mathematics

MDPI

Article

Generalized Linear Models with Covariate Measurement Error and Zero-Inflated Surrogates

Ching-Yun Wang [1,*], Jean de Dieu Tapsoba [2], Catherine Duggan [1] and Anne McTiernan [1]

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Center, P.O. Box 19024, Seattle, WA 98109-1024, USA; cduggan@fredhutch.org (C.D.); amctiern@fredhutch.org (A.M.)
[2] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, P.O. Box 19024, Seattle, WA 98109-1024, USA; jtapsoba@fredhutch.org
* Correspondence: cywang@fredhutch.org

Abstract: Epidemiological studies often encounter a challenge due to exposure measurement error when estimating an exposure–disease association. A surrogate variable may be available for the true unobserved exposure variable. However, zero-inflated data are encountered frequently in the surrogate variables. For example, many nutrient or physical activity measures may have a zero value (or a low detectable value) among a group of individuals. In this paper, we investigate regression analysis when the observed surrogates may have zero values among some individuals of the whole study cohort. A naive regression calibration without taking into account a probability mass of the surrogate variable at 0 (or a low detectable value) will be biased. We developed a regression calibration estimator which typically can have smaller biases than the naive regression calibration estimator. We propose an expected estimating equation estimator which is consistent under the zero-inflated surrogate regression model. Extensive simulations show that the proposed estimator performs well in terms of bias correction. These methods are applied to a physical activity intervention study.

Keywords: measurement error; surrogate; zero-inflated data

MSC: 62E20; 62F10; 62J12

Citation: Wang, C.-Y.; Tapsoba, J.d.D.; Duggan, C.; McTiernan, A. Generalized Linear Models with Covariate Measurement Error and Zero-Inflated Surrogates. *Mathematics* **2024**, *12*, 309. https://doi.org/10.3390/math12020309

Academic Editor: Chin-Shang Li

Received: 30 November 2023
Revised: 5 January 2024
Accepted: 15 January 2024
Published: 17 January 2024

1. Introduction

In biomedical research, regression analysis is an important tool to understand associations between disease outcomes and risk factors. In practice, however, a risk factor may not be measured precisely. This problem is often called covariate measurement error [1–3]. We consider an example when a biomarker is a risk factor for a disease outcome. In practice, the biomarker may have seasonal, daily, or even hourly variation, and a single measurement is prone to a covariate measurement error from instrumentation or human error. Hence, an average of an infinite number of the biomarker measurements during a specified period of time is, therefore, a more meaningful covariate variable than the average of a few observed measurements. However, in practice it is not feasible to make such measurements, and thus studies often rely on single measures at a specific time point with associated measurement error.

Physical activity and nutrient intake are important risk factors for disease incidence and mortality. However, physical activity and nutrient intake data may be measured with errors since they are generally self-report data. This issue is important since measurement error in diet or physical activity may have an attenuation effect on the regression coefficients of exposures in the range of approximately 20% to 50% [4–6]. That is, an odds ratio of 1.5 from diet or physical activity may be reduced to the range of 1.22 to 1.38 due to measurement errors in these measures. In addition, an important challenge in this research is that some physical activity or dietary data may have a zero value, such as 0 metabolic

equivalent (MET) hours per week from moderate or vigorous physical activity or 0 alcohol intake. One MET is defined as the amount of oxygen consumed while at rest per kilogram of body weight [7]. A 3 MET activity expends three times the energy used by the body at rest. Hence, if a person does a 3 MET activity for 4 h in a week, he or she has done 12 MET hours of physical activity in a week. A naive method without taking into account measurement error may lead to biased effect estimation in regression analysis, and the bias is attenuation in most (but not all) cases [8]. A standard bias correction for measurement error without taking into account a subset of individuals with zero exposure value may be biased in the effect estimation.

One motivating example of our methodology research is covariate measurement error associated with the measurement of physical activity in the APPEAL study (A Program Promoting Exercise and Active Lifestyles; APPEAL: Clinicaltrials.gov NCT00668161) [9]. APPEAL was a year long randomized controlled trial of moderate-to-vigorous intensity exercise vs. control (no exercise) among 202 healthy, sedentary adults recruited between 2001 and 2004 primarily through physician practices, and randomized to an exercise program (n = 100) or a control group (n = 102). The trial was designed to test the effects of exercise on biomarkers of colon cancer and other physiologic and psychosocial outcomes. Numerous case-control and cohort studies have found an inverse association between physical activity and risk of colon cancer [10]. Physical activities are commonly quantified by determining the energy expenditure in kilocalories or by using the MET of the activity. A question of interest is whether there is an association between physical activity via MET-hours/week and c-reactive protein, a biomarker of inflammation, with elevated levels of CRP associated with risk of developing colon cancer. The true average of MET-hours/week is an unobserved variable that is the average of an infinite number of MET-hours/week scores. However, in practice it is not possible to obtain this measure and, thus, the true average of MET-hours/week scores cannot be observed.

In the motivating example given above, two methodology challenges are involved. The first challenge is regression analysis with covariate measurement error, which is due to physical activity (MET-hours/week). The observed error-prone variable is typically called a surrogate variable for the true but unobserved exposure. The second challenge is the zero-inflated surrogate model because some individuals may have zero MET-hours/week. The zero-inflated surrogate issue in some similar research examples is also called truncation of the observed surrogates. In our problem, the second challenge (zero-inflated surrogate modeling) is added to the first challenge (covariate measurement error). Methods for covariate measurement error have been well developed. For example, regression calibration (RC) for covariate measurement error is to replace an error-prone covariate by its conditional expectation given the observed covariates [11]. In linear regression, the RC estimator is a consistent estimator for regression coefficients (Buonaccorsi, 2010, Chapter 5) [12]. However, for logistic and Cox regression, it is known that it is not consistent (Carroll, et al., 2006, Chapter 4) [2]. There is further research on refinement of RC for logistic and Cox regression [13,14]. Another general approximation approach for covariate measurement error is the simulation extrapolation (SIMEX) approach [15,16]. An advantage of SIMEX is that it has the advantage of being easy to implement. There are methods to address the situation when the surrogate variables may be truncated (which is in general the same as zero-inflated surrogate modeling). Tooze et al. investigated a likelihood approach for repeated measures data with clumping at zero [17]. When the observed exposure variables are truncated by a lower limit, the estimation of the disease–exposure association due to measurement error and truncation may not always be attenuation [18].

As discussed above, there is relatively limited research that addresses the issue of measurement error when some individuals may have a zero value (or lower limit) in the observed surrogates. The main objective of the paper is to develop and apply methods to adjust for measurement error in generalized linear models when the observed surrogates may be truncated at a low value (such as 0) among some individuals. The paper is organized

as follows: In Section 2, we describe the statistical models for the problem of interest, and discuss the bias issue when we apply a naive RC estimator without taking into account the zero-inflated surrogates. In Section 3, we study a regression calibration estimator for this problem. In Section 4, we propose a maximum likelihood estimator via expected estimating equations for this problem. In Section 5, the results from simulation studies are presented. In Section 6, we apply the methods to the APPEAL study data. We discuss the advantages and limitations of the proposed EEE estimator in Section 7. Concluding remarks are given in Section 8.

2. Statistical Models and Naive RC Estimator

We assume that the total sample size of the study cohort is n. The regression model of interest is the generalized linear model. Let Y_i be the response variable, X_i be the unobserved true covariate (dietary intake or physical activity) that cannot be precisely measured, and $\boldsymbol{Z}_i$ be the vector of covariates which is available for all individuals, i, $i = 1, \ldots, n$. For simplicity of presentation, the true unobserved exposure X is assumed to be a scalar throughout this paper. The main interest is to estimate the vector of regression coefficients $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \boldsymbol{\beta}_2')'$ in the followingregression model:

$$\mathrm{E}(Y_i|X_i, \boldsymbol{Z}_i) = g(\beta_0 + \beta_1 X_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i), \tag{1}$$

where $g(\cdot)$ is a specified function. Model (1) contains many important regression models. For example, $g(u) = u$ in linear regression, while $g(u) = (1 + e^{-u})^{-1}$ in logistic regression. The goal of the research is to develop valid estimation methods for the regression coefficients $\boldsymbol{\beta}$. For the true unobserved covariate X_i, we assume that there are k_i non–negative surrogate variables $W_{ij}, j = 1, \ldots, k_i$ such that $W_{ij} = \max(c, W_{ij}^*)$, where c is a detection limit, $W_{ij}^* = X_i + U_{ij}$, in which U_{ij} is an independent measurement error with $E(U_{ij}) = 0$. Let η_{ij} be the indicator function for a positive W_{ij} value, that is, $\eta_{ij} = I[W_{ij} > c]$. In a covariate measurement error problem when the surrogates are not truncated, replicates $W_{ij}, j = 1, \ldots, k_i$, are used to estimate the measurement error variance where k_i is the number of replicates. We use notation $\tilde{W}_i$ for $(W_{i1}, \ldots, W_{ik_i})$, $\tilde{W}_i^*$ for $(W_{i1}^*, \ldots, W_{ik_i}^*)$, and $\tilde{\eta}_i$ for $(\eta_{i1}, \ldots, \eta_{ik_i})$.

To understand the RC estimator, we consider a special linear regression case that $Y_i = \beta_0 + \beta_1 X_i + e_i$, where e_i is a mean-zero random residual term. Assume $W_{ij}^* = X_i + U_{ij}$, $j = 1, \ldots, k$, then it is easily seen that $\mathrm{E}(Y_i|\tilde{W}_i^*) = \beta_0 + \beta_1 \mathrm{E}(X_i|\tilde{W}_i)$. From this argument, it is seen that under the special linear regression case above, replacing an unobserved true X_i with $\mathrm{E}(X_i|\tilde{W}_i^*)$ will lead to a consistent estimator. This method is often called the RC estimator [2]. In this case, $\mathrm{E}(Y_i|\tilde{W}_i^*)$ is the calibration function. We may also use $\mathrm{E}(Y_i|\overline{W}_i^*)$, where $\overline{W}_i^* = \sum_{j=1}^k W_{ij}^*/k$, as the calibration function to replace the unobserved X_i. If replicates $W_{ij}^*, j = 1, \ldots, k_i$ are from a normal distribution, then $\mathrm{E}(Y_i|\overline{W}_i^*) = \mathrm{E}(Y_i|\tilde{W}_i^*)$ [14]. Let μ_x and σ_x denote the mean and standard deviation of any random variable X, respectively. Calculation of the conditional expectation of the unobserved exposure given the surrogates can be obtained based on a bivariate normal assumption such that

$$E(X_i|\overline{W}_i^*) = \boldsymbol{\mu}_x + \sigma_x^2 \left(\sigma_x^2 + \sigma_u^2/k\right)^{-1} \left(\overline{W}_i^* - \boldsymbol{\mu}_x\right).$$

Therefore, $\mathrm{E}(Y_i|\overline{W}_i^*) = \beta_0 + \boldsymbol{\beta}_1^* \overline{W}_i^*$, then $\boldsymbol{\beta}_1^* = \{\sigma_x^2 \left(\sigma_x^2 + \sigma_u^2/k\right)^{-1}\}\boldsymbol{\beta}_1$. From this calculation, a naive estimator using $\overline{W}_i^*$ as a replacement for X_i will have an attenuation effect. When $\boldsymbol{Z}$ is in the model, a standard RC estimator is to replace X_i with $\mathrm{E}(X_i|\overline{W}_i^*, \boldsymbol{Z}_i)$. This can be done by a multivariate-normal assumption with a conditional mean formula similar to the formula given above. However, a more practical approach is via a semiparametric RC approach by assuming a working regression model of $\mathrm{E}(W_{ij}^*|W_{ij'}^*, \boldsymbol{Z}_i) = \alpha_0 + \alpha_1 W_{ij'}^* + \boldsymbol{\alpha}_2' \boldsymbol{Z}_i$,

where $j \neq j' = 1, \ldots, k$, and $(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2')'$ is the vector of regression coefficients. This semi-parametric RC estimator does not assume a multivariate normality assumption of the observed surrogates and covariates [19,20].

However, in our problem, the observed W_{ij} is different from W_{ij}^* if $W_{ij}^* < c$. Using W_{ij} data will likely overestimate μ_x, but underestimate σ_x, and σ_u since $W_{ij} = c$ if $W_{ij}^* < c$. For linear regression with truncated surrogates, standard RC may be biased because $\mathrm{E}(X_i|\overline{W}_i)$ will be different from $\mathrm{E}(X_i|\overline{W}_i^*)$. One naive approach is to use the observed W_{ij} as W_{ij}^*, without taking into consideration the truncated surrogates, to calculate the RC estimator. We call this estimator a naive RC (NRC) estimator. As discussed above, the NRC estimator is biased even when the main regression model is linear. The asymptotic variance of the NRC estimator can be obtained by a sandwich variance estimator where the vector of the estimating equations is obtained by stacking the estimating equations for $\boldsymbol{\beta}$ and the nuisance parameters involved in the calculation of the calibration function $\mathrm{E}(X_i|\tilde{W}_i^*, \boldsymbol{Z}_i)$ (but noting that the NRC estimator assumes $\tilde{W}_i$ is the same as $\tilde{W}_i^*$). However, if there are many covariates in the modeling of the calibration function, then it will be computationally easier to use bootstrap variance estimation to obtain the standard errors.

3. Regression Calibration for Zero-Inflated Surrogates

The NRC estimator described in the previous section does not take into account zero values due to truncation. Now, we consider calibration based on truncate surrogates due to zero values. To understand the method, we first consider a linear regression model $Y_i = \beta_0 + \beta_1 X_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i + e_i$, where e_i has mean 0, and is independent of X_i and $\boldsymbol{Z}_i$. Then, $\mathrm{E}(Y_i|\tilde{W}_i, \boldsymbol{Z}_i) = \beta_0 + \beta_1 \mathrm{E}(X_i|\tilde{W}_i, \boldsymbol{Z}_i) + \boldsymbol{\beta}_2' \boldsymbol{Z}_i$. That is, replacing X_i with $\mathrm{E}(X_i|\tilde{W}_i, \boldsymbol{Z}_i)$ in the regression analysis may be a valid approach. Let $\widehat{X}_i \equiv \mathrm{E}(X|\tilde{W}, \boldsymbol{Z})$. The estimating equation for the RC estimator can be expressedas

$$\sum_{i=1}^{n} (1, \widehat{X}_i, \boldsymbol{Z}_i')' \{Y_i - (\beta_0 + \beta_1 \widehat{X}_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i)\} = \mathbf{0}. \tag{2}$$

Hence, when Y_i given $(X_i, \boldsymbol{Z}_i)$ is linear, we have the following result:

Proposition 1. *Assume the surrogate variables $W_{ij}^*, j = 1, \ldots, k_i$ may be truncated by a lower limit, and the truncation indicator $\tilde{\eta}_i$ is independent of Y_i given $(X_i, \boldsymbol{Z}_i)$. If $Y_i = \beta_0 + \beta_1 X_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i + e_i$, where e_i has mean 0, and is independent of X_i and $\boldsymbol{Z}_i$. Then the RC estimator solving (2) is a consistent estimator of β.*

The proof of Proposition 1 is given in Appendix A. We note that because of the surrogate assumption, the measurement errors U_{ij} and e_i are independent, which is needed to ensure that estimating Equation (2) is unbiased. Hence, for linear regression with zero-inflated surrogates, the RC estimator is consistent. However, when the mean function of Y_i given $X_i, \boldsymbol{Z}_i$ is not linear, the RC estimator may be biased since the expectation of the estimating score will no longer be zero. For logistic regression, $\mathrm{pr}(Y_i = 1|X_i, \boldsymbol{Z}_i) = H(\beta_0 + \beta_1 X_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i)$, where $H(u) = \{1 + \exp(-u)\}^{-1}$ is the logistic function. Although the RC estimator is not consistent, the RC estimator can be considered as an improved estimator of the NRC estimator described in the last section. The calibration function can be calculated based on the likelihood function. We use notation $\mathcal{L}(X)$ to denote a likelihood function for any random variable X, and $\mathcal{L}(Y|X)$ to denote a conditional likelihood function of Y given X, for any two random variables X and Y. Generally, the conditional calibration function can be calculated by the following:

$$\mathrm{E}\{X_i|\tilde{W}_i, \boldsymbol{Z}_i\} = \frac{\int_x x \prod_j \{\mathcal{L}(W_{ij}|X_i = x, \boldsymbol{Z}_i)\}^{\eta_{ij}} \{\mathcal{L}(W_{ij} = c|X_i = x, \boldsymbol{Z}_i)\}^{1-\eta_{ij}} \mathcal{L}(\boldsymbol{Z}_i, X_i = x) dx}{\int_x \prod_j \{\mathcal{L}(W_{ij}|X_i = x, \boldsymbol{Z}_i)\}^{\eta_{ij}} \{\mathcal{L}(W_{ij} = c|X_i = x, \boldsymbol{Z}_i)\}^{1-\eta_{ij}} \mathcal{L}(\boldsymbol{Z}_i, X_i = x) dx}. \tag{3}$$

In (3), we note that $\mathcal{L}(W_{ij} = c|X_i = x, \mathbf{Z}_i) = \mathcal{L}(U_{ij} \leq c - x)$. From the argument given above, the RC estimator can be obtained by replacing an unobserved X_i by $\mathrm{E}\{X_i|\tilde{W}_i, \mathbf{Z}_i\}$ based on (3). The asymptotic variance of the RC estimator can be obtained by a stacked sandwich estimator that is similar to the one for the NRC estimator described in the last section, or by bootstrap variance estimation.

4. Expected Estimating Equation Estimator

We now develop another approach to this problem via the maximum likelihood (ML) estimation. We first take a different viewpoint linking the ML estimation and the conditional expectation of the *full data estimating equation*, namely, the estimating equation when there is no measurement error. The full data likelihood, $\mathcal{L}(Y_i|X_i, \mathbf{Z}_i)$, is the likelihood function of Y_i given $(X_i, \mathbf{Z}_i)$. The full data estimating equation for $\boldsymbol{\beta}$ can be expressed as $\sum_{i=1}^{n} \boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta}) = 0$, in which $\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})$ is the derivative of $\log\{\mathcal{L}(Y_i|X_i, \mathbf{Z}_i)\}$ with respect to $\boldsymbol{\beta}$. Because the true X_i is not observed, the full data estimating equation can not be directly applied to the data. With the observed data, the estimating score will be from the likelihood of Y_i given $\mathbf{Z}_i$ and W_i, denoted by $\mathcal{L}(Y_i|\mathbf{Z}_i, W_i)$. If the distribution of $(\tilde{W}_i, X_i, \mathbf{Z}_i)$ does not involve β, then

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(Y_i|\tilde{W}_i, \mathbf{Z}_i) &= \frac{(\partial/\partial\boldsymbol{\beta}) \int_x \mathcal{L}(Y_i|X_i, \mathbf{Z}_i)\mathcal{L}(\tilde{W}_i|X_i = x, \mathbf{Z}_i)\mathcal{L}(X_i = x, \mathbf{Z}_i)dx}{\mathcal{L}(Y_i, \tilde{W}_i, \mathbf{Z}_i)} \\ &= \mathrm{E}\{\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(Y_i|X_i = x, \mathbf{Z}_i)|Y_i, \tilde{W}_i, \mathbf{Z}_i\}.\end{aligned}$$

From the equations given above, the likelihood-based score of the observed data can be obtained by the conditional expectation of the likelihood-based score of the full data given the observed data. That is, the estimating score for an individual can be expressed as $\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})|Y_i, \tilde{W}_i, \mathbf{Z}_i\}$, which is the observed data estimating score. The ML estimator can be obtained from the idea of expected estimating equations [21]. Therefore, the ML estimator can be obtained by solving

$$\sum_{i=1}^{n} \mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})|Y_i, \tilde{W}_i, \mathbf{Z}_i\} = 0. \tag{4}$$

In general, $\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})$ does not need to be the full data likelihood-based estimating score. It can be any estimating equation that satisfies $\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})\} = 0$. For example, it can be a weighted estimating equation of the ML estimator. The estimator solving (4) is the expected estimating equation (EEE) estimator for $\boldsymbol{\beta}$. Let Equation (4) be denoted by $\boldsymbol{S}(\boldsymbol{\beta}, X, \mathbf{Z}) = 0$. Let the EEE estimator be denoted by $\widehat{\boldsymbol{\beta}}_{eee}$. The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{eee}$ can be presented as the following result:

Proposition 2. *Assume Y_i given $(X_i, \mathbf{Z}_i)$ follows (1), and the surrogate variables $W_{ij}^*, j = 1, \ldots, k_i$ may be truncated by a lower limit, and the truncation indicator $\tilde{\eta}_i$ is conditionally independent of Y_i given $(X_i, \mathbf{Z}_i)$. Assume $\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})$ is any estimating equation that satisfies $E\{\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})\} = 0$. The EEE estimator solving (4) is consistent for $\boldsymbol{\beta}$. Furthermore, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{eee} - \boldsymbol{\beta})$ is asymptotically normal with mean 0 and asymptotic variance given in Appendix A.*

The proof of Proposition 2 is given in Appendix A. The EEE in (4) can be calculated by the following:

$$\begin{aligned}&\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i, \boldsymbol{\beta})|Y_i, \tilde{W}_i, \mathbf{Z}_i\} \\ = \; &\frac{\int_x \boldsymbol{\phi}(Y_i, X_i, \mathbf{Z}_i)\mathcal{L}(Y_i|X_i = x, \mathbf{Z}_i)\{\prod_{j=1}^{k_i} \mathcal{L}(W_{ij}|X_i = x, \mathbf{Z}_i)\}\mathcal{L}(\mathbf{Z}_i, X_i = x)dx}{\int_x \mathcal{L}(Y_i|X_i = x, \mathbf{Z}_i)\{\prod_{j=1}^{k_i} \mathcal{L}(W_{ij}|X_i = x, \mathbf{Z}_i)\}\mathcal{L}(\mathbf{Z}_i, X_i = x)dx},\end{aligned}$$

where $\mathcal{L}(W_{ij}|X_i = x, \boldsymbol{Z}_i) = \{\mathcal{L}(W_{ij}|X_i = x, \boldsymbol{Z}_i)\}^{\eta_{ij}}\{\mathcal{L}(W_{ij} = c|X_i = x, \boldsymbol{Z}_i)\}^{1-\eta_{ij}}$. The asymptotic variance of the EEE estimator solving (4) for $\boldsymbol{\beta}$ can be obtained by a sandwich variance estimator. The vector of the estimating equations is obtained by stacking two sets of estimating equations. The first set is the estimating equations for $\boldsymbol{\beta}$ and the second set is the nuisance parameters involved in the conditional distribution of Y_i given $(\boldsymbol{Z}_i, \tilde{W}_i)$. However, bootstrap variance estimation is another approach to obtain the standard errors of the EEE estimator.

5. Simulation Study

We conducted a simulation study to examine the finite sample performance of the NRC, RC, and EEE estimators with the naive estimator that used $\overline{W}_i$ for X_i. In Table 1, we illustrate the situation when the regression model is linear and the observed surrogates may have a zero value among some individuals. That is, the observed surrogates were truncated at $c = 0$ in the simulations. In this table, each individual's true covariate is X_i. We first generated X_i, $i = 1, \ldots, n$, from a normal distribution, where the sample size was $n = 500$, and $n = 1000$, respectively. We generated two replicates W_{i1}^* and W_{i2}^* for the unobserved X_i. With $\mu_x = 1.5, \sigma_x = 1$, and $\sigma_u = 0.707$. The percent of non–zero W_{ij} was $\overline{\eta} = 89\%$; 11% of W_{ij} was truncated at 0. We also considered the situation when $\sigma_u = 1$, 1.5, and $\sqrt{3}$, respectively, in which the percent of non-zero covariates were $\overline{\eta} = 86\%$, 80%, and 77%, respectively. The outcomes were generated based on linear regression with coefficients $\beta_0 = 0.5$ and $\beta_1 = 1$, and the residuals were from a standard normal distribution. In Tables 1–4, "bias" was obtained from the average of the biases of the regression coefficients estimates of the 500 simulation replicates, "SD" was the sample standard deviation of the estimates, and "ASE" was the average of the estimated standard errors of the estimates. The 95% confidence interval coverage probabilities (CP) were also obtained. The standard errors of the estimates were obtained from sandwich variance estimation. From the result of Table 1, the NRC estimator was not much better than the naive estimator. The reason for limited improvement from the NRC over the naive estimator was because of truncated W values. The RC and EEE estimators were consistent with limited biases under this setting, and hence, they were better than the naive and NRC estimators. Under this setting, the RC and EEE were very comparable.

Table 1. Simulation study for linear regression with truncated surrogates.

		Naive	**NRC**	**RC**	**EEE**	**Naive**	**NRC**	**RC**	**EEE**
				$n = 500$				$n = 1000$	
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 0.707, \overline{\eta} = 89\%$									
$\beta_0 = 0.5$	Bias	0.134	−0.230	−0.002	0.003	0.133	−0.228	−0.003	0.002
	SD	0.093	0.117	0.103	0.103	0.064	0.080	0.072	0.071
	ASE	0.093	0.117	0.106	0.106	0.066	0.083	0.075	0.074
	CP	0.684	0.486	0.972	0.962	0.460	0.180	0.954	0.966
$\beta_1 = 1$	Bias	−0.126	0.107	0.004	0.000	−0.127	0.103	0.001	−0.002
	SD	0.050	0.068	0.060	0.060	0.035	0.047	0.043	0.042
	ASE	0.049	0.068	0.061	0.061	0.035	0.048	0.043	0.043
	CP	0.270	0.658	0.958	0.954	0.056	0.446	0.956	0.960
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1, \overline{\eta} = 86\%$									
$\beta_0 = 0.5$	Bias	0.301	−0.349	−0.007	−0.006	0.299	−0.343	−0.005	−0.004
	SD	0.096	0.161	0.133	0.132	0.067	0.109	0.091	0.091
	ASE	0.095	0.162	0.136	0.136	0.068	0.113	0.095	0.095
	CP	0.122	0.404	0.960	0.952	0.002	0.106	0.966	0.962
$\beta_1 = 1$	Bias	−0.252	0.154	0.006	0.006	−0.252	0.147	0.003	0.002
	SD	0.050	0.096	0.080	0.079	0.035	0.066	0.056	0.056
	ASE	0.049	0.096	0.082	0.082	0.035	0.067	0.057	0.057
	CP	0.002	0.674	0.952	0.958	0.000	0.424	0.948	0.958

Table 1. *Cont.*

		Naive	NRC	RC	EEE	Naive	NRC	RC	EEE
				$n = 500$				$n = 1000$	
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1.5, \overline{\eta} = 80\%$									
$\beta_0 = 0.5$	Bias	0.556	−0.652	−0.035	0.033	0.558	−0.616	−0.018	−0.019
	SD	0.101	0.341	0.244	0.241	0.070	0.217	0.156	0.157
	ASE	0.098	0.325	0.230	0.229	0.069	0.220	0.157	0.158
	CP	0.000	0.462	0.962	0.942	0.000	0.104	0.960	0.960
$\beta_1 = 1$	Bias	−0.445	0.263	0.023	0.022	−0.447	0.241	0.011	0.012
	SD	0.048	0.197	0.152	0.150	0.033	0.126	0.097	0.099
	ASE	0.047	0.188	0.144	0.144	0.033	0.128	0.099	0.099
	CP	0.000	0.846	0.960	0.942	0.000	0.558	0.952	0.954
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = \sqrt{3}, \overline{\eta} = 77\%$									
$\beta_0 = 0.5$	Bias	0.655	−0.839	−0.057	−0.051	0.657	−0.769	−0.024	−0.025
	SD	0.101	0.609	0.323	0.307	0.070	0.302	0.197	0.198
	ASE	0.098	0.466	0.300	0.296	0.069	0.302	0.198	0.229
	CP	0.000	0.634	0.956	0.922	0.000	0.150	0.956	0.950
$\beta_1 = 1$	Bias	−0.519	0.327	0.038	0.034	−0.522	0.287	0.015	0.015
	SD	0.046	0.286	0.204	0.195	0.033	0.170	0.126	0.127
	ASE	0.045	0.263	0.191	0.189	0.032	0.170	0.126	0.148
	CP	0.000	0.972	0.956	0.918	0.000	0.716	0.948	0.930

NOTE: Naive is an estimator that uses the average of two replicates as the covariate, NRC is the naive RC estimator described in Section 2, RC is the RC estimator that uses $\mathrm{E}(X|\tilde{W})$ as the covariate, and EEE is the expected estimating equation estimator described in Section 4.

Table 2. Simulation study for linear regression with truncated surrogates; misspecified distribution for covariate X or measurement error.

		Naive	NRC	RC	EEE	Naive	NRC	RC	EEE
				$n = 500$				$n = 1000$	
X is from a mixture of two normal distributions and the error is normal									
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 0.707, \overline{\eta} = 91\%$									
$\beta_0 = 0.5$	Bias	0.209	−0.096	0.041	0.036	0.204	−0.101	0.037	0.032
	SD	0.081	0.099	0.097	0.097	0.061	0.074	0.073	0.073
	ASE	0.084	0.105	0.103	0.103	0.060	0.074	0.072	0.073
	CP	0.300	0.878	0.940	0.946	0.074	0.720	0.900	0.916
$\beta_1 = 1$	Bias	−0.160	0.038	−0.020	−0.018	−0.158	0.041	−0.018	−0.016
	SD	0.045	0.058	0.057	0.057	0.033	0.043	0.042	0.042
	ASE	0.046	0.061	0.059	0.060	0.032	0.043	0.042	0.042
	CP	0.060	0.920	0.946	0.950	0.002	0.848	0.928	0.928
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1, \overline{\eta} = 86\%$									
$\beta_0 = 0.5$	Bias	0.341	−0.199	0.051	0.036	0.336	−0.204	0.050	0.034
	SD	0.084	0.132	0.123	0.125	0.063	0.098	0.090	0.091
	ASE	0.086	0.139	0.130	0.131	0.061	0.098	0.091	0.092
	CP	0.024	0.734	0.928	0.946	0.000	0.460	0.902	0.920
$\beta_1 = 1$	Bias	−0.268	0.074	−0.024	−0.017	−0.265	0.076	−0.024	−0.017
	SD	0.045	0.078	0.075	0.076	0.033	0.058	0.054	0.055
	ASE	0.046	0.082	0.078	0.079	0.033	0.058	0.055	0.055
	CP	0.000	0.892	0.938	0.950	0.000	0.744	0.916	0.932

Table 2. *Cont.*

		Naive	NRC	RC	EEE	Naive	NRC	RC	EEE
		$n = 500$				$n = 1000$			
X is normal and the error is from a modified chi-square distribution									
	$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1, \overline{\eta} = 87\%$								
$\beta_0 = 0.5$	Bias	0.384	−0.278	0.082	0.088	0.385	−0.275	0.085	0.091
	SD	0.095	0.169	0.134	0.134	0.067	0.118	0.094	0.094
	ASE	0.093	0.163	0.129	0.129	0.066	0.115	0.091	0.091
	CP	0.012	0.614	0.870	0.850	0.000	0.322	0.816	0.792
$\beta_1 = 1$	Bias	−0.295	0.125	−0.038	−0.040	−0.293	0.125	−0.038	−0.040
	SD	0.052	0.101	0.081	0.081	0.036	0.070	0.056	0.056
	ASE	0.050	0.097	0.078	0.078	0.036	0.069	0.055	0.055
	CP	0.000	0.764	0.898	0.890	0.000	0.594	0.880	0.882
X is normal and the error is from a mixture of two normal distribution									
	$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1, \overline{\eta} = 84\%$								
$\beta_0 = 0.5$	Bias	0.376	−0.431	0.024	−0.024	0.380	−0.418	−0.018	−0.018
	SD	0.096	0.196	0.162	0.162	0.069	0.136	0.107	0.107
	ASE	0.096	0.198	0.160	0.161	0.068	0.139	0.112	0.112
	CP	0.030	0.402	0.954	0.958	0.000	0.114	0.954	0.958
$\beta_1 = 1$	Bias	−0.311	0.183	0.013	0.013	−0.314	0.175	0.009	0.009
	SD	0.048	0.116	0.098	0.098	0.033	0.080	0.066	0.066
	ASE	0.049	0.118	0.098	0.099	0.035	0.082	0.068	0.068
	CP	0.000	0.724	0.950	0.950	0.000	0.430	0.954	0.956

NOTE: See the footnote of Table 1 for notation.

We considered non-normal X in Table 2 to investigate if the estimators were sensitive to the normality assumption in the calculation. We also examined the sensitivity of the estimators to misspecification of the measurement error distribution. On the upper portion of Table 2, the unobserved X was generated from a mixture of two normal distributions; one with mean 2.5 and variance 1, and the other with mean 1 and variance 0.25, and the mixture percentages were (1/3, 2/3). The result from the upper portion of the table was similar to that of Table 1, except that there were small biases from the RC and EEE estimators. We found that the RC and EEE showed small biases when the unobserved exposure had a skewed distribution, but the bias was not too large in general. Nevertheless, the RC and EEE estimators were still better than the NRC and naive estimators under this situation. On the lower portion of Table 2, we considered the situation when X was normal but measurement error was from a location/scale-transformed chi-squared distribution and a mixture of two normal distributions, respectively. The specification of the mixture of two normal distributions was the same as the mixture of normal distributions given above. The location/scale-transformed chi-squared distribution has mean 0 and variance σ_u^2 after a chi-squared random variable was location/scale-transformed. From the sensitivity analysis, the RC and EEE estimators were not sensitive to mild violation due to a mixture of normal distributions since the biases were considered small. However, the biases may be sensitive to violation of the normality assumption while the true distribution was very skewed, as for chi-squared distributions. The biases were moderate, rather than small, when the errors were from chi-squared distributions.

Table 3. Simulation study for logistic regression with truncated surrogates.

		Naive	NRC	RC	EEE	Naive	NRC	RC	EEE
		$n = 500$				$n = 1000$			
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 0.707, \overline{\eta} = 89\%$									
$\beta_0 = 0$	Bias	0.065	−0.190	−0.010	−0.010	0.063	−0.190	−0.012	−0.012
	SD	0.191	0.234	0.203	0.208	0.136	0.169	0.147	0.150
	ASE	0.181	0.224	0.193	0.199	0.128	0.158	0.136	0.140
	CP	0.922	0.836	0.938	0.944	0.892	0.766	0.936	0.942
$\beta_1 = \ln(2)$	Bias	−0.080	0.083	−0.008	0.007	−0.079	0.083	−0.006	0.008
	SD	0.122	0.154	0.133	0.142	0.085	0.109	0.094	0.100
	ASE	0.115	0.147	0.126	0.134	0.082	0.104	0.089	0.095
	CP	0.868	0.914	0.928	0.930	0.788	0.874	0.936	0.944
$\beta_0 = 0$	Bias	0.069	−0.340	−0.014	−0.013	0.065	−0.341	−0.018	−0.016
	SD	0.207	0.266	0.219	0.232	0.148	0.189	0.159	0.169
	ASE	0.197	0.254	0.210	0.223	0.139	0.179	0.148	0.156
	CP	0.930	0.706	0.950	0.948	0.900	0.518	0.928	0.928
$\beta_1 = \ln(3)$	Bias	−0.116	0.146	−0.035	0.015	−0.114	0.145	−0.034	0.014
	SD	0.159	0.205	0.165	0.190	0.111	0.141	0.115	0.132
	ASE	0.149	0.191	0.155	0.178	0.106	0.135	0.109	0.125
	CP	0.848	0.884	0.920	0.940	0.766	0.836	0.920	0.942
$\mu_x = 1.5, \sigma_x^2 = 1, \sigma_u^2 = 1, \overline{\eta} = 86\%$									
$\beta_0 = 0$	Bias	0.175	−0.276	−0.014	−0.015	0.171	−0.277	−0.017	−0.016
	SD	0.186	0.277	0.222	0.230	0.135	0.203	0.166	0.172
	ASE	0.177	0.267	0.214	0.223	0.125	0.188	0.150	0.156
	CP	0.824	0.800	0.938	0.948	0.700	0.672	0.934	0.940
$\beta_1 = \ln(2)$	Bias	−0.173	0.108	−0.014	0.011	−0.171	0.109	−0.012	0.012
	SD	0.113	0.178	0.146	0.162	0.081	0.128	0.106	0.117
	ASE	0.108	0.171	0.140	0.155	0.076	0.121	0.098	0.109
	CP	0.610	0.914	0.948	0.946	0.404	0.856	0.926	0.940
$\beta_0 = 0$	Bias	0.232	−0.487	−0.028	−0.023	0.225	−0.487	−0.031	−0.023
	SD	0.204	0.333	0.249	0.269	0.146	0.236	0.183	0.199
	ASE	0.193	0.314	0.238	0.259	0.136	0.221	0.167	0.181
	CP	0.754	0.642	0.946	0.952	0.626	0.398	0.924	0.922
$\beta_1 = \ln(3)$	Bias	−0.273	0.175	−0.056	0.023	−0.270	0.174	−0.055	0.021
	SD	0.148	0.240	0.183	0.227	0.104	0.166	0.129	0.162
	ASE	0.138	0.222	0.171	0.213	0.098	0.156	0.120	0.148
	CP	0.488	0.892	0.900	0.946	0.230	0.824	0.902	0.940

NOTE: See the footnote of Table 1 for notation.

In Table 3, the data were generated similarly to those in Table 1 but the main model was logistic regression such that $\text{pr}(Y_i = 1|X_i) = H(\beta_0 + \beta_1 X_i)$, where the regression coefficients were $\boldsymbol{\beta} = (0, \ln(2))$ and $\boldsymbol{\beta} = (0, \ln(3))$, respectively. The findings were similar to those from Table 1 for the situation when $\boldsymbol{\beta} = (0, \ln(2))$. The biases of the RC and EEE estimators were very small. Although RC is not consistent, it may have limited biases if the relative risk parameter is small to moderate, such as $\beta_1 = \ln(1.5)$ or $\beta_1 = \ln(2)$ when the exposure's standard deviation is about 1. However, when $\beta_1 = \ln(3)$, the biases of the RC estimator were larger than those of the EEE estimator. The reason is that the RC estimator's bias will increase if the relative risk parameter is large. The findings are typically similar to those for measurement error in longitudinal data and survival analysis with covariate measurement error [20,21].

In Table 4, we investigated the situation when both X and Z were included in a linear regression model. We first generated $X_i, i = 1, \ldots, n$ and two replicates W_{i1} and W_{i2} in the same way as those in Table 1. Covariate Z_i, $i = 1, \ldots, n$, were generated via $Z_i = \rho X_i/\sigma_x + \sqrt{1-\rho^2} V_i/\sigma_z$, where V_i were from $N(0, \sigma_z^2)$ and independent from X_i, $\sigma_z^2 = 1$ and $\rho = 0.2$. The outcomes were generated via $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$, where $\beta_0 = 0.5$, $\beta_1 = 1$ and $\beta_2 = -1$, The residuals e_i, $i = 1, \ldots, n$, were generated from a standard normal random variable which was independent of X_i and Z_i. The findings were mostly similar to those from Table 1. That is, the naive and NRC estimators had large biases while the RC and EEE estimators were consistent with limited biases.

Table 4. Simulation study for linear regression model with truncated surrogates; covariates are X and Z.

		Naive	RC	CRC	EEE	Naive	RC	CRC	EEE
		$n = 500$				$n = 1000$			
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 0.707, \overline{\eta} = 89\%$									
$\beta_0 = 0.5$	Bias	0.137	−0.225	−0.006	−0.001	0.134	−0.224	−0.001	0.005
	SD	0.095	0.122	0.109	0.110	0.065	0.082	0.074	0.073
	ASE	0.093	0.117	0.106	0.106	0.066	0.083	0.075	0.074
	CP	0.694	0.504	0.938	0.930	0.454	0.226	0.946	0.944
$\beta_1 = 1$	Bias	−0.137	0.094	0.004	0.001	−0.136	0.093	0.001	−0.003
	SD	0.051	0.071	0.071	0.065	0.033	0.048	0.044	0.043
	ASE	0.050	0.069	0.064	0.063	0.036	0.049	0.044	0.044
	CP	0.204	0.742	0.940	0.938	0.020	0.538	0.954	0.956
$\beta_2 = -1$	Bias	0.042	0.042	−0.004	−0.004	0.049	0.049	0.002	0.003
	SD	0.052	0.052	0.053	0.053	0.036	0.036	0.038	0.037
	ASE	0.050	0.050	0.050	0.050	0.035	0.035	0.036	0.036
	CP	0.852	0.852	0.938	0.938	0.704	0.704	0.942	0.942
$\mu_x = 1.5, \sigma_x = 1, \sigma_u = 1, \overline{\eta} = 86\%$									
$\beta_0 = 0.5$	Bias	0.300	−0.347	−0.016	−0.016	0.298	−0.338	−0.005	−0.004
	SD	0.098	0.170	0.142	0.143	0.067	0.114	0.095	0.094
	ASE	0.095	0.162	0.136	0.136	0.067	0.113	0.095	0.094
	CP	0.110	0.406	0.944	0.944	0.006	0.132	0.956	0.954
$\beta_1 = 1$	Bias	−0.264	0.138	0.011	0.011	−0.264	0.132	0.004	0.002
	SD	0.051	0.099	0.087	0.087	0.033	0.068	0.060	0.059
	ASE	0.049	0.096	0.083	0.083	0.035	0.068	0.058	0.058
	CP	0.000	0.732	0.944	0.948	0.000	0.518	0.958	0.958
$\beta_2 = -1$	Bias	0.070	0.070	−0.005	−0.006	0.076	0.076	0.002	0.002
	SD	0.054	0.054	0.059	0.059	0.038	0.038	0.042	0.042
	ASE	0.052	0.052	0.053	0.054	0.037	0.037	0.038	0.038
	CP	0.736	0.736	0.934	0.938	0.464	0.464	0.922	0.920

NOTE: Naive is an estimator that uses the average of two replicates as the covariate, RC is the usual RC estimator that uses $\mathrm{E}(X|\tilde{W}, Z)$ as the covariate, CRC is a conditional RC estimator that uses $\mathrm{E}(X|\tilde{W}, Z, \eta)$ as the covariate, EEE is the expected estimating equation estimator described.

6. Analysis of APPEAL Data

The design of the APPEAL study was briefly reviewed in the Introduction. In this section, we are interested in investigating the association between physical activity measured via MET hours per week and CRP. The outcome variable of interest is the CRP value at baseline. In the APPEAL study, MET hours per week and other data including biomarkers were collected at both baseline and 12 months (end of study). In the control group who did not receive the exercise intervention, physical activity levels did not change significantly between baseline and 12 months. Hence, it seems reasonable to assume that the two MET-hours/week scores at baseline and 12 months in the control group (n = 102) can be treated as replicates. The MET-hours/week data for the exercise intervention group at 12 months were not included in the analysis as the MET-hours/week value changed significantly for study participants randomized to the exercise intervention between baseline and 12 months. As such, these values cannot be treated as replicates. The MET-hours/week scores at baseline and 12 months are surrogate variables (replicates, control arm only) for an unobserved true MET-hours/week score of an individual (unobserved underlying average of a period of time). The true unobserved average MET-hours/week variable is a variable to measure the actual physical activity which cannot be observed. In addition to MET-hours/week, age at baseline was another covariate in the regression analysis.

We first investigated an association between MET-hours/week and CRP at baseline. A scatterplot and a fitted kernel smoother of MET-hours/week and CRP at baseline are shown in the upper portion of Figure 1. The lower portion of Figure 1 is the scatterplot and a fitted kernel smoother of log(MET+1) and log(CRP) at baseline. We excluded 26 individuals with missing data and outliers (defined as values larger than median + 3× interquartile range) for CRP. Hence, a total of 176 individuals are included in the data analysis. The percentage of non-zero log(MET+1) at baseline is 67%, and 68% at 12 month. In our regression analysis, we used the log-transformed data since the transformed data were less skewed.

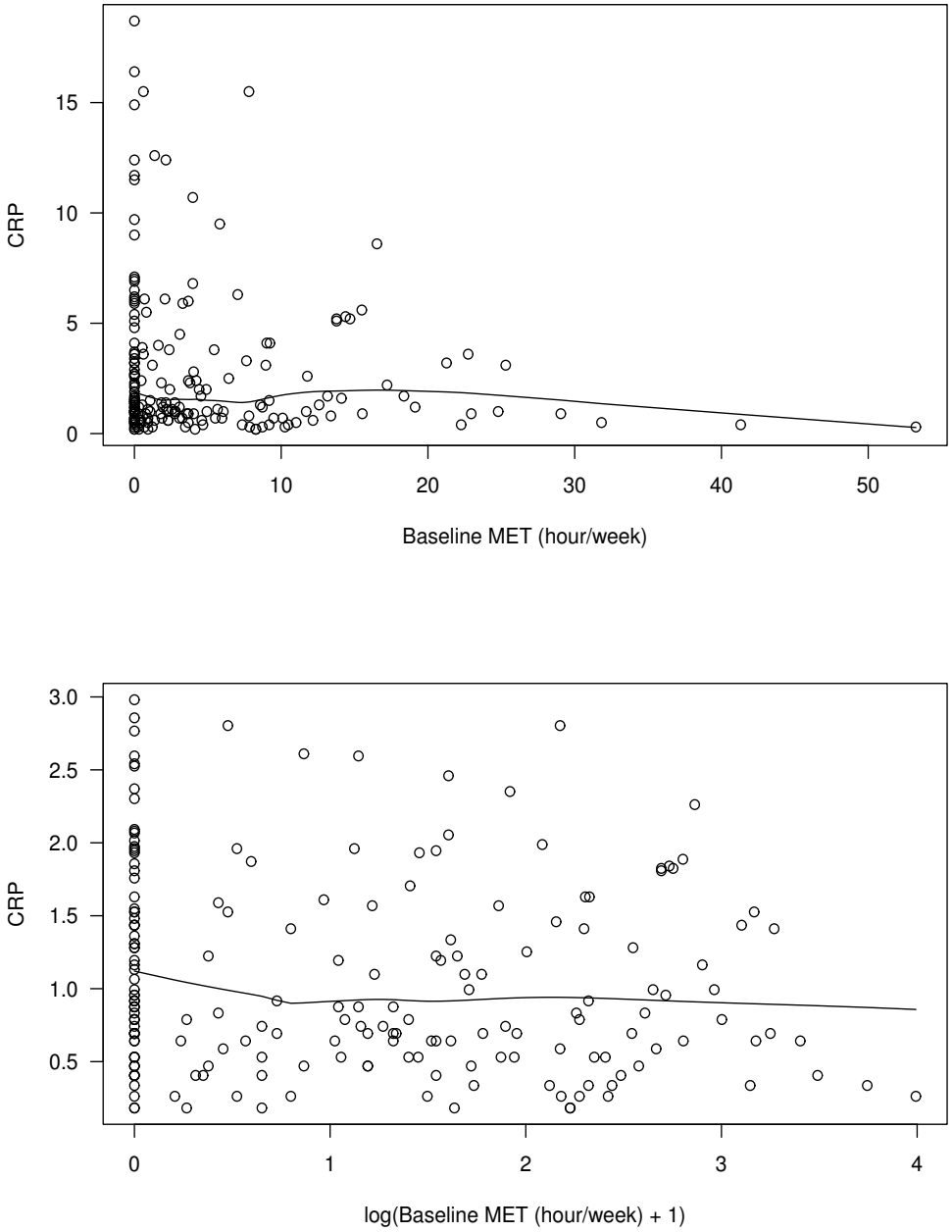

Figure 1. Upper: CRP versus MET; **Lower**: log(CRP) versus log(MET+1). The lines were obtained from fitting lowess smoothers.

In this section, the data analysis involved applying our methods to the regression association for the effects of physical activity (MET-hours/week) and age on CRP. The data application here is primarily for the purpose of a demonstration of our new methods. The regression coefficients were estimated based on the naive, RC, CRC, and EEE estimators. The results are given in Table 5. All the four estimators showed that MET was negatively associated with the inflammatory marker CRP; but not significant.

From the naive estimator, when the log(MET+1) score increased by 1 h/week, the CRP, on average, decreased by about 0.07 mg/L. From the NRC, RC, and EEE estimates, when the log(MET+1) score increased by 1 h/week, the CRP, on average, decreased by about 0.1 mg/L. It was observed that the standard errors from the NRC, RC, and EEE estimates were larger than those from the naive estimates. This was a general phenomenon of a bias–efficiency trade-off that has been reported in the measurement error literature, and is consistent with the findings from our simulations. Furthermore, all the four estimates

demonstrated a significant effect of age on CRP. On average, an increase of 10 years in age was associated with an increase of approximately 0.15 mg/L in log(CRP).

Table 5. Analysis results of data from the APPEAL study.

		Naive	NRC	RC	EEE
Intercept	β_0	0.259	0.345	0.299	0.282
	SE	0.360	0.377	0.367	0.364
log(MET+1)	β_1	−0.067	−0.136	−0.107	−0.098
	SE	0.045	0.098	0.071	0.062
Age	β_2	0.015	0.015	0.014	0.015
	SE	0.006	0.006	0.007	0.007
Nuisance parameters					
	μ_x		1.258	0.925	0.927
	SE		0.100	0.160	0.161
	σ_x^2		0.447	0.976	0.987
	SE		0.145	0.337	0.330
	σ_u^2		0.910	1.674	1.671
	SE		0.130	0.293	0.292

Note: See the footnote of Table 1 for notation. The percentages of non-zero log(1+MET) were 66.7% and 67.8% at baseline and 12 months among the participants in the control group, respectively. The total sample size in the analysis was 176.

7. Discussion

In the paper, we propose an EEE estimator for generalized linear models with covariate measurement error when the surrogate variables may have zero values among a subset of individuals. Our work is applicable to the situation for more applications when an exposure may be truncated. Our numerical studies show that RC is better than the naive estimator and NRC estimator in general, but it may be biased under some situations. Overall, the EEE estimator has smaller biases. There is a trade-off between bias and efficiency. The EEE has a larger SE due to this. One limitation of the proposed EEE estimator is that it may be biased if the likelihood function of the exposure variable is misspecified. Our simulation results demonstrate that the biases are moderate if the exposure distribution is not too skewed. Future research is needed to develop a non-parametric approach that does not require the exposure variable distribution [22].

In addition to physical activity or dietary data, biomarker measurements are important for the early detection and monitoring of disease progression. Our methods developed in this paper can be applied to biomarker data. When a biomarker is truncated due to a detection limit, decisions are required concerning how to handle values at or below the threshold in order to avoid biasing the parameter estimates. However, biomarkers are often measured with errors for many reasons, such as imperfect laboratory conditions, analytic variability of the assay, or temporal variability within individuals. The statistical modeling of zero-inflated surrogates in this paper can be applied to the situation when biomarker data are truncated due to a detection limit. Further research is needed if longitudinal biomarker, physical activity, or dietary data, are available over time [23–25].

8. Conclusions

We have developed an EEE approach for regression analysis with covariate measurement error when the surrogates may be truncated. One limitation of our proposed EEE estimator is that it is not consistent if the covariate distribution or the measurement error distribution is misspecified. In our simulations, the covariates and measurement errors are from normal distributions. Our simulation results demonstrate that if the misspecification is not too extreme, then the bias is typically small. Hence, if the covariates are skewed, then an appropriate (such as a logarithmic) transformation of the data may reduce the skewness of the data. Then the proposed EEE estimator may work well with likely minimal biases.

Author Contributions: Conceptualization, C.-Y.W. and A.M.; investigation, C.-Y.W. and J.d.D.T.; methodology, C.-Y.W. and J.d.D.T.; writing—original draft, C.-Y.W.; writing—review and editing, C.-Y.W., J.d.D.T., C.D. and A.M. All authors read and agreed to the published version of the manuscript.

Funding: This research was partially supported by US National Institute of Health grants CA235122 (Wang), HL130483 (Wang), CA77572 (McTiernan), CA239168 (Wang, Tapsoba, Duggan and McTiernan), a Breast Cancer Research Foundation award BCRF-23-107 (Wang, Tapsoba, Duggan and McTiernan), and a travel award from the Mathematics Research Promotion Center of the National Science Council of Taiwan (Wang).

Data Availability Statement: The data that support the findings of this study are not available for public access at this moment, but can be requested from the APPEAL study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs of Propositions 1 and 2

Proof of Proposition 1. Based on a standard surrogate assumption, the measurement errors U_{ij} and e_i are independent. Also, the truncation indicator $\tilde{\eta}_i$ is independent of e_i. Hence, $\mathrm{E}(e_i|\tilde{W}_i, \boldsymbol{Z}_i) = 0$. The unbiasedness of the estimating Equation (2) of the RC estimator can be obtained by calculating the expectation of the estimating score for individual i,

$$\begin{aligned} & \mathrm{E}\Big[(1, \widehat{X}_i, \boldsymbol{Z}_i')'\{Y_i - (\beta_0 + \beta_1 \widehat{X}_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i)\Big] \\ = \; & \mathrm{E}\Big((1, \widehat{X}_i, \boldsymbol{Z}_i')'\mathrm{E}\Big[\{Y_i - (\beta_0 + \beta_1 \widehat{X}_i + \boldsymbol{\beta}_2' \boldsymbol{Z}_i)\}|\tilde{W}_i, \boldsymbol{Z}_i\Big]\Big) \\ = \; & \mathbf{0}. \end{aligned}$$

Hence, for linear regression with zero-inflated surrogates, the RC estimator is consistent. □

Proof of Proposition 2. We note that $\boldsymbol{\phi}(Y_i, X_i, \boldsymbol{Z}_i, \boldsymbol{\beta})$ is an estimating score that satisfies $\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \boldsymbol{Z}_i, \boldsymbol{\beta})\} = 0$. We note that

$$\mathrm{E}\big[\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \boldsymbol{Z}_i, \boldsymbol{\beta})|Y_i, \tilde{W}_i, \boldsymbol{Z}_i\}\big] = E\{\phi(Y_i, X_i, \boldsymbol{Z}_i, \boldsymbol{\beta})\} = 0.$$

Hence, estimating Equation (4) for the EEE estimator is unbiased. We now develop the asymptotic distribution of the EEE estimator. Let the estimating score of the EEE estimator for the ith participant $\mathrm{E}\{\boldsymbol{\phi}(Y_i, X_i, \boldsymbol{Z}_i, \boldsymbol{\beta})|Y_i, \tilde{W}_i, \boldsymbol{Z}_i\}$ be denoted by $\psi(Y_i, \tilde{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta})$. Let $\boldsymbol{G}(\boldsymbol{\beta}) = -\mathrm{E}\{\partial \boldsymbol{\psi}(Y, \tilde{W}, \boldsymbol{Z}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\}$. By a Taylor expansion of the estimating equation at the true $\boldsymbol{\beta}$, and under some regularity conditions, it can be shown that

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_{eee} - \boldsymbol{\beta}) = \boldsymbol{G}^{-1}(\boldsymbol{\beta}) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \tilde{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}) + o_p(1),$$

Hence, it is seen that $n^{1/2}(\widehat{\boldsymbol{\beta}}_{eee} - \boldsymbol{\beta})$ is asymptotically normal with mean 0 and variance

$$\{\boldsymbol{G}(\boldsymbol{\beta})\}^{-1} n^{-1} [\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \tilde{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}) \{\boldsymbol{\psi}(Y_i, \tilde{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta})\}'] \{\boldsymbol{G}^{-1}(\boldsymbol{\beta})\}',$$

□

References

1. Fuller, W.A. *Measurement Error Models*; John Wiley & Sons: New York, NY, USA, 1987.
2. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement Error in Nonlinear Models, A modern Perspective*, 2nd ed.; Chapman and Hall: London, UK, 2006.
3. Yi, G.Y. *Statistical Analysis with Measurement Error or Misclassification, Strategy, Methods and Application*; Springer: New York, NY, USA, 2017.
4. Freedman, L.S.; Carroll, R.J.; Wax, Y. Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *Am. J. Epidemiol.* **1991**, 134, 310–320. [CrossRef]
5. Kipnis, V.; Subar, A.F.; Midthune, D.; Freedman, L.S.; Ballard-Barbash, R.; Troiano, R.; Bingham, S.; Schoeller, D.A.; Schatzkin, A.; Carroll, R.J. The structure of dietary measurement error: Results of the OPEN biomarker study. *Am. J. Epidemiol.* **2003**, *158*, 14–21. [CrossRef]

6. Kipnis, V.; Midthune, D.; Buckman, D.W.; Dodd, K.W.; Guenther, P.M.; Krebs-Smith, S.M.; Subar, A.F.; Tooze, J.A.; Carroll, R.J.; Freedman, L.S. Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* **2009**, 65, 1003–1010. [CrossRef]
7. Jette, M.; Sidney, K.; Blumchen, G. Metabolic equiva-lents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clin Cardiol.* **1990**, *13*, 555–565. [CrossRef] [PubMed]
8. Carroll, R.J.; Galindo C.D. Measurement Error, Biases, and the Validation of Complex Models for Blood Lead Levels in Children. *Environ. Health Perspect.* **1998**, *106*, 1535–1539. [CrossRef] [PubMed]
9. McTiernan, A.; Yasui, Y.; Sorensen, B.; Irwin, M.L.; Morgan, A.; Rudolph, R.E.; Surawicz, C.; Lampe, J.W.; Ayub, K.; Potter, J.D.; Lampe, P.D. Effect of a 12-month exercise intervention on patterns of cellular proliferation in colonic crypts: A randomized controlled trial. *Cancer Epidemiol. Biomarkers Prev.* **2006**, *15*, 1588–1597. [CrossRef]
10. Slattery, M.L.; Potter, J.; Caan, B.; Edwards, S.; Coates, A.; Ma, K.N.; Berry, T.D. Energy balance and colon cancer-—beyond physical activity. *Cancer Res.* **1997**, *57*, 75–80. [PubMed]
11. Prentice, R.L. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **1982**, *69*, 331–342. [CrossRef]
12. Buonaccorsi, J. *Measurement Error: Models, Methods, and Applications*; Hapman and Hall/CRC: Boca Raton, FL, USA, 2010.
13. Tsiatis, A.A.; DeGruttola, V.; Wulfsohn, M.S. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 count in patients with AIDS. *J. Am. Stat. Assoc.* **1995**, *90*, 27–37. [CrossRef]
14. Wang, C.Y.; Wang, N.; Wang, S. Regression analysis when covariates are regression parameters of a random effect model for observed longitudinal measurements. *Biometrics* **2000,** *56*, 487–495. [CrossRef]
15. Cook, J.; Stefanski, L.A. A simulation extrapolation method for parametric measurement error models. *J. Amer. Statist. Assoc.* **1994**, *89*, 1314–1328. [CrossRef]
16. Stefanski, L.A.; Cook, J.R. Simulation-Extrapolation: The Measurement Error Jackknife. *J. Am. Stat. Assoc.* **1995**, *90*, 1247–1256. [CrossRef]
17. Tooze, J.A.; Grunwald, G.K.; Jones, R.H. Analysis of repeated measures data with clumping at zero. *Stat. Methods Med. Res.* **2002**, *11*, 341–355. [CrossRef]
18. Richardson, D.B.; Ciampi, A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am. J. Epidemiol.* **2003**, *15*, 355–363. [CrossRef]
19. Wang, C.Y.; Cullings, H.; Song, X.; Kopecky, K.J. Joint nonparametric correction estimation for excess relative risk regression in survival analysis. *J. Roy. Statist. Soc. Ser. B* **2017**, *79*, 1583–1599. [CrossRef]
20. Wang, C.Y.; Song, X. Semiparametric regression calibration for general hazard models in survival analysis with covariate measurement error; surprising performance under linear hazard. *Biometrics* **2021**, *77*, 561–572. [CrossRef]
21. Wang, C.Y.; Huang, Y.; Chao, E.C.; Jeffcoat M.K. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorably missing data. *Biometrics* **2008**, *64*, 85–95. [CrossRef] [PubMed]
22. Huang, Y.H.; Hwang, W.H.; Chen, F.Y. Differential measurement errors in zero-truncated regression models for count data. *Biometrics* **2011**, *67*, 1471–1480. [CrossRef] [PubMed]
23. Tsiatis, A.A.; Davidian, D. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **2001**, *88*, 447–458. [CrossRef]
24. Tsiatis, A.A.;Davidian, M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **2004**, *14*, 809–834.
25. Tooze, J.A.; Kipnis, V.; Buckman, D.W.; Carroll, R.J.; Freedman, L.S.; Guenther, P.M.; Krebs-Smith, S.M.; Subar, A.F.; Dodd, K.W. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. *Stat. Med.* **2010**, *29*, 2857–2868. [CrossRef]