A Bivariate Measurement Error Model for Semicontinuous and Continuous Variables: Application to Nutritional Epidemiology

Victor Kipnis,[1,*] **Laurence S. Freedman,**[2] **Raymond J. Carroll,**[3] **and Douglas Midthune**[1]

[1]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland
[2]Information Management Services, Inc., Rockville, Maryland and Gertner Institute for Epidemiology and Health
Policy Research, Sheba Medical Center, Tel Hashomer, Israel
[3]Department of Statistics, Texas A&M University, College Station, Texas
*email: kipnisv@mail.nih.gov

SUMMARY. Semicontinuous data in the form of a mixture of a large portion of zero values and continuously distributed positive values frequently arise in many areas of biostatistics. This article is motivated by the analysis of relationships between disease outcomes and intakes of episodically consumed dietary components. An important aspect of studies in nutritional epidemiology is that true diet is unobservable and commonly evaluated by food frequency questionnaires with substantial measurement error. Following the regression calibration approach for measurement error correction, unknown individual intakes in the risk model are replaced by their conditional expectations given mismeasured intakes and other model covariates. Those regression calibration predictors are estimated using short-term unbiased reference measurements in a calibration substudy. Since dietary intakes are often "energy-adjusted," e.g., by using ratios of the intake of interest to total energy intake, the correct estimation of the regression calibration predictor for each energy-adjusted episodically consumed dietary component requires modeling short-term reference measurements of the component (a semicontinuous variable), and energy (a continuous variable) simultaneously in a bivariate model. In this article, we develop such a bivariate model, together with its application to regression calibration. We illustrate the new methodology using data from the NIH-AARP Diet and Health Study (Schatzkin et al., 2001, *American Journal of Epidemiology* **154**, 1119–1125), and also evaluate its performance in a simulation study.

KEY WORDS: Bivariate modeling; Episodically consumed dietary components; Measurement error; Nutritional epidemiology; Regression calibration; Semicontinuous variables

1. Introduction

Semicontinuous data in the form of a mixture of a large portion of zero values and continuously distributed positive values frequently arise in many areas of biostatistics. This article is motivated by the analysis of relationships between disease outcomes and usual, i.e., long-term average, intakes of episodically consumed dietary components that are not typically eaten daily by nearly everyone. The latter is true of many foods, e.g., fish, red meat, whole grains, dark green vegetables, orange vegetables, etc, and some nutrients, e.g., vitamin A or B-12.

An important aspect of dietary studies is that true usual diet is unobservable and evaluated by dietary-assessment instruments with some degree of measurement error. Due to relatively low cost and convenience of administration, large epidemiologic studies of diet and disease commonly employ food frequency questionnaires (FFQs) that enquire about the participants recent dietary intake, usually over the past year. It has long been recognized that FFQs contain considerable measurement error, both random and systematic, which may substantially bias estimated diet–disease relationships and reduce the statistical power to detect a dietary effect (Freudenheim and Marshall, 1988). Moreover, at least in theory, conventional statistical tests to detect an effect may become invalid (do not control the Type I error rate) in the presence of error-prone covariates (Carroll et al., 2006). It is therefore important to adjust the analysis of diet–disease associations for FFQ measurement error.

The most popular method of adjusting for measurement error in nutritional epidemiology has become regression calibration, which substitutes for the unknown true dietary exposures their conditional expectations (i.e., best mean squared error predictors) given their observed values and other covariates in the disease model (Carroll et al., 2006). In practice, regression calibration predictors are estimated in a calibration substudy that includes (often repeated) short-term reference measurements assumed to be unbiased for individual true usual intakes.

Currently, large cohort studies are designed with a calibration substudy that includes as reference measurements reported intakes from a 24-hour dietary recall (24HR) that queries all foods and their amounts eaten on the previous day, or sometimes a multiple-day food diary. The working assumption is that such short-term measurements are unbiased for true usual intake, even though they may fall somewhat short of this ideal. Although, the methodology in this article is developed for any short-term reference instrument, it is exemplified using the 24HR.

Short-term reference measurements of episodically consumed dietary components are semicontinuous variables with

excess zeros and positive values with a usually skewed to the right distribution. Even if otherwise precise, as a measure of usual intake they contain substantial within-person measurement error due to daily variation of intake. Modeling such reference data and using them to compute regression calibration predictors is, therefore, challenging.

In the literature, a semicontinuous variable is commonly specified as the result of two distinct, although generally correlated, processes: one determines whether the variable takes positive or zero values; the other specifies its positive value. Such a two-part model was first introduced by Cragg (1971), who suggested using a logit or probit regression to specify the probability of positive values and a linear regression to specify log-transformed positive values. The two-part model was extended to longitudinal data by Olsen and Schafer (2001) and Tooze, Grunwald, and Jones (2002) by considering mixed effects regressions in both parts of the model and allowing random effects to be correlated. Recently, a new methodology, called the NCI method, has further extended the two-part model by considering the Box–Cox transformation of the positive values to achieve more flexibility in addressing skewness and by allowing for measurement error (Tooze et al., 2006; Kipnis et al., 2009).

The application of such models to nutritional epidemiology presents an additional challenge. In many cases, regression calibration may be applied to each error-prone covariate in the regression model one by one. But this may not always apply to dietary risk models. To better understand the effect of dietary composition, dietary exposures are usually "energy-adjusted," e.g., by using ratios of the intake of interest to total energy intake, called "densities." Since intakes of many dietary components and total energy are generally correlated, the correct application of regression calibration for each energy-adjusted episodically consumed dietary component requires modeling short-term reference measurements of the component (a semicontinuous variable) and energy (a continuous variable) simultaneously in a bivariate model.

In this article, we develop such a model, together with its application to regression calibration. Although the theoretical development underlying this model has never been published, some applications of the model and its multivariate extension to surveillance were presented in articles by Zhang et al. (2011a and 2011b). Both articles referenced the preprint by Kipnis et al. (an early version of the present article) for model justification and concentrated on fitting the models using the Markov Chain Monte Carlo (MCMC) technique. Here, we fully describe the methodology, together with its application to nutritional epidemiology, using the maximum likelihood approach. In Section 2, we discuss risk models in nutritional epidemiology. We point out that conventional regression models with linear predictors may not always produce a good fit and suggest, as a simple generalization, models with linear predictors after appropriate covariate transformations. In Section 3, we develop the bivariate measurement error model, and in Section 4, we describe its implementation in the analysis of diet–cancer relationships. In Section 5, we illustrate the method using data from the NIH-AARP Diet and Health Study (Schatzkin et al., 2001). We demonstrate that covariate transformation may improve unsatisfactory fit of conventional Cox regression and lead to substantially differ-

ent estimates of the exposure effect. In Section 6, we present the results of a simulation study to evaluate the performance of the proposed method in finite samples. We conclude with discussion in Section 7.

2. Risk Models

The most commonly used risk models in nutritional epidemiology are the generalized linear regression (especially logistic regression) and the proportional hazards Cox regression. Below, a health outcome is denoted by Y, a vector of true usual intakes of dietary components of interest by $\mathbf{T} = (T_1, \ldots, T_K)^t$, and a vector of other exactly measured covariates by $\mathbf{Z} = (Z_1, \ldots, Z_L)^t$. As explained above, some components of $\mathbf{T}$ include energy-adjusted intakes. A risk model can be written generally as

$$r(Y|\mathbf{T}, \mathbf{Z}) = \eta(\mathbf{T}, \mathbf{Z}; \boldsymbol{\alpha}),$$

where $r(Y|\mathbf{T}, \mathbf{Z})$ is the risk function for outcome Y and $\eta = \eta(\mathbf{T}, \mathbf{Z}; \boldsymbol{\alpha})$ is a predictor based on covariates $\mathbf{T}, \mathbf{Z}$ and parameters $\boldsymbol{\alpha}$.

In logistic regression, Y takes values of 0 (no event) or 1 (event), and the risk function is the logit of the probability of event, $r(Y|\mathbf{T}, \mathbf{Z}) = \log \frac{P(Y=1|\mathbf{T}, \mathbf{Z})}{1 - P(Y=1|\mathbf{T}, \mathbf{Z})}$. In standard Cox regression, Y is time t to event since entry into the study, and the risk function is the log ratio $r(Y|\mathbf{T}, \mathbf{Z}) = \log \frac{h(t|\mathbf{T}, \mathbf{Z})}{h_0(t)}$ of the hazard function $h(t|\mathbf{T}, \mathbf{Z})$ to the baseline hazard $h_0(t)$.

Conventional risk models specify the predictor $\eta(\mathbf{T}, \mathbf{Z}; \boldsymbol{\alpha})$ as a linear function of covariates. This is convenient but does not always provide a good fit. To alleviate this problem, especially if covariates have skewed distributions (typical of dietary exposures), epidemiologists sometimes transform covariates using log or power transformations. In the spirit of this approach, in this article, we will consider more flexible risk models

$$r(Y|\mathbf{T}, \mathbf{Z}) = \alpha_0 + \boldsymbol{\alpha}_{\mathbf{T}}^t \mathbf{T}^*(\boldsymbol{\lambda}_{\mathbf{T}}) + \boldsymbol{\alpha}_{\mathbf{Z}}^t \mathbf{Z}^*(\boldsymbol{\lambda}_{\mathbf{Z}}), \quad (1)$$

$\mathbf{T}^*(\boldsymbol{\lambda}_{\mathbf{T}}) = (T_1^*)(\lambda_{T1} \ldots T_K^*(\lambda_{TK}))^t$, $\mathbf{Z}^*(\boldsymbol{\lambda}_{\mathbf{Z}}) = (Z_1^*(\lambda_{Z_1}) \ldots Z_L^*(\lambda_{Z_L}))^t$, $\boldsymbol{\alpha}_T^t = (\alpha_{T_1} \ldots \alpha_{T_K})$, and $\boldsymbol{\alpha}_Z^t = (\alpha_{Z_1} \ldots \alpha_{Z_L})$, with predictors that are linear over transformed covariates using the Box–Cox family of transformations

$$v^*(\lambda) = g(v; \lambda) = \begin{cases} (v^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(v), & \text{if } \lambda = 0 \end{cases} \quad (2)$$

(Box and Cox, 1964). Note that $\alpha_0 = 0$ in the Cox regression.

The slope α_{T_k} in model (1) represents the effect of dietary component $T_k, k = 1, \ldots, K$. Due to the Box–Cox transformation, this effect is not a simple function of the additive or multiplicative change in exposure (as it would be if intake were expressed on the absolute or log scale, respectively) but depends on the baseline comparator value as follows: the effect of changing exposure T_k from T_{k0} to T_{k1} on an outcome Y_i is given by $\alpha_{T_k} \left\{ T_{k1}^*(\lambda_{Tk}) - T_{k0}^*(\lambda_{Tk}) \right\}$.

For person $i = 1, \ldots, n$ in the study, denote FFQ-measured intakes by $\mathbf{Q}_i$. These are the error-prone observed values of $\mathbf{T}_i$. Assume that $\mathbf{Q}_i$ has nondifferential measurement error with respect to Y_i, i.e., that Y_i and $\mathbf{Q}_i$ are independent given

true intakes $\mathbf{T}_i$ and covariates $\mathbf{Z}_i$. Denote by $\mathbf{X}_i$ a vector of covariates $\mathbf{Q}_i, \mathbf{Z}_i$ or their monotonic transformations. Regression calibration states that approximately (exactly for a linear regression risk model)

$$r(Y_i|\mathbf{X}_i) \approx \alpha_0^* + \boldsymbol{\alpha}_T^t E[\mathbf{T}_i^*(\boldsymbol{\lambda_T})|\mathbf{X}_i] + \boldsymbol{\alpha}_Z^t \mathbf{Z}_i^*(\boldsymbol{\lambda_Z}) \qquad (3)$$

For nonlinear risk models, the approximation in (3) is derived under the working hypothesis that $\boldsymbol{\alpha}_T^t \mathrm{var}[\mathbf{T}^*(\boldsymbol{\lambda_T})|\mathbf{X}]\boldsymbol{\alpha}_T$ is small (Carroll et al., 2006); it is usually very good in most applications to nutritional epidemiology due to relatively small relative risks.

As noted above, the regression calibration predictor $E[\mathbf{T}_i^*(\boldsymbol{\lambda_T})|\mathbf{X}_i]$ is estimated using repeated short-term reference measurements $\mathbf{R}_{ij}$ for individual $i, i = 1, \ldots, m$, and repeat measure $j, j = 1, \ldots, J_i$, in a calibration substudy. We will use the notation $E(\bullet|i)$ below to indicate that the expectation is conditional on information for the ith individual. Under the assumption that the reference measurements are unbiased for individual true usual intakes, i.e.,

$$E(\mathbf{R}_{ij}|i) = \mathbf{T}_i, \qquad (4)$$

the regression calibration predictor is given by $E[\mathbf{T}_i^*(\boldsymbol{\lambda_T})|\mathbf{X}_i] = E\big[\mathbf{g}\{E(\mathbf{R}_{ij}|i); \boldsymbol{\lambda_T}\}\big]|\mathbf{X}_i\big]$, and its estimation requires modeling of $\mathbf{R}_{ij}$.

3. Bivariate Measurement Error Model

We use subscripts "C" and "E" to denote intakes of the episodically consumed component of interest and total energy, respectively. We assume that individual usual intakes T_{Ci} and T_{Ei} are strictly positive continuous variables. This is natural for energy intake; for an episodic component, this assumption implies the absence of never-consumers. For the reference measurements, we assume that energy is always reported as positive, i.e., $R_{Eij} > 0$. However, for an episodically consumed component, we allow that $R_{Cij} = 0$ for any finite number of short-term periods (days).

For person $i = 1, \ldots, n$ in the main study, the observed data consist of the outcome variable Y_i and vector $\mathbf{X}_i$ of the covariates. Continuous covariates may be monotonically transformed to approximate normality thereby improving fit of the measurement error model described below. In a calibration substudy, the data additionally contain repeat short-term reference measurements $(R_{Cij}, R_{Eij}), i = 1, \ldots, m, j = 1, \ldots, J_i$. In models below, generically, $\boldsymbol{\beta}$ denotes population-level covariate effects (fixed effects), $\mathbf{u}_i$ denotes random effects representing the part of the within-subject mean not explained by the covariates, and $\boldsymbol{\varepsilon}_{ij}$ denotes within-subject random errors representing longitudinal variation.

A bivariate measurement error model of a semicontinuous (the episodically consumed component R_{Cij}) and a continuous (total energy R_{Eij}) variables requires combining the two-part model for the episodic component with a one-part model for total energy, leading to a three-part model. Reflecting the nature of short-term dietary intakes, such a model needs to

satisfy the following requirements:

(i) for each individual i on any single day j, energy intake may be correlated with both the fact of reported consumption of the dietary component, and the amount consumed (on a consumption day);

(ii) for each individual i, the probability to consume the dietary component may be correlated with the usual amount consumed on a consumption day, and both may be correlated with usual energy intake.

With the first requirement in mind, we specify a binary indicator I_{Cij} of reference consumption during a specific period j as follows. Let $I_{Cij} = I[R_{Cij} > 0]$, where $I[x]$ is the indicator function. We model I_{Cij} as resulting from dichotomizing a continuous latent variable R_{C1ij}, i.e.,

$$I_{Cij} \equiv I_{C1ij}, I_{C1ij} = I[R_{C1ij} > 0],$$
$$R_{C1ij} = \boldsymbol{\beta}_{C1X}^{\mathrm{t}}\mathbf{X}_i + u_{C1i} + \varepsilon_{C1ij}, \qquad (5)$$

$u_{C1i} \sim \mathrm{Normal}\,(0; \sigma_{u_{C1}}^2)$ and $\{\varepsilon_{C1ij}\} \overset{iid}{\sim} \mathrm{Normal}\,(0; \sigma_{\varepsilon_{C1}}^2)$ are independent of each other and of $\mathbf{X}_i$. For identifiability, $\sigma_{\varepsilon_{C1}}^2$, has to be fixed, and without loss of generality we set $\sigma_{\varepsilon_{C1}}^2 = 1$ (Catalano and Ryan, 1992). Note that model (5) is equivalent to specifying the probability of consumption on any given day using the mixed effects probit regression. The advantage of specification (5) is that it allows for correlation of latent variable ε_{C1ij} with its counterpart in the model for energy intake (see below).

In the second part of our model, we follow the second part of the two-part model in Kipnis et al. (2009) and specify the Box–Cox transformed reference measurements during a consumption period as a mixed effects linear model. Define $R_{C2ij} = (R_{Cij}|R_{Cij} > 0)$ as the positive part of the reported intake R_{Cij} (if $R_{Cij} = 0$ the value of R_{C2ij} is irrelevant). We assume that there is a Box–Cox transformation with parameter λ_{R_C} such that

$$R_{C2ij}^*(\lambda_{R_C}) = \boldsymbol{\beta}_{C2}^{\mathrm{t}}\mathbf{X}_i + u_{C2i} + \varepsilon_{C2ij}, \qquad (6)$$

$u_{C2i} \sim \mathrm{Normal}(0; \sigma_{u_{C2}}^2)$ and $\{\varepsilon_{C2ij}\} \overset{iid}{\sim} \mathrm{Normal}(0; \sigma_{\varepsilon_{C2}}^2)$ are independent of each other and of $\mathbf{X}_i$.

In the third part of our model, reference energy intake is specified similarly to submodel (6) as follows. We assume that there is a Box–Cox transformation with parameter λ_{R_E} such that

$$R_{Eij}^*(\lambda_{R_E}) = \boldsymbol{\beta}_E^{\mathrm{t}}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}, \qquad (7)$$

$u_{Ei} \sim \mathrm{Normal}(0; \sigma_{u_E}^2)$ and $\{\varepsilon_{Eij}\} \overset{iid}{\sim} \mathrm{Normal}(0; \sigma_{\varepsilon_E}^2)$ are independent of each other and of $\mathbf{X}_i$.

The random effects $(u_{C1i}, u_{C2i}, u_{Ei})$ in model (5)–(7) are allowed to be mutually correlated to satisfy requirement (ii) above. To satisfy requirement (i), within-person errors ε_{Eij} in submodel (7) are allowed to be correlated with their counterparts ε_{C1ij} and ε_{C2ij} in submodels (5)–(6). We, however, assume that $\mathrm{cov}(\varepsilon_{C1ij}, \varepsilon_{C2ij}) = 0$, so that marginally R_{Cij} and

R_{Eij} follow the two-part model (5)–(6) and one-part model (7), respectively.

The underlying three-part model (5)–(7) can be expressed more succinctly as the following bivariate measurement error model for reference measurements R_{Cij} and R_{Eij}:

$$R_{Cij} = I[\boldsymbol{\beta}_{C1}^{\text{t}}\mathbf{X}_i + u_{C1i} + \varepsilon_{C1ij} > 0]$$
$$\times g_{R_C}^{-1}(\boldsymbol{\beta}_{C2}^{\text{t}}\mathbf{X}_i + u_{C2i} + \varepsilon_{C2ij}; \lambda_{R_C})$$
$$R_{Eij} = g_{R_E}^{-1}(\boldsymbol{\beta}_{E}^{\text{t}}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}; \lambda_{R_E}) \qquad (8)$$

where $\mathbf{u}_i=(u_{C1i}, u_{C2i}, u_{Ei})^t \sim \text{Normal}(0, \boldsymbol{\Sigma}_u)$, with an unstructured variance-covariance matrix $\boldsymbol{\Sigma}_u$, and $\{\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{C1ij}, \varepsilon_{C2ij}, \varepsilon_{Eij})^t\} \sim \text{Normal}(0, \boldsymbol{\Sigma}_\varepsilon)$, with a structured variance-covariance matrix equal to

$$\boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} 1 & 0 & \sigma_{\varepsilon_E}\rho_{\varepsilon_{C1}\varepsilon_E} \\ 0 & \sigma_{\varepsilon_{C2}}^2 & \sigma_{\varepsilon_{C2}}\sigma_{\varepsilon_E}\rho_{\varepsilon_{C2}\varepsilon_E} \\ \sigma_{\varepsilon_E}\rho_{\varepsilon_{C1}\varepsilon_E} & \sigma_{\varepsilon_{C2}}\sigma_{\varepsilon_E}\rho_{\varepsilon_{C2}\varepsilon_E} & \sigma_{\varepsilon_E}^2 \end{bmatrix}, \qquad (9)$$

and vectors $\mathbf{X}_i$, $\boldsymbol{\varepsilon}_{ij}$, and $\mathbf{u}_i$ are mutually independent. All model (including transformation) parameters may be estimated by maximizing likelihood function derived in Web Appendix A.

4. Fitting the Risk Model Using Regression Calibration

Our methodology combines a pseudo-likelihood approach with regression calibration and may be implemented in three steps. First, the bivariate measurement error model (8)–(9) is fitted in the calibration substudy using the maximum likelihood. Second, for any given transformation parameters $(\lambda_{T_D}, \lambda_{T_E})$, regression calibration predictors are calculated in the main study as follows. Under assumption (4), true usual intakes are given by integrating R_{Cij} and R_{Eij} over the distributions of ε_{C2ij} and ε_{Eij}, respectively, i.e.,

$$T_{Ci} = E(R_{Cij}|\mathbf{X}_i, u_{C1i}, u_{C2i}) = \Phi_1(\boldsymbol{\beta}_{C1}^{\text{t}}\mathbf{X}_i + u_{C1i}; 1)$$
$$\times \int g_{R_C}^{-1}(\boldsymbol{\beta}_{C2}^{\text{t}}\mathbf{X}_i + u_{C2i} + \varepsilon_{C2ij}; \lambda_{R_C})d\Phi_1(\varepsilon_{C2ij}; \sigma_{\varepsilon_{C2}}^2);$$
$$T_{Ei} = E(R_{Eij}|\mathbf{X}_i, u_{Ei})$$
$$= \int g_{R_E}^{-1}(\boldsymbol{\beta}_{E}^{\text{t}}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}; \lambda_{R_E})d\Phi_1(\varepsilon_{Eij}; \sigma_{\varepsilon_E}^2),$$

where $\Phi_q(\xi, \Sigma)$ denotes the q-dimensional normal cumulative distribution function of a random vector ξ with mean zero and covariance matrix Σ. Let $u_{Ci} = (u_{C1i}, u_{C2i})^t \sim \text{N}(0; \Sigma_{u_C})$. The regression calibration predictors of transformed true density $T_{Di} = T_{Ci}/T_{Ei}$, and energy intake T_{Ei} are given by

and

$$E[T_{Ei}^*(\lambda_{T_E})|\mathbf{X}_i]$$
$$= \int g_{T_E}\left\{\int g_{R_E}^{-1}(\boldsymbol{\beta}_{E}^{\text{t}}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}; \lambda_{R_E})d\Phi_1(\varepsilon_{Eij}; \sigma_{\varepsilon_E}^2); \lambda_{T_E}\right\}$$
$$\times d\Phi_1(u_{Ei}; \sigma_{u_E}^2), \qquad (11)$$

respectively. These regression calibration predictors are then estimated in the main study by substituting the estimated in the calibration substudy model parameters into equations (10)–(11) and numerically integrating the resulting expressions.

Third, the risk model, with regression calibration predictors as dietary exposures, is fitted in the main study by the maximum likelihood (partial likelihood in case of Cox regression). Finally, transformation parameters $(\lambda_{T_D}, \lambda_{T_E})$ may be estimated by exploring a grid of their possible values, fitting the risk models with corresponding calibration predictors and choosing transformations that maximize the corresponding profile likelihood.

The standard errors of the estimated parameters in the risk model may be calculated using the asymptotic theory outlined in Web Appendix B. Alternatively, they may be estimated using the bootstrap approach as we do in our examples in Section 5 below. In implementing the bootstrap, the whole fitting procedure (see steps 1–3 above) should be applied to each bootstrap replication to take into account all the variation due to each step.

The described approach may be carried out in SAS using NLMIXED procedure to fit the measurement error model and PHREG or LOGISTC procedures to fit the Cox regression or logistic regression risk models, respectively. The SAS code for fitting the measurement error model and estimating regression calibration predictors is provided in Supplemental Materials.

5. Application to NIH-AARP Diet and Health Study

We apply the above methods to analyses of two diet–cancer relationships based on data from the NIH-AARP Diet and Health study (Schatzkin et al., 2001). This is a prospective cohort of 567,169 participants (340,148 men and 227,021 women) designed to investigate relationships between diet and cancer. Study participants completed a baseline FFQ on usual intake of 124 different foods and food groups. In addition, a calibration substudy was conducted on 2053 participants (1034 men, 1019 women), who were asked to complete, in addition to the FFQ, two unannounced 24-hour dietary recalls on nonconsecutive days.

NIH-AARP study investigators have reported small but statistically significant associations between orange vegetable intake and lung cancer incidence in men (Wright et al., 2008), and between red meat intake and lung cancer incidence in men (Tasevska et al., 2009). These reports were based on analyses

$$E[T_{Di}^*(\lambda_{T_D})|\mathbf{X}_i] = \int g_{T_D}\left\{\frac{\Phi_1(\boldsymbol{\beta}_{C1}^{\text{t}}\mathbf{X}_i + u_{C1i}; 1) \times \int g_{R_C}^{-1}(\boldsymbol{\beta}_{C2}^{\text{t}}\mathbf{X}_i + u_{C2i} + \varepsilon_{C2ij}; \lambda_{R_C})d\Phi_1(\varepsilon_{C2ij}; \sigma_{\varepsilon_{C2}}^2)}{\int g_{R_E}^{-1}(\boldsymbol{\beta}_{E}^{\text{t}}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}; \lambda_{R_E})d\Phi_1(\varepsilon_{Eij}; \sigma_{\varepsilon_E}^2)}; \lambda_{T_D}\right\}d\Phi_2(\mathbf{u}_{Ci}; \Sigma_{u_C}) \qquad (10)$$

Table 1
24-hour dietary recall reported consumption of red meat and orange vegetables in men[a]; NIH-AARP Diet and Health Study

	Red meat (g/day)	Orange vegetables (cups/day)
Mean reported intake (s.e.)	82.8 (2.3)	0.14 (0.01)
Mean amount on consumption days (s.e.)	117.7 (2.6)	0.32 (0.01)
Mean probability to consume on any single day (s.e.)	0.70 (0.01)	0.44 (0.01)
Percentage of subjects who consumed food:		
0 out of 2 days	14.7	33.5
1 out of 2 days	29.9	44.9
2 out of 2 days	55.4	21.6

[a]Based on the 886 males in the calibration substudy who completed two 24-hour dietary recalls.

of FFQ-reported diet that were not adjusted for measurement error. We used regression calibration based on the developed methodology to adjust these associations for measurement error in the FFQ-reported diet. The analysis included 279,149 men (4110 incident lung cancers) in the AARP main cohort, and 986 men in the calibration substudy.

5.1. Diet–Disease Risk Model

In these analyses, T_{Ci} denotes true usual red meat intake (g/day) or orange vegetable intake (cups/day), T_{Ei} true usual energy intake (kcal/day), and $T_{Di} = (T_{Ci}/T_{Ei}) \times 1000$ denotes true red meat density (g/1000 kcal) or orange vegetable density (cups/1000 kcal). Let $(Q_{Ci}, Q_{Ei}, Q_{Di} = (Q_{Ci}/Q_{Ei}) \times 1000)$ and $(R_{Cij}, R_{Eij}, R_{Dij} = (R_{Cij}/R_{Eij}) \times 1000)$ denote, respectively, the FFQ- and 24HR-reported values and calculated densities of the corresponding dietary components. In accordance with (3), we considered the Cox proportional hazards model

$$h(t|\mathbf{T}_i, \mathbf{Z}_i) = h_0(t)\exp\left\{\alpha_{T_D} g(T_{Di}; \lambda_{T_D}) + \alpha_{T_E} g(T_{Ei}; \lambda_{T_E}) + \alpha_Z^t \mathbf{Z}_i\right\}$$
(12)

where $\mathbf{T}_i = (T_{Di}, T_{Ei})^t$ and $\mathbf{Z}_i$ is a vector of the following covariates: age at entry into the study (continuous) and smoking status (never smoked; quit ≥ 10 years ago; quit 1–9 years ago; current smoker or quit <1 year ago). The original analyses (Wright et al., 2008; Tasevska et al., 2009) included more covariates, but to simplify the example we included only the most important risk factors for lung cancer.

5.2. Regression Calibration

In regression calibration, we considered Box–Cox transformed FFQ variables, so that

$$\mathbf{X}_i = \left(g(Q_{Ci}; \lambda_{Q_C}), g\left(Q_{Ei}; \lambda_{Q_E}\right), \mathbf{Z}_i^t\right)^t$$

The Box–Cox parameters λ_{Q_C} and λ_{Q_E} were chosen among those in the interval $[0, 1]$ to minimize the Kolmogorov test statistic for normality in the main study. A small percentage of subjects reported zero intakes of orange vegetables (0.4%) and red meat (0.05%) on the FFQ. We, therefore, added a relatively small (with respect to the corresponding distribution) value to the FFQ measurements before choosing a transformation. For red meat, we added 1 g/day and for orange vegetables 0.01cups/day. This allowed us to consider the log transformation and also to remove the leverage of zero values in regression models.

After fitting the bivariate model (8)–(9) in the calibration substudy, for any given values of transformation parameters λ_{T_D} and λ_{T_E}, the calibration predictors $E[g(T_{Di}; \lambda_{T_D})|\mathbf{X}_i]$ and $E[g(T_{Ei}; \lambda_{T_E})|\mathbf{X}_i]$ were calculated by numerical integration according to expressions (10)–(11) for each subject in the main study and used as covariates, together with $\mathbf{Z}_i$, in risk model (12). Standard errors for the estimated parameters were calculated using the bootstrap method with 200 replications.

Following the usual practice, we first fitted the risk model using dietary exposures on the original ($\lambda_{T_D} = \lambda_{T_E} = 1$) as well as the log ($\lambda_{T_D} = \lambda_{T_E} = 0$) scales and, for each set of transformations, assessed the proportional hazards assumption of the Cox model and the risk model fit using methods based on cumulative martingale residuals (Lin, Wei, and Ying, 1993) and implemented in the SAS PHREG procedure (version 9.2, SAS Institute). In the case that neither model produced an acceptable fit, we planned to explore a grid of possible Box–Cox transformation parameters ($\lambda_{T_D}, \lambda_{T_E}$) in the interval $[-1, 1]$ and to choose the transformation based on the maximization of the corresponding profile likelihood.

In our examples, we compared the results of our proposed method to those without correction for measurement error, in which case, for any given transformation parameters ($\lambda_{Q_D}, \lambda_{Q_E}$), risk model (12) was fitted to $g(Q_{Di}; \lambda_{Q_D})$, $g(Q_{Ei}; \lambda_{Q_E})$, and $\mathbf{Z}_i$. Note that for an a priori chosen scale, such as the original or the log, we considered $\lambda_{Q_D} = \lambda_{T_D}, \lambda_{Q_E} = \lambda_{T_E}$. If based on maximizing the profile likelihood, the transformation parameters chosen for the uncorrected risk analysis and those after regression calibration would generally differ.

Table 1 summarizes the distributions of 24HR-reported orange vegetable and red meat intakes for men in the calibration substudy. Both food groups are episodically consumed, although red meat is consumed more frequently than orange vegetables; on average for a single day, nonzero consumption of red meat is reported on 70% of the 24HRs, compared to only 44% for orange vegetables.

Table 2 displays the estimated parameters in the bivariate model with dietary exposures being i) red meat and energy intake; and ii) orange vegetable and energy intake. As the Table shows, for both food groups, FFQ-reported intake is related to both the probability to consume and the mean amount consumed on consumption days. It also follows that current smokers are more likely to eat red meat and less likely to eat orange vegetables than never smokers. In addition, the variances of within-person errors $\boldsymbol{\varepsilon}_{ij}$ are larger than the variances of the random effects $\mathbf{u}_i$, and the within-person errors for en-

Table 2

Estimated bivariate model parameters for red meat and orange vegetables; NIH-AARP Diet and Health Study

Parameter	Red meat (g/day) Estimate (s.e.)[a]	Orange vegetables (cups/day) Estimate (s.e.)
Box–Cox parameter for food intake	0.376 (0.023)	0.173 (0.025)
Box–Cox parameter for energy intake	0.272 (0.050)	0.267 (0.051)
Fixed effects in part I of food model:		
FFQ-reported food intake	0.212 (0.018)	0.239 (0.039)
FFQ-reported energy intake	−0.184 (0.040)	−0.034 (0.029)
Age at baseline	0.003 (0.009)	0.007 (0.006)
Former smoker, quit ≥ 10 years ago	−0.013 (0.086)	−0.076 (0.079)
Former smoker, quit 1–9 years ago	−0.168 (0.148)	−0.208 (0.116)
Current smoker or quit < 1 year ago	0.260 (0.156)	−0.305 (0.116)
Fixed effects in part II of food model:		
FFQ-reported food intake	0.325 (0.061)	0.249 (0.037)
FFQ-reported energy intake	−0.222 (0.145)	−0.112 (0.029)
Age at baseline	−0.052 (0.028)	−0.005 (0.007)
Former smoker, quit ≥ 10 years ago	0.674 (0.355)	0.072 (0.070)
Former smoker, quit 1–9 years ago	1.230 (0.552)	0.086 (0.129)
Current smoker or quit < 1 year ago	1.280 (0.466)	−0.105 (0.132)
Fixed effects in energy model:		
FFQ-reported food intake	−0.010 (0.029)	−0.248 (0.104)
FFQ-reported energy intake	0.571 (0.085)	0.600 (0.070)
Age at baseline	−0.046 (0.014)	−0.043 (0.014)
Former smoker, quit ≥ 10 years ago	0.199 (0.194)	0.171 (0.184)
Former smoker, quit 1–9 years ago	0.045 (0.294)	0.022 (0.294)
Current smoker or quit < 1 year ago	−0.216 (0.293)	−0.367 (0.290)
Random effects:		
$\mathrm{Var}(u_{C1})$	0.370 (0.101)	0.087 (0.056)
$\mathrm{Var}(u_{C2})$	3.940 (1.059)	0.057 (0.049)
$\mathrm{Var}(u_E)$	3.165 (0.308)	2.929 (0.306)
$\mathrm{Corr}(u_{C1}, u_{C2})$	0.562 (0.229)	0.333 (0.559)
$\mathrm{Corr}(u_{C1}, u_E)$	0.171 (0.109)	0.060 (0.291)
$\mathrm{Corr}(u_{C2}, u_E)$	0.307 (0.104)	0.564 (0.270)
Whithin-person errors:		
$\mathrm{Var}(\varepsilon_{C2})$	18.765 (1.144)	0.773 (0.064)
$\mathrm{Var}(\varepsilon_E)$	4.847 (0.265)	4.538 (0.263)
$\mathrm{Corr}(\varepsilon_{C1}, \varepsilon_E)$	0.223 (0.043)	0.055 (0.049)
$\mathrm{Corr}(\varepsilon_{C2}, \varepsilon_E)$	0.281 (0.036)	−0.057 (0.064)

[a]Standard errors are calculated using the bootstrap method.

ergy intake, ε_{Eij}, are statistically significantly correlated with ε_{C1ij} and ε_{C2ij} for red meat intake, but not for orange vegetable intake. This is to be expected since, unlike orange vegetables, red meat contributes a substantial amount of energy.

5.3. Red Meat Intake and Lung Cancer Risk

Table 3 presents the results of an analysis of red meat density intake and lung cancer risk. The risk model did not fit with dietary exposures on the original scale for the unadjusted analysis (goodness-of-fit $p = 0.041$), although the regression calibration method produced an acceptable fit ($p = 0.136$). With dietary exposures on the log scale, both the unadjusted and the regression calibration methods produced a good fit. The p-values for the tests of the proportional hazards assumption were greater than 0.05 for all covariates in both models, indicating no serious violation of the assumption. The estimated hazard ratios shown in the table correspond to a change in red meat density intake from 10 to 60 g/1000 kcal, intakes that are close to the 10th and slightly smaller than 90th percentiles,

respectively, of the population distribution. As expected, the unadjusted estimates of the hazard ratio for red meat density are closer to 1.0 than are the measurement error-corrected estimates. It is interesting to note that for the regression calibration method, where the model fit using either the original or log-transformed scales was acceptable, the estimated hazard ratios were very similar.

5.4. Orange Vegetable Intake and Lung Cancer Risk

Table 4 presents estimated hazard ratios for lung cancer when orange vegetable density intake changes from 0.08 to 0.10 cups/1000 kcal, which are again close to the 10th and slightly smaller than 90th percentiles of the population distribution, respectively. When checking goodness-of-fit, the p-values for the tests of the proportional hazards assumption were greater than 0.05 for all covariates in both unadjusted and adjusted for measurement error models, indicating no serious violation of the assumption. However, the goodness-of-fit tests indicated that both adjusted for measurement error and un-

Table 3

NIH-AARP Diet and Health study: Red meat intake and lung cancer risk in men; estimated hazard ratios[a] for red meat (g/1000 kcal) intake

Measurement error correction method	Estimated log hazard ratio (s.e.)[b]	Estimated hazard ratio (95% CI)[b]	Goodness-of-fit p-value[c]
Untransformed intakes:			
No correction	0.225 (0.035)	1.252 (1.169, 1.341)	0.041
Regression calibration	0.392 (0.077)	1.480 (1.273, 1.721)	0.136
Log transformed intakes:			
No correction	0.241 (0.041)	1.273 (1.174, 1.379)	0.438
Regression calibration	0.389 (0.080)	1.476 (1.261, 1.726)	0.524

[a] Cox proportional hazards model adjusted for total energy intake, age and smoking status; hazard ratio compares intake of 60 g/1000 kcal to 10 g/1000 kcal.
[b] Standard errors and 95% confidence intervals are calculated using the bootstrap method.
[c] P-value for testing that the functional form of red meat or energy intake is correctly specified; test based on cumulative martingale residuals.

adjusted risk models with dietary exposures on the original scale did not fit the data ($p = 0.002$ and $p < 0.001$, respectively). When we considered the log scale for dietary exposures, both models produced a more acceptable fit. As in the previous analysis, the unadjusted estimate of the hazard ratio was closer to 1.0 than the estimate adjusted for measurement error. In this example, choosing a more appropriate scale for dietary exposures produced substantially different results compared to the original scale where the model fit was very poor.

6. Simulation Study

We carried out two simulation studies to evaluate the performance of regression calibration using the bivariate episodic component and energy model in finite samples, and to compare it to unadjusted results. The two simulation studies were based on the examples given in the previous section. For simplicity, the risk models were chosen to be logistic regressions with only the corresponding food group density and energy intakes as covariates. For each simulation study, we performed

two sets of 200 simulations. The first set was based on the logistic regression with true linear predictor on the original scale, and the second on the log scale. For each simulation, we generated two independent datasets. The first was a main study with 200,000 subjects, while the second was a calibration substudy with 1000 subjects. The details of the simulations are provided in Web Appendix C. For each simulation, we assumed that the correct scale for energy intake was known and chose λ_{T_D} by considering different values between -1 and 1 with step 0.1, fitting the logistic regression risk model with calibrated covariates and choosing λ_{T_D} that maximized the corresponding profile likelihood.

Tables 5 and 6 show the results of the first and second simulation studies, respectively. Regression calibration based on the bivariate model performed well. The increase in standard errors of the estimated odds ratios was more than compensated for by the negligible biases, producing estimates with smaller root mean squared errors (RMSEs) than the unadjusted method. Note that when the true linear predictor was on the original scale, regression calibration led to a better

Table 4

NIH-AARP Diet and Health study: Orange vegetable intake and lung cancer risk in men; estimated hazard ratios[a] for orange vegetable (cups/1000 kcal) intakes

Measurement error Correction Method	Estimated log hazard ratio (s.e.)[b]	Estimated hazard ratio (95% CI)[b]	Goodness-of-fit p-value[c]
Untransformed intakes:			
No correction	-0.076 (0.021)	0.927 (0.889, 0.966)	< 0.001
Regression calibration	-0.265 (0.086)	0.767 (0.648, 0.908)	0.002
Log transformed intakes:			
Orange vegetable intake			
no correction	-0.180 (0.030)	0.835 (0.788, 0.886)	0.264
Regression calibration	-0.376 (0.085)	0.687 (0.581, 0.811)	0.052

[a] Cox proportional hazards model adjusted for total energy intake, age and smoking status; hazard ratio compares intake of 0.10 cups/1000 kcal to 0.02 cups/1000 kcal.
[b] Standard errors and 95% confidence intervals are calculated using the bootstrap method.
[c] P-value for testing that the functional form of orange vegetable intake is correctly specified; test based on cumulative martingale residuals.

Table 5

Simulation results; logistic regression of disease on red meat density and total energy intake; simulated means, standards deviations and root mean squared errors (RMSE) of estimated log odds ratios (OR) comparing density of 60–10 g/1000 kcal

Sim	Measurement error correction method	Mean λ	Mean log OR (s.e.)	Standard deviation	RMSE
1	True parameters	1	0.4		
	No correction	0.44	0.240 (0.002)	0.027	0.162
	Regression calibration	0.82	0.410 (0.005)	0.058	0.059
2	True parameters	0	0.4		
	No correction	0.06	0.222 (0.002)	0.027	0.180
	Regression calibration	0.07	0.386 (0.004)	0.063	0.065

estimated transformation parameter λ than the unadjusted method, but when the true linear predictor was on the log scale, the scales chosen by each method were on average similar and close to the true scale.

7. Discussion

The presented bivariate measurement error model addresses major challenges for simultaneous modeling of longitudinal error-prone semicontinuous (short-term reference measurements of an episodically consumed component) and continuous (short-term reference measurements of energy intake) variables. During a given short-term period, energy intake is allowed to be correlated with the binary indicator of episodic component consumption as well as with its positive intake if it is consumed. Allowing correlations among random effects in the model induces correlations between usual probability to consume an episodic component, usual consumption amount, and usual energy intake. Bivariate modeling of short-term reference intakes of an episodically consumed dietary component and energy leads to a rigorous application of regression calibration when the risk model contains energy-adjusted dietary exposures. As we demonstrate in applications to real data from the NIH-AARP Diet and Health Study, conventional risk models with linear predictors may fail to produce a good fit to the data. A generalization to predictors that are linear over appropriately Box–Cox transformed covariates generally improves model fit. In our analysis of orange vegetables and lung cancer, the conventional Cox regression did not fit the data. The log transformation of the exposure not only led to

an acceptable model fit, but also produced a substantially different estimated hazard ratio. The results of our simulations indicate that the developed methodology works well in finite samples, providing unbiased or nearly unbiased estimated exposure effects.

The presented bivariate model is easily extendable to handle longitudinal (time-dependent) covariates, and to include random slopes in addition to random intercepts. In addition, one could apply an enhanced regression calibration by including additional covariates in the measurement error model (5)–(7). As shown in Kipnis et al. (2009), under the condition that those covariates are related to true dietary exposure but, given truth, are not related to the outcome, the enhanced regression calibration may lead to more efficient estimates of the regression coefficients in the risk model. Note that additional covariates may differ for each part of the episodic component model as well as the model for energy intake.

The considered risk models may be also generalized. For example, it is straightforward to fit models of general form such as $r(Y|\mathbf{T}, \mathbf{Z}) = \alpha_0 + \boldsymbol{\alpha}^T g(\mathbf{T}, \mathbf{Z}; \boldsymbol{\theta})$ for any function g. Regression calibration simply substitutes $E\{g(\mathbf{T}, \mathbf{Z}; \boldsymbol{\theta})|\mathbf{Z}, \mathbf{Q}\}$ for $g(\mathbf{T}, \mathbf{Z}; \boldsymbol{\theta})$, with the expectation calculated only once by numerical integration.

As mentioned above, regression calibration in nonlinear risk models provides an approximate adjustment for measurement error, and the excellent quality of this approximation under rare disease and/or relatively small exposure effect assumption may deteriorate if the assumption is violated. To reduce potential bias in Cox regression, there have recently been sug-

Table 6

Simulation results; logistic regression of disease on orange vegetable density and total energy intake; simulated means, standards deviations and root mean squared errors (RMSE) of estimated log odds ratios (OR) comparing density of 0.10–0.02 cups/1000 kcal

Sim	Measurement error Correction method	Mean λ	Mean log OR (s.e.)	Standard deviation	RMSE
1	True parameters	1	−0.4		
	No correction	0.42	−0.241 (0.001)	0.019	0.160
	Regression calibration	0.89	−0.408 (0.005)	0.072	0.072
2	True parameters	0	−0.4		
	No correction	−0.03	−0.202 (0.001)	0.018	0.199
	Regression calibration	0.00	−0.415 (0.006)	0.080	0.081

gested the use of risk set calibration (Xie, Wang, and Prentice, 2001) or a computationally less excessive follow-up time regression calibration (Zhao and Prentice, 2014). Applying our approach with those methodologies is rather straightforward. It requires fitting the measurement error model in the calibration substudy to those still at risk within a given time interval (for follow-up calibration) or every time there is an event in the main study. In our applications, since cancer is a rare disease, the results would be very close to "usual" regression calibration as was demonstrated by simulations in both articles by Xie et al., 2001 and Zhao and Prentice, 2014.

In our applications to real data as well as in simulations, we considered risk models with only two dietary exposures, the density of an episodically consumed dietary component of interest and energy. In practice, risk models often contain energy-adjusted intakes of several dietary components as well as energy intake. Using the bivariate model, regression calibration may be applied to each pair of the dietary component and energy intakes one by one. Although not fully efficient, such a procedure does not jeopardize asymptotic properties of the estimated exposure effects. The only additional issue arises with respect to the final estimate of calibrated energy intake, since it may be estimated somewhat differently in each bivariate model. One option is to use the calibrated predictor with the greatest variance because that would produce the most efficient estimate of energy effect on health outcome (Kipnis et al., 2009). Another option may be based on estimated calibrated energy intake from the marginal energy model.

In general, the most efficient regression calibration adjustment for measurement error would be based on a multivariate modeling of all dietary components and energy simultaneously. Such a multivariate model is specified as a straightforward extension of the suggested bivariate model. Each episodically consumed dietary component is specified using the two-part model (5)–(6), and each regularly consumed dietary component (including energy) is specified using model (7). All random effects in the model are allowed to be correlated with an unstructured covariance matrix $\Sigma_{\mathbf{u}}$. The within-person errors have a structured covariance matrix $\Sigma_{\boldsymbol{\varepsilon}}$ where, for each episodically consumed component C_k, $\sigma^2_{\varepsilon_{Ck1}} = 1$ and $\sigma_{\varepsilon_{Ck1}\varepsilon_{Ck2}} = 0$, and other elements are unrestricted.

Although, the extension to a multivariate model with several semicontinuous and continuous variables is straightforward in theory, in practice it leads to difficulties in maximum likelihood estimation, since the likelihood quickly becomes very complicated, involving multivariate normal distribution functions. As a result, available software, such as the SAS NLMIXED procedure, can handle only a limited number of random effects. Recently, Zhang et al. (2011b) suggested an MCMC method of fitting such a multivariate model. This technique holds promise for more efficient application of regression calibration to risk models that include several, possibly energy-adjusted, episodically and daily consumed dietary components.

8. Supplementary Materials

Web Appendices A-C, referenced in Sections 2, 4, and 5, as well as the SAS programs generating example data and imple-

menting the proposed method, are available with this article at the *Biometrics* website on Wiley Online Library.

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.

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651–658.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **39**, 829–844.

Freudenheim, J. L. and Marshall, J. R. (1988). The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. *Nutrition and Cancer* **11**, 243–250.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., et al. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* **65**, 1003–1010.

Lin, D. Y, Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

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.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., et al. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-American Association of Retired Persons Diet and Health Study. *American Journal of Epidemiology* **154**, 1119–1125.

Tasevska, N., Sinha, R., Kipnis, V., Subar, A. F., Leitzmann, M. F., Hollenbeck, A. R., et al. (2009). A prospective study of meat, cooking methods, meat mutagens, heme iron, and lung cancer risks. *American Journal of Clinical Nutrition* **89**, 1884–1894.

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., et al. (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.

Wright, M. E., Park, Y., Subar, A. F., Freedman, N. D., Albanes, D., Hollenbeck, A. R., et al. (2008). Intakes of fruit, vegetables, and specific botanical groups in relation to lung cancer risk in the NIH-AARP Diet and Health study. *American Journal of Epidemiology* **168**, 1024–1034.

Xie S. X., Wang C. Y., Prentice R. L. (2001). A risk set calibration method for failure time regression by using a covariate

reliability sample. *Journal of the Royal Statistical Society, Series B* **63**, 855–870.

Zhang, S., Midthune, D., Perez, A., Buckman, D. W., Kipnis, V., Freedman, L. S., et al. (2011a). Fitting a bivariate measurement error model for episodically consumed dietary components. *International Journal of Biostatistics*, **7**, 1.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., et al. (2011b). A new multivariate measurement error model with zero-inflated dietary data,

and its application to dietary assessment. *Annals of Applied Statistics*, **5**(2B), 1456–1487.

Zhao S. and Prentice R. L. (2014). Covariate measurement error correction methods in mediation analysis with failure time data. *Biometrics*, **70**, 835–844.