Published in final edited form as:

Lifetime Data Anal. 2010 January ; 16(1): 136–153. doi:10.1007/s10985-009-9124-6.

Measurement error correction for the cumulative average model in the survival analysis of nutritional data: application to Nurses' Health Study

Weiliang Qiu and Bernard Rosner

Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

Weiliang Qiu: ; Bernard Rosner: stbar@channing.harvard.edu

Abstract

The use of the cumulative average model to investigate the association between disease incidence and repeated measurements of exposures in medical follow-up studies can be dated back to the 1960s (Kahn and Dawber, *J Chron Dis* 19:611–620, 1966). This model takes advantage of all prior data and thus should provide a statistically more powerful test of disease-exposure associations. Measurement error in covariates is common for medical follow-up studies. Many methods have been proposed to correct for measurement error. To the best of our knowledge, no methods have been proposed yet to correct for measurement error in the cumulative average model. In this article, we propose a regression calibration approach to correct relative risk estimates for measurement error. The approach is illustrated with data from the Nurses' Health Study relating incident breast cancer between 1980 and 2002 to time-dependent measures of calorie-adjusted saturated fat intake, controlling for total caloric intake, alcohol intake, and baseline age.

Keywords

Measurement error; Regression calibration; Nutritional data

1 Introduction

Kahn and Dawber (1966) used a cumulative average model to make use of all known cholesterol values prior to the occurrence of an event to investigate the development of coronary heart disease over a short time period in relation to sequential biennial measures of cholesterol in the Framingham Heart Study. The cumulative average model has been also studied by several researchers (e.g., Wu and Ware 1979; Cupples et al. 1988; D'Agostino et al. 1990) using data from the Framingham Heart Study. Several authors (e.g., Hu et al. 1997; Kim et al. 2006) applied the cumulative average model to the Nurses' Health Study.

The cumulative average model takes advantage of all prior data and thus should provide a statistically more powerful test of an association of cumulative exposure (Willett 1998, p. 333). However, measurement error in covariates is not unusual for medical follow-up studies. For example, nutritional intake measured with instruments such as a food frequency questionnaire

Correspondence to: Bernard Rosner, stbar@channing.harvard.edu.

Electronic supplementary material The online version of this article (doi:10.1007/s10985-009-9124-6) contains supplementary material, which is available to authorized users.

(FFQ) or a 24-h recall often have substantial systematic measurement error, in addition to random measurement error. Unlike random measurement error, systematic measurement error cannot be reduced by obtaining replicate measures of diet at different points in time and averaging the responses. Ignoring systematic measurement error could cause biased inference. Hence correction for systematic measurement error is needed for the cumulative average model.

Many methods have been proposed to correct for systematic measurement error (see Carroll et al. 2006, for an overview). However, most existing methods (e.g., Rosner et al. 1989, 1990; Spiegelman et al. 1997) considered measurement error correction based on covariates measured at only one time point. Recently, several authors, such as, Rosner et al. (2008) and Yi (2008) proposed methods for longitudinal data. To the best of our knowledge, no measurement correction methods have been proposed for the cumulative average model. In this article, we propose a regression calibration approach based on the method described in Rosner et al. (1990) to correct measurement error for the cumulative average relative risk estimates. The approach is illustrated with data from the Nurses' Health Study relating incident breast cancer between 1980 and 2002 to calorie-adjusted saturated fat intake, controlling for total caloric intake, alcohol intake, and baseline age.

The remainder of the article is structured as follows. In Sect. 2, we briefly describe the Nurses' Health Study data sets to motivate the cumulative average model. In Sect. 3, we then introduce the cumulative average model. In Sect. 4, we propose the measurement error correction method for the cumulative average model for the case where both main study data set and validation study data set are available at each time point. We conduct a simulation study to evaluate the performance of the measurement error correction in this setting. In Sect. 5, we describe how to correct measurement error based on the cumulative average model when validation study data are not available at some time points. The analysis results for the Nurses' Health Study (NHS) data based on the proposed measurement error correction method are presented in Sect. 6. A discussion is given in Sect. 7. Technique details are given in Web Appendices.

2 Nurses' Health Study data sets

The NHS is a continuing epidemiologic cohort study established at the Channing Laboratory, Brigham and Women's Hospital in 1976. Initially, the NHS investigated the potential long term consequences of the use of oral contraceptives. Around 120,000 married registered nurses (aged 30–54 in 1976) were selected to be followed prospectively. In 1980, diet information were collected via food frequency questionnaire (FFQ), which measures how often an individual eats particular types of food in the previous year. Subsequent diet information were collected in 1984, 1986, and every four years since.

It is well-known that nutrient intakes measured by FFQ have substantial measurement error. To correct for measurement error, two validation/calibration studies were conducted in 1980 and 1986 for two subsets of participants (Willett et al. 1985, 1988). During 1980, 173 participants were asked to fill out 4 weeks of diet record (DR) approximately 3 months apart over 1 year. At the end of the year, a second FFQ was administered which could be directly compared with the average of 28 days of DR. The 1986 validation study consisted of 190 participants, 92 of whom were also in the 1980 validation study.

We are interested in assessing the possible association between the incidence of breast cancer and calorie-adjusted saturated fat intake, controlling for total caloric intake, alcohol intake, and baseline age.

Calorie-adjusted saturated fat intake is defined as percent of energy due to saturated fat ($=100 \times 9 \times \text{saturated fat intake (g)} / \text{total caloric intake (kcal)}$). Age in 1980 is categorized in five

categories: 34–39, 40–44, 45–49, 50–54, and 55+. Calorie-adjusted saturated fat, total caloric intake, and alcohol intake are measured with error while age is measured without error. FFQ questionnaires were collected from the main NHS studies in 1980, 1984, 1986, 1990, 1994, and 1998.

Some descriptive statistics concerning the main and validation study populations are given in Table 1.

For the main studies, the average % calories due to saturated fat decreased over time. The total caloric intake increased from 1980 to 1984, and then remained about the same from 1984 to 1998. This is probably because the 1980 questionnaire had 61 food items while questionnaires from 1984 on had over 120 food items. Mean alcohol intake decreased from 1980 to 1998, except for a slight increase from 1980 to 1984.

Pearson correlations between FFQ intake at repeated surveys over 18 years are presented in Table 2.

As expected, the correlations between intake of the same nutrient over time decreases as the difference between time points increases, although the rate of decrease is small. There is a low correlation (0.05–0.16) between calorie-adjusted saturated fat intake and total caloric intake when measured at the same time. The cross correlation between calorie-adjusted saturated fat intake at time t_1 and total caloric intake at time t_2 ranges from -0.01 to 0.08 . The correlation between calorie-adjusted saturated fat intake and alcohol intake is slightly negative and the correlation between total caloric intake and alcohol intake is slightly positive. The correlation between the same nutrient measured in 1980 and that measured in 1984 is smaller than the correlation between the same nutrient measured over any other two consecutive surveys. This is partly because the FFQ included around 60 more food items starting from 1984, resulting in more stable estimates of nutrient intake over time.

We used four different approaches to assess the possible association between the incidence of breast cancer and calorie-adjusted saturated fat intake, controlling for total caloric intake, alcohol intake, and baseline age. All four approaches make use of the Cox proportional hazards regression model.

With the first approach, we related exposures in 1998 to breast cancer incidence from 1998 to 2002 (1181 cases). With the second approach, we related the average exposure from 1980 to 1998 to breast cancer incidence from 1998 to 2002 (1351 cases). With the third approach, we related exposure in 1980 to breast cancer incidence from 1980 to 2002 (5672 cases). With the fourth approach, we first related average exposure during each of 6 time periods (1980, 1980–1984–1980–1986–1980–1990–1980–1994–1980–1998) to breast cancer incidence over the corresponding succeeding time period (1980–1984–1984–1986–1986–1990–1990–1994–1994–1998–1998–2002) with a total of 5672 cases over 1980–2002, and then pooled the six estimates.

Compared to the first three approaches, the fourth approach make full use of all available information and hence can dampen the effects of random errors more effectively. The fourth approach was mentioned in the literature as early as 1960s (e.g., Kahn and Dawber 1966) and was referred to as the cumulative average model. In this article, we consider the measurement error correction problem for the cumulative average model. We briefly describe this model in the next section.

3 Cumulative average model

The idea of the cumulative average model described in Kahn and Dawber (1966) is to first treat each time interval as a mini follow-up study, then to pool observations over all intervals to examine the short-term development of disease (D'Agostino et al. 1990). Wu and Ware (1979) implemented the cumulative average model using pooled logistic regression. In addition, Cupples et al. (1988), and Kim et al. (2006) implemented the cumulative average model using Cox proportional hazards regression. In this article, we will also use the Cox proportional hazards regression model.

Let $\mathbf{X}_{it}$ be $k_1 \times 1$ vector corresponding to k_1 exposure variables (e.g. calorie-adjusted saturated fat intake, total caloric intake, and alcohol intake measured via DR) which represent the true levels of exposures for the i -th subject at time t . Let $\mathbf{U}_{it}$ be a $k_2 \times 1$ vector of variables measured without error (e.g. baseline age) for the i -th subject at time t . The period between exams t and $t+1$ will be called the t -th observation period. For the t -th observation period, a Cox proportional hazards regression model can be fitted to relate disease hazard to cumulative average exposures measured up to time t :

$$h_c(t|\bar{\mathbf{X}}_{it}, \bar{\mathbf{U}}_{it}) = h_{c0}(t) \exp[\beta_{c1}^* \bar{\mathbf{X}}_{it} + \beta_{c2}^* \bar{\mathbf{U}}_{it}], \quad i=1, \dots, N, \quad t=1, \dots, T, \quad (1)$$

where $h_c(t|\bar{\mathbf{X}}_{it}, \bar{\mathbf{U}}_{it})$ is the hazard function during observation period t , $h_{c0}(t)$ is the baseline hazard function at time t , $\bar{\mathbf{X}}_{it} = \sum_{\ell=1}^t \mathbf{X}_{i\ell} / n_t$, $\bar{\mathbf{U}}_{it} = \sum_{\ell=1}^t \mathbf{U}_{i\ell} / n_t$, n_t is the number of subjects free of disease up to time t , and T is the end point of the study (e.g. $T = 6$ for NHS). Most standard statistical software can provide parameter estimates for the Cox proportional hazards regression model.

Denote $\widehat{\beta}_c^{*(t)} = (\widehat{\beta}_{c1}^{*(t)}, \widehat{\beta}_{c2}^{*(t)})$ as the parameter estimates for Model (1). A *pooled* estimate of regression coefficients up to time T for the cumulative average model is

$$\widehat{\beta}_{c,\ell,T}^* = \sum_{t=1}^T w_{\ell}^{*(t)} \widehat{\beta}_{c,\ell}^{*(t)} = (\mathbf{w}_{\ell,T}^*)' \widehat{\gamma}_{\ell,T}^*, \quad (2)$$

where $\widehat{\beta}_{c,\ell}^{*(t)}$ is the ℓ -th element of the $1 \times (k_1 + k_2)$ vector $\widehat{\beta}_c^{*(t)}$, $\ell = 1, \dots, k \equiv k_1 + k_2$, and the $T \times 1$ vector $\mathbf{w}_{\ell,T}^* = (w_{\ell}^{*(1)}, w_{\ell}^{*(2)}, \dots, w_{\ell}^{*(T)})' = (\mathbf{\Omega}_{\ell,T}^*)^{-1} \mathbf{1}_T / [\mathbf{1}_T' (\mathbf{\Omega}_{\ell,T}^*)^{-1} \mathbf{1}_T]$, where $\mathbf{1}_T$ is the $T \times 1$ vector of ones, $\mathbf{\Omega}_{\ell,T}^*$ is the conditional variance-covariance matrix of the $T \times 1$ vector $\widehat{\gamma}_{\ell,T}^* = (\widehat{\beta}_{c,\ell}^{*(1)}, \dots, \widehat{\beta}_{c,\ell}^{*(T)})'$ given $\mathbf{W}^*$, $\mathbf{W}^*$ is the matrix of exposures $\mathbf{X}^{(t)}$ and $\bar{\mathbf{U}}^{(t)}$ up to time t , $t = 1, \dots, T$, and the symbol $'$ represents the matrix/vector transpose operation.

The Lagrange multiplier method can be used to derive the weight vector $\mathbf{w}_{\ell,T}^*$ which minimizes the variance of the overall estimator $\sum_{t=1}^T v^{*(t)} \widehat{\beta}_{\ell}^{*(t)}$, subject to the conditions $\sum_{t=1}^T v^{*(t)} = 1$, and $v^{*(t)} > 0$, $t = 1, \dots, T$ (see Web Appendix A). The variance of $\widehat{\beta}_{c,\ell,T}^*$ given $\mathbf{W}^*$ is

$$\text{Var}(\widehat{\beta}_{c,\ell,T}^* | \mathbf{W}^*) = 1 / [\mathbf{1}_T' (\mathbf{\Omega}_{\ell,T}^*)^{-1} \mathbf{1}_T]. \quad (3)$$

4 Measurement error correction method

It is usually impossible or expensive to directly measure X_{it} on a large number of subjects. Instead, we usually observe *surrogate* exposure variables which are subject to measurement error. Let Z_{it} be a $k_1 \times 1$ vector corresponding to the k_1 surrogate exposure variables (e.g. calorie-adjusted saturated fat intake, total caloric intake, and alcohol intake measured via FFQ) for the i -th subject at time t . The Cox proportional hazards regression for the surrogate variables is

$$h_c(t|\bar{Z}_{it}, \bar{U}_{it}) = h_{c0}(t) \exp[\beta_{c1} \bar{Z}_{it} + \beta_{c2} \bar{U}_{it}], \quad i=1, \dots, N, \quad (4)$$

where $\bar{Z}_{it} = \sum_{m=1}^t Z_{im}/n_t$. Denote $\hat{\beta}_c^{(t)} = (\hat{\beta}_{c1}^{(t)}, \hat{\beta}_{c2}^{(t)})$ as the parameter estimates for Model (4).

The parameters $\beta_c^{(t)} = (\beta_{c1}^{(t)}, \beta_{c2}^{(t)})$ for the surrogate regression model (4) might be attenuated relative to the true regression parameters $\beta_c^{*(t)} = (\beta_{c1}^{*(t)}, \beta_{c2}^{*(t)})$ due to measurement error (see the simulation study below). Many methods (e.g. Prentice 1982; Spiegelman et al. 1997; Li and Lin 2003; Yi and He 2006; Carroll et al. 2006; Yi and Lawless 2007) have been proposed to correct relative risk estimates for measurement error in a survival data analysis setting. However, to the best of our knowledge, no methods have been proposed for measurement error correction of cumulative average exposures.

We apply the regression calibration method described in Rosner et al. (1990) to handle the cumulative average exposure case. The regression calibration method described in Rosner et al. (1990) first models the relation between true exposure and surrogate exposure, usually based on a validation study. It then replaces the true exposures by the conditional expectation of the true exposures conditioned on the surrogate exposures in the model relating the true exposures to disease. Next the regression coefficients of the revised model are compared with those of the model relating the surrogate exposures to disease. The comparison finally leads to the corrected regression coefficients. The regression calibration method assumes that a nondifferential error mechanism holds (i.e. $Pr(Z|X, D) = Pr(Z|X)$, where $D = 1$ means diseased, $D = 0$ means not-diseased). In our example, it is reasonable to assume a nondifferential error mechanism since both Z values and X values are observed before any breast cancer occurs.

We assume a multivariate regression relating X_{it} , Z_{it} , and $\bar{U}_{it}$ in the validation data:

$$\bar{X}_{it} = \alpha_c + \Lambda_{c1} \bar{Z}_{it} + \Lambda_{c2} \bar{U}_{it} + e_{c,it}, \quad i=1, \dots, n_t, \quad (5)$$

where Λ_{c1} and Λ_{c2} are $k_1 \times k_1$ and $k_1 \times k_2$ matrices, $e_{c,it}$ is multivariate normally distributed with a $k_1 \times 1$ mean vector $\mathbf{0}$ and a $k_1 \times k_1$ covariance matrix Σ_{ce} .

By substituting expected values $E[X_{it} | Z_{it}, \bar{U}_{it}]$ obtained from Model (5) for X_{it} in Model (1) and then comparing it with Model (4), we can show that to an excellent degree of approximation (see the simulation study below):

$$\beta_c^* = \beta_c \Lambda_c^{-1}, \quad (6)$$

where $\beta_c^* = (\beta_{c1}^*, \beta_{c2}^*)$ and $\beta_c = (\beta_{c1}, \beta_{c2})$ are $1 \times k$ vectors, Λ_c is a $k \times k$ matrix given by

$$\Lambda_c = \begin{pmatrix} \Lambda_{c1} & \Lambda_{c2} \\ \mathbf{0}_{k_2 \times k_1} & \mathbf{I}_{k_2 \times k_2} \end{pmatrix} \quad (7)$$

and $k = k_1 + k_2$.

Using the results from Web Appendix B, we have

$$\widehat{\text{Cov}}(\widehat{\beta}_{c,j_1}^*, \widehat{\beta}_{c,j_2}^*) = \left(\mathbf{A}_c' \widehat{\Sigma}_{\beta_c} \mathbf{A}_c \right)_{j_1, j_2} + \widehat{\beta}_c \widehat{\Sigma}_{\mathbf{A}_{c,j_1,j_2}} \widehat{\beta}_c', \quad (8)$$

where $\mathbf{A}_c = \Lambda_c^{-1}$,

$$\widehat{\Sigma}_{\mathbf{A}_{c,j_1,j_2}} = w \cdot [\mathbf{A}_{c1} \widehat{\Sigma}_{ce} \mathbf{A}_{c1}'] \quad (9)$$

is the $k \times k$ variance-covariance matrix relating the elements in the j_1 -th and j_2 -th columns of $\mathbf{A}_c$ (denoted by $\mathbf{a}_{cj_1}$ and $\mathbf{a}_{cj_2}$, respectively), $w = \left[\mathbf{a}_{cj_1}' \mathbf{D} (\overline{\mathbf{W}}_t' \overline{\mathbf{W}}_t)^{-1} \mathbf{D}' \mathbf{a}_{cj_2} \right]$ is a scalar, the $k \times k_1$ matrix $\mathbf{A}_{c1}$ is the first k_1 columns of $\mathbf{A}_c$, $\Sigma_{ce} = (\text{Cov}(X_{t,r}, X_{t,s}) | \mathbf{W}_t)_{r,s=1,\dots,k_1} = \text{Cov}(e_{ct,r}, e_{ct,s})_{r,s=1,\dots,k_1}$, $\mathbf{D} = (\mathbf{0}_k \times \mathbf{1}, \mathbf{I}_k \times k)$, $\overline{X}_{t,r}$ is the average true value of the r -th variable measured with error, and

$$\overline{\mathbf{W}}_t = \begin{pmatrix} 1 & \overline{\mathbf{Z}}_{1t}' & \overline{\mathbf{U}}_{1t}' \\ \vdots & \vdots & \vdots \\ 1 & \overline{\mathbf{Z}}_{n_t t}' & \overline{\mathbf{U}}_{n_t t}' \end{pmatrix},$$

where $\mathbf{W}_t$ is obtained from the validation study of size n_t at time t . Using the results from Web Appendix C, we can obtain the variance $\widehat{\beta}_{c,t,T}^*$ in Eq. 3 of the *pooled* estimate of regression coefficients in Eq. 2 up to time T .

To evaluate the validity of the approximation in Eq. 6, we conducted a simulation study based on the 1980 and 1986 surveys. In the simulation study, we consider three exposures: calorie-adjusted saturated fat intake (denoted as Z_1 measured by FFQ, X_1 measured by DR), total caloric intake (denoted as Z_2 measured by FFQ, X_2 measured by DR), and baseline age (regarded as a continuous variable and denoted by U). We generated 1,000 simulated data sets. Each data set consists of a main study and a corresponding validation study at 2 time points. Each main study contains 10,000 subjects. Each validation study contains 200 subjects.

To take into account of (a) the correlations among different exposures at the same time point, (b) the correlations of the same exposure measured at different times, and (c) the cross correlations among different exposures measured at different times, we assume that the random vector $\mathbf{W} = (Z_{1,t_1}, Z_{2,t_1}, Z_{1,t_2}, Z_{2,t_2}, U)'$ is normally distributed with mean vector $\boldsymbol{\mu}_c$ and covariance matrix Σ_c . We regard the moment estimates of $\boldsymbol{\mu}_c$ and Σ_c obtained based on the 1980 and 1986 main studies as the true values of $\boldsymbol{\mu}_c$ and Σ_c in the simulations study.

For each data set, we first generated random multivariate normal vectors $W_1, \dots, W_{10,000}$ in the main study. We then computed the average nutrient intakes $\bar{Z}_{1,k} = (Z_{1,t1,k} + Z_{1,t2,k})/2$ and $\bar{Z}_{2,k} = (Z_{2,t1,k} + Z_{2,t2,k})/2$, $k = 1, \dots, 10,000$. We next generated random error vectors $e_{c,k}$, $k = 1, \dots, 10,000$, from the multivariate normal distribution $N(\mathbf{0}, \Sigma_{ce})$, where Σ_{ce} is the sample covariance matrix of the residuals of the multivariate linear model (5) based on the 1980 and 1986 validation studies. We next computed $\mathbf{X}_k = (X_{1,k}, X_{2,k})'$ based in Eq. 5, where the parameters α_c , Λ_1 , and Λ_2 are the estimated regression coefficients based on the 1980 and 1986 validation studies. In the same way, we generated data for the validation study ($k = 1, \dots, 200$) using the same μ_c and Σ_{ce} as in the main study.

We next generated survival times based on the Weibull distribution with hazard function $h(t|\delta, \eta(\mathbf{X}, U)) = \delta t^{\delta-1} \eta^{-\delta}$, where δ is the shape parameter and $\eta(\mathbf{X}, U)$ is the scale parameter. We assume that the logarithm of the scale parameter η is a linear function of $\mathbf{X}$ and U :

$\log(\eta) = \alpha_0 + \theta'_1 \bar{\mathbf{X}} + \theta'_2 U$. Then, the Weibull distribution has proportional hazards. The log hazard ratios for $\mathbf{X}$ and U are $\beta_{c1}^* = -\delta\theta'_1$ and $\beta_{c2}^* = -\delta\theta'_2$, respectively. In the simulation study, the true values of the parameters α_0 and δ are the estimates obtained by fitting the Weibull regression to the FFQ data $(\mathbf{Z}, U)$ based on 1980 and 1986 main studies. We considered several values for $\exp(\beta_{c1,1}^*) = 0.7, 1.0, 1.5$ corresponding to a protective, null and deleterious effect of saturated fat. We set $\exp(\beta_{c1,2}^*) = 1$ since there usually is no association between total calories and breast cancer, and set $\exp(\beta_{c2,1}^*) = 1.1$ allowing for a 10% increase in breast cancer incidence per year, where $\beta_{c1,1}^* = \%$ calories due to saturated fat, $\beta_{c1,2}^* =$ total caloric intake, and $\beta_{c2,1}^* =$ age at baseline.

Once we generated the survival times, we choose the 10% quantile of the survival times as the censoring time so that the specified survival rate 90% is achieved.

Finally, we apply the measurement error correction method we proposed for the cumulative average model to each simulated data set.

The percent biases and coverage probabilities of un-corrected and corrected estimates are shown in Table 3. The full simulation results are given in Web Appendix D.

The simulation results show that (i) corrected estimates of HRs have small percent bias and have appropriate coverage both for $HR = 1.0$ and $HR \neq 1.0$; (ii) If the true HR is 1.0, the uncorrected estimates have small percent bias and have appropriate coverage; (iii) If the true HR is not equal to 1.0, the un-corrected estimates for the exposure (Z) subject to measurement error have a large percent bias towards the null and low coverage. (iv) If the true HR is not equal to 1.0, the un-corrected estimates for the exposure (U) without measurement error have small percent bias, but low coverage; (v) For the exposure (U) without measurement error, corrected estimates have smaller percent bias than un-corrected estimates. In summary, the corrected estimates in Eq. 6 have small bias and appropriate coverage, while the un-corrected estimates tend to be attenuated towards 1.0, resulting in large bias and low coverage.

5 Analysis of NHS data

We could not apply the proposed method directly to NHS data sets because validation studies are not available at some time points. Hence, we cannot estimate Λ_c directly.

We propose an indirect method to overcome this difficulty. From (5), we have

$$\begin{aligned}\text{Cov}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{Z}}_{it}) &= \Lambda_{c1} \text{Var}(\bar{\mathbf{Z}}_{it}) + \Lambda_{c2} \text{Cov}(\bar{\mathbf{U}}_{it}, \bar{\mathbf{Z}}_{it}) \\ \text{Cov}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{U}}_{it}) &= \Lambda_{c1} \text{Cov}(\bar{\mathbf{Z}}_{it}, \bar{\mathbf{U}}_{it}) + \Lambda_{c2} \text{Var}(\bar{\mathbf{U}}_{it}).\end{aligned}\quad (10)$$

Thus, from (10)

$$(\Lambda_{c1}, \Lambda_{c2}) = \mathbf{B}_1 \mathbf{B}_2^{-1}, \quad (11)$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ are $k_1 \times k$ and $k \times k$ matrices given by

$$\begin{aligned}\mathbf{B}_1 &= [\text{Cov}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{Z}}_{it}), \text{Cov}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{U}}_{it})] \\ \mathbf{B}_2 &= \begin{bmatrix} \text{Var}(\bar{\mathbf{Z}}_{it}) & \text{Cov}(\bar{\mathbf{Z}}_{it}, \bar{\mathbf{U}}_{it}) \\ \text{Cov}(\bar{\mathbf{U}}_{it}, \bar{\mathbf{Z}}_{it}) & \text{Var}(\bar{\mathbf{U}}_{it}) \end{bmatrix}.\end{aligned}$$

We can estimate $\mathbf{B}_2$ from the main study where

$$\widehat{\text{Var}}(\bar{\mathbf{Z}}_{it}) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{Z}_{it_1}, \mathbf{Z}_{it_2}), \widehat{\text{Var}}(\bar{\mathbf{U}}_{it}) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{U}_{it_1}, \mathbf{U}_{it_2}), \widehat{\text{Cov}}(\bar{\mathbf{Z}}_{it}, \bar{\mathbf{U}}_{it}) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{Z}_{it_1}, \mathbf{U}_{it_2}).$$

To estimate $\mathbf{B}_1$, we assume a stationary covariance structure, whereby the covariance between the true exposures $\mathbf{X}_{t_1}$ at time point t_1 and the surrogate exposures $\mathbf{Z}_{t_2}$ at time point t_2 (or the covariates measured without error $\mathbf{U}_{t_2}$ at time point t_2) do not depend on the starting time point, but on the duration $|t_2 - t_1|$, i.e., $\text{Cov}(\mathbf{X}_{it_1}, \mathbf{Z}_{it_2}) = \text{Cov}(\mathbf{X}_{i1}, \mathbf{Z}_{i,|t_1-t_2|+1})$, $\text{Cov}(\mathbf{X}_{it_1}, \mathbf{U}_{it_2}) = \text{Cov}(\mathbf{X}_{i1}, \mathbf{U}_{i,|t_1-t_2|+1})$. The assumption is reasonable based on the NHS data (see Web Appendix E). Hence,

$$\begin{aligned}\widehat{\text{Cov}}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{Z}}_{it}) &= \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{X}_{it_1}, \mathbf{Z}_{it_2}) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{X}_{i1}, \mathbf{Z}_{i,|t_1-t_2|+1}), \\ \widehat{\text{Cov}}(\bar{\mathbf{X}}_{it}, \bar{\mathbf{U}}_{it}) &= \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{X}_{it_1}, \mathbf{U}_{it_2}) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{\text{Cov}}(\mathbf{X}_{i1}, \mathbf{U}_{i,|t_1-t_2|+1}).\end{aligned}\quad (12)$$

Thus, we can estimate β_c from (4), Λ_c from (10) to (12), and β_c^* from (6) whereby

$$\widehat{\beta}_c^* = \widehat{\beta}_c \widehat{\Lambda}_c^{-1}. \quad (13)$$

Equations 8 and 9 indicate that the calculation of the covariance matrix $\text{Cov}(\widehat{\beta}_c^*)$ involves the estimation of Σ_{ce} , i.e., $\text{Cov}(\bar{X}_{t,r}, \bar{X}_{t,s} | \bar{\mathbf{W}}_t)$. We have

$$\text{Cov}(\bar{X}_{t,r}, \bar{X}_{t,s} | \bar{\mathbf{W}}_t) = \frac{1}{t^2} \sum_{t_1, t_2=1}^t \text{sd}(X_{t_1,r} | \mathbf{W}_t) \text{sd}(X_{t_2,s} | \mathbf{W}_t) \text{corr}(X_{t_1,r}, X_{t_2,s} | \mathbf{W}_t), \quad (14)$$

where $sd(X_{t,r}|\mathbf{W}_t)$ is the standard deviation of the r -th true exposure at time point t given the surrogate exposures $\mathbf{Z}_t$ and variables measured without error $\mathbf{U}_t$ at time point t (c.f. Model (5)).

We assume the availability of longitudinal validation study data where $\mathbf{X}_t$, $\mathbf{Z}_t$ and $\mathbf{U}_t$ are available for at least a subset of subjects at 2 points in time. We also assume stationarity, whereby $sd(X_{t,r}|\mathbf{W}_t)$ is the same for all t and is denoted by $sd(X_r|\tilde{\mathbf{W}})$. The estimates of the conditional standard deviations for calorie-adjusted saturated fat intake, total caloric intake, and alcohol intake are 0.33, 0.31, and 0.18, respectively, based on the 1980 validation study, and are 0.32, 0.41, 0.22, respectively, based on the 1986 validation study. Hence the stationarity assumption is reasonable. Finally, we also assume a compound symmetric correlation structure whereby $corr(X_{t_1,r}, X_{t_2,s}|\mathbf{W}_t)$ is the same for all $t_1 \neq t_2$ and is denoted by $(\rho_{rs}|\tilde{\mathbf{W}})$. Since we have only two validation study data sets, we could not check this assumption. However, the correlations and cross-correlations of surrogate intake (Z_{it}) in Table 2 decreases only slightly over time. This gives some indication that the compound symmetric correlation assumption is reasonable for the NHS data sets.

Note that $sd(X_r|\tilde{\mathbf{W}})$ can be estimated from the cross-sectional validation study database of size n_t based on the residual mean square from the regression of $\mathbf{X}_t$ on $\mathbf{Z}_t$, $\mathbf{U}_t$. Also $(\rho_{rs}|\tilde{\mathbf{W}})$ can be estimated from the longitudinal validation study database of $n_{t_1 t_2}$ subjects seen at both times t_1 and t_2 obtained by correlating the residual $(X_{t_1,r}|\mathbf{Z}_{t_1}, \mathbf{U}_{t_1})$ and the residual $(X_{t_2,s}|\mathbf{Z}_{t_2}, \mathbf{U}_{t_2})$. Thus, we estimate $Cov(X_{t,r}, X_{t,s}|\mathbf{W}_{it})$ from

$$\begin{aligned} \widehat{Cov}(\bar{X}_{t,r}, \bar{X}_{t,s}|\bar{\mathbf{W}}_t) &= \frac{1}{t^2} \sum_{t_1, t_2=1}^t \widehat{sd}(X_{t_1,r}|\mathbf{Z}_{t_1}, \mathbf{U}_{t_1}) \widehat{sd}(X_{t_2,s}|\mathbf{Z}_{t_2}, \mathbf{U}_{t_2}) (\widehat{\rho}_{rs}|\mathbf{Z}_t, \mathbf{U}_t) \\ &= \frac{1}{t} \widehat{sd}(X_r|\tilde{\mathbf{W}}) \widehat{sd}(X_s|\tilde{\mathbf{W}}) \left[\widehat{\rho}_{rs}^{(1)} + (t-1) (\widehat{\rho}_{rs}|\tilde{\mathbf{W}}) \right], \end{aligned} \quad (15)$$

where $\widehat{\rho}_{rs}^{(1)} = corr(X_{t,r}, X_{t,s}|\mathbf{W}_t)$ and $\widehat{\rho}_{rs} = corr(X_{t_1,r}, X_{t_2,s}|\mathbf{W}_{t_1}, \mathbf{W}_{t_2})$, $t_1 \neq t_2$. Hence

$$\widehat{\Sigma}_{ce} = \frac{1}{t} \left(\frac{\widehat{\Sigma}^{(1)} + \widehat{\Sigma}^{(3)}}{2} \right) + \frac{t-1}{t} \left(\frac{\widehat{\Sigma}^{(1,3)} + \widehat{\Sigma}^{(3,1)}}{2} \right), \quad (16)$$

where $\Sigma^{(1)}$ and $\Sigma^{(3)}$ are the covariance matrices for the k_1 variables measured with error at time point 1 (1980) and 3 (1986), respectively, the (r, s) -th cell of the matrix $\Sigma^{(1,3)}$ is the covariance between the r -th variable measured with error at time point 1 and the s -th variable measured with error at time point 3, and the (r, s) -th cell of the matrix $\Sigma^{(3,1)}$ is the covariance between the r -th variable measured with error at time point 3 and the s -th variable measured with error at time point 1. $\Sigma^{(1)}$ can be estimated based on the 1980 validation study with size $n_1 = 173$, $\Sigma^{(3)}$ can be estimated based on the 1986 validation study with size $n_3 = 190$, and $\Sigma^{(1,3)}$ and $\Sigma^{(3,1)}$ can be estimated based on the longitudinal validation study consisting of $n_{1,3} = 92$ subjects in both 1980 and 1986 validation studies.

Thus, from (9) and (16) we can estimate $\Sigma \mathbf{A}_{c,j_1,j_2}$ based on the cross-sectional validation study databases of sizes n_1 and n_3 and the longitudinal validation study database of size $n_{1,3}$, respectively. We can also estimate $\Sigma \hat{\beta}_c$ from the main study database of size N . Therefore, using (8), we can estimate the variance-covariance matrix of $\hat{\beta}_c^*$ given by $Cov(\hat{\beta}_{c,j_1}^*, \hat{\beta}_{c,j_2}^*), j_1, j_2 = 1,$

..., k , and can estimate 95% confidence limits for $\exp(\beta_{c,\ell}^*)$ based on the Wald method given by $\exp\left[\widehat{\beta}_{c,\ell}^* \pm 1.96 \sqrt{\text{Var}(\widehat{\beta}_{c,\ell}^*)}\right]$.

6 Results

We apply the proposed measurement error correction method described in (6), (8), (9), (15), and (16) to Nurses' Health Study to assess the possible association between the incidence of breast cancer and calorie-adjusted saturated fat intake (time-dependent), controlling for total caloric intake (time-dependent), alcohol intake (time-dependent), and baseline age. We present four different models.

In *Model 1*, we relate exposures in 1998 to breast cancer incidence from 1998 to 2002 (1181 cases) using the measurement error correction method in Rosner et al. (1990). In *Model 2*, we relate cumulative average exposures from 1980 to 1998 to breast cancer incidence from 1998 to 2002 (1351 cases) using Eq. 13 without any pooling and assessing outcome over only one time period (1998 to 2002). In *Model 3*, we relate baseline exposure in 1980 (without updating) to breast cancer incidence from 1980 to 2002 (5672 cases) using the measurement error correction method in Rosner et al. (1990). Finally, in *Model 4*, we relate cumulative average exposure from 1980 to 1998 to breast cancer incidence from 1980 to 2002 (5672 cases) using the pooled estimate in Eq. 2.

The point and interval estimates of hazard ratios (both un-corrected and corrected estimates) for *Model 1–Model 4* are shown in Table 4.

Since 1 g and 1 kcal are small increments for nutrient intake, we followed Kim et al. (2006) and have expressed the hazard ratios in increments of 5% energy for calorie-adjusted saturated fat intake, 800 kcal for total caloric intake, and 25 g for alcohol intake, which approximately correspond to increments between the 10th and 90th percentiles of the distribution of FFQ intake.

For all four models, we observe that (1) the corrected point estimates of the regression coefficients are further away from 1 than the uncorrected estimates (In other words, the point estimates are attenuated if we do not correct for measurement error), (2) the corrected 95% CIs are wider than the uncorrected 95% CIs, (3) total caloric intake has no significant effect on the development of breast cancer, and (4) alcohol intake and age in 1980 have a positive association with the incidence of breast cancer. These four observations are consistent with previous results (e.g. Rosner et al. 1989, 1990; Spiegelman et al. 1997). The observation that calorie-adjusted saturated fat intake has a slightly protective, but not significant, effect on the development of breast cancer for *Model 1* and *Model 3* is also consistent with previous results. However, for both the corrected and uncorrected estimates in *Model 2* and *Model 4* (the cumulative average model), calorie-adjusted saturated fat intake has a significant protective effect on the incidence of breast cancer.

Confidence interval widths for saturated fat intake tend to be smaller based on cumulative average intake vs. baseline intake. However, this is not true for alcohol or total caloric intake. Finally, as expected, confidence interval widths are narrower for *Models 3* and *4* based on 5672 events over 22 years than *Models 1* and *2* based on 1181 and 1351 events over 4 years. The number of events and person-years are larger for *Model 2* than *Model 1* because some subjects ($n = 20507$, 170 events) did not return the 1998 questionnaire, but did return at least one previous questionnaire.

7 Discussion

The proposed regression calibration model for cumulative average intake is based on a continuous representation of nutrient intake. However, nutrient intake is often represented in terms of quantiles to avoid making assumptions of a linear relationship between outcome and exposure. We have previously considered measurement error models where nutrient intake is represented as an ordinal variable (Rosner 1996). More work is needed to extend this model to the cumulative average setting such as in Eqs. 1, 4, and 5. Similar issues apply to extending misclassification models for nominal categorical exposure variables (e.g., Greenland 1980) to the cumulative average setting.

A possible extension for the cumulative average model is to use a more general weighted average $\sum_{j=1}^K w_j \mathbf{X}_{t_j}$ of a vector of covariates $\mathbf{X}$ measured over time with more weight given to reported intake on the most recent surveys. A possible weight w_j at time point t_j is $\exp[-\lambda(t_K - t_j)] / \sum_{\ell=1}^K \exp[-\lambda(t_K - t_\ell)]$, where t_K is the current time point, $t_j, j < K$ are previous time points when the FFQ was administered. When $\lambda = 0$, the weighted average gives equal weight to all previous surveys and is equivalent to a cumulative average. When $\lambda \gg 0$, the weighted average gives most weight to the most recent survey. It is also possible to allow $\lambda < 0$ in which case reported intake at distant time points is given more weight which might be a reasonable way to introduce a lag effect. Indeed, for the “baseline” model (Model 3) relating disease incidence to baseline exposures (e.g., nutritional intakes obtained from the 1980 FFQ questionnaire in NHS), $\lambda \ll 0$. This generalized cumulative average model is also an extension of Wu and Ware’s (1979, Sect. 3, Eq. 3) hierarchical model. We can generalize the measurement error correction method proposed in this article to this generalized cumulative average model.

Another assumption is that the diet record provides an unbiased estimate of true intake with no correlated error between DR and FFQ measurements. To account for correlated error in a cumulative average setting one would also need simultaneous biomarker measurements as part of the validation study data obtained on at least two points in time (Rosner et al. 2008).

It would be conceptually straightforward to work out the likelihood method. However, the likelihood method involves multiple integrations that integrate out the unobserved true exposures and hence is computationally intensive (Crouch and Spiegelman 1990). Hence, it would be prohibitively expensive to use likelihood method in a data set of the size of the Nurses’ Health Study. The simulation study in Sect. 4 indicates that the regression calibration method produces approximately unbiased estimates and approximately correct coverage under conditions similar to those seen in Nurses’ Health Study data.

In conclusion, we have presented a method for correcting for measurement error with nutrient intake measured on a continuous scale over multiple surveys. To use this methodology, one needs at least two validation studies that are concurrent with the main study at two different points in time with some subjects participating in both validation studies. It of course would be desirable to have validation study data at additional time points to further validate some of the assumptions mentioned above.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge the support of NHS program project CA87969 and also CA50597 in performing this work. We thank the editors and referees for their invaluable comments and suggestions. We acknowledge programming support of Rong Chen.

References

- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement error in nonlinear models: a modern perspective. Vol. 2. Chapman & Hall; Boca Raton, FL: 2006.
- Crouch EA, Spiegelman D. The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp(-t^2) dt$: application to logistic-normal models. *J Am Stat Assoc* 1990;85:464–469.
- Cupples LA, D'Agostino RB, Anderson K, Kannel WB. Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study. *Stat Med* 1988;7:205–218. [PubMed: 3353604]
- D'Agostino RB, Lee M-LT, Belanger AJ, Cupples LA, Anderson K, Kannel W. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med* 1990;9:1501–1515. [PubMed: 2281238]
- Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980;112:564–569. [PubMed: 7424903]
- Hu FB, Stampfer MJ, Manson JE, Rimm E, Colditz GA, Rosner BA, Hennekens CH, Willett WC. Dietary fat intake and the risk of coronary heart disease in women. *N Engl J Med* 1997;337:1491–1499. [PubMed: 9366580]
- Kahn HA, Dawber TR. The development of coronary heart disease in relation to sequential biennial measures of cholesterol in the Framingham Study. *J Chron Dis* 1966;19:611–620. [PubMed: 5966010]
- Kim EHJ, Willett WC, Colditz GA, Hankinson SE, Stampfer MJ, Hunter DJ, Rosner B, Holmes MD. Dietary fat and risk of postmenopausal breast cancer in a 20-year follow-up. *Am J Epidemiol* 2006;164:990–997. [PubMed: 16968865]
- Li Y, Lin X. Functional inference in frailty measurement error models for clustered survival data using the simex approach. *J Am Stat Assoc* 2003;98:191–203.
- Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982;69:331–342.
- Rosner BA. Measurement error models for ordinal exposure variables measured with error. *Stat Med* 1996;15:293–303. [PubMed: 8643887]
- Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989;8:1051–1069. [PubMed: 2799131]
- Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol* 1990;132:734–745. [PubMed: 2403114]
- Rosner B, Michels KB, Chen Y-H, Day NE. Measurement error correction for nutritional exposures with correlated measurement error: use of the method of triads in a longitudinal setting. *Stat Med* 2008;27:3466–3489. [PubMed: 18416440]
- Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr* 1997;65(suppl):1179S–1186S. [PubMed: 9094918]
- Willett, W. Nutritional epidemiology. Vol. 2. Oxford University Press; New York: 1998.
- Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol* 1985;122:51–65. [PubMed: 4014201]
- Willett WC, Sampson L, Browne ML, Stampfer MJ, Rosner B, Hennekens CH, Speizer FE. The use of a self-administered questionnaire to assess diet four years in the past. *Am J Epidemiol* 1988;127:188–199. [PubMed: 3337073]
- Wu M, Ware JH. On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics* 1979;35:513–521. [PubMed: 486684]

- Yi GY. A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics* 2008;9:501–512. [PubMed: 18199691]
- Yi GY, He W. Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. *Commun Stat A Theory Methods* 2006;35:1539–1554.
- Yi GY, Lawless JF. A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *J Stat Plan Infer* 2007;137:1816–1828.

Table 1

Descriptive statistics for the main and validation study population

Variable	Year:	Main study					Validation study		
		80	84	86	90	94	98	80	86
n		88999	70508	66112	65999	67931	64171	173	190
Age		46.5 ± 7.1 ^a	50.5 ± 7.1	52.7 ± 7.0	56.5 ± 7.0	60.4 ± 7.0	64.2 ± 7.0	45.9 ± 7.4	52.2 ± 7.2
Breast cancer	Cases	606 ^b	310	788	882	1046	1181	0	0
	Incidence rate	170 ^c	220	298	334	385	460	0	0
% Calorie due to sat. fat (%)	DR							13.8 ± 2.1	12.1 ± 2.4
	FFQ	15.6 ± 3.6	12.5 ± 2.6	11.7 ± 2.6	10.7 ± 2.5	9.7 ± 2.8	9.3 ± 2.7	14.4 ± 3.4	11.0 ± 2.3
Total calories (kcal/day)	DR							1620 ± 323	1635 ± 322
	FFQ	1562 ± 484	1730 ± 501	1752 ± 498	1736 ± 488	1727 ± 492	1710 ± 507	1372 ± 482	1823 ± 522
Alcohol (g/day)	DR							9.0 ± 9.7	9.4 ± 13.6
	FFQ	6.4 ± 10.6	6.9 ± 11.2	6.2 ± 10.7	5.1 ± 9.5	5.0 ± 9.1	5.0 ± 9.0	9.0 ± 12.3	7.1 ± 10.3

^aMean±1SD

^bAmong the 88,999 women who filled out the 1980 FFQ, 606 developed breast cancer from 1980 to 1984. The number of cases for the other time periods are defined similarly

^cPer 10⁵ person-years among women who filled out the 1980 FFQ and were followed from 1980 to 1984. Incidence rates for other years are defined similarly

Table 2

Correlation between FFQ nutrient intakes over time in main studies

Saffat					
	80	84	86	90	94
80	1.00	0.42	0.38	0.31	0.27
84		1.00	0.58	0.45	0.39
86			1.00	0.52	0.44
90				1.00	0.58
94					1.00
98					1.00
Calorie					
	80	84	86	90	94
80	1.00	0.49	0.46	0.42	0.40
84		1.00	0.63	0.56	0.52
86			1.00	0.60	0.55
90				1.00	0.60
94					1.00
98					1.00
Alco					
	80	84	86	90	94
80	1.00	0.74	0.70	0.63	0.59
84		1.00	0.79	0.70	0.65
86			1.00	0.74	0.69
90				1.00	0.75
94					1.00
98					1.00

Calorie						
	80	84	86	90	94	98
Satfat						
80	0.16	0.01	-0.00	-0.01	-0.01	-0.01
84		0.05	0.02	0.01	0.01	0.02
86			0.05	0.03	0.02	0.03
90				0.05	0.06	0.06
94					0.09	0.08
98						0.07
Alco						
	80	84	86	90	94	98
Satfat						
80	-0.14	-0.08	-0.08	-0.06	-0.06	-0.05
84		-0.16	-0.12	-0.10	-0.09	-0.08
86			-0.14	-0.09	-0.09	-0.08
90				-0.09	-0.06	-0.06
94					-0.05	-0.03
98						-0.04
Alco						
	80	84	86	90	94	98
Calorie						
80	0.10	0.04	0.04	0.03	0.02	0.02
84		0.11	0.05	0.05	0.04	0.04
86			0.10	0.05	0.04	0.04
90				0.09	0.05	0.04
94					0.10	0.05
98						0.11

Table 3

Percent bias and coverage based on 1,000 simulated data sets

True HR ^a		Uncorrected			Corrected			
H R ₁	H R ₂	H R ₃	β ₁	β ₂	β ₃	$\hat{\beta}_1^{\square}$	$\hat{\beta}_2^{\square}$	$\hat{\beta}_3^{\square}$
Percent bias ^b								
0.7	1.0	1.1	19.0	2.8	0.6	1.3	2.5	−0.0
1.0	1.0	1.1	0.3	0.4	−0.0	0.7	5.2	−0.0
1.5	1.0	1.1	−17.1	−3.1	−0.7	1.5	3.7	0.0
Coverage ^c								
0.7	1.0	1.1	20.1	94.0	75.6	96.8	97.5	95.4
1.0	1.0	1.1	95.1	95.2	93.3	95.9	96.8	93.8
1.5	1.0	1.1	13.3	91.2	67.2	96.4	96.5	94.9

Each data set contains data for 10,000 subjects in the main study and 200 subjects in the validation study. The mortality rate is set to be 0.1

^aThe 1st exposure is calorie-adjusted saturated fat intake; the 2nd exposure is total caloric intake; the 3rd exposure is baseline age. The first 2 exposures are measured with error, while the 3rd exposure is measured without error.

^b% bias:=100%[exp($\hat{\beta}$)/exp(β^*)-1]

^cCoverage probability=percentage of $B = 1,000$ samples for which the 95 percent confidence interval for $H R\{\exp[\hat{\beta}^* - 1.96se(\hat{\beta}^*)], \exp[\hat{\beta}^* + 1.96se(\hat{\beta}^*)]\}$ based on each respective estimator included the true parameter value $H R = \exp(\beta^*)$

Table 4
Comparison of the measurement error correction models based on Nurses' Health Study^a

Model	Follow-up period	#Cases	Independent variables								
			Nutritional intake			Age (years) in 1980 ^b					
			S ^c	C ^c	A ^c	40-44	45-49	50-54	55+		
Model 1	98-02	1181									
	U ^d		HR	0.96	0.99	1.37	1.06	1.17	1.24	1.31	
			95% CI	0.86-1.06	0.90-1.09	1.20-1.57	0.89-1.28	0.98-1.40	1.04-1.49	1.08-1.58	
	C ^d		HR	0.89	1.01	1.35	1.07	1.17	1.17	1.31	
Model 2	98-02	1351									
	U ^d		OR	0.79	1.07	1.36	1.05	1.13	1.14	1.09	
			95% CI	0.70-0.88	0.96-1.19	1.20-1.55	0.88-1.24	0.96-1.34	0.96-1.35	0.92-1.31	
	C ^d		OR	0.61	1.07	1.59	0.97	1.08	1.08	0.96	
Model 3	80-02	5672									
	U ^d		OR	1.00	0.99	1.21	1.26	1.48	1.65	1.80	
			95% CI	0.96-1.03	0.94-1.03	1.14-1.28	1.15-1.38	1.36-1.62	1.51-1.80	1.65-1.97	
	C ^d		OR	0.98	0.99	1.33	1.26	1.48	1.64	1.79	
Model 4	80-02	5672									
	U ^d		OR	0.88-1.10	0.82-1.20	1.21-1.47	1.15-1.39	1.34-1.63	1.49-1.80	1.62-1.97	
			HR	0.90	1.03	1.22	1.25	1.48	1.64	1.78	
	C ^d		OR	0.85-0.94	0.98-1.08	1.14-1.29	1.14-1.37	1.35-1.61	1.50-1.78	1.63-1.94	

Model 1: single measure of the exposures based on the 1998 survey only

Model 2: average of 6 exposures based on 1980, 1984, 1986, 1990, 1994, 1998 surveys

Model 3: single measure of the exposures based on the 1980 survey only

Model 4: cumulative average of the exposures based on 1980, 1984, 1986, 1990, 1994, 1998 surveys.

Example 1980 exposure is used to predict cases from 1980 to 1984. Average of 1980 and 1984 exposures is used to predict cases from 1984 to 1986, etc.

^a

based on 88,999 women ascertained in 1980 of whom 5672 developed breast cancer from 1980 to 2002.

^b

Reference group is age 34–39 in 1980

^c

S: % calories due to saturated fat intake (unit = 5%); C: total caloric intake (unit=800kcal/day); A: alcohol intake (unit = 25g/day)

^d

U: uncorrected; C: corrected