Published in final edited form as:

Stat Med. 2016 May 10; 35(10): 1676–1688. doi:10.1002/sim.6812.

Methods to Adjust for Misclassification in the Quantiles for the Generalized Linear Model with Measurement Error in Continuous Exposures

Ching-Yun Wang¹, Jean De Dieu Tapsoba¹, Catherine Duggan¹, Kristin L Campbell², and Anne McTiernan¹

¹Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109-1024, U.S.A

²Department of Physical Therapy, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

SUMMARY

In many biomedical studies, covariates of interest may be measured with errors. However, frequently in a regression analysis the quantiles of the exposure variable are often used as the covariates in the regression analysis. Because of measurement errors in the continuous exposure variable, there could be misclassification in the quantiles for the exposure variable. Misclassification in the quantiles could lead to bias estimation in the association between the exposure variable and the outcome variable. Adjustment for misclassification will be challenging when the gold standard variables are not available. In this paper, we develop two regression calibration estimators to reduce bias in effect estimation. The first estimator is normal likelihood-based. The second estimator is linearization-based, and it provides a simple and practical correction. Finite sample performance is examined via a simulation study. We apply the methods to a 4-arm randomized clinical trial that tested exercise and weight loss interventions in women aged 50–75 years.

Keywords

Measurement error; Misclassification; Regression calibration

1 Introduction

Covariate measurement error occurs in many research studies when the covariate variable is continuous. Covariate measurement error can result in misclassification if the exposure variable is analyzed as quantiles. One motivating example of our methodology research is covariate measurement error associated with measurement of aerobic fitness in the Nutrition and Exercise in Women (NEW) study, conducted from 2005 to 2009. The NEW study is a 12-month randomized, controlled trial using a 4-arm design to compare the effect of three

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

lifestyle-change interventions (dietary weight loss, moderate-to-vigorous intensity aerobic exercise, or both interventions combined) versus control (no lifestyle change) [1][2]. Ancillary studies have evaluated the effects of physical activity on disease-related biomarkers, such as C-reactive protein (CRP); see Imayama, et al. [3]. VO_2max is a measure of aerobic fitness or exercise capacity through determination of the maximum rate of oxygen consumption as measured during incremental exercise, typically on a motorized treadmill. Inflammation markers have been shown to be inversely associated with VO_2max (Kullo *et al.* [4]). In regression analysis, the quantiles of VO_2max are often used as a covariate. Because of biological variation and factors related to measurement of VO_2max (e.g., individual's effort, variations in testing protocols), VO_2max may be associated with errors, and hence the quantiles for VO_2max may be misclassified. Misclassification of exposure can cause bias; see Gustafson [5] for a review.

When there exists a subset that contains all covariates, measured without error, then methods for missing covariates generally can be applied to the misclassification in covariate problem. This subset is often called the complete-case set, or the validation set; see for example Qi *et al.* [6] and Natarajan [7]. Many important methods for covariate measurement errors have been well reviewed by Carroll *et al.* [8] and Buonacorsi [9]. However, overall there are considerably fewer methods for correcting for covariate misclassification when a validation set does not exist. A significant challenge in this situation is that the misclassification probabilities could not be identified without additional information. Other than the availability of a validation set, the misclassification probabilities could be identified if multiple surrogate variables are available for the true unobserved misclassified covariates; see for example Kosinski and Flanders [10], Huang and Bandeen-Roche [11], and Wang and Song [12]. In general, multiple surrogate variables are required to be conditionally independent given the latent covariates.

Our main goal in this paper is to develop methods to correct for measurement error in a continuous exposure variable when their quantiles are used as the covariates. Flegal *et al.* [13] observed that misclassification from the dichotomization of a continuous error-prone exposure variable is *differential*. By differential misclassification, the misclassification probability of the error-prone binary variable given the true binary variable is not independent of the outcome variable. Under the assumption that the outcome variable is linear regression of the true unobserved continuous exposure, Gustafson and Le [14] developed expressions for the effect of misclassification from the dichotomization of a continuous error-prone exposure variable. However, it is a restriction to assume that the outcome variable is either linear or logistic regression of the true unobserved continuous exposure. In some applications the outcome variable may not be linearly associated with the underlying unobserved continuous exposure variable. For example, among women with breast cancer, Duggan *et al.* [15] found that the risk of all-cause mortality is higher among women in the first and fifth quintiles of insulin-like growth factor.

In this paper we propose two regression calibration estimators for misclassification in the quantiles due to measurement error in their continuous exposure variable. In the development of our methodology, we do not need to assume a validation subset with gold standard measurements for the exposure variable, which is the case in Natarajan [7], Lyles

[16], and Kuchenhoff *et al.* [17]. Furthermore, we do not need to assume that the outcome variable is linear (Gustafson and Le [14] and Natarajan [7]) or logistic regression (Dalen *et al.* [18]) of the true unobserved continuous exposure variable. In our main analysis, the outcome variable is a generalized linear model of the true unobserved quantiles (such as median-split binary covariates, quartiles, or quintiles), but the outcome variable may not have a linear association with the unobserved continuous exposure variable. Section 2 describes the regression models for our problem. A likelihood-based approach is described. However, the likelihood-based approach is in general not practical due to additional model assumptions and it often requires technical computations. This leads us to the development of a normality-based regression calibration (NRC) estimator (Section 3). The NRC estimator is *consistent* under a normality assumption when the primary model is linear. An estimator for a vector of parameters of interest (say denoted by β) is consistent if it converges to the true β when the sample size increases (to infinity). To understand whether the effect of misclassification results in *attenuation* bias under some situations, we further develop a linearization-based regression calibration (LRC) estimator in Section 4. By attenuation bias, the effect of the categorical covariate is under-estimated (or biased towards 0) when treating the misclassified covariate as the true covariate. The LRC estimator can be treated as a simple replacement estimator for NRC. Although it does not perform as well as NRC in general, the LRC estimator has a simple correction for attenuation interpretation. Finite sample performance of the proposed NRC and LRC estimators is investigated in Section 5. We apply the methods to real data in Section 6. Some concluding remarks are given in Section 7.

2 Statistical Models

Assume that the study cohort consists of n subjects. For $i = 1, \dots, n$, let X_i be a continuous exposure variable that can not be precisely measured, such as physical activity data. In the analysis, its quantiles are often used as the covariate of interest. Given that a variable split by quantiles can be presented by binary variables in regression analysis, it is generally sufficient for us to first consider a dichotomous variable as a covariate in regression. In the Discussion section, we will describe the details to adjust for misclassification in the quantiles (such as quartiles or quintiles). Let X_i^* be a dichotomous variable of X_i with a cutoff point x_0 such that $X_i^* = I[X_i \geq x_0]$, where $I[\cdot]$ is an indicator function. For example, if x_0 is the median of X_i then X_i^* is a median-split binary covariate. In the development of the method, we assume X_i^* is the covariate of interest. Let μ_x be the mean of X_i and σ_x^2 be the variance of X_i . We will need to estimate μ_x and σ_x^2 in the development of the proposed methods to be described later.

We assume that a surrogate variable W for X is available and that follows the classical additive measurement error model

$$W_{ij} = X_i + U_{ij}, \text{ where } E(U_{ij}|X_i) = 0,$$

for $i = 1, \dots, n, j = 1, \dots, k_i$ in which U_{ij} is the unobserved error. The use of replicates for W is for the purpose of identifying the measurement error variance (σ_u^2). The model given

above includes the situation when a reliability sample for W is available. A reliability sample with $k_j = 2$ is often conducted in a study to assess the within subject variation of W . For example, in the Women's Health Initiative (WHI), investigators were interested in the association between sedentary activity and mortality [19], and the data from a subset of WHI participants were used as reliability analysis of physical activity data. Let W_{ij}^* be binary such that $W_{ij}^* = 1$ if $W_{ij} \geq x_0$ and 0 otherwise (for $j = 1, \dots, k_i$). In the data analysis without an adjustment, W^* is used as a binary covariate. Let $\mathbf{Z}$ be the vector of other observed covariates (other than W) such as age and gender. The outcome variable Y_i , $i = 1, \dots, n$ follows a generalized linear model such that

$$E(Y_i | X_i^*, \mathbf{Z}_i) = g(\beta_0 + \beta_1 X_i^* + \beta_2' \mathbf{Z}_i), \quad (1)$$

where $g(\cdot)$ is a specified function, and our goal is to estimate $(\beta_0, \beta_1, \beta_2')' \equiv \boldsymbol{\beta}$. For example, $g(u) = u$ in linear regression, $g(u) = \exp(u)$ in Poisson regression, while $g(u) = (1 + e^{-u})^{-1}$ in logistic regression.

The problem of interest is a combination of measurement error and misclassification. It is a measurement error problem since X is not available while its surrogate W is available. It is a misclassification problem since the true covariate X^* is not available while its surrogate W^* is available. Although the problem is a mixture of measurement error and misclassification, it has not been well addressed in the literature (Gustafson [5]; Carroll *et al.* [8]; Buonacorsi [9]). To address this issue, we first consider an approach using estimating equations. It is easily seen that if X^* were known, then an estimating equation for $\boldsymbol{\beta}$ is

$\sum_{i=1}^n (1, X_i^*, \mathbf{Z}_i')' \{Y_i - g(\beta_0 + \beta_1 X_i^* + \beta_2' \mathbf{Z}_i)\} = 0$. This estimating equation is normally called the *full data estimating equation*; that is, the estimating equation for the case when all the data are observed without measurement error. Hence when X^* is not available, a consistent estimating approach is to take conditional expectation of the estimating equation of the full data estimating equation given the observed data (Wang *et al.* [20], Wu *et al.* [21], Yi *et al.* [22]). For simplicity of notation, unless otherwise stated $\tilde{W}_i$ denotes $(W_{i1}, \dots, W_{ik_i})$. The likelihood score-based expected estimating equation (EEE) for estimating $\boldsymbol{\beta}$ can be written as

$$\sum_{i=1}^n E \left[\begin{pmatrix} 1 \\ X_i^* \\ \mathbf{Z}_i \end{pmatrix} \left\{ Y_i - g(\beta_0 + \beta_1 X_i^* + \beta_2' \mathbf{Z}_i) \right\} | Y_i, \tilde{W}_i, \mathbf{Z}_i \right] = 0. \quad (2)$$

The conditional expectation given in (2) can be calculated by assuming the distribution of X and measurement error U . For example, let $f(V)$ denote the probability density function of any random variable V , then $f(Y | \tilde{W}, \mathbf{Z}) = \sum_{x^*=0}^1 f(Y | x^*, \tilde{W}, \mathbf{Z}) f(x^* | \tilde{W}, \mathbf{Z})$ and that

$$f(X^* = 1 | W, \mathbf{Z}) = \frac{\int_x I[x \geq x_0] f(W|x) f(x, \mathbf{Z}) dx}{\int_x f(W|x) f(x, \mathbf{Z}) dx}.$$

3 Normality-based Regression Calibration Method

The EEE approach discussed in the previous section would require calculation via the use of likelihood functions. In this section, we will propose a new NRC estimator. The NRC estimator is a replacement estimator that replaces an unobserved X^* by a quantity that can be obtained by simple calculation.

3.1 NRC for Linear Regression

In this subsection, we consider linear regression such that

$E(Y_i|X_i^*, \mathbf{Z}_i) = \beta_0 + \beta_1 X_i^* + \beta_2' \mathbf{Z}_i$. Under this model, it is seen that

$E(Y_i|X_i, \mathbf{Z}_i) = \beta_0 + \beta_1 I[X_i \geq x_0] + \beta_2' \mathbf{Z}_i$. However, like X_i^* , X_i is not observed and hence the mean model will need to be further developed. By simple calculation, we have

$E(Y_i|\bar{W}_i, \mathbf{Z}_i) = \beta_0 + \beta_1 \text{pr}(X_i \geq x_0|\bar{W}_i, \mathbf{Z}_i) + \beta_2' \mathbf{Z}_i$, where $\bar{W}_i = \sum_{j=1}^{k_i} W_{ij}/k_i$. From this association, the regression coefficients can be consistently estimated if we replace the unobserved X_i^* by $\text{pr}(X_i \geq x_0|\bar{W}_i, \mathbf{Z}_i)$. The conditional probability in general can be well estimated under a multivariate normality assumption of $(X_i, \bar{W}_i, \mathbf{Z}_i)$. When $(X_i, \bar{W}_i, \mathbf{Z}_i)$ is multivariate-normal, it is easily seen that

$$\begin{aligned} \text{pr}(X_i \geq x_0|\bar{W}_i, \mathbf{Z}_i) &= \text{pr}\left(\frac{X_i - E(X_i|\bar{W}_i, \mathbf{Z}_i)}{\{\text{var}(X_i|\bar{W}_i, \mathbf{Z}_i)\}^{1/2}} \geq \frac{x_0 - E(X_i|\bar{W}_i, \mathbf{Z}_i)}{\{\text{var}(X_i|\bar{W}_i, \mathbf{Z}_i)\}^{1/2}}|\bar{W}_i, \mathbf{Z}_i\right) \\ &= \Phi\left(\frac{(\sigma_x^2, \sum_{xz}) \begin{pmatrix} \sigma_x^2 + (\sigma_u^2/k_i) & \sum_{xz} \\ \sum_{xz} & \sum_z \end{pmatrix}^{-1} \begin{pmatrix} \bar{W}_i - \mu_x \\ \mathbf{Z}_i - \mu_z \end{pmatrix} - (x_0 - \mu_x)}{\left\{ \sigma_x^2 - (\sigma_x^2, \sum_{xz}) \begin{pmatrix} \sigma_x^2 + (\sigma_u^2/k_i) & \sum_{xz} \\ \sum_{xz} & \sum_z \end{pmatrix}^{-1} \begin{pmatrix} \sigma_x^2 \\ \sum_{xz} \end{pmatrix} \right\}^{1/2}}\right), \end{aligned} \quad (3)$$

where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution, μ_z is the mean of $\mathbf{Z}$, $\sum_{xz} = \text{cov}(X_i, \mathbf{Z}_i)$ and $\sum_z = \text{cov}(\mathbf{Z}_i)$. The NRC estimator is to replace X^* by the conditional probability given above in (3). The (nuisance) parameters involved in (3) can be consistently estimated since replicates for W are available (see the Appendix). Let $\hat{X}_i^*$ be the conditional probability $\text{pr}(X \geq x_0|\bar{W}, \mathbf{Z})$ in (3), which is often called the *calibration function*. When there is no $\mathbf{Z}$ or when X and $\mathbf{Z}$ are independent, the replacement variable (3) for the unobserved X_i^* can be written as

$$\hat{X}_i^* = \Phi\left(\frac{\sigma_x (\bar{W}_i - \mu_x)}{(\sigma_u/\sqrt{k_i}) \sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}} + \frac{\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}(\mu_x - x_0)}{\sigma_x (\sigma_u/\sqrt{k_i})}\right). \quad (4)$$

However, the expression still does not provide a simple correction for attenuation in our problem. That is, from the development of this section, we can not see directly a method to correct the bias from the naive estimator that uses covariates $(\bar{W}_i^*, \mathbf{Z}_i)$. The argument given above has shown that when Y given $(X^*, \mathbf{Z})$ is linear, a consistent estimator for β can be obtained by using covariates $(\hat{X}_i^*, \mathbf{Z})$. In the Appendix, we provide details on how to obtain

the standard error of the NRC estimator. The idea is based on stacking the estimating equations for β and the nuisance parameters discussed above, and then the asymptotic variance can be obtained by a sandwich estimator. The asymptotic variance has taken into consideration the uncertainty due to the replacement of X_i^* with $\hat{X}_i^*$.

3.2 NRC for Logistic Regression

In this subsection, we consider logistic regression such that

$E(Y_i|X_i^*, Z_i) = H(\beta_0 + \beta_1 X_i^* + \beta_2' Z_i)$, where $H(v) = \{1 + \exp(-v)\}^{-1}$. The mean function in logistic regression is not linear. We note that by a Taylor expansion,

$$E\{H(\beta_0 + \beta_1 X_i^* + \beta_2' Z_i) | \bar{W}_i, Z_i\} \approx H\{\beta_0 + \beta_1 E(X_i^* | \bar{W}_i, Z_i) + \beta_2' Z_i\} + 0.5 \beta_1^2 H^{(2)}\{\beta_0 + \beta_1 E(X_i^* | \bar{W}_i, Z_i) + \beta_2' Z_i\} E\{X_i^* - E(X_i^* | \bar{W}_i, Z_i)\}^2 | \bar{W}_i, Z_i\},$$

where $H^{(2)}(v) = \frac{d^2}{dv^2} H(v)$. If the second term of the above equation is relatively small, then we have the following approximation:

$E(Y_i | \bar{W}_i, Z_i) \approx H\{\beta_0 + \beta_1 E(X_i^* | \bar{W}_i, Z_i) + \beta_2' Z_i\}$. Hence, following a similar approximation for $\beta_0 + \beta_1 E(X_i^* | \bar{W}_i, Z_i)$ and the approximation given in the previous subsection, we have that

$$E(Y_i | \bar{W}_i, Z_i) \approx H\{\beta_0 + \beta_1 \hat{X}_i^* + \beta_2' Z_i\},$$

where $\hat{X}_i^*$ is from (3) if the calibration function is a function of both $\bar{W}_i$ and Z_i , or from (4) if the calibration function is independent of Z_i .

4 Linearization-based Regression Calibration Method

The proposed NRC approach discussed in the previous section has nice performance in general, and is consistent under a multivariate normality assumption. In this section, we aim at developing a new approximation approach that can provide an interpretation of the correction for attenuation from the naive estimator.

4.1 LRC for Linear Regression

We first consider linear regression in the subsection such that $E(Y|X^*, Z) = \beta_0 + \beta_1 X^* + \beta_2' Z$.

Under this model, we have that $E(Y|X, Z) = \beta_0 + \beta_1 I[X \geq x_0] + \beta_2' Z$. For any fixed Z , we let

$\beta_0 + \beta_2' Z$ be denoted by γ_0 . The idea of an approximation will be to approximate $Y = \gamma_0 + \beta_1 I[X \geq x_0]$ by a straight line as a function of Y and X . Figure 1 provides a visual demonstration of the linearization. The idea is that among the points with $X < x_0$, most points are in the range $(\mu_X - 2\sigma_X, x_0)$. Hence, the straight line to approximate the step function would pass through $((x_0 + \mu_X - 2\sigma_X)/2, \gamma_0)$. Similarly, among the points with $X \geq x_0$, most points are in the range $(x_0, \mu_X + 2\sigma_X)$. Therefore, the straight line to approximate the step function would also pass through $((x_0 + \mu_X + 2\sigma_X)/2, \gamma_0 + \beta_1)$. This indicates that we have that $E(Y|X, Z) \approx \theta_0 + (\beta_1/2\sigma_X) X + \beta_2' Z$, where $\theta_0 = \beta_0 + (\beta_1/2) - ((x_0 + \mu_X)\beta_1/(4\sigma_X))$

is a new intercept. Note that because X_i is not observed, the above model can not be directly applied to data analysis. From this approximation, we can further express the conditional expectation of Y given the observed $\overline{W}$, Z as:

$$E(Y|\overline{W}, Z) \approx \theta_0 + (\beta_1/2\sigma_x)E(X|\overline{W}, Z) + \beta_2'Z. \quad (5)$$

The above approximation (5) has an idea similar to the RC approach for linear model (Carroll et al. [8], Chapter 3), but our problem is that the unobserved covariate in our model is X^* (binary covariate split by median), not X . From the above approximation, we can correct for measurement error in X , or misclassification in X^* . This is normally done by assuming a bivariate-normal model for the joint distribution of X , U . If

$E(X|\overline{W}, Z) = E(X|\overline{W})$, then we can further express the above association by:

$$E(Y_i|\overline{W}_i, Z_i) \approx \theta_0^* + \frac{\beta_1\sigma_x}{2\{\sigma_x^2 + (\sigma_u^2/k_i)\}}\overline{W}_i + \beta_2'Z_i,$$

where $\theta_0^* = \beta_0 + (\beta_1/2) - ((x_0 + \mu_x)/(4\sigma_x))\beta_1 + [\mu_x\sigma_u^2/\{2k_i\sigma_x(\sigma_x^2 + (\sigma_u^2/k_i))\}]\beta_1$. Note that for a given Z_i , the next step for the approximation is to estimate $E(\overline{W}_i|\overline{W}_i^*)$. If W_i is approximately normally distributed, then by some calculation we have that

$E(\overline{W}_i|\overline{W}_i^*) \approx c(x_0) + 1.6\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}\overline{W}_i^*$, where $c(x_0)$ depends on x_0 . For example, if x_0

is the median, then $c(x_0) \approx \mu_x - 0.8\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}$. If x_0 is the first quartile, then

$c(x_0) \approx \mu_x - 1.2\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}$. If x_0 is the third quartile, then

$c(x_0) \approx \mu_x - 0.4\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}$.

Hence, by further simple calculations, we can obtain the following approximation:

$$E(Y_i|\overline{W}_i^*, Z_i) \approx \theta_0^+ + 0.8\frac{\sigma_x}{\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}}\beta_1\overline{W}_i^* + \beta_2'Z_i, \quad (6)$$

where

$$\theta_0^+ = \beta_0 + (\beta_1/2) - \{(x_0 + \mu_x)/(4\sigma_x)\}\beta_1 + [\mu_x\sigma_u^2/\{2k_i\sigma_x(\sigma_x^2 + (\sigma_u^2/k_i))\}]\beta_1 + \{\sigma_x\beta_1c(x_0)\}/[2\{\sigma_x^2 + (\sigma_u^2/k_i)\}]$$

. From the above approximation, our LRC estimator is to replace X_i^* by the following calibration function

$$\hat{X}_i^* = 0.8\frac{\sigma_x}{\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}}\overline{W}_i^*. \quad (7)$$

From the approximation given above and if $E(X|\overline{W}, Z) \approx E(X|\overline{W})$, then naively using the observed quantiles for a variable that is error-prone would likely lead to an attenuation bias. Under the linear model, a correction for attenuation method for β_1 can be obtained based on the approximation given above. By (6), if there are k replicates for W for all subjects ($k_i = k$

for all i), then the attenuation factor is $0.8\sigma_x / \sqrt{\sigma_x^2 + (\sigma_u^2/k)}$. That is, let $\hat{\beta}_{N,1}$ be the naive estimate of β_1 by using the observed quantile $\overline{W}_i^*$, then the LRC estimator for β_1 can be obtained by dividing the naive estimator $\hat{\beta}_{N,1}$ by the attenuation factor $0.8\sigma_x / \sqrt{\sigma_x^2 + (\sigma_u^2/k)}$. However, the standard error of the LRC estimator can not be obtained by the use of the attenuation factor directly. It can be obtained by a sandwich variance estimator where the vector of the estimating equations is obtained by stacking the estimating equations for β and the nuisance parameters discussed above (see the Appendix).

4.2 LRC for Logistic Regression

We consider logistic regression such that $E(Y_i|X_i^*, Z_i) = H(\beta_0 + \beta_1 X_i^* + \beta_2' Z_i)$. As discussed in the previous section, by a Taylor expansion, if the disease is rare, then $E(Y_i|X_i^*, Z_i)$ can be approximated by $H\{\beta_0 + \beta_1 E(X_i^*|\overline{W}_i, Z_i) + \beta_2' Z_i\}$. Hence, following an approximation that is similar to that for $\beta_0 + \beta_1 I[X_i \geq x_0]|\overline{W}_i, Z_i)$ given in (6) of the previous subsection, we have that

$$E(Y_i|W_i^*, Z_i) \approx H\{\theta_0^+ + 0.8 \frac{\sigma_x}{\sqrt{\sigma_x^2 + (\sigma_u^2/k_i)}} \beta_1 W_i^* + \beta_2' Z_i\}.$$

Hence, if we naively use the observed quantiles for a variable that is error-prone, then the naive estimator would likely have an attenuation bias. If there are k replicates for W for all subjects ($k_i = k$ for all i), then the attenuation factor is $0.8\sigma_x / \sqrt{\sigma_x^2 + (\sigma_u^2/k)}$. Let $\hat{\beta}_{N,1}$ be the naive estimate of β_1 by using the observed dichotomized $\overline{W}_i^*$. To adjust for bias, the LRC estimator for β_1 can be obtained by dividing the naive estimator $\hat{\beta}_{N,1}$ by the attenuation factor $0.8\sigma_x / \sqrt{\sigma_x^2 + (\sigma_u^2/k)}$.

5 Simulation Study

We conducted a simulation study to compare the naive estimator and the proposed NRC and LRC estimators. The naive estimator applied the usual regression analysis by replacing the unobserved X_i^* by $\overline{W}_i^*$. The NRC estimator replaced X_i^* by (4), and the LRC estimator replaced X_i^* by $0.8\sigma_x / \sqrt{\sigma_x^2 + (\sigma_u^2/k_i)} \overline{W}_i^*$ as in (7), and then we corrected the intercept by adjusting for the difference between θ_0^+ and β_0 . In Table 1, we considered the situation when Y_i given X_i^* was linear. We first generated covariates X_{ij} , $j = 1, \dots, n$ from a standard normal distribution. The true binary X_i^* was 1 if $X_i > \mu_x$, or 0 otherwise. Then Y_i , $i = 1, \dots, n$, were generated by $Y_i = \beta_0 + \beta_1 X_i^* + e_i$ where e_i was from a standard normal distribution. The surrogate variables W_{ij} , $j = 1, \dots, k_i$, were generated by $W_{ij} = X_i + U_{ij}$ where U_{ij} was normal with mean 0 and variance $\sigma_u = 0.5$ and 1.0 respectively. There were 50% of the individuals with $k_i = 2$, and 50% with $k_i = 1$. This was like a situation when a reliability sample was available with replicates for W . The binary surrogate W_{ij}^* was 1 if $W_{ij} > \mu_x$, or 0 otherwise. The sample size in the simulation was $n = 500$, and $n = 1000$, respectively. In the tables, the

“biases” were calculated by taking the average of $\hat{\beta} - \beta$ from 500 replicates, “SD” denotes the sample standard deviation of the estimators, “ASE” denotes the average of the estimated standard errors of the estimators. The 95% confidence interval coverage probabilities are also included. All the parameters are given in the tables.

From Table 1, it was seen that the naive estimator had large biases. The NRC estimator performed very well since it was a consistent estimator when X was normal and the measurement error was normal. The LRC estimator was not as good as the NRC estimator in terms of bias and efficiency. The LRC’s performance was in general acceptable when $\sigma_u = 0.5$, but the bias was larger when $\sigma_u = 1.414$. If we consider the relative bias versus the true slope, then for $\sigma_u = 1.414$ the relative bias was about 50% from the naive estimator while about 10% from the LRC estimator. Hence, although the LRC estimator was not consistent (such as NRC) it did serve well in terms of bias correction.

The data used for Table 2 were generated similarly to these for Table 1 except that the distribution of X was skewed. The underlying exposure X variables were generated from a mixture of two normals with means $(1/5^{1/2}, -2/5^{1/2})$, variances $(4/5, 1/5)$, and the mixture percentages were $(2/3, 1/3)$. The purpose of the simulation setup was to understand if the proposed NRC and LRC estimators were sensitive to the normality assumption of X . We note that when X was not normal, the NRC estimator was no longer consistent. From Table 2, it was seen that the performance of NRC and LRC was still satisfactory.

In Table 3, we studied simulations when the regression model was logistic such that $\text{pr}(Y_i|X_i^*) = \{1 + \exp(-\beta_0 - \beta_1 X_i^*)\}^{-1}$, where $\beta_0 = 0$ and $\beta_1 = \ln(3)$. The variables $X_i, W_i, X_i^*, \bar{W}_i^*$ were all generated similarly to those in Table 1. The results under logistic regression were very similar to those in Table 1 (linear regression). The NRC estimator did the best in terms of bias correction. The LRC estimator was not as good as NRC but it was considered as a satisfactory estimator with a good interpretation in correction for attenuation.

The data used for Table 4 were generated similarly to those for Table 3 except that the distribution of X was skewed. The underlying exposure X variables were generated from a mixture of two normals with means $(1/5^{1/2}, -2/5^{1/2})$, variances $(4/5, 1/5)$, and the mixture percentages were $(2/3, 1/3)$. The simulation set was aimed to understand if the proposed NRC and LRC estimators were sensitive to the normality assumption of X . From the results in Table 4, it was seen that the performance of NRC and LRC was still satisfactory.

6 Data Analysis of the NEW Trial

The NEW study was briefly described in the introduction. In the data analysis in this section, we investigated the association between CRP and VO_2max . We primarily considered the baseline data except that we used the VO_2max values at the 12th month from the control group as the 2nd measurement of VO_2max replicates. That is, the VO_2max values at baseline and 12-month from the control group ($n = 87$) were used as a reliability subsample to estimate the measurement error variance. We did not use the 12-month VO_2max values from the three intervention groups since the values were slightly higher than their baseline values.

In our analysis the unit of VO_2max was $\text{ml}/(\text{kg} \cdot \text{min})$, and hence the intervention effect on weight change has also contributed to the increase of VO_2max among the intervention individuals. In this data application, we were interested in the effect of median-split binary VO_2max and body mass index (BMI) on CRP. BMI is a measure of relative size based on the mass and height of an individual, calculated by dividing their weight in kilograms by the square of their height in meters.

Among the 439 individuals of the study, we excluded one individual with missing baseline CRP and 32 individuals who had CRP values higher than 10 mg/L which could be due to acute illness [1]. In addition, we also excluded three individuals with baseline VO_2max less than 10 $\text{ml}/(\text{kg} \cdot \text{min})$ which was a threshold typically required to complete basic activities of daily living [23]. As a result, 403 individuals of the NEW study were included. We first examined an association between VO_2max and CRP at baseline. The upper portion of Figure 2 showed the scatter plot and a fitted kernel smoother of VO_2max and CRP at baseline. To understand the measurement error property of VO_2max , the lower portion of Figure 2 was for VO_2max at baseline and 12-month from the control group. The 12-month VO_2max measurements in the control group in general were not systematically different from the baseline measurements. The differences were likely due to variations in physical conditions (or efforts) at different time points, and can be treated as measurement errors.

The application in this section was to apply our methods to the regression association for the effects of median-split VO_2max and BMI on CRP. Our work was focused on an application of our statistical methodology, and hence the clinical implication of the application here should be further investigated in a clinical trial research. We estimated the regression coefficients using the naive, NRC, and LRC estimators. The results were given in Table 5. All the three estimators showed that VO_2max was a significant risk factor for inflammatory marker CRP; but with different magnitudes. From the naive estimator, individuals with higher than mean VO_2max had on average about 0.45 mg/L lower CRP than the individuals with lower than mean VO_2max . From the NRC and LRC estimator, the difference in CRP between the 2 median-split groups was increased to either 0.87 mg/L or 0.68 mg/L. The effects of VO_2max from NRC and LRC estimators were larger than the corresponding naive estimator, which can be explained by the correction for attenuation discussed in Section 4. The standard errors from NRC and LRC estimates were both larger than the naive estimates, which is similar to a general phenomenon of bias-efficiency trade-off that was seen in measurement error literature. In addition, all the three estimates showed a similar significant effect of BMI on CRP that on average an increase of 1 kg/m^2 in BMI is associated with an increase of about 0.16 mg/L in CRP.

7 Discussion

Most of the details of the methods in the paper are based on dichotomization of an error-prone exposure variable. The methods can be applied to the situations when the covariates are q -quantiles of an error-prone exposure variable. For example, $q = 4$ for quartiles and $q = 5$ for quintiles. For simplicity, we consider linear regression in which the q -quantiles are covariates. We can write the regression model as $Y_i = \beta_0 + \sum_{j=1}^{q-1} \beta_j I[X_i \geq Q_j] + e_i$, in which

Q_j is the j^{th} q-quantile. If $q = 4$, then in the model, Q_1 is the first quartile, Q_2 is the median, and Q_3 is the third quartile. In this model, if we denote $I(X_j \leq Q_j)$ by $X_{i,j}^*$, then the methods in Section 3 (NRC) and Section 4 (LRC) can be applied to estimate $X_{i,j}^*$. Extension of the modeling of quantiles as covariates from linear regression to logistic regression is straightforward by using the approaches in Sections 3.2 and 4.2. For survival analysis with censored outcomes, the methods could be extended, but more detailed research is needed. Hence, further research is warranted for the methodology for survival analysis.

We have proposed two bias correction methods in regression analysis when quantiles for the error-prone exposure variable are used as the covariates. Although the proposed methods may not be consistent, the estimators are valid for the analysis of real data. For example, in biomedical research, analysis results are often presented by comparing the risk at different ranges (such as quartiles) of the exposures. The proposed estimators have limited bias especially for data when the exposure effect is not too large, which is often the case for environmental factors. Hence, the methods will likely have good methodological contributions for biomedical data.

In many studies, a calibration sample (or reliability sample) is often used to estimate the within subject and between subject variations (Guo *et al.* [24], White and Xie [25]). If there are two replicates in the calibration sample then $k_j = 2$ for individuals in the calibration sample, while $k_j = 1$ for the rest of the cohort. The statistical model for W in Section 2 can be treated as an internal reliability sample; say with $k_j = 2$. The simulation study and data analysis can be considered as the setup with an internal reliability sample. If in an application, a reliability sample is not a subset of the main cohort, then it is an external reliability sample. The main difference in analysis between internal and external reliability sample is in the standard error estimation of the estimators. For an internal reliability sample, there is a need to adjust for uncertainty due to the estimation of the measurement error parameters; just like what has been done in the paper. For an external reliability sample, there is no need to adjust for uncertainty due to estimation of the measurement error parameters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by National Institutes of Health grants AI112341 (Wang) CA53996 (Wang), ES017030 (Wang and Tapsoba), HL121347 (Wang, Tapsoba, Duggan and McTiernan), CA 105204 (Wang, Duggan, Campbell and McTiernan), CA 116847 (Wang, Duggan, Campbell and McTiernan), the Breast Cancer Research Foundation (Wang, Duggan and McTiernan) and a travel award from the Mathematics Research Promotion Center of National Science Council of Taiwan (Wang).

Appendix: Estimating Equations and Standard Error Estimation for NRC and LRC

The NRC or LRC estimator can be obtained by solving a set of estimating equations, and the asymptotic variance can be established by a sandwich estimator [26]. Recall that the mean function of Y given (X_i^*, Z_i) is $g(\beta_0 + \beta_1 X_i^* + \beta_2' Z_i)$. Let the replacement value (calibration function) of X_i^* for NRC or LRC be denoted by $\widehat{X}_i^*$. When X and Z are independent (or when Z is not in the main regression model), μ_x , σ_x and σ_u are the nuisance parameters involved in the calculation of $\widehat{X}_i^*$ via (4) or (7). The estimating equations for β and $(\mu_x, \sigma_x, \sigma_u)$ can be expressed as below.

$$\begin{cases} \sum_{i=1}^n \left(\frac{1}{\widehat{X}_i^* Z_i} \right) \{Y_i - g(\beta_0 + \beta_1 \widehat{X}_i^* + \beta_2' Z_i)\} = 0; \\ \sum_{i=1}^n (\overline{W}_i - \mu_x) = 0; \\ \sum_{i=1}^n \left\{ \sum_{j=1}^{k_i} (W_{ij} - \overline{W}_i)^2 - (k_i - 1) \sigma_u^2 \right\} = 0; \\ \sum_{i=1}^n \left\{ (\overline{W}_i - \mu_x)^2 - \sigma_x^2 - \sigma_u^2 / k_i \right\} = 0. \end{cases}$$

If in case X and Z are dependent, then the calibration function $E(X^* | \overline{W}, Z)$ (as in (3) of Section 3) is slightly more complicated since there are a few more nuisance parameters to be estimated [26]. See also Wang, et al. [20, Section 3.2] for a similar sandwich covariance estimator.

References

1. Foster-Schubert KE, Alfano CM, Duggan CR, Xiao L, Campbell KL, Kong A, Bain C, Wang CY, Blackburn G, McTiernan A. Effect of exercise and diet, alone or combined, on weight and body composition in overweight-to-obese post-menopausal women. *Obesity*. 2012; 20:1628–1638. [PubMed: 21494229]
2. Campbell KL, Foster-Schubert KE, Alfano CM, Wang C, Wang CY, Duggan CR, Mason C, Imayama I, Kong A, Xiao L, Bain CE, Blackburn GL, Stanczyk FZ, McTiernan A. Reduced-calorie dietary weight loss, exercise, and sex hormones in postmenopausal women: randomized controlled trial. *Journal of Clinical Oncology*. 2012; 30:2314–2326. [PubMed: 22614972]
3. Imayama I, Ulrich CM, Alfano CM, Wang C, Xiao L, Wener MH, Campbell KL, Duggan CR, Foster-Schubert KE, Kong A, Mason CE, Wang CY, Blackburn GL, Bain CE, Thompson HJ, McTiernan A. Effects of a caloric restriction weight loss diet and exercise on inflammatory biomarkers in overweight/obese postmenopausal women: a randomized controlled trial. *Cancer Research*. 2012; 72:2314–2326. [PubMed: 22549948]
4. Kullo IJ, Khaleghi M, Hensrud DD. Markers of inflammation are inversely associated with VO2 max in asymptomatic men. *Journal of Applied Physiology*. 2007; 102:1374–1379. [PubMed: 17170204]
5. Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall; New York: 2004.
6. Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *J Amer Statist Assoc*. 2005; 100:1250–1263.

7. Natarajan L. Regression calibration for dichotomized mismeasured predictors. *International Journal of Biostatistics*. 2009; 5(1):1143.
8. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models, A modern Perspective*. second. Chapman and Hall; London: 2006.
9. Buonaccorsi, J. *Measurement Error: Models, Methods, and Applications*. Hapman and Hall/CRC; Boca Raton: 2010.
10. Kosinski A, Flanders WD. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach. *Stat Med*. 1999; 18:2795–2808. [PubMed: 10521867]
11. Huang GH, Bandeen-Roche K. Building an identifiable latent variable model with covariate effects on underlying and measured variables. *Psychometrika*. 2004; 69:5–32.
12. Wang CY, Song X. Expected Estimating Equations via EM for Proportional Hazards Regression with Covariate Misclassification. *Biostatistics*. 2013; 14:351–365. [PubMed: 23178735]
13. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Amer J Epi*. 1991; 134:1233–1244.
14. Gustafson P, Le Nhu D. Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*. 2002; 4:878–87. [PubMed: 12495142]
15. Duggan C, Wang CY, Neuhausser M, Xiao L, Smith AW, Reding K, Baumgartner RN, Baumgartner KB, Bernstein L, Ballard-Barbash R, McTiernan A. Associations of insulin-like growth factor and insulin-like growth factor binding protein-3 with mortality in women with breast cancer. *International Journal of Cancer*. 2013; 132:1191–1200. [PubMed: 22847383]
16. Lyles R. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*. 2002; 58:1034–1036. [PubMed: 12495160]
17. Kchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*. 2006; 62:85–96. [PubMed: 16542233]
18. Dalen I, Buonaccorsi JP, Sexton JA, Laake P, Thoresen M. Correction for misclassification of a categorized exposure in binary regression using replication data. *Stat Med*. 2009; 28:3386–3410. [PubMed: 19757445]
19. Seguin RA, Buchner D, Lui J, Messina C, Manson J, Moreland L, Allison M, Wang CY, Patel M, LaCroix AZ. Sedentary Behavior and Mortality in Older Women: The Womens Health Initiative Observational and Extension Studies. *American Journal of Preventive Medicine*. 2014; 46:122–135. [PubMed: 24439345]
20. Wang CY, Huang Y, Chao EC, Jeffcoat MK. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*. 2008; 64:85–95. [PubMed: 17608787]
21. Wu L, Liu W, Hu XJ. (2010) Joint Inference on HIV Viral Dynamics and Immune Suppression in Presence of Measurement Errors. *Biometrics*. 2010; 66:327–335. [PubMed: 19673859]
22. Yi GY, Liu W, Wu L. Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*. 2011; 67:67–75. [PubMed: 20528858]
23. Shephard RJ. Maximal oxygen intake and independence in old age. *Br J Sports Med*. 2009; 43:342–346. [PubMed: 18403414]
24. Guo Y, Little RJ, McConnell DS. On using summary statistics from an external calibration sample to correct for covariate measurement error. *Epidemiology*. 2012; 23:165–174. [PubMed: 22157312]
25. White MT, Xie SX. Adjustment for measurement error in evaluating diagnostic biomarkers by using an internal reliability sample. *Stat Med*. 2013; 32:4709–4725. [PubMed: 23765915]
26. Wang CY. Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error. *Statistics and Probability Letters*. 1999; 45:371–378.

Figure 1.
Approximation of a step function with a jump at x_0 in the range of $(\mu_x - 2\sigma_x, \mu_x + 2\sigma_x)$.

Figure 2.

The top figure was for baseline VO_{2max} and CRP ($n = 403$); the bottom figure was for VO_{2max} at baseline and 12-month (from the control group, $n = 87$). The curve in each scatter plot was from a kernel smoother. The 12-month VO_{2max} measurements in the control group in general were not systematically different from the baseline measurements. The differences were likely due to variations in physical conditions (or efforts) at different time points, and can be treated as measurement errors.

Table 1Simulation study when Y given X^* was linear and X was normal

		$n = 500$			$n = 1000$		
		Naive	NRC	LRC	Naive	NRC	LRC
$\sigma_u = 0.5$							
β_0	Bias	0.127	0.000	-0.007	0.126	-0.003	-0.009
	SD	0.068	0.077	0.082	0.051	0.055	0.058
	ASE	0.067	0.074	0.080	0.047	0.054	0.057
β_1	CP	0.508	0.958	0.944	0.232	0.950	0.952
	Bias	-0.259	-0.005	0.009	-0.255	0.004	0.016
	SD	0.096	0.118	0.131	0.069	0.078	0.089
	ASE	0.094	0.120	0.128	0.067	0.080	0.091
	CP	0.222	0.954	0.944	0.060	0.946	0.960
$\sigma_u = 1.0$							
		Naive	NRC	LRC	Naive	NRC	LRC
β_0	Bias	0.222	-0.002	0.041	0.221	-0.004	0.040
	SD	0.069	0.091	0.092	0.047	0.065	0.068
	ASE	0.069	0.088	0.096	0.048	0.064	0.067
β_1	CP	0.088	0.960	0.920	0.002	0.946	0.900
	Bias	-0.430	0.005	-0.083	-0.443	0.009	-0.080
	SD	0.097	0.151	0.159	0.070	0.112	0.118
	ASE	0.097	0.156	0.162	0.068	0.104	0.114
	CP	0.006	0.944	0.922	0.000	0.946	0.868

NOTE: The naive estimator replaced X^* by W^* ; NRC was the normality-based RC estimator, and LRC was the linearization-based RC estimator. The regression parameters were $\beta = (0.5, 1)$. The nuisance parameters were $\mu_X = 0$, $\sigma_X = 1$. The results were obtained from 500 replicates.

Table 2

Simulation study when Y given X^* was linear and X was mixture-normal

		$n = 500$			$n = 1000$		
		Naive	NRC	LRC	Naive	NRC	LRC
$\sigma_u = 0.5$							
β_0	Bias	0.096	-0.019	-0.047	0.098	-0.017	-0.044
	SD	0.066	0.075	0.078	0.047	0.052	0.056
	ASE	0.065	0.072	0.077	0.046	0.051	0.054
	CP	0.688	0.930	0.896	0.402	0.918	0.872
β_1	Bias	-0.213	0.018	0.071	-0.217	0.013	0.066
	SD	0.094	0.114	0.128	0.066	0.078	0.090
	ASE	0.094	0.111	0.127	0.066	0.079	0.090
	CP	0.400	0.936	0.910	0.102	0.956	0.884
$\sigma_u = 1.0$							
β_0	Bias	0.180	-0.050	0.014	0.181	-0.049	0.012
	SD	0.071	0.096	0.101	0.046	0.063	0.064
	ASE	0.067	0.089	0.094	0.047	0.063	0.066
	CP	0.242	0.894	0.934	0.038	0.884	0.958
β_1	Bias	-0.409	0.046	-0.023	-0.404	0.053	-0.015
	SD	0.098	0.155	0.166	0.067	0.107	0.112
	ASE	0.096	0.151	0.161	0.068	0.107	0.114
	CP	0.006	0.930	0.926	0.000	0.928	0.962

NOTE: The naive estimator replaced X^* by W^* ; NRC was the normality-based RC estimator, and LRC was the linearity-based RC estimator. The regression parameters were $\beta = (0.5, 1.0)$. The nuisance parameters were $\mu_X = 0$, $\sigma_X = 1$. The results were obtained from 500 replicates.

Table 3Simulation study when Y given X was logistic and X was normal

		$n = 500$			$n = 1200$		
		Naive	NRC	LRC	Naive	NRC	LRC
$\sigma_u = 0.5$							
β_0	Bias	0.143	0.005	0.000	0.032	-0.007	-0.013
	SD	0.118	0.135	0.141	0.094	0.108	0.113
	ASE	0.127	0.146	0.152	0.090	0.103	0.108
	CP	0.820	0.968	0.968	0.682	0.936	0.948
β_1	Bias	-0.313	-0.029	-0.028	-0.296	-0.010	-0.006
	SD	0.181	0.233	0.245	0.142	0.181	0.193
	ASE	0.190	0.240	0.259	0.134	0.169	0.183
	CP	0.616	0.962	0.958	0.430	0.938	0.946
$\sigma_u = 1.0$							
β_0	Bias	0.219	-0.032	0.017	0.225	-0.014	0.032
	SD	0.134	0.181	0.188	0.086	0.121	0.120
	ASE	0.128	0.174	0.177	0.090	0.122	0.124
	CP	0.598	0.944	0.940	0.292	0.954	0.952
β_1	Bias	-0.476	0.032	-0.069	-0.498	0.014	-0.111
	SD	0.198	0.315	0.326	0.132	0.215	0.218
	ASE	0.188	0.307	0.313	0.133	0.215	0.219
	CP	0.296	0.946	0.936	0.032	0.954	0.926

NOTE: The naive estimator replaced X^* by W^* ; NRC was the normality-based RC estimator, and LRC was the linearity-based RC estimator. The regression parameters were $\beta = (0, \ln(3))$. The nuisance parameters were $\eta_X = 0$, $\sigma_X = 1$. The results were obtained from 500 replicates.

Table 4Simulation study when Y given X was logistic and X was mixture-normal

		$n = 500$			$n = 1200$		
		Naive	NRC	LRC	Naive	NRC	LRC
$\sigma_u = 0.5$							
β_0	Bias	0.091	-0.030	-0.063	0.102	-0.024	-0.049
	SD	0.131	0.143	0.155	0.088	0.098	0.104
	ASE	0.124	0.140	0.149	0.087	0.099	0.105
	CP	0.868	0.948	0.928	0.780	0.952	0.914
β_1	Bias	-0.313	-0.029	-0.028	-0.262	-0.006	-0.040
	SD	0.181	0.233	0.245	0.129	0.156	0.175
	ASE	0.190	0.240	0.259	0.135	0.166	0.183
	CP	0.616	0.962	0.958	0.484	0.968	0.952
$\sigma_u = 1.0$							
β_0	Bias	0.184	-0.058	-0.025	0.181	-0.063	-0.025
	SD	0.128	0.169	0.178	0.087	0.119	0.119
	ASE	0.126	0.171	0.175	0.089	0.120	0.123
	CP	0.692	0.954	0.954	0.442	0.916	0.946
β_1	Bias	-0.455	0.032	-0.037	-0.464	0.026	-0.054
	SD	0.190	0.300	0.318	0.130	0.214	0.216
	ASE	0.188	0.303	0.313	0.133	0.212	0.219
	CP	0.332	0.966	0.960	0.056	0.952	0.946

NOTE: The naive estimator replaced X^* by W^* ; NRC was the normality-based RC estimator, and LRC was the linearity-based RC estimator. The regression parameters were $\beta = (0, \ln(3))$. The nuisance parameters were $\eta_X = 0$, $\sigma_X = 1$. The results were obtained from 500 replicates.

Table 5

Modelling the relationship between VO₂max (ml/(kg • min)) and CRP (mg/L) using naive, NRC and LRC in the NEW trial (n = 403). The outcome variable was CRP. The covariate VO₂max was the true underlying binary median-split variable that may be misclassified due to measurement error in the continuous VO₂max.

	Naive	NRC	LRC
$\hat{\beta}_{0(\text{intercept})}$	-2.072	-1.447	-1.940
SE	0.955	1.094	1.037
$\hat{\beta}_{1(\text{VO}_2\text{max})}$	-0.446	-0.874	-0.684
SE	0.220	0.366	0.336
$\hat{\beta}_{2(\text{BMI})}$	0.166	0.153	0.166
SE	0.029	0.032	0.031
$\hat{\mu}_x$		24.960	24.960
SE		0.254	0.254
$\hat{\sigma}_x$		3.132	3.132
SE		0.131	0.131
$\hat{\sigma}_u$		3.512	3.512
SE		0.003	0.003

Note: The naive estimator replaced X^* by W^* , NRC was the normality-based RC estimator, and LRC was the linearity-based RC estimator.