Biometrics 68, 397–407
June 2012
DOI: 10.1111/j.1541-0420.2011.01690.x

Hazard Ratio Estimation for Biomarker-Calibrated Dietary Exposures

Pamela A. Shaw[1,*] and Ross L. Prentice[2,]**

[1]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, U.S.A.

[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

[*]*email:* shawpa@niaid.nih.gov

[**]*email:* rprentic@fhcrc.org

Summary. Uncertainty concerning the measurement error properties of self-reported diet has important implications for the reliability of nutritional epidemiology reports. Biomarkers based on the urinary recovery of expended nutrients can provide an objective measure of short-term nutrient consumption for certain nutrients and, when applied to a subset of a study cohort, can be used to calibrate corresponding self-report nutrient consumption assessments. A nonstandard measurement error model that makes provision for systematic error and subject-specific error, along with the usual independent random error, is needed for the self-report data. Three estimation procedures for hazard ratio (Cox model) parameters are extended for application to this more complex measurement error structure. These procedures are risk set regression calibration, conditional score, and nonparametric corrected score. An estimator for the cumulative baseline hazard function is also provided. The performance of each method is assessed in a simulation study. The methods are then applied to an example from the Women's Health Initiative Dietary Modification Trial.

Key words: Biomarkers; Conditional score; Cox regression; Measurement error; Nonparametric corrected score; Nutrition assessment; Regression calibration.

1. Introduction

International reviews of diet and chronic disease data report many possible diet–disease associations, but few that are firmly established (World Cancer Research Fund/American Institute for Cancer Research, 2007). Studies of the same association in different populations, or using differing dietary assessment methods, often yield conflicting results. There are now cohorts where diet is assessed with both food frequency and food record data. In these contexts a positive association between dietary fat and breast cancer (Bingham et al., 2003, Freedman et al., 2006) and an inverse association between dietary fiber and colorectal cancer (Dahm et al., 2010) were found when food records were used, but these associations were not apparent when food frequency questionnaire (FFQ) data were substituted. These reports strongly suggest that the measurement error properties of the dietary assessment methods used need to be assessed and accommodated if reliable diet and disease associations are to be obtained.

Some nutritional epidemiology observational studies have attempted to address the measurement error issue by using one self-report assessment to calibrate another. However, a basic requirement for a reference estimate to be used to calibrate (or correct) another assessment is that of independent measurement errors. Because measurement errors for two self-report assessments may be strongly positively correlated, recent efforts have instead focused on biomarkers of nutrient consumption. For example, a doubly labeled water technique (Schoeller, 1988) provides a reliable assessment of short-term total energy expenditure, while urinary nitrogen yields a good biomarker of short-term protein expenditure (Bingham and Cummings, 1985). These recovery biomarkers (Kaaks, 1997) also provide estimates of consumption among weight-stable persons. Available biomarker study data document important systematic bias in relation to body mass index for energy, protein, and percentage of energy from protein for self-reported food frequency data (Heitman and Lissner, 1995, Subar et al., 2003, Neuhouser et al., 2008).

Building on the work of several authors (Prentice, 1996; Carroll et al., 1998, Jiang et al., 2001, Kipnis et al., 2001), Prentice et al. (2002) proposed a general measurement error model for self-reported dietary intake that incorporates location and scale bias terms that may depend on observed covariates. This model also allows the measurement error variance to depend on observed subject characteristics. Sugar, Wang, and Prentice (2007) considered this model for covariate measurement error and developed methods for odds ratio estimation (logistic regression model). Many potential applications, however, involve time-to-event outcomes. Here we consider hazard ratio (Cox model) estimators that accommodate this general error structure to relate nutrient consumption to a time-to-disease outcome.

Several methods have been proposed for Cox regression with mismeasured covariates, under a classical additive measurement model. These include regression calibration

(Prentice, 1982, Wang et al., 1997), risk set regression calibration (RRC) (Xie, Wang, and Prentice, 2001), parametric, semiparametric and nonparametric likelihood procedures (Hu, Tsiatis, and Davidian, 1998), conditional score (CS) (Tsiatis and Davidian, 2001), parametric corrected score (Nakamura, 1992), and nonparametric corrected score (NP) procedures (Huang and Wang, 2000, 2006; Hu and Lin, 2002, 2004; Gorfine, Hsu, and Prentice, 2004; Song and Huang, 2005). There has also been consideration of more general error models (Hu and Lin, 2002, Liao et al., 2011) under the assumption that a validation subsample is available, where the covariate of interest is precisely measured. In our setting, the precisely measured covariate is not obtainable.

Here our interest focuses on a study cohort with available self-report data and a biomarker subsample. Our applications to energy and protein consumption in relation to cancer (Prentice et al., 2009) and to cardiovascular disease incidence (Prentice et al., 2011) indicate that these assessments can be much improved by using such study subject characteristics as body mass index, age, and ethnicity to augment the food frequency self-report assessments. The disease occurrence rates are low (<5%) and censoring rates are high in the context of these studies, configurations under which bias in the regression calibration estimators is expected to be negligible. In this article, we extend RRC, CS, and NP procedures to the measurement model of Prentice et al. (2002) that allows for these types of dependencies. We evaluate and compare the performance of these procedures in simulation study and application.

2. Methods

Let X_i be the covariate of interest measured with error, a nutritional intake in our setting, and Z_i a vector of precisely measured covariates, for subjects $i = 1, \ldots, n$. The usual proportional hazards model for the continuous failure time $T_i \geq 0$ is assumed. The hazard rate $\lambda_i(t)$ for individual i at time t is given by $\lambda_0(t)\exp(\beta_1 X_i + \boldsymbol{\beta}_2' Z_i)$, where $\lambda_0(t)$ is an arbitrary baseline hazard function. Assume the right-censoring time C is independent of T given (X, Z). Let $N_i(t)$ denote the counting process for observed events, $Y_i(t) = I\{U_i = \min(T_i, C_i) \geq t\}$ be the at-risk indicator at time t, and $\Delta_i = I(T_i \leq C_i)$. A finite follow-up interval $[0, M]$ is assumed.

2.1 *Measurement Error Model*

Instead of observing X_i, k_i replicates of a self-reported Q_{ij} are observed, where Q_{ij} follows the general measurement error model (Prentice et al., 2002)

$$Q_{ij} = \delta_0 + \delta_1 X_i + \delta_2' Z_i + \delta_3' Z_i X_i + \gamma_i + \xi_{ij}, \qquad (1)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k_i$. Here, ξ_{ij} is mean-zero random error and γ_i is a mean-zero random effect that allows errors in repeat assessments for subject i to be correlated. The δ parameters determine the systematic bias of the assessment, including scale and location bias dependent on Z_i. A variance of the form $ae^{b'V_i}$ is considered for γ_i to allow the subject-specific error variance to depend on V_i, the categorical components of Z_i. Note, more generally that V_i could be a categorical characteristic derived from Z_i. For regression calibration, there may be continuous and discrete components of Z_i. For the conditional and corrected score methods that follow, all components of Z_i that impact the scale bias (i.e., have a nonzero δ_3 coefficient) are assumed to be discrete. The error ξ_{ij} is assumed independent of the other random variables on the right side of (1).

For $i = 1, \ldots, n$, we also assume there are κ_i replicates of an additional covariate W_{ij}, for $j = 1, \ldots, \kappa_i$, a biomarker that obeys the classical measurement error model

$$W_{ij} = X_i + \epsilon_{ij}. \qquad (2)$$

Importantly, the mean-zero error ϵ_{ij} is assumed independent of X_i and other terms on the right side of (1). Typically, due to expense, the biomarker would only be measured on a random subset of subjects, called the biomarker subset. Let R_i be the indicator that subject i is in the biomarker subset.

We assume that $(N_i, Y_i, X_i, \mathbf{Z}_i, R_i, \gamma_i, k_i, \kappa_i, \epsilon_{i1}, \ldots, \epsilon_{k_i}, \xi_{i1}, \ldots, \xi_{\kappa_i})$ are independent and identically distributed random vectors and that $(R_i,\ k_i,\ \kappa_i)$ may depend on observed baseline covariates $\boldsymbol{Z}_i$, but is otherwise independent of all other random variables in the survival and error models. With the additional assumptions that $P(R = 1) > 0$ and $P(V = v) > 0$ for all $v \in \{v|V = v\}$, one has $\rho_{1v} = P(R = 1, V = v) > 0$ from the independence of R and V. With this result and the strong law of large numbers, we have the necessary regularity that $n_{1v}/n \to \rho_{1v}$ as $n \to \infty$ for all v, where n_{1v} is the number of individuals with $R = 1$ and $V = v$. If $P(R = 1) < 1$, similar regularity holds for the nonbiomarker subset $(R = 0)$.

For ease of notation, X and Z are specialized to be univariate, with Z being categorical, and we assume all n individuals have k replicates of Q and all individuals in the biomarker subset have κ observations of W. The error model nuisance parameters are estimated separately using method of moments (Sugar et al., 2007). To ensure identifiability of all parameters in the measurement error model, it is enough that only a random subset of the biomarker cohort has $\kappa > 1$ replicates of W and a random subset of the main cohort has $k > 1$ replicates of Q. Importantly, for regression calibration, only single measures of Q and W are needed. The variance of γ in (1) is assumed to follow the model $\Sigma_\gamma = ae^{bZ}$.

2.2 *Risk Set Regression Calibration*

In this section, we extend the RRC estimator of Xie et al. (2001) to the generalized measurement error model (Section 2.1). For this model, the unobserved X_i are estimated separately depending on membership in the biomarker subset.

For an observed failure time t, define

$$\widehat{X}_i(t) = \begin{cases} \widehat{E}\{X_i | Y_i(t) = 1, W_{i\cdot}, Q_{i\cdot}, Z_i\} & \text{if} \quad R_i = 1, \\ \widehat{E}\{X_i | Y_i(t) = 1, Q_{i\cdot}, Z_i\} & \text{if} \quad R_i = 0, \end{cases} \qquad (3)$$

where $W_{i\cdot} = \kappa^{-1}\sum_{j=1}^{\kappa} W_{ij}, Q_{i\cdot} = k^{-1}\sum_{j=1}^{k} Q_{ij}$, and $\widehat{}$ denotes estimate. Sugar et al. (2007) discuss a class of $n^{\frac{1}{2}}$-consistent estimators for the nuisance parameters pertaining to the same measurement error model considered here. Explicit moment plug-in estimates for the nuisance parameters are given in the Supplementary Materials. The nuisance parameters can also be estimated by performing linear regression of the biomarker on the self-report and other observed covariates in

the measurement error model, as discussed in the appendix of Neuhouser et al. (2008). This method requires no replicates of the error-prone W or Q.

The RRC estimator is found by solving the following estimating equation for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$

$$n^{-1}\sum_{i=1}^{n}\int_0^M \left[\{\widehat{X}_i(t), Z_i\}' - \frac{\sum_{j=1}^{n} Y_j(t)\{\widehat{X}_j(t), Z_i\}'\exp\{\beta_1\widehat{X}_j(t) + \beta_2 Z_i\}}{\sum_{j=1}^{n} Y_j(t)\exp\{\beta_1\widehat{X}_j(t) + \beta_2 Z_i\}} \right] dN_i(t) = 0. \tag{4}$$

The ordinary regression calibration (RC) estimator is found by a similar equation to (4), only $\widehat{X}_j$ is estimated only once (at $t = 0$) instead of being reestimated for each risk set. The RC and RRC estimators are generally not consistent for the true $\boldsymbol{\beta}$, even if X_i is normally distributed, as the distribution for $X_i|\{Y_i(t) = 1,\ W_{i\cdot},\ Q_{i\cdot},\ Z_i\}$ is not normal typically (Prentice, 1982). In the classical measurement error setting, regression calibration provides an estimate for $\boldsymbol{\beta}$ with little asymptotic bias, provided there is small to moderate $\boldsymbol{\beta}$ and failure probabilities are small (Prentice, 1982, Xie et al., 2001). In many settings this simple estimator substantially eliminates the naive estimator bias and has good efficiency. Issues of bias will be explored for the proposed RRC estimator using simulation studies. Regularity conditions sufficient for asymptotic normality are listed in the Supplementary Materials.

2.3 *Conditional Score*

Stefanski and Carroll (1987) developed the CS estimator for generalized linear models. In the CS approach, a joint probability model for the mismeasured covariates and the response variable Y is assumed, and the unobserved covariates are treated as parameters. The CS estimating equation is obtained by conditioning the derived estimating equation on the sufficient statistics for the unobserved covariates, the X_i in our setting. Tsiatis and Davidian (2001) adapted this approach to the partial likelihood score, assuming the mismeasured covariates follow a linear mixed effects model with classical normal measurement error. Here, their CS method is extended to the generalized error model described above.

First consider an individual in the biomarker subset. Assuming normally distributed errors, one can condition the likelihood of $\{dN_i(t),\ Q_i,\ W_i\}$ given $\{X_i,\ Z_i,\ Y_i(t) = 1\}$ on the statistic

$$\zeta_i = \frac{\beta_1\Sigma_{e_i}\Sigma_{\epsilon_i}\, dN_i(t) + \Sigma_{\epsilon_i}(\delta_1 + \delta_3 Z_i)(Q_{i\cdot} - \delta_0 - \delta_2 Z_i) + \Sigma_{e_i} W_{i\cdot}}{\Sigma_{e_i} + \Sigma_{\epsilon_i}(\delta_1 + \delta_3 Z_i)^2},$$

where $\Sigma_{\epsilon_i} = \Sigma_\epsilon/\kappa$ is the variance of the error in $W_{i\cdot}$ and $\Sigma_{e_i} = \mathrm{var}(Q_{i\cdot}|X_i, Z_i) = ae^{bZ_i} + \Sigma_\xi/k$. The resulting conditional intensity

$$\lim_{dt\to 0} dt^{-1}\mathrm{P}\{dN_i(t) = 1|\zeta_i, Z_i, Y_i(t)\} = \lambda_0(t)\exp\left\{\beta_1\zeta_i - \frac{\beta_1^2\Sigma_{e_i}\Sigma_{\epsilon_i}/2}{\Sigma_{e_i} + \Sigma_{\epsilon_i}(\delta_1 + \delta_3 Z_i)^2} + \beta_2 Z_i\right\} Y_i(t),$$

does not depend on the unobserved X_i. Similarly, for a nonmember of the biomarker cohort, conditioning on the statistic $\zeta_i = \frac{\beta_1\Sigma_{e_i}}{(\delta_1+\delta_3 Z_i)^2} dN_i(t) + (\delta_1 + \delta_3 Z_i)^{-1}(Q_{i\cdot} - \delta_0 - \delta_2 Z_i)$ gives the conditional intensity

$$\lim_{dt\to 0} dt^{-1}\mathrm{P}\{dN_i(t) = 1|\zeta_i, Z_i, Y_i(t)\} = \lambda_0(t)\exp\left\{\beta_1\zeta_i - \frac{\beta_1^2\Sigma_{e_i}/2}{(\delta_1 + \delta_3 Z_i)^2} + \beta_2 Z_i\right\} Y_i(t).$$

As in the classical measurement error case, it can be shown ζ_i in both cases above is of the form $\zeta_i = \widetilde{W}_{i\cdot} + \beta_1\widetilde{\Sigma}_i dN_i(t)$, where $\widetilde{W}_{i\cdot}|X_i$ has mean X_i and variance $\widetilde{\Sigma}_i$. For the biomarker cohort, $\widetilde{W}_{i\cdot}$ is a weighted average of the self-report measure $Q_{i\cdot}$, recentered and rescaled so that it is unbiased for X_i at the true value of the nuisance parameters, and the biomarker $W_{i\cdot}$, where the weights are inversely proportional to the error variance in these two variables. For the nonmembers of the biomarker cohort, $\widetilde{W}_{i\cdot}$ is simply the recentered and rescaled $Q_{i\cdot}$. Now define $E_{0i}(t;\boldsymbol{\beta},\boldsymbol{\phi}) = \exp(\beta_1\zeta_i - \beta_1^2\widetilde{\Sigma}_i/2 + \beta_2 Z_i)Y_i(t)$, where $\boldsymbol{\phi}$ is the vector of error model nuisance parameters. Proceeding in a manner similar to Tsiatis and Davidian (2001), the estimating equation for β_1 is given by

$$\sum_{z\in\{Z\}}\sum_{i=1}^{n_z}\int_0^M \left\{\zeta_{zi} - \frac{\sum_{j=1}^{n_z}\zeta_{zj}E_{0j}(t,\boldsymbol{\beta},\widehat{\boldsymbol{\phi}})}{\sum_{j=1}^{n_z}E_{0j}(t,\boldsymbol{\beta},\widehat{\boldsymbol{\phi}})}\right\} dN_i(t) = 0, \tag{5}$$

where for the $Z = z$ stratum, n_z denotes the number of individuals and subscript zi denotes the ith member. This CS equation reduces to the ordinary stratified Cox partial likelihood score equation when there is no measurement error, i.e., when $\boldsymbol{\delta} = (0, 1, 0, 0)$ and $\Sigma_\xi = \Sigma_\epsilon = 0$. As discussed in Section 2.1, a plug-in estimate for the vector of the measurement error nuisance parameters can be estimated separately. Details of this derivation and regularity conditions for equation (5) are provided in the Supplementary Materials.

A second estimator $\widehat{\beta}_{cs}^{w}$ using a CS approach can be obtained by taking a weighted combination of two estimating functions: (1) the CS function for $\boldsymbol{\beta}$ of Tsiatis and Davidian (2001), which assumes classical measurement error, using only the biomarker data W_{ij}, and (2) the left-hand side of the proposed CS equation (5), using only the self-report data Q_{ij}. For the latter CS, the biomarker data are still needed to estimate the nuisance parameters. Subject to normality and regularity conditions, every weighted average of these

two CS estimating equations would be consistent for $\boldsymbol{\beta}$. One could choose the weight w that minimizes the variance of $\widehat{\boldsymbol{\beta}}_{cs}^{w}$. This strategy is evaluated in the simulation study that follows. Note, however, this approach may not be practical if the biomarker subsample includes few uncensored failure times.

2.4 *Nonparametric Corrected Score*

The idea behind the corrected score approach for consistent estimation with mismeasured covariates is to derive the necessary adjustment to the estimating equation with the error-prone covariate so that it has the same expected value as the desired estimating equation with the true covariate and outcome of interest. Nakamura (1992) and Buzas (1998) developed a parametric corrected score for Cox regression. Huang and Wang (2000, 2006) developed NP scores for Cox regression assuming classical measurement error. The estimator of Huang and Wang (2006) requires replicate mismeasured covariates only on a subset and is extended here to the error model in Section 2.1.

Define $\widetilde{X}_i$ to be the main instrument $Q_{i\cdot}$ recentered and rescaled by the nuisance parameters $\boldsymbol{\delta} = (\delta_0, \delta_1, \delta_2, \delta_3)$ from the error model in equation (1), i.e., $\widetilde{X}_i(\boldsymbol{\delta}) = (Q_{i\cdot} - \delta_0 - \delta_2 Z_i)/(\delta_1 + \delta_3 Z_i)$. At the true parameter value $\boldsymbol{\delta}^0 = (\delta_{00}, \delta_{10}, \delta_{20}, \delta_{30})$, the variable $\widetilde{X}_i$ is composed of X_i plus an error term. That is $\widetilde{X}_i(\boldsymbol{\delta}^0) = X_i + (\gamma_i + \xi_{i\cdot})/(\delta_{10} + \delta_{30} Z_i) = X_i + \nu_i$, where ν_i given $Z_i = z$ has zero mean and variance $(\Sigma_{\gamma_{i\cdot}} + \Sigma_{\xi_{i\cdot}})/(\delta_{10} + \delta_{30} z)^2$. With this transformed covariate $\widetilde{X}_i(\boldsymbol{\delta})$, one can adapt the corrected score approach of Huang and Wang (2006). For consistent estimation, the method of Huang and Wang (2000, 2006) requires there to be individuals with at least two "error-prone" measures observed, which are conditionally independent given X_i and whose errors are independent of X_i and the at-risk process. The distribution of the error in $\widetilde{X}_i$ depends on Z_i so if Z_i, either through correlation with X_i or independently, is associated with the hazard, then the error ν_i will be correlated with both X_i and $I\{Y_i(t) = 1\}$. Assuming discreteness and conditioning on the value of Z_i, ν_i is independent of X_i, the failure time, and ϵ_i. Thus by stratifying the partial likelihood score on Z, a technique similar to Huang and Wang (2006) can be applied to achieve consistency.

To derive the corrected score, first note at $\boldsymbol{\delta} = \boldsymbol{\delta}^0$ the solution to the following estimating equation based only on individuals in the biomarker subset:

$$\sum_{z\in\{Z\}}\sum_{i=1}^{n_z}\int_0^M R_{zi}\left[\widetilde{X}_{zi}(\boldsymbol{\delta}) - \frac{\sum_{j=1}^{n_z} Y_{zj}(t) W_{zj}\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}{\sum_{j=1}^{n_z} Y_{zj}(t)\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}\right] dN_{zi}(t) = 0,$$

is consistent for β_1; where for the $Z = z$ stratum, n_z denotes the number of individuals, and subscript zi denotes the ith member. This equation can be rewritten as

$$\sum_{z\in\{Z\}}\sum_{i=1}^{n_z}\int_0^M R_{zi} \times \left[\widetilde{X}_{zi}(\boldsymbol{\delta}) - \frac{\sum_{j=1}^{n_z} Y_{zj}(t)\widetilde{X}_{zj}\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}{\sum_{j=1}^{n_z} Y_{zj}(t)\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}} + \frac{\sum_{j=1}^{n_z} Y_{zj}(t)\{\widetilde{X}_{zj}(\boldsymbol{\delta}) - W_{zj}\}\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}{\sum_{j=1}^{n_z} Y_{zj}(t)\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}\right] dN_{zi}(t) = 0.$$

This suggests the following corrected score equation based on the entire cohort:

$$\sum_{z\in\{Z\}}\sum_{i=1}^{n_z}\int_0^M \left[\widetilde{X}_{zi}(\boldsymbol{\delta}) + \widehat{D}_z(\boldsymbol{\theta}, t) - \frac{\sum_{j=1}^{n_z} Y_{zj}(t)\widetilde{X}_{zj}(\boldsymbol{\delta})\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}{\sum_{j=1}^{n_z} Y_{zj}(t)\exp\{\beta_1\widetilde{X}_{zj}(\boldsymbol{\delta})\}}\right] dN_{zi}(t) = 0, \tag{6}$$

where

$$\widehat{D}_z(\boldsymbol{\theta}, t) = \frac{\sum_{i=1}^{n_z} Y_{zj}(t) R_{zi}\{\widetilde{X}_{zi}(\boldsymbol{\delta}) - W_{zi}\}\exp\{\beta_1\widetilde{X}_{zi}(\boldsymbol{\delta})\}}{\sum_{i=1}^{n_z} Y_{zj}(t) R_{zi}\exp\{\beta_1\widetilde{X}_{zi}(\boldsymbol{\delta})\}},$$

and $\boldsymbol{\theta} = (\beta_1, \boldsymbol{\delta})$. The estimate of $\mathrm{D}_z(\boldsymbol{\theta}, t)$ is a nonparametric moment estimator using data from individuals in the biomarker subcohort with $Z_i = z$. If the value of the nuisance parameter $\boldsymbol{\delta}$ is not known, a separate moment estimator can again be used as a plug-in. Notably, a subset of individuals with at least one measure of both W and Q at risk at time t is all that is necessary to estimate $\widehat{\mathrm{D}}_z(\boldsymbol{\theta}, t)$. The solution $\widehat{\beta}_{np}$ to equation (6) is referred to as the NP estimator.

As was done for the CS approach in Section 2.3, a second potentially more efficient nonparametric estimator $\widehat{\beta}_{np}^{w}$ can be obtained by taking a weighted average of the above score equation (6) and the nonparametric score equation for classical measurement error (Huang and Wang, 2000, 2006) based on the biomarker data alone. The weight w can be chosen to minimize the sample variance of $\widehat{\beta}_{np}^{w}$.

3. Estimation of Cumulative Baseline Hazard Function

For the assumed Cox model, the Breslow estimator for the cumulative baseline hazard is

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{dN(u)}{\sum_{i=1}^{n} Y_j(u)\exp(\widehat{\beta}_1 X_j + \widehat{\beta}_2 Z_j)} = \sum_{T_i \le t} \frac{\Delta_i}{\sum_{j \in \mathcal{R}_i} \exp(\widehat{\beta}_1 X_j + \widehat{\beta}_2 Z_j)}.$$

Huang and Wang (2000) provided a nonparametric consistent estimator for Λ_0 under classical measurement error using a representation of this estimator as a functional of empirical processes. This estimator, unlike their $\widehat{\boldsymbol{\beta}}$, requires additional assumptions of mean zero and symmetric error. Making these assumptions for ϵ only in (2), we extend their estimator to accommodate error model (1). We adopt their notation, using $\widehat{\mathcal{E}}$ to denote the sample empirical mean and $I(U = \min(T, C) \ge u)$ in place of $Y(u)$ to highlight the connection between their estimator and equation (7) below. For notational simplicity, assume two repeat measures of Q on everyone ($k_i = 2$). One approach to estimating Λ_0 involves stratifying on values of Z, so that approximately $\lambda_{z0}(t) = \lambda_0(t)\exp(\beta_2 Z)$, where Z denotes a representative Z-value in stratum z. A consistent estimator for $\Lambda_{z0}(t)$ for stratum $Z = z$ is

$$\begin{aligned} &\hat{\Lambda}_{z0}^{np}(t; \widehat{\beta}_1; \widehat{\boldsymbol{\delta}}) = \\ &(\widehat{\mathcal{E}}[I(Z = z)R\exp\{\widehat{\beta}_1(W^{(1)} - W^{(2)})/2\}])^{-1} \\ &\times \widehat{\mathcal{E}}(I(Z = z)R\exp[\widehat{\beta}_1\{\widetilde{X}(\widehat{\boldsymbol{\delta}}) - (W^{(1)} + W^{(2)})/2\}]) \\ &\times \int_0^t \frac{d\widehat{\mathcal{E}}\{I(Z = z)\Delta I(U \le u)\}}{\widehat{\mathcal{E}}[I(Z = z)\exp\{\widehat{\beta}_1\widetilde{X}(\widehat{\boldsymbol{\delta}})\}I(U \ge u)]}, \end{aligned} \quad (7)$$

where $\widehat{\beta}_1$ is the solution to (6). Stratification is useful, as in (6), because the error in $\widetilde{X}_{ij}$ depends on values of Z. For the RC and RRC estimators, this also leads to the convenient overall estimator of $\Lambda_0, \widehat{\Lambda}_0(t) = \sum_{z \in \{Z\}} \int_0^t \exp(-\widehat{\beta}_2 z) n_z(u) n(u)^{-1} \widehat{\Lambda}_{z0}^{np}(du; \widehat{\beta}_1, \widehat{\boldsymbol{\delta}})$, where $n_z(u)$ and $n(u)$ denote the risk set size in stratum $Z = z$ and the overall risk set size, respectively, at time u. Equation (7) relies only on $\widetilde{X}$ derived from the error model and a suitable estimate of $\boldsymbol{\beta}$, and it can be used in conjunction with any of the hazard ratio estimators described above. Details showing consistency are provided in the Supplementary Materials.

4. Simulation Study

Through simulation, the relative performance of the risk set regression calibration (RRC), CS, and NP are studied. Properties of these estimators will be compared to the true method, Cox regression on the unobserved true exposure X_i; the naive method, Cox regression on the error-prone $Q_{i\cdot}$; and RC. For the CS and NP methods, we examine the performance of the weighted estimator described in Sections 2.3 and 2.4 compared with the classical measurement error versions of these estimators based on the biomarker data alone. We compare performance for different scenarios that vary the magnitude of the relative risk parameter β, the random and systematic subject-specific measurement error nuisance parameters, and the assumed covariate and error distributions. We also consider versions of these estimators that ignore the dependence of subject-specific error variance on the observed covariate Z. Standard errors for the error-correction estimators are estimated using a bootstrap procedure.

For this simulation study, the cohort size is set at 1000 and the randomly selected biomarker subset at 250. Individuals have two copies of the main instrument Q_{ij}; biomarker cohort members have two copies of the biomarker W_{ij}. Let $\boldsymbol{\eta} = (\delta_0, \delta_1, \delta_2, \delta_3, a, b)$, from the measurement error model (1). The scenario $\boldsymbol{\eta} = (0, 0.9, -0.2, -0.3, 0.5, \log 2)$ represents a moderate amount of subject-specific error, with 10% scale bias for $Z_i = 0$ and 40% for $Z_i = 1$. The variance for γ_i, the subject-specific random effect term in (1), is allowed to vary between 0.5 ($Z_i = 0$) and 1 ($Z_i = 1$), whereas the variances of ϵ_{ij} and ξ_{ij} are fixed at 0.5. The scenario $\boldsymbol{\eta} = (0, 0.5, -0.2, -0.2, 0.5, \log 2)$ represents strong subject-specific bias, with 50% scale bias for $Z_i = 0$ and 70% for $Z_i = 1$; variances for error terms γ_i, ϵ_{ij} and ξ_{ij} are kept as before. Results are presented for $\beta = \log 1.5$, $\log 3$. Note that $\log 3$ is quite extreme, with a hazard ratio of 3 for a unit increase in a standard normal exposure variable. It is included here because the regression calibration estimator is known to perform less well in Cox regression for large β (Xie et al., 2001). Survival data are generated with an exponential distribution with unit rate and a fixed censoring time of $t_0 = 1$, resulting in roughly 40% censoring. Much larger censoring rates, with associated smaller biases for regression calibration procedures, attend the type of application that motivated this research.

Mean bias, bootstrap standard deviation (BSD), empirical standard deviation (across simulations), root mean squared error (RMSE), and empirical coverage probability for the bootstrap 95% confidence intervals (CIs) are provided. Bootstrap estimates are based on 100 bootstrap samples and the empirical results are based on 1000 simulations.

Table 1 presents the results under normal covariate and error distributions. X_i and Z_i are generated from a bivariate normal with zero mean, unit variance, and $\rho = 0.5$.Z_i is converted to a binary indicator variate for being above the median. The upper left of Table 1 shows the results for $\beta = \log 1.5$ and moderate systematic error. The naive estimate has a bias of -0.217 (54% reduction from target) and a smaller standard error than for the estimate based on the true exposure, leading to 0% coverage for a nominal 95% CI. For this scenario, with moderate error and $\beta = \log 1.5$, all of the measurement error corrected methods had small biases and came close to the nominal 95% coverage. The nonparametric estimators had the largest sample standard error. For the strong systematic error scenario and $\beta = \log 1.5$ (top right), results were similar. The bias for the RRC is somewhat lower than for RC, but the RC estimator had the smallest mean-squared error of all the methods.

The lower half of Table 1 shows the results for the same error settings, with $\beta = \log 3$. As expected, bias of the regression calibration estimators increased, but RRC had less bias and better RMSE than RC. For the larger β, CS had the smallest bias, good nominal coverage, and nearly the same RMSE

Table 1

Simulation results for the general measurement error model with Gaussian subject-specific and random error. For 1000 simulated data sets, the mean bias, empirical standard deviation (SD), bootstrap standard deviation (BSD), RMSE, and estimated 95% coverage probability (CP) are given for $\beta = \log 1.5, \log 3$.

	Moderate subject-specific bias					Strong subject-specific bias				
$\beta = \log 1.5$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.000	0.041	0.041	0.041	96.5	0.000	0.041	0.041	0.041	96.5
Naive	−0.217	0.032	0.032	0.219	0.0	−0.265	0.037	0.037	0.268	0.0
RC	−0.014	0.063	0.062	0.065	92.9	−0.016	0.069	0.069	0.071	93.3
RRC	−0.012	0.068	0.073	0.069	95.4	−0.013	0.073	0.073	0.074	93.4
CS B	0.010	0.099	0.102	0.099	95.8	0.010	0.099	0.102	0.099	95.8
CS W	0.011	0.091	0.096	0.092	96.5	0.010	0.102	0.105	0.102	95.6
NP B	0.014	0.123	0.128	0.124	96.9	0.014	0.123	0.128	0.124	96.9
NP W	0.012	0.119	0.123	0.120	96.7	0.007	0.122	0.126	0.122	95.7
$\beta = \log 3$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.001	0.050	0.051	0.050	95.4	0.001	0.050	0.051	0.050	95.4
Naive	−0.678	0.035	0.034	0.679	0.0	−0.806	0.039	0.038	0.806	0.0
RC	−0.190	0.095	0.094	0.213	46.2	−0.219	0.098	0.097	0.240	38.5
RRC	−0.120	0.107	0.119	0.161	79.7	−0.146	0.107	0.108	0.181	69.9
CS B	0.035	0.182	0.198	0.186	97.4	0.035	0.182	0.198	0.186	97.4
CS W	0.023	0.171	0.183	0.172	96.5	0.025	0.180	0.195	0.181	97.6
NP B	0.070	0.255	0.289	0.265	96.6	0.070	0.255	0.289	0.265	96.6
NP W	0.056	0.246	0.275	0.252	96.3	0.041	0.262	0.276	0.265	95.2

True: Cox regression with true X; Naive: Cox regression with unadjusted Q; RC: ordinary regression calibration; RRC: risk set regression calibration; CS B: CS from Tsiatis and Davidian (2001) using biomarker W data only; CS W: weighted combination of CS using Q only and W only; NP B: NP equation from Huang and Wang (2000) using only W; NP W: weighted combination of NP using Q only and W only.

as RRC. The RRC estimator had the lowest mean-squared error for the large β for both subject-specific error scenarios, although it had appreciable bias and poorer coverage for the larger value of β than either the CS or NP estimator.

The CS and NP estimators had variances that were appreciably larger than the other estimators. It is noteworthy for this setting, with relatively large error in Q_{ij} compared to W_{ij} and a substantial number of events observed in the biomarker subset, the NP and CS estimators based on the complete data had modest to no gain in efficiency over their counterparts, CS B and NP B, based on the biomarker subset alone. Sugar et al. (2007) also observed the CS estimator in the context of logistic regression with the same general measurement error model was highly variable, particularly in presence of strong subject-specific bias. For classical measurement error, the nonparametric corrected estimator has been observed to have numerical instability and problems due to multiple roots (Song and Huang, 2005, Carroll et al., 2006), with these problems getting worse as the measurement error variance increases (Song and Huang, 2005). In the case of multiple roots, the root closest to the RC estimate was selected. These simulations suggest that for a similar setting of moderate sample sizes, potentially large and normal error, the RRC and CS methods perform better overall, with CS preferred for very extreme β values.

To explore robustness, a similar set of simulations were repeated with skewed distributions. The systematic and random error terms in W_{ij} and Q_{ij} were generated from a unit exponential distribution, reflected about zero to create left skewness and offset by its mean to create mean-zero errors. The same bivariate normal distribution was used to create X and Z as above, and then both were exponentiated to create skewed log-normal random variables. Results are shown in Table 2. As expected, the regression calibration estimators, which rely on approximate normality, have more bias particularly for the larger β. For the extreme β, however, the RRC estimator was much less affected. The CS methods, which rely on Gaussian error for consistency, had noticeably larger bias and relatively larger variance. The weighted CS estimator did not improve on the CS estimator based on the biomarker alone, likely due to the larger amount of error and skewness in Q. The displayed CS W estimator had a weight of 0.1 for the score using information from Q, having the smallest variance among the (nontrivial) decile weights. The skewness had a larger impact on the relative performance of the RC, RRC, and CS estimators (in terms of bias and variance) for the larger β. The performance of the NP estimator, as expected, was unaffected by the skewness in the distributions, with little small sample bias, good nominal coverage, and nearly or the smallest RMSE for all scenarios. The RRC estimator, even with large β and strong systematic error, had reasonable coverage and bias less than 15%.

Table 3 compares the proposed estimators to those incorrectly based on the error model (1) without dependence of $\text{var}(\gamma_i)$ on Z_i. Scenarios are the same as in Table 1, except the impact of Z in $\text{var}(\gamma_i)$ was increased with $b = \log 4$. It is interesting to note that the misspecified RC and RRC estimators have similar bias but increased standard errors compared to their correctly specified versions. The misspecified CS estimator contains increased bias and standard errors, with similar coverage. The NP estimator, because it uses rescaling by Z_i,

Table 2

Simulation results for the general measurement error model with skewed distributions for the model covariates as well as the subject-specific and random error. For 1000 simulated data sets, the mean bias, empirical standard deviation (SD), bootstrap standard deviation (BSD), RMSE, and estimated 95% coverage probability (CP) are given for $\beta = \log 1.5, \log 3$.

	Moderate subject-specific bias					Strong subject-specific bias				
$\beta = \log 1.5$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.002	0.020	0.020	0.020	94.6	0.002	0.020	0.020	0.020	94.6
Naive	−0.082	0.022	0.022	0.085	3.8	−0.072	0.042	0.034	0.083	45.4
RC	−0.025	0.099	0.079	0.102	86.9	−0.045	0.089	0.069	0.099	88.7
RRC	0.016	0.096	0.121	0.097	93.7	0.044	0.076	0.071	0.088	82.1
CS B	0.012	0.048	0.050	0.050	94.9	0.012	0.048	0.050	0.050	94.9
CS W	0.028	0.061	0.060	0.067	93.5	0.054	0.075	0.076	0.092	88.4
NP B	0.011	0.057	0.061	0.058	96.2	0.011	0.057	0.061	0.058	96.2
NP W	0.008	0.055	0.061	0.056	96.6	0.009	0.056	0.060	0.057	96.4
$\beta = \log 3$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.003	0.038	0.037	0.038	94.5	0.003	0.038	0.037	0.038	94.5
Naive	−0.491	0.032	0.029	0.492	0.0	−0.590	0.051	0.038	0.592	0.0
RC	−0.343	0.193	0.152	0.394	44.4	−0.458	0.157	0.122	0.485	0.7
RRC	−0.077	0.197	0.201	0.211	94.3	−0.116	0.167	0.161	0.203	90.9
CS B	0.136	0.160	0.184	0.210	99.0	0.136	0.160	0.184	0.210	99.0
CS W	0.138	0.226	0.225	0.265	94.9	0.112	0.231	0.249	0.257	98.0
NP B	0.068	0.210	0.243	0.221	96.6	0.068	0.210	0.242	0.221	96.6
NP W	0.059	0.195	0.234	0.204	97.3	0.064	0.209	0.237	0.219	96.6

True: Cox regression with true X; Naive: Cox regression with unadjusted Q; RC: ordinary regression calibration; RRC: risk set regression calibration; CS B: CS from Tsiatis and Davidian (2001) using biomarker W data only; CS W: weighted combination of CS using Q only and W only; NP B: NP equation from Huang and Wang (2000) using only W; NP W: weighted combination of NP using Q only and W only.

Table 3

Simulation study comparing the proposed estimators with misspecified versions of the measurement error variance, under the general measurement error model with Gaussian distributions for the model covariates as well as for the subject-specific and random error. For 1000 simulated data sets, the mean bias, empirical standard deviation (SD), bootstrap standard deviation (BSD), RMSE, and estimated 95% coverage probability (CP) are given for $\beta = \log 1.5, \log 3$.

	Moderate subject-specific bias					Strong subject-specific bias				
$\beta = \log 1.5$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.000	0.041	0.041	0.041	96.5	0.000	0.041	0.041	0.041	96.5
Naive	−0.262	0.028	0.028	0.263	0.0	−0.308	0.031	0.031	0.309	0.0
RC	−0.014	0.064	0.063	0.066	93.1	−0.016	0.069	0.070	0.071	93.2
RC-M	−0.008	0.071	0.071	0.071	95.0	−0.011	0.080	0.082	0.081	95.6
RRC	−0.014	0.069	0.074	0.070	94.7	−0.014	0.073	0.073	0.074	93.1
RRC-M	0.007	0.077	0.077	0.077	95.9	−0.004	0.083	0.084	0.083	96.2
CS W	0.014	0.098	0.102	0.099	96.0	0.008	0.105	0.107	0.105	95.5
CS W-M	−0.040	0.101	0.111	0.109	94.3	−0.020	0.153	0.146	0.155	93.9
NP W	0.008	0.119	0.125	0.120	96.6	0.000	0.123	0.127	0.123	95.5
$\beta = \log 3$	Bias	SD	BSD	RMSE	CP	Bias	SD	BSD	RMSE	CP
True	0.001	0.050	0.051	0.050	95.4	0.001	0.050	0.051	0.050	95.4
Naive	−0.790	0.031	0.029	0.791	0.0	−0.901	0.032	0.031	0.901	0.0
RC	−0.193	0.096	0.095	0.216	45.9	−0.220	0.099	0.097	0.241	38.0
RC-M	−0.196	0.102	0.101	0.221	47.6	−0.210	0.105	0.105	0.235	46.4
RRC	−0.131	0.106	0.118	0.169	77.8	−0.150	0.107	0.108	0.184	68.2
RRC-M	−0.088	0.110	0.109	0.141	83.8	−0.157	0.105	0.106	0.189	63.2
CS W	0.020	0.177	0.189	0.178	96.8	0.026	0.178	0.195	0.180	97.8
CS W-M	0.061	0.253	0.249	0.261	95.6	0.067	0.191	0.219	0.202	97.6
NP W	0.044	0.246	0.277	0.250	96.1	0.023	0.259	0.278	0.260	94.7

True: Cox regression with true X; Naive: Cox regression with unadjusted Q; RC: ordinary regression calibration; RRC: risk set regression calibration; CS W: weighted combination of CS using Q only and W only; NP W: weighted combination of the nonparametric estimator using Q only and W only. Misspecified error-correction estimators are noted by a −M.

induces and adjusts for error variance that depends on Z; so misspecification is not possible.

5. Women's Health Initiative (WHI) Example

The WHI Dietary Modification (DM) Trial followed 48,835 women for an average of 8.1 years and examined whether a low-fat dietary pattern intervention could lower the risk of breast and colorectal cancer (Women's Health Initiative Study Group, 1998). Prentice et al. (2006) reported a non-significant reduction in breast cancer of 9% (logrank p=0.07) for the intervention compared to the control (usual diet) arm. There was no suggested reduction for colorectal cancer (Beresford et al., 2006). An important question is whether the equivocal breast cancer finding is from a lack of efficacy or a lack of adherence to the diet, but actual diet is not obtainable. Instead, the primary tool for measuring diet was a self-reported FFQ, an instrument known to be subject to both random and subject-specific reporting errors.

The DM trial included a Nutritional Biomarker Substudy (NBS) that collected self-reported intake along with several objective biomarkers on 544 weight-stable women randomly selected at a representative set of 12 of the 40 participating clinical centers. The NBS protocol included the doubly labeled water recovery marker for total energy consumption (Schoeller, 1988). There were 110 women recruited from early NBS enrollees who had repeat biomarker measures, allowing the general measurement error model to be identifiable. Further details of the NBS study and an analysis of the measurement error in the WHI dietary instruments were reported by Neuhouser et al. (2008), who found BMI to a strong determinant of subject-specific bias in this cohort. In this illustrative example, we fit the measurement error in equation (1) with possible dependence on obesity status (BMI $\geq$ 30) and apply the developed methods to provide error-adjusted estimates of the risk of breast cancer associated with total energy consumption. Because the baseline FFQ was used to determine eligibility in the DM trial by requiring a minimum of 32% estimated calories from fat, the baseline for this analysis was taken as 1 year after enrollment, at which time another FFQ was obtained. We analyze data from the usual diet (control) arm.

5.1 *Results*

There were 25,803 women in the DM control group included in this analysis, 884 of whom developed breast cancer following the 1 year FFQ collection. The estimate (95% CI) for the breast cancer hazard ratio associated with a 20% increase in energy intake was: 1.00 (0.97, 1.04) for the naive estimate, 1.24 (1.03, 1.48) for RC, 1.23 (1.03, 1.48) for RRC, 1.30 (0.80, 2.10) for CS and 1.43 (0.95, 2.15) for NP. Note the hazard ratio for a fractional increase in intake is constant under the Cox model applied, because the log-hazard ratio was assumed to be a linear function of log consumption. A 20% increase is roughly the difference between the third and first quartile of energy consumption, as measured by the recovery biomarker (2268 versus 1869 calories). The RC and RRC estimates are nearly identical, because for this relatively rare disease with most censoring at the planned study termination, the distribution for $E(X|W, Q, Z)$ changes very little across risk sets. The NP hazard ratio estimate is slightly larger than the regression calibration estimates, and its 95% CI is much wider. The CS estimate is the most variable of the error-corrected estimates and had some numerical problems, with skewness in the bootstrap estimates and more than 4% of the bootstrap iterations failing to find a root to the score equation.

To help interpret the above estimates, a simulation study was built on the observed WHI data. The simulation cohort had 25,000 individuals, with 540 in the biomarker sample and 110 in its reliability subset (both randomly selected). The self-reported and biomarker values for log energy were well approximated by Gaussian distributions in the NBS (Figure 1, Supplementary Materials). BMI and X were generated as multivariate normal variates on the log scale. Roughly 25% of the simulated cohort were obese. In the fitted error model (1), δ_3 was estimated to be nearly zero, so the model was applied without this term. The variance of ϵ was 30% of the total variance in W. The variance of the subject-specific bias plus random error terms was extreme, about 95% the total variance of Q. Survival time was generated according to a proportional hazards model dependent on log energy consumption with baseline survival an exponential distribution with overall event rate of about 3%.

Table 4 shows results for $\beta = 0$ and 1.25; i.e., the β for which a 20% increase in consumption leads to a hazard ratio of 1 and 1.26, respectively. None of the estimators showed evidence of bias under the null. The naive estimator under $\beta = 1.25$ had extreme bias, nearly 95% of the true value, and small standard error, leading to 0% empirical coverage. The RC and RRC estimators performed well for $\beta = 1.25$, having no detectable bias. There were some numerical problems for the CS and NP estimators, with more than 15% of the simulations failing to find a solution. These estimators were quite variable and had some small sample bias, although the coverage probabilities were close to the nominal 95%. Using the empirical mean and SD across simulations, the estimated hazard ratio (empirical 95% CI) for a 20% increase in intake is: 1.01 (0.98, 1.05) for the naive method, 1.25 (0.98, 1.60) for RC, 1.25 (0.98, 1.60) for RRC, 1.18 (0.69, 2.03) for CS, 1.23 (0.73, 2.08) for NP. The CIs for the regression calibration estimators are considerably narrower than for the other two error-corrected estimators. The CS estimator had the largest variance. This simulation study suggests for this example, where there was a sizable biomarker subset, that each error correction method is providing approximately unbiased estimates of the risk associated with energy intake, with RC or RRC preferred due to their comparatively better efficiency. Another advantage of regression calibration is that standard software can be used to find the parameter estimates, and replicates of neither the biomarker nor self-report are required.

6. Discussion

In this work, three methods for hazard ratio estimation were extended for the setting where the exposure of interest was measured with subject-specific bias and random error in the main cohort and additionally with classical measurement error on a subset. We also provided an estimator of the cumulative baseline hazard function. This error model is more flexible than others considered previously for these methods and does not rely on a validation subset. The extensions provided here allow for a more flexible error structure and also

Table 4

Results of simulations designed to emulate hazard ratio estimation for energy in relation to breast cancer in the WHI, using 500 simulated data sets. The mean bias, empirical standard deviation (SD), RMSE, type I error (α) for $\beta = 0$, and estimated 95% coverage probability (CP) for $\beta = 1.25$ are given.

$\beta = 0$	Bias	SD	RMSE	α	$\beta = 1.25$	Bias	SD	RMSE	CP
True	0.004	0.296	0.296	0.046	True	0.006	0.302	0.302	95.4
Naive	−0.008	0.105	0.105	0.056	Naive	−1.181	0.106	1.189	0.0
RC	−0.005	0.709	0.709	0.050	RC	−0.006	0.691	0.691	96.2
RRC	−0.005	0.709	0.709	0.050	RRC	−0.006	0.690	0.691	96.2
CS	−0.052	1.507	1.507	0.052	CS	−0.318	1.510	1.543	93.1
NP	0.014	1.528	1.528	0.042	NP	−0.217	1.471	1.487	95.0

True: Cox regression with true X; Naive: Cox regression with unadjusted Q; RC: ordinary regression calibration; RRC: risk set regression calibration; CS: conditional score; NP: nonparametric corrected score.

accommodate sources of intake assessment errors that vary across individuals.

The relocation and scaling parameters (the δ's in (1)) are crucial measurement model generalizations; the allowance for correlation between replicate error-prone measurements (Q) is also important for some of the estimation procedures (conditional and nonparametric scores), while allowing the random effect variance to depend on Z may often be less important (Table 3). The RRC estimator is a straightforward adaption of its counterpart for classical measurement error. The conditional and nonparametric scores required more detailed calculations and rely on a subtle, but necessary, stratification of the Cox model for proper error correction. RRC is an approximate method that typically incorporates some asymptotic bias. For consistency, the proposed CS estimator relies on a normality assumption for the error terms, but does not need a distributional assumption for unobserved true covariate. The nonparametric score method made no distributional assumptions for the error terms or the unobserved covariate.

Despite its lack of technical consistency, RRC had the smallest mean-squared error in nearly all simulation scenarios considered, often by a considerable margin. Bias for both regression calibration estimators was more noticeable for extreme β, particularly when covariates and error terms had skewed distributions; however, risk set regression reduced the bias considerably. Under Gaussian error, the CS method had little small sample bias, maintained nominal 95% coverage, and with large β and stronger subject-specific error, had close to the smallest mean-squared error amongst the estimators. The CS estimator, however, was not robust to departures from normality, with bias in the presence of skewness increasing for larger β. The nonparametric method was robust to departures from normality with little small sample bias and good nominal coverage, and this was true for the more extreme β and subject-specific bias. Typically, however, the nonparametric estimates had substantially larger variance than regression calibration.

The mean–variance tradeoff between robust nonparametric or semiparametric methods and efficiency of parametric approaches is a familiar one for estimation. In light of the relative success of regression calibration, it is of interest to consider other approximate methods for this setting. Hu et al. (1998) presented a semiparametric likelihood approach for Cox regression with classical covariate measurement error that uses flexible distributional assumptions on the unobserved covariate and in some settings performed better than regression calibration. Alternatively, increasing the number of moments used for regression calibration and transforming data to improve the normal approximation could also be useful, and computationally less burdensome than the available likelihood approaches. As illustrated above, simulation studies can guide the practitioner's choice of an appropriate method.

In the WHI example, the association under study appeared moderate in size, the disease relatively rare, the measurement error substantial, and the number of disease events fairly large. The simulations based on this example suggest these methods provide quite adequate estimates of the hazard ratio. These simulations also demonstrated that for the naive analysis, the error in the self-reported data was large enough to cause extreme bias and obscure a clinically relevant association between caloric intake and disease incidence. This finding is consistent with Prentice et al. (2009) who reported for this cohort that several cancer endpoints had no association with self-reported energy intake, whereas with calibrated energy intake there were associations of public health importance. Variation in the self-report intake variable was largely due to measurement error in this setting. This likely contributed to the numerical problems seen for conditional and nonparametric score estimators.

Because of the large number of nuisance parameters in the measurement error model, it is difficult to do an exhaustive exploration of the relative performance of these methods. For settings with error properties much different than those studied here, further study may be needed to understand which method is most appropriate. Some interpretational challenges arise with any of the estimation procedures considered here if one or more covariates (Z_i) determining the subject-specific bias in (1) are also important mediators of the association between X_i and the hazard ratio. See Prentice and Huang (2011) for further discussion of this issue, and for data analysis options. The censoring mechanism could also impact the relative performance of the methods. Regression calibration methods work best when censoring times tend to be short, as longer follow-up will lead to X distributions that depart more extensively from baseline in risk sets at later failure times. CS and nonparametric estimators on the other hand, which do not rely on distributional assumptions for the unobserved covariate X, are unlikely to be much affected by variations

in the censoring distributions. The development of estimation procedures that can accommodate departures from independent censoring, for example through inverse censoring probability weighting, would also be of interest in the context of our measurement model.

The bootstrap variance estimator was studied for the proposed estimators. Robust variance estimators have been developed in the case of classical measurement error for each of the methods studied (Wang, 1999; Tsiatis and Davidian, 2001; Xie et al., 2001; Huang and Wang, 2006) and a similar approach could be taken for the proposed methods. A general discussion of sandwich estimators in the context of measurement error is provided by Carroll et al. (2006, Appendix A.6). For ease of implementation, the bootstrap estimator is preferred.

Across all scenarios studied, the nonparametric score method typically maintained the smallest bias and best nominal coverage. The risk set regression estimator had good relative performance, in terms of maintaining the smallest mean-squared error. The numerical stability and ease of implementation in standard software makes regression calibration attractive for settings with substantial measurement error. In settings where a very large hazard ratio is expected, along with potentially skewed distributions for the involved covariates, the nonparametric score estimator may be preferred. Simulation can be used to explore properties of the estimators for the data structure in a given application, and in particular, explore which method has the best numerical performance, given the observed error structure in the data and the expected size of the hazard ratio for the exposure of interest.

7. Supplementary Materials

The Supplementary Materials referenced in Sections 2, 3, and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This work was partially supported by the National Institute of Allergy and Infectious Diseases, grants CA53996 and CA119171 from the National Cancer Institute, and by contract N01-WH22110 from the National Heart, Lung and Blood Institute. The authors thank the WHI investigators for access to the data used to illustrate the methods presented here. A list of WHI investigators can be found at `www.whiscience.org`. The WHI program is funded via contract by the National Heart, Lung and Blood Institute, National Institutes of Health.

References

Beresford, S. A. A., Johnson, K. C., Ritenbaugh, C., Lasser, N. L., Snetselaar, L., Black, H. R., Anderson, G. L., Assaf, A. R., Bassford, T., Bowen, D., Brunner, R., Brzyski, R., Caan, B., Chlebowski, R., Gass, M., Harrigan, R. C., Hays, J., Heber, D., Heiss, G., Hendrix, S., Howard, B., Hsia, J., Hubbell, F., Jackson, R., Kotchen, J., Kuller, L., LaCroix, A., Lane, D., Langer, R., Lewis, C., Manson, J., Margolis, K., Mossavar-Rahmani, H., Ockene, J., Parker, L., Perri, M., Phillips, L., Prentice, R., Robbins, J., Rossouw, J., Sarto, G., Stefanick, M., Van Horn, L., Vitolins, M., Wactawski-Wende, J., Wallace, R., and Whitlock, E. (2006). Low-fat dietary pattern and risk of colorectal cancer: The women's health initiative randomized controlled dietary modification trial. *Journal of the American Medical Association* **295,** 643–654.

Bingham, S. A. and Cummings, J. H. (1985). Urine nitrogen as an independent validatory measure of dietary intake: A study of nitrogen balance in individuals consuming their normal diet. *American Journal of Clinical Nutrition* **42,** 1276–1289.

Bingham, S. A., Luben, R., Welch, A., Wareham, N., Khaw, K. T., and Day, N. (2003). Are imprecise methods obscuring a relation between fat and breast cancer? *Lancet* **362,** 212–214.

Buzas, J. S. (1998). Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference* **67,** 247–257.

Carroll, R. J., Freedman, L. S., Kipnis, V., and Li, L. (1998). A new class of measurement error models, with applications to dietary data. *Canadian Journal of Statistics* **26,** 467–477.

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.

Dahm, C. C., Keogh, R. H., Spencer, E. A., Greenwood, D. C., Key, T. J., Fentiman, I. S., Shipley, M. J., Brunner, E. J., Cade, J. E., Burley, V. J., Mishra, G., Stephen, A. M., Kuh, D., White, I. R., Luben, R., Lentjes, M. A. H., Khaw, K. T., and Rodwell (Bingham), S. A. (2010). Dietary fiber and colorectal cancer: A nested case-control study using food diaries. *Journal of the National Cancer Institute* **102,** 614–626.

Freedman, L. S., Potischman, N., Kipnis, V., Midthune, D., Schatzkin, A., Thompson, F. E., Troiano, R. P., Prentice, R., Patterson, R., Carroll, R., and Subar, A. F. (2006). A comparison of two dietary instruments for evaluating the fat-breast cancer relationship. *International journal of Epidemiology* **35,** 1011–1021.

Gorfine, M., Hsu, L., and Prentice, R. (2004). Nonparametric correction for covariate measurement error in a stratified Cox model. *International journal of Epidemiology* **5,** 75–87.

Heitman, B. and Lissner, L. (1995). Dietary underreporting by obese individuals—is it specific or non-specific. *British Medical Journal* **311,** 986–989.

Hu, C. and Lin, D. Y. (2002). Cox regression with covariate measurement error. *Scandinavian Journal of Statistics* **29,** 637–655.

Hu, C. and Lin, D. Y. (2004). Semi-parametric failure time regression with replicates of mismeasured covariates. *Journal of the American Statistical Association* **99,** 105–117.

Hu, P., Tsiatis, A., and Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* **54,** 1407–1419.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric correction approach. *Journal of the American Statistical Association* **95,** 1209–1219.

Huang, Y. and Wang, C. Y. (2006). Errors-in-covariates effect on estimating functions: Additivity in limit and nonparametric correction. *Statistica Sinica* **16,** 861–881.

Jiang, W., Kipnis, V., Midthune, D., and Carroll, R. (2001). Parameterization and inference for nonparametric regression problems. *Journal of the Royal Statistical Society, Series B* **63,** 583–591.

Kaaks, R. (1997). Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: Conceptual issues. *American Journal of Clinical Nutrition* **65,** 1232S–1239S.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A., and Carroll, R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology* **153,** 394–403.

Liao, L., Zucker, D. M., Li, Y., and Spiegelman, D. (2011). Survival analysis with error-prone time-varying covariates: A risk set calibration approach. *Biometrics* **67,** 50–58.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics* **48,** 829–838.

Neuhouser, M. L., Tinker, L., Shaw, P. A., Schoeller, D., Bingham, S. A., Van Horn, L., Beresford, S. A. A., Caan, B., Thompson, C., Satterfield, S., Kuller, L., Heiss, G., Smit, E., Sarto, G., Ockene, J., Stefanick, M. L., Assaf, A., Runswick, S., and Prentice, R. L. (2008). Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. *American Journal of Epidemiology* **167,** 1247–1259.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342.

Prentice, R. L. (1996). Measurement error and results from analytic epidemiology: Dietary fat and breast cancer. *Journal of the National Cancer Institute* **88,** 1738–1747.

Prentice, R. L. and Huang, Y. (2011). Measurement error modeling and nutritional epidemiology association analyses. *Canadian Journal of Statistics*, Early online access Epub: 2011 Jul 27, DOI: 10.1002/cjs.10116.

Prentice, R. L., Sugar, E., Wang, C. Y., Neuhouser, M., and Patterson, R. (2002). Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. *Public Health Nutrition* **5,** 977–984.

Prentice, R. L., Caan, B., Chlebowski, R. T., Patterson, R., Kuller, L. H., Ockene, J. K., Margolis, K. L., Limacher, M. C., Manson, J. E., Parker, L. M., Paskett, E., Phillips, L., Robbins, J., Rossouw, J. E., Sarto, G. E., Shikany, J. M., Stefanick, M. L., Thomson, C. A., Van Horn, L., Vitolins, M. Z., Wactawski-Wende, J., Wallace, R. B., Wassertheil-Smoller, S., Whitlock, E., Yano, K., Adams-Campbell, L., Anderson, G. L., Assaf, A. R., Beresford, S. A., Black, H. R., Brunner, R. L., Brzyski, R. G., Ford, L., Gass, M., Hays, J., Heber, D., Heiss, G., Hendrix, S. L., Hsia, J., Hubbell, F. A., Jackson, R. D., Johnson, J. M., Kotchen, K. C., LaCroix, A. Z., Lane, D. S., Langer, R. D., Lasser, N. L., and Henderson, M. M. (2006). Low-fat dietary pattern and risk of invasive breast cancer: The women's health initiative randomized controlled dietary modification trial. *Journal of the American Medical Association* **295,** 629–642.

Prentice, R. L., Shaw, P. A., Bingham, S., Beresford, S. A. A., Caan, B., Neuhouser, M. L., Patterson, R. E., Stefanick, M. L., Satterfield, S., Thomson, C. A., Snetselaar, L., Thomas, A., and Tinker, L. F. (2009). Biomarker-calibrated energy and protein consumption and increased risk among postmenopausal women. *American Journal of Epidemiology* **169,** 977–989.

Prentice, R. L., Huang, Y., Kuller, L., Tinker, L., Van Horn, L., Stefanick, M., Sarto, G., Ockene, J., and Johnson, K. (2011). Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. *Epidemiology* **22,** 170–179.

Schoeller, D. A. (1988). Measurement of energy expenditure in free-living humans by using doubly labeled water. *Journal of Nutrition* **118,** 1278–1289.

Song, X. and Huang, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics* **61,** 702–714.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for general linear measurement-error models. *Biometrika* **74,** 703–716.

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D., Bingham, S., Sharbaugh, C., Trabulsi, J., Runswick, S., Ballard-Barbash, R., Sunshine, J., and Schatzkin, A. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *American Journal of Epidemiology* **158,** 1–13.

Sugar, E. A., Wang, C. Y., and Prentice, R. L. (2007). Logistic regression with exposure biomarkers and flexible measurement error. *Biometrics* **63,** 143–151.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88,** 447–458.

Wang, C. Y. (1999). Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error. *Statistics & Probability Letters* **45,** 371–378.

Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics* **53,** 131–145.

Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials* **19,** 61–109.

World Cancer Research Fund/American Institute for Cancer Research. (2007). *Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective*. Washington, DC: American Institute for Cancer Research.

Xie, S. X., Wang, C. Y., and 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.

Received October 2010. *Revised September* 2011.
Accepted September 2011.