method) for modeling such food intakes reported on two or more 24-hour recalls (24HRs) and demonstrate its use to estimate the distribution of the food's usual intake in the general population. In this article, we propose an extension of this method to predict individual usual intake of such foods and to evaluate the relationships of usual intakes with health outcomes. Following the regression calibration approach for measurement error correction, individual usual intake is generally predicted as the conditional mean intake given 24HR-reported intake and other covariates in the health model. One feature of the proposed method is that additional covariates potentially related to usual intake may be used to increase the precision of estimates of usual intake and of diet-health outcome associations. Applying the method to data from the Eating at America's Table Study, we quantify the increased precision obtained from including reported frequency of intake on a food frequency questionnaire (FFQ) as a covariate in the calibration model. We then demonstrate the method in evaluating the linear relationship between log blood mercury levels and fish intake in women by using data from the National Health and Nutrition Examination Survey, and show increased precision when including the FFQ information. Finally, we present simulation results evaluating the performance of the proposed method in this context.

KEY WORDS: Dietary measurement error; Dietary survey; Episodically consumed foods; Excess zero models; Food frequency questionnaire; Fish; Individual usual intake; Mercury; Nonlinear mixed models; Regression calibration; 24-hour recall.

1. Introduction

U.S. national nutritional surveys traditionally have used the 24-hour recall (24HR) to collect information on food intake as the primary assessment instrument (Dwyer et al., 2003). The main purposes of such surveys are to estimate the distribution of usual (that is, average long-term) intake of nutrients and foods in the population, and to monitor such intakes over time. Another important purpose is to relate individual usual intakes to health outcomes such as blood pressure.

Even assuming unbiasedness of the 24HR, there has been concern over its use for assessing intake of foods that are not typically consumed every day. Many consumers of such

episodically consumed foods report zero intake on the 24HR if the report happens to be on a nonconsumption day. Consequently, with typically only one or two administrations of a 24HR in surveys, usual intake of such foods is difficult to estimate. For this reason, an additional instrument, a food frequency questionnaire (FFQ) that queries frequency of consumption over the past year, was included in the National Health and Nutrition Examination Survey (NHANES) conducted in 2003–2006 (Subar et al., 2006). Although the FFQ leads to biased reporting of intake of energy (Kipnis et al., 2003) and therefore of at least some foods, it might nevertheless provide valuable information together with the 24HR to improve estimates of usual intake.

Dodd et al. (2006) reviewed the methods for estimating distributions of usual intake. Tooze et al. (2006) proposed a new method, called the NCI method, to handle nonnegative data with excess zeros that occur in 24HR reports on episodically consumed foods, and demonstrated its use for estimating distributions of usual intakes. Generalizing the two-part modeling approach to longitudinal semicontinuous observations (Olsen and Schafer, 2001; Tooze, Grunwald, and Jones, 2002) to include latent variables, the NCI method uses a two-part nonlinear mixed effects *measurement error* model with correlated random effects, where both parts may incorporate covariates, including other dietary-assessment instruments, such as a FFQ. In this article, we propose an extension of the NCI method to estimate an individual's usual intake of episodically consumed foods using 24HR data with covariate information. The method may fill a void in analyzing relationships between usual intake and health outcomes for these foods.

All dietary assessment methods based on self-report, including the 24HR, fail to measure true usual intake precisely. The measurement error distorts diet-health outcome relationships, often attenuating them. A popular method of correcting for measurement error is regression calibration (Carroll et al., 2006), which uses, in place of the unknown usual intake, its best mean square error (MSE) predictor, that is, its estimated conditional expectation given the observed 24HRs and other covariates in the health outcome model.

In this article, we derive this conditional expectation for episodically consumed foods. In Section 2, we describe the measurement error model and derive the corresponding regression calibration predictor. The approach allows conditioning on additional covariates related to intake, which may increase the precision of estimates of usual intake and diet-health outcome associations. In Section 3, using data from the Eating at America's Table Study (EATS), we quantify the increased precision obtained from including a FFQ report as a covariate. In Section 4, we demonstrate the method for evaluating the relationship between blood mercury and fish intake based on NHANES data and show increased precision of the estimated association from incorporating the FFQ. In Section 5, we present simulations to evaluate the finite-sample performance of the method. Section 6 contains discussion.

2. Measurement Error and Regression Calibration Models

For individual i on day j , $i = 1, \dots, n$; $j = 1, \dots, J_i$; let T_{ij} and R_{ij} denote true intake of a nutrient or food and the corresponding 24HR, respectively. We define true individual usual intake T_i as the within-person expectation of true daily intake, $T_i = E(T_{ij} | i)$. The notation $E(\bullet | i)$ indicates that the expectation is conditional on the i th individual. Let Y_i denote a health outcome possibly related to T_i through the regression model

$$E(Y_i | T_i, \mathbf{Z}_i) = m(\alpha_0 + \alpha_T T_i + \alpha_z^t \mathbf{Z}_i), \quad (1)$$

where $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^t$ is a vector of covariates measured without error, and m^{-1} is a link function. For example, $m(v) \equiv v$ leads to linear regression, while $m(v) \equiv H(v) = 1/(1 + e^{-v})$ to logistic regression. Our main interest is estimation of the dietary effect

$$\Delta_T = m^{-1}\{E(Y_i | T_1, \mathbf{Z}_i)\} - m^{-1}\{E(Y_i | T_0, \mathbf{Z}_i)\}, \quad (2)$$

when dietary intake changes from T_0 to T_1 . For example, for linear regression, expression (2) represents a change in the mean outcome, while, for logistic regression, a log odds ratio.

Let $\mathbf{X}_i = (\mathbf{Z}_i^t, \mathbf{C}_i^t)^t$ be a vector of covariates related to usual intake. It generally includes covariates \mathbf{Z}_i in the health outcome model (1) and some additional factors $\mathbf{C}_i = (C_{i1}, \dots, C_{iq})^t$ that are related to usual intake T_i given \mathbf{Z}_i , but unrelated to outcome Y_i given T_i and \mathbf{Z}_i . Although \mathbf{C}_i formally follows the definition of instrumental variables, its use here is different.

Regression calibration requires evaluation of the best MSE predictor $E(T_i | \mathbf{R}_i, \mathbf{X}_i)$ given the reported intakes $\mathbf{R}_i = (R_{i1}, \dots, R_{iJ_i})^t$ and covariates \mathbf{X}_i for individual i . Assuming that error in \mathbf{R}_i is nondifferential with respect to Y_i , i.e., the conditional distribution of Y_i given (T_i, \mathbf{X}_i) is independent of \mathbf{R}_i , using predictor $E(T_i | \mathbf{R}_i, \mathbf{X}_i)$ in model (1) in place of the unknown T_i retains the same (e.g., linear regression), or approximately the same (e.g., logistic regression), regression coefficient α_T . When predictor $E(T_i | \mathbf{R}_i, \mathbf{X}_i)$ is consistently estimated, regression calibration yields a consistent (approximately consistent for most nonlinear models) estimate of α_T . It is required that \mathbf{X}_i include all covariates \mathbf{Z}_i in the health outcome model (1). We show theoretically in Web Appendix A and through examples in the main text that inclusion of additional predictors \mathbf{C}_i in \mathbf{X}_i can improve both the MSE of predicted usual intake and the precision of the estimated coefficient $\hat{\alpha}_T$.

Before introducing the measurement error model for episodically consumed foods, we present the model for nutrients and foods consumed daily.

2.1 Statistical Model for Daily Consumed Nutrients or Foods

2.1.1 *Classical measurement error model.* Following convention, the 24HRs are assumed unbiased for individual usual intake, i.e.,

$$R_{ij} = T_i + \varepsilon_{ij}, \quad E(\varepsilon_{ij} | i) = 0, \quad (3)$$

where within-person random errors ε_{ij} reflect the daily variation in an individual's intake and other sources of random error. The model requires that ε_{ij} be independent of T_i and each other and have a constant variance σ_ε^2 . In addition, it is often assumed that $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Assume the linear regression of T_i on covariates $\mathbf{X}_i = (X_{i1}, \dots, X_{ik})$, $k = p + q$, i.e.,

$$T_i = \beta_0 + \beta_X^t \mathbf{X}_i + u_i, \quad u_i \sim \text{Normal}(0; \sigma_u^2). \quad (4)$$

From equations (3)–(4), the 24HR-reported intake follows the linear mixed measurement error model

$$R_{ij} = \beta_0 + \beta_X^t \mathbf{X}_i + u_i + \varepsilon_{ij}. \quad (5)$$

In addition to the fixed effect population-level parameters $\beta = (\beta_0, \beta_X^t)^t$, model (5) includes the random effect u_i representing person-specific deviations of usual intake from the population profile defined by covariates \mathbf{X}_i , and within-person random error ε_{ij} . Two or more 24HRs on a number of individuals are required to distinguish between- and within-person variation and uniquely estimate all model parameters.

Evaluation of the regression calibration predictor $E(T_i | \mathbf{R}_i, \mathbf{X}_i)$ is a well-known procedure in the theory of mixed models (McCulloch and Searle, 2001). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \sigma_u^2, \sigma_\varepsilon^2)^t$ be the vector of parameters in model (5) and f denote a probability density function. Because, from equation (4), T_i is a function of \mathbf{X}_i and u_i , denoted as $T_i = \mathfrak{T}(X_i, u_i; \boldsymbol{\theta})$, we have

$$\hat{T}_i(\boldsymbol{\theta}) \equiv E(T_i | \mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta}) = \int \mathfrak{T}(\mathbf{X}_i, u_i; \boldsymbol{\theta}) f(u_i | \mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta}) du_i, \quad (6)$$

where, according to Bayes' theorem

$$f(u_i | \mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta}) = \frac{f(\mathbf{R}_i | \mathbf{X}_i, u_i; \boldsymbol{\theta}) f(u_i | \mathbf{X}_i; \boldsymbol{\theta})}{\int f(\mathbf{R}_i | \mathbf{X}_i, u_i; \boldsymbol{\theta}) f(u_i | \mathbf{X}_i; \boldsymbol{\theta}) du_i}. \quad (7)$$

When parameters in $\boldsymbol{\theta}$ are estimated by fitting model (5) to the data, predicted usual intake $\hat{T}_i(\hat{\boldsymbol{\theta}})$ is known as the Empirical Bayes' (EB) estimator, which for the linear model (5) is also known as the best linear unbiased predictor (BLUP), given by the weighted average

$$\hat{T}_i(\hat{\boldsymbol{\theta}}) = \hat{w}_i \bar{R}_i + (1 - \hat{w}_i) (\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_X^t X_i), \quad (8)$$

of the mean of the J_i reported intakes, \bar{R}_i , and the covariate predictor $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_X^t X_i$, with weights $\hat{w}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / J_i}$ and $1 - \hat{w}_i$, respectively (McCulloch and Searle, 2001).

This methodology is already known and was previously suggested for classical measurement error correction (e.g., Whittemore, 1989; Tsiatis, DeGruttola, and Wulfsohn, 1995), but is applicable only to reported intakes that follow the classical error model on the original scale.

2.1.2 Classical error model on transformed scale. Often, within-person random error in the 24HR reported intake is dependent on the individual mean and has a skewed distribution, violating the classical error model assumptions. The most common fix has been to monotonically transform the intakes R_{ij} to values $R_{ij}^* = g(R_{ij})$ that more closely follow the classical model with normally distributed error

$$\begin{aligned} \hat{T}_i(\boldsymbol{\theta}) &\equiv E \left\{ g * (\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i, \lambda_R) | \mathbf{R}_i^*, \mathbf{X}_i; \boldsymbol{\theta} \right\} \\ &= \frac{\int g * (\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i, \lambda_R) \prod_{j=1}^{J_i} \left\{ \frac{1}{\sigma_\varepsilon} \phi \left(\frac{g(R_{ij}, \lambda_R) - \beta_0 - \boldsymbol{\beta}_X^t \mathbf{X}_i - u_i}{\sigma_\varepsilon} \right) \right\} \frac{1}{\sigma_u} \phi(u_i / \sigma_u) du_i}{\int \prod_{j=1}^{J_i} \left\{ \frac{1}{\sigma_\varepsilon} \phi \left(\frac{g(R_{ij}, \lambda_R) - \beta_0 - \boldsymbol{\beta}_X^t \mathbf{X}_i - u_i}{\sigma_\varepsilon} \right) \right\} \frac{1}{\sigma_u} \phi(u_i / \sigma_u) du_i}, \end{aligned} \quad (13)$$

(Eckert, Carroll, and Wang, 1997). In most cases, it may be achieved using the Box-Cox family of transformations (Box and Cox, 1964)

$$g(v, \lambda) = (v^\lambda - 1) / \lambda. \quad (9)$$

We assume that such a transformation exists and that, on the transformed scale, we have

$$\begin{aligned} R_{ij}^* &\equiv g(R_{ij}, \lambda_R) \\ &= E\{g(R_{ij}, \lambda_R) | i\} + \varepsilon_{ij}, \varepsilon_{ij} \sim \text{Normal}(0, \sigma_\varepsilon^2), \end{aligned}$$

where ε_{ij} are independent of the individual mean $\mu_i^* = E(R_{ij}^* | i)$ and of each other. Assuming the regression of μ_i^*

on \mathbf{X}_i is linear with a normally distributed regression error u_i , i.e.,

$$\mu_i^* = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i, u_i \sim \text{Normal}(0; \sigma_u^2),$$

the reported intakes follow the *nonlinear* mixed effects measurement error model

$$g(R_{ij}, \lambda_R) = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij}. \quad (10)$$

Following convention (Dodd et al., 2006), we continue to assume that 24HRs are unbiased for true individual intake *on the original scale*, so the usual intake of person i is:

$$\begin{aligned} T_i &\equiv E(R_{ij} | \mathbf{X}_i, u_i) \\ &= E \left\{ g^{-1} (\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij}, \lambda_R) | \mathbf{X}_i, u_i \right\} \\ &\approx g * (\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i, \lambda_R), \end{aligned} \quad (11)$$

where, from Taylor's expansion,

$$g * (v, \lambda_R) = g^{-1}(v, \lambda_R) + \frac{1}{2} \sigma_\varepsilon^2 \frac{\partial^2 \{g^{-1}(v, \lambda_R)\}}{\partial v^2}. \quad (12)$$

Following equation (11), the best predictor of individual usual intake is given by

$$E(T_i | \mathbf{R}_i, \mathbf{X}_i) \approx E \left\{ g * (\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i, \lambda_R) | \mathbf{R}_i, \mathbf{X}_i \right\}.$$

When g is the identity function, this conditional expectation reduces to the BLUP (8). For general g , it is different and needs to be evaluated according to equations (6)–(7). Because u_i and ε_{ij} are independent and normally distributed, and because conditioning on $(\mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta})$ is the same as conditioning on $(\mathbf{R}_i^*, \mathbf{X}_i; \boldsymbol{\theta}) \equiv \{g(\mathbf{R}_i, \lambda_R), \mathbf{X}_i; \boldsymbol{\theta}\}$, $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}_X^t, \sigma_\varepsilon^2, \sigma_u^2, \lambda_R)^t$, we have:

$$\begin{aligned} f(\mathbf{R}_i^* | \mathbf{X}_i, u_i; \boldsymbol{\theta}) &= \prod_{j=1}^{J_i} \left\{ \frac{1}{\sigma_\varepsilon} \phi \left(\frac{g(R_{ij}) - \beta_0 - \boldsymbol{\beta}_X^t \mathbf{X}_i - u_i}{\sigma_\varepsilon} \right) \right\}, \\ f(u_i | \mathbf{X}_i; \boldsymbol{\theta}) &= \frac{1}{\sigma_u} \phi(u_i / \sigma_u), \end{aligned}$$

so that

where ϕ is the standard normal distribution density. Following the EB approach, the integrals in expression (13) are evaluated by substituting in $\hat{\boldsymbol{\theta}}$ after fitting model (10) to the data.

As far as we know, this method has not previously been described and should be useful in itself for intakes of nutrients or foods that are daily consumed. It is also an intermediate step to the estimation of usual intake for episodically consumed foods.

2.2 Statistical Model for Episodically Consumed Foods

2.2.1 The measurement error model. Generalizing the approach developed at Iowa State University (Nusser, Fuller, and Guenther, 1997) to deal with episodically consumed foods,

Tooze et al. (2006) considered two components of usual intake in the NCI method. The first is the individual *probability* to consume a food on a given day, $p_i = P(T_{ij} > 0 | i)$. The second is the usual intake *amount* on a consumption day, $A_i = E(T_{ij} | i; T_{ij} > 0)$. It follows that usual intake is

$$T_i \equiv E(T_{ij} | i) = p_i A_i, \quad (14)$$

the product of the probability to consume and the usual amount on consumption days.

To specify a measurement error model in this case requires modifying the assumptions. Following Tooze et al. (2006), we assume that (i) a food is reported on the 24HR as consumed on a certain day if and only if it *was* consumed on that day,

$$\begin{aligned} \hat{T}_i(\boldsymbol{\theta}) &\equiv E(T_i | \mathbf{R}_i, \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}) \\ &\approx \frac{\int H(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i}) g^* (\beta_{20} + \boldsymbol{\beta}_{X2}^t \mathbf{X}_{2i} + u_{2i}, \lambda_R) f\{g(\mathbf{R}_i, \lambda_R) | \mathbf{u}_i, \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}\} f(\mathbf{u}_i | \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}) du_{1i} du_{2i}}{\int f\{g(\mathbf{R}_i, \lambda_R) | \mathbf{u}_i, \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}\} f(\mathbf{u}_i | \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}) du_{1i} du_{2i}}, \end{aligned} \quad (20)$$

so that $P(R_{ij} > 0 | i) = P(T_{ij} > 0 | i) \equiv p_i$ and (ii) the 24HR is unbiased for true usual intake on consumption days, $E(R_{ij} | i; R_{ij} > 0) = A_i$. From this it follows that overall the 24HR is unbiased for true usual intake

$$E(R_{ij} | i) = p_i A_i = T_i. \quad (15)$$

Following the NCI method, we consider a two-part measurement error model for the 24HR. In the first part, we model the consumption probability as the mixed effects logistic regression

$$P(R_{ij} > 0 | i) \equiv p_i = H(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i}), j = 1, \dots, J_i, \quad (16)$$

where $u_{1i} \sim \text{Normal}(0; \sigma_{u1}^2)$ is independent of \mathbf{X}_{1i} . In addition to fixed effect population-level parameters $\boldsymbol{\beta}_1 = (\beta_{10}, \boldsymbol{\beta}_{X1}^t)^t$, the model includes the random effect u_{1i} allowing an individual's consumption probability to deviate from the population profile defined by \mathbf{X}_{1i} . In the second part, the measurement error model for the positive reported intake is the same as equation (10), except that it relates only to consumption days. Specifically, we assume that the Box-Cox transformed positive intake follows the nonlinear mixed effects measurement error model

$$g(R_{ij}, \lambda_R | R_{ij} > 0) = \beta_{20} + \boldsymbol{\beta}_{X2}^t \mathbf{X}_{2i} + u_{2i} + \varepsilon_{ij}, \quad (17)$$

where u_{2i} and ε_{ij} are independent of $(\mathbf{X}_{1i}, \mathbf{X}_{2i})$, each other, and are normally distributed.

The two parts of the model are linked in two ways. First, the random effects, u_{1i} and u_{2i} , may be correlated, so that

$$\begin{aligned} \mathbf{u}_i &= (u_{1i}, u_{2i})^t = \text{Normal}(0, \Sigma_u), \\ \Sigma_u &= \begin{pmatrix} \sigma_{u1}^2 & \rho_{u1, u2} \sigma_{u1} \sigma_{u2} \\ \rho_{u1, u2} \sigma_{u1} \sigma_{u2} & \sigma_{u2}^2 \end{pmatrix}. \end{aligned} \quad (18)$$

Second, both parts of the model may share common covariates among the components of \mathbf{X}_{1i} and \mathbf{X}_{2i} , also inducing correlation between probability and amount.

In our model, the probability of consumption for an individual may be arbitrarily small, but is always positive. The model therefore allows for any finite number of days with zero intakes, but does not incorporate never-consumers, if they exist. We discuss this further in Section 6.

2.2.2 Regression calibration model. According to equation (15), usual intake T_i on the original scale, is

$$\begin{aligned} T_i &= P(R_{ij} > 0 | i) E\{g^{-1}(R_{ij}, \lambda_R) | i; R_{ij} > 0\} \\ &\approx H(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i}) g^* (\beta_{20} + \boldsymbol{\beta}_{X2}^t \mathbf{X}_{2i} + u_{2i}, \lambda_R), \end{aligned} \quad (19)$$

where g^* is defined by equation (12). As before, we follow formulas (6)–(7) to obtain:

where

$$\begin{aligned} &f\{g(\mathbf{R}_i, \lambda_R) | \mathbf{u}_i, \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}\} \\ &= \prod_{j=1}^{J_i} \left[\frac{\{\exp(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i})\}^{I(R_{ij} > 0)}}{1 + \exp(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i})} \right. \\ &\quad \left. \times \left\{ \frac{1}{\sigma_\varepsilon} \phi \left(\frac{g(R_{ij}, \lambda_R) - \beta_{20} - \boldsymbol{\beta}_{X2}^t \mathbf{X}_{2i} - u_{2i}}{\sigma_\varepsilon} \right) \right\}^{I(R_{ij} > 0)} \right]; \end{aligned}$$

$$\begin{aligned} f(\mathbf{u}_i | \mathbf{X}_{1i}, \mathbf{X}_{2i}; \boldsymbol{\theta}) &\equiv f(\mathbf{u}_i; \boldsymbol{\theta}) = \frac{1}{\sigma_{u1} \sqrt{1 - \rho_{u1, u2}^2}} \\ &\times \phi \left(\frac{u_{1i} - \rho_{u1, u2} (\sigma_{u1} / \sigma_{u2}) u_{2i}}{\sigma_{u1} \sqrt{1 - \rho_{u1, u2}^2}} \right) \frac{1}{\sigma_{u2}} \phi \left(\frac{u_{2i}}{\sigma_{u2}} \right), \end{aligned}$$

$$\boldsymbol{\theta} = (\beta_{10}, \boldsymbol{\beta}_{X1}^t, \beta_{20}, \boldsymbol{\beta}_{X2}^t, \rho_{u1, u2}, \sigma_{u1}, \sigma_{u2}, \lambda_R)^t.$$

Following the EB approach, the integrals in equation (20) may be evaluated using adaptive Gaussian quadrature (Liu and Pierce, 1994) by substituting in the maximum likelihood estimates of parameters $\boldsymbol{\theta}$ after simultaneously fitting models (16)–(18) to the data.

When usual intake has a skewed distribution with high leverage points, one might transform it to a more appropriate scale before relating to a health outcome in model (1). Assuming that such transformation is given by $g(T_i, \lambda_T)$, the only change in the methodology is in substituting

$$\begin{aligned} g(T_i, \lambda_T) &\approx g\{H(\beta_{10} + \boldsymbol{\beta}_{X1}^t \mathbf{X}_{1i} + u_{1i}) \\ &\quad g^* (\beta_{20} + \boldsymbol{\beta}_{X2}^t \mathbf{X}_{2i} + u_{2i}, \lambda_R), \lambda_T\} \end{aligned} \quad (21)$$

instead of T_i in formula (20).

3. Contribution of FFQ Data

Because the FFQ report is related to true food intake and is independent of the health outcome given true intake and other covariates in model (1), it may be included as an additional covariate (a component of vector \mathbf{C}_i) in the calibration model.

In this section, we quantify the contribution of the FFQ to the prediction of individual usual intake, using the EATS data (Subar et al., 2001). We considered 965 respondents who successfully completed four 24HRs and a FFQ.

We assumed a simple univariate model relating a transformed food intake to either a hypothetical continuous health outcome (linear regression) or to a dichotomous outcome (logistic regression). For a given food, we fit models (16)–(18) to the data, estimated the distribution of usual intake by applying the NCI method (Tooze et al., 2006), found the best Box–Cox transformation of true intake to approximate normality, and finally predicted individual usual intake on the transformed scale. We assumed the same set of additional covariates for both parts of the model and considered three different sets. The first set was empty; the second contained age, body mass index (BMI), and education (no college, some college, and college graduate); the third set additionally contained the FFQ report. We Box–Cox transformed FFQ positive values to improve linearity and homoscedasticity of model (17), and used an indicator variable for zero FFQ reports.

Finally, we calculated the variances V_1 , V_2 , V_3 of the predicted usual intakes corresponding to each of the three scenarios described above. We quantified the contribution of the additional set of covariates (age, BMI, education) by the ratio

V_2/V_1 , and the additional contribution of the FFQ as V_3/V_2 . As we prove in Web Appendix A, the larger the ratios V_2/V_1 and V_3/V_2 are, the larger is the contribution of the covariate(s) to the precision of the predictor and the greater is the precision of estimated exposure effects in the health outcome model.

We present results of our analyses in Table 1, separately for men and women, for five selected food groups: dark green vegetables, tomatoes and tomato products, fruit, milk and milk products, and fish. The variance ratios show that the FFQ contributed considerably more to the estimation of usual intake than did the other covariates, but the size of contribution varied substantially across foods. For dark green vegetables, tomatoes (for women), and fish intake, the increase in efficiency due to the FFQ was large and ranged from 21% (dark green vegetables for women) to 137% (dark green vegetables for men). For other foods, the increase in efficiency from the FFQ was more modest, ranging from 3% (fruit for men) to 13% (tomatoes for men).

4. Relationship between Fish Intake and Mercury Level in Women in NHANES

To further illustrate our methods, we examined the relationship between fish intake and blood mercury levels in women of

Table 1
Predicting individual intake in EATS using the proposed method: contribution of additional covariates in the calibration model

Food group	Gender	Percentage with 1 or more 24HR > 0	Percentage with 2 or more 24HR > 0	Additional covariates in calibration model	Predictor variance	Ratio V_2/V_1	Ratio V_3/V_2
Dark green vegetables	Male	48.0	19.5	None	0.1087	1.28	
				Age, education, BMI	0.1394		
	Age, education, BMI, FFQ	0.3307	2.37				
	None	0.2079					
Female	56.6	26.9	None	0.2917	1.40		
			Age, education, BMI	0.3539		1.21	
Tomatoes	Male	98.2	90.3	None	0.0918		1.03
				Age, education, BMI	0.0941		
	Age, education, BMI, FFQ	0.1064	1.13				
	None	0.0203					
Female	98.2	86.5	None	0.0250	1.23		
			Age, education, BMI	0.0372		1.49	
Fruit	Male	98.9	92.6	None	1.0009		1.02
				Age, education, BMI	1.0185		
	Age, education, BMI, FFQ	1.0472	1.03				
	None	0.5727					
Female	97.5	91.9	None	0.5800	1.01		
			Age, education, BMI	0.6044		1.04	
Milk	Male	98.9	91.9	None	0.4821		1.00
				Age, education, BMI	0.4827		
	Age, education, BMI, FFQ	0.5002	1.04				
	None	0.3950					
Female	97.5	89.8	None	0.3950	1.00		
			Age, education, BMI	0.4246		1.07	
Fish	Male	54.1	20.3	None	0.2069		1.18
				Age, education, BMI	0.2438		
	Age, education, BMI, FFQ	0.4672	1.92				
	None	0.2096					
Female	48.0	14.8	None	0.2348	1.12		
			Age, education, BMI	0.3590		1.53	
Age, education, BMI, FFQ							

child-bearing age by using NHANES data. This subject is important because it has been shown that high levels of mercury can increase complications of pregnancy (Xue et al., 2007). We analyzed 1605 females, aged 12–49 years, who participated in the 2003–2004 round of NHANES, and provided at least one 24HR, one FFQ, and a blood sample for measurement of serum mercury. Among these participants, 1206 (75%) reported no fish consumption on either 24HR, 342 (21%) reported fish consumption on one of the two days, and 57 (3.6%) reported consuming fish on both days. In view of the large proportion of nonconsumption days, one might expect that the FFQ, with 1553 (96.8%) women reporting fish consumption over the past year, would add useful information.

We examined the linear regression of log serum mercury level ($\mu\text{g/l}$) on usual fish intake (oz/day), obtained after a suitable Box–Cox transformation. We compared three regressions, one where individual fish intake was represented by the average of the 24HRs (the “naïve” analysis), and the other two that used the regression calibration to adjust for measurement error in the 24HR using the proposed method. In the first calibration model, individual fish intake was predicted using age, race (White, African American, other), and education (no college, some college, college graduate) as additional covariates (components of vector \mathbf{C}_i). In the second, fish intake was predicted with the same three additional covariates plus the FFQ frequencies that were handled the same way as described above for the EATS example. Standard errors (SEs) of the estimated regression slopes were estimated using the balanced repeated replication method (Wolter, 1995). Results are presented in Table 2.

The naïve analysis indicates a clear relationship between serum mercury and fish intake (Wald’s $z = 0.33/0.038 = 8.7$), but quantifies the effect as a 39% increase ($\exp(0.33)$) in serum mercury between those who consume an average of 0.1 oz of

fish per day and those who consume 1 oz per day. (These consumption levels were approximately the 10th and 90th percentiles in the population.) Such a small increase might only be of modest public health concern. The proposed method estimates the increase in serum mercury to be approximately fivefold ($\exp(1.58)$) to sevenfold ($\exp(1.97)$) with/without the FFQ in the calibration model, respectively, values that certainly would warrant concern. Note that the addition of the FFQ increases efficiency in the estimated effect approximately twofold yielding a SE of 0.38 compared with 0.54 for the regression calibration without the FFQ.

To assess model fit, we applied an informal graphical approach, as described in Web Appendix B. The results (Web Figures 1–4) did not exhibit obvious model misspecification.

5. Simulation Results

We conducted a simulation study to evaluate the performance of the proposed method in a finite sample. We designed our simulation to mimic the investigation of the relationship between serum mercury levels and usual fish intake in women in NHANES, described in Section 4. In the simulation, the true relationship was specified as the simple linear regression

$$Y_i = -1.35 + 0.8T_i^* + \delta_i, \quad (22)$$

where Y represented log mercury concentration ($\mu\text{g/l}$), and T^* represented Box–Cox transformed usual fish intake (oz/day) with parameter $\lambda_T = 0.17$, chosen to improve the linearity and homoscedasticity of the model. The coefficient of 0.8 leads to a true 1.52 increase in log mercury level between persons who consume 0.1 oz of fish compared with 1 oz per day, similar to the 1.58 estimated from the NHANES data (Table 2). We used $\delta_i = \text{Normal}(0, 1.07)$ so that the linear regression (22) would explain 25% of the total variation of Y .

The details of the simulations are provided in Web Appendix C. Three different sets of additional covariates (components of vector \mathbf{C}_i) were used in the calibration model: (a) empty set; (b) age, BMI, and education; and (c) age, BMI, education, and Box–Cox transformed FFQ report. Table 3 presents the overall results of applying this procedure to 250 simulated data sets.

The results show that, due to measurement error, the naïve approach using the average of the 24HRs grossly underestimates the true value, as expected by theory. On average, estimates based on the proposed method have negligible bias,

Table 2

Estimated difference in log mercury ($\mu\text{g/l}$) between women with an average of 0.1 oz and 1.0 oz of fish per day in NHANES

Estimate of fish intake used	Estimated difference (SE)
Average 24HR (naïve)	0.33 (0.038)
Proposed method without FFQ	1.97 (0.54)
Proposed method with FFQ	1.58 (0.38)

Table 3

Estimating the increase in log mercury level between consumers of an average of 0.1 oz and 1.0 oz of fish per day: empirical results based on 250 simulations regressing log mercury level on estimated usual fish intake (true model: $Y_i = -1.35 + 0.8T_i^ + \delta_i$)*

Increase in log mercury	Estimated usual fish intake			
	Average 24HR (naïve)	Intercept only	Proposed method Age, BMI, and education	Proposed method Age, BMI, education, and FFQ
True value	1.524	1.524	1.524	1.524
Mean of the 250 estimates	0.213	1.546	1.490	1.508
(SE of mean)	(0.001)	(0.018)	(0.014)	(0.009)
Empirical SE of estimate	0.020	0.288	0.226	0.145
Root MSE	1.311	0.289	0.229	0.146

although, again as expected by theory, their precision is poorer than that of the naïve estimate. Importantly, the precision improves with the inclusion of additional covariates for predicting an individual’s usual intake. The estimate based on demographic covariates and the FFQ report is four times more efficient ($[0.018/0.009]^2$) than the estimate based on no covariates and 2.42 times more efficient than the estimate based on the demographic covariates only. The latter effect is equivalent to reducing the sample size of the study by approximately 60% ($[1 - 1/2.42] \times 100\%$), illustrating the potential gains from including the FFQ in the prediction of an individual’s usual intake.

6. Discussion

We have presented a method of predicting an individual’s usual intake of an episodically consumed food and relating it to a health outcome. The method is based on regression calibration prediction applied to short-term repeat observations of intake that contain measurement error and excess zeros, under two important assumptions. First, the fact of short-term consumption is assumed to be correctly classified. Second, the reported intake on consumption days is assumed unbiased for true intake. In our method, information from the main dietary instrument may be combined with that from another longer-term, presumably less precise and even biased, report using an auxiliary instrument. We have demonstrated, through real data and simulations, that the gain from combining two instruments may be substantial, with increases in the precision of the predicted usual intake and of the estimated diet-health outcome relationship.

In our applications, the main instrument was a 24HR and the auxiliary instrument a FFQ. Unfortunately, the assumption of unbiasedness of the main instrument does not strictly apply to the 24HR. Recent biomarker studies (Kipnis et al., 2003) have shown that, for total energy, the 24HR also involves systematic error related to true usual intake. Such biases in reporting energy intake indicate bias also in the reporting of at least some energy-contributing foods. On the other hand, these same studies confirmed that the bias in 24HR reports is considerably less than that in FFQs. Thus, in the absence of any accurate biomarker for most foods and nutrients, using the 24HR in our proposed method may provide the best available approximation.

Our method appears to fill a gap in the analytic tools of nutritional epidemiologists estimating food and health outcome associations. Use of 24HRs alone is known to be problematic when there is a large number of zero values, whereas use of the FFQ alone is marred by the large reporting biases of this instrument. Our examples have demonstrated that the proposed method is feasible to implement and produces nearly unbiased estimates of associations of intakes of episodically consumed foods with health outcomes. The method outperformed the “naïve” approach even without the FFQ in the calibration model, giving an estimate with a much reduced MSE. However, use of the FFQ greatly increased the precision of the estimate.

As shown in Section 3, use of the FFQ will not have a large impact for all foods. Probably the most important factor that determines the impact of the FFQ is the overall probability to

consume the food on a given day. For foods with a relatively low probability of consumption (e.g., fish and dark green vegetables in Table 1), the FFQ will most likely provide a larger increase in efficiency. However, a larger sample size (or, alternatively, more repeat 24HRs) is required to obtain reliable estimates of the model parameters when the consumption probability is very low. This is because a substantial number of individuals with at least two consumption days are needed to estimate properly the within-person variance in the second part of the model. In our NHANES example, there were 57 women (out of 1605) who consumed fish on both days. We would not expect reliable fits for very rarely consumed foods (e.g., organ meats or yogurt in NHANES) with considerably fewer than 50 individuals with two positive intakes and indeed we have encountered some convergence problems in simulations of such cases.

In our two-part model, the first part specifies the probability of the point mass at zero, and the second part *conditionally* models the continuous variable given that it is positive. Another potential approach to modeling semicontinuous data with measurement error was proposed by Li, Shao, and Palta (2005). It is based on the sample selection model that posits an underlying continuous variable censored by a random mechanism. Using our notation, true long-term and reported intakes are specified as $T_i = \max(0, V_i)$ and $R_{ij} = \max(0, V_i + \varepsilon_{ij})$, respectively, with the underlying variable $V_i = \beta_0 + \beta_X^t \mathbf{X}_i + u_i$. The use of the same linear function of covariates and the same random effect to specify the censoring mechanism and the positive observations makes this model less flexible than ours. Its advantage is formal modeling of never-consumers.

Our two-part model assumes that each food is ultimately consumed by all individuals, so that $T_i > 0$. This derives from specifying the random effect in the probability part as a continuous variable. In a similar situation, Olsen and Schafer (2001) suggested a two-part mixture for the distribution of this random effect, where the status of a “teetotaler” is specified by a latent class classification variable, but did not provide any details of fitting such a model.

We considered adding a third part to our model, which specifies for each person the probability to be a never-consumer by using fixed-effect logistic regression on a vector of covariates \mathbf{X}_{3i} . We have fitted this model to the data on fish intake in EATS among 515 women, including 30 who reported zero intakes on the FFQ. An indicator variable of whether fish consumption was reported on the FFQ was used as a covariate in \mathbf{X}_{3i} . In a simulation study similar to the one described in Section 5 (but this time simulating never-consumers), we investigated cases where the number of 24HRs was 2, 4, or 6. With only two 24HRs, the model fit was unstable in 64 out of 250 simulated data sets, although the problem disappeared when we increased the number of 24HRs to four or more. Modeling never-consumers is an area for further research, but, with only two 24HRs, the two-part model seems the most feasible approach.

Our methodology is suitable for analysis of a particular food and its relationship with a health outcome that involves no other dietary factors. An extension to a multivariate case with several foods and nutrients requires conditioning in

formula (20) on potentially correlated random effects for all considered dietary factors simultaneously and is another area for future research.

Although we concentrated on dietary surveys, the proposed method can also be applied to cohort studies of associations between episodically consumed foods and disease. Currently, most such studies use a FFQ as the main dietary-assessment instrument, while a more precise short-term reference instrument is available only in a calibration substudy. In such cases, the regression calibration is based on estimating

$$\hat{T}_i(\boldsymbol{\theta}) \equiv E(T_i | \mathbf{X}_i; \boldsymbol{\theta}) = \int \tau(\mathbf{X}_i; \boldsymbol{\theta}) f(u_i | \mathbf{X}_i; \boldsymbol{\theta}) du_i,$$

which involves conditioning on the FFQ and other covariates, but not on the 24HR (and therefore random effects) as in formulas (6)–(7). This simplifies the method and, more importantly, allows its application to a multivariate case with several foods and nutrients by considering regression calibration of each dietary factor, one at a time.

In the future, as automated 24HRs become available, our methodology could combine multiple administrations of this instrument with the FFQ to achieve more precise results.

7. Supplementary Materials

Web Appendices A–C, Web Figures 1–4, and NHANES example data, referenced in Sections 2, 4, and 5, as well as the SAS program implementing the proposed method are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

ACKNOWLEDGEMENTS

R.J.C.'s research was supported by a grant from the National Cancer Institute (CA57030) and by Award KUS-CI-016-04, made by King Abdullah University of Science and Technology.

REFERENCES

- Box G. E. P. and Cox D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211–252.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. Boca Raton, Florida: Chapman and Hall CRC Press.
- Dodd, K., Guenther, P. M., Freedman, L. S., Subar, A. F., Kipnis, V., Midthune, D., Tooze, J. A., and Krebs-Smith, S. M. (2006). Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. *Journal of the American Dietetic Association* **106**, 1640–1650.
- Dwyer, J., Picciano, M. F., Raiten, D. J., and Members of the Steering Committee. (2003). Collection of food and dietary supplement intake data: What we eat in America—NHANES. *The Journal of Nutrition* **133**, 590S–600S.
- Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics* **53**, 262–272.
- Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology* **158**, 14–21.
- Li, L., Shao, J., and Palta, M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics* **61**, 824–830.
- Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika* **81**, 624–629.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997). Estimating usual dietary intake distributions: Adjusting for measurement error and nonnormality in 24-hour food intake data. In *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. Deleeuw, C. Dippo, N. Schwartz, and D. Trewin (eds), 670–689. New York: Wiley.
- Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.
- Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology* **154**, 1089–1099.
- Subar, A. F., Dodd, K. W., Guenther, P. M., Kipnis, V., Midthune, D., McDowell, M., Tooze, J. A., Freedman, L. S., and Krebs-Smith, S. M. (2006). The Food Propensity Questionnaire (FPQ): Concept, development and validation for use as a covariate in model to estimate usual food intake. *Journal of the American Dietetic Association* **106**, 1556–1563.
- Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data clumping at zero. *Statistical Methods in Medical Research* **11**, 341–355.
- Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Carroll, R. J., and Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *Journal of the American Dietetic Association* **106**, 1575–1587.
- Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.
- Whittemore A. S. (1989). Errors-in-variables regression using Stein estimates. *The American Statistician* **43**, 226–228.
- Wolter, K. M. (1995). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Xue, F., Holzman, C., Rahbar, M. H., Trosko, K., and Fischer, L. (2007). Maternal fish consumption, mercury levels, and risk of preterm delivery. *Environmental Health Perspectives* **115**, 42–47.

Received January 2008. Revised November 2008.

Accepted November 2008.