Grace Y. Yi¹ / Ying Yan² / Xiaomei Liao³ / Donna Spiegelman⁴

Parametric Regression Analysis with Covariate Misclassification in Main Study/Validation Study Designs

¹ Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, E-mail: yyi@uwaterloo.ca

² Department of Statistical Science, School of Mathematics, Sun Yat-sen University, Guangzhou, China

³ Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

⁴ Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA; Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, E-mail: stdls@channing.harvard.edu

Abstract:

Measurement error and misclassification have long been a concern in many fields, including medicine, administrative health care data, epidemiology, and survey sampling. It is known that measurement error and misclassification may seriously degrade the quality of estimation and inference, and should be avoided whenever possible. However, in practice, it is inevitable that measurements contain error for a variety of reasons. It is thus necessary to develop statistical strategies to cope with this issue. Although many inference methods have been proposed in the literature to address mis-measurement effects, some important issues remain unexplored. Typically, it is generally unclear how the available methods may perform relative to each other. In this paper, capitalizing on the unique feature of discrete variables, we consider settings with misclassified binary covariates and investigate issues concerning covariate misclassification; our development parallels available strategies for handling measurement error in continuous covariates. Under a unified framework, we examine a number of valid inferential procedures for practical settings where a validation study, either internal or external, is available besides a main study. Furthermore, we compare the *relative* performance of these methods and make practical recommendations.

Keywords: efficiency, estimating function, external validation study, internal validation study, likelihood method, measurement error, misclassification

DOI: 10.1515/ijb-2017-0002

Received: January 3, 2017; **Revised:** October 30, 2018; **Accepted:** November 6, 2018

1 Introduction

Mis-measurement has long been a concern in various fields, most notably in clinical and epidemiological studies. It is well documented that ignoring mis-measurement in inferential procedures can lead to seriously biased results. For example, in the simple linear regression model where a covariate is subject to an additive error, the estimate of the slope may be, attenuated towards zero if the error in the covariate is ignored in the analysis [1]. In other models, measurement error may inflate estimates of model parameters, change the sign of estimated treatment effects, change the model structures, and mask the true relationship between the response variable and covariates [2, 3]. Measurement error has effects that vary from problem to problem. Sometimes, these effects are drastic and must be incorporated in the analysis. Other times, these effects may be less impactful. While it is difficult to characterize their complex nature, measurement error effects are often influenced by multiple factors, including the nature of the measurement error process, the form of the response model, the variability of covariates as well as, the correlation between covariates and the analysis method.

It is difficult to list all the references here since the literature is vast. To name a few, see discussions by Stefanski and Carroll [4], Rosner, Willett and Spiegelman [5], Rosner, Spiegelman and Willett [6, 7], Cook and Stefanski [8], Wang and Davidian [9], Lin and Carroll [10], Spiegelman, Rosner and Logan [11], Huang and Wang [12], Liang and Wang [13], Spiegelman, Zhao and Kim [14], Zucker and Spiegelman [15, 16], Sugar, Wang and Prentice [17], Yi [18], Yi, Liu and Wu [19], Yi, Ma and Carroll [20], and Yi et al. [21], among many others.

Grace Y. Yi is the corresponding author.

© 2018 Walter de Gruyter GmbH, Berlin/Boston.

While mis-measurement may equally arise from either response or covariate variables, research on covariate mis-measurement, especially for continuous covariates, has dominated the literature. When multiple options are possible, researchers often struggle to choose a sensible method that suits their study. Although the merits of available methods have been demonstrated for specific problems individually, it is generally unclear how these methods would perform *relatively* if they are applied to the same problem; this is partially due to the complexity of simultaneously formulating all the methods under a unified, convenient framework for error-prone continuous covariates. However, such comparisons are possible if error-prone covariates are discrete.

Although there has been research on covariate misclassification (e.g. [15, 16, 22–27]), a comparison of methods for handling covariate misclassification has not been systematically conducted. In this paper, we consider a unified framework for handling error-prone discrete covariates and explore several valid inferential procedures for settings where a validation study is available alongside a main study. We consider cases where the validation study is either internal or external separately [28, 29]. Furthermore, we compare the *relative* performance of those methods and make practical recommendations.

Our investigation leverages the unique features of discrete covariates and contrasts the methods that have been developed for accounting for measurement error in continuous covariates. This research is partly motivated by the Nurses' Health Study [30] and the Health Professionals' Follow-up Study (HPFS) [31]. In the Nurses' Health Study (NHS), one objective is to understand risk factors for distal colon cancer incidence, including the role of calcium intake which is subject to mis-measurement. In the Health Professionals' Follow-up Study (HPFS), there is interest in questions such as what the effect of carbohydrate intake is on risk of type 2 diabetes incidence (T2D), and what the effect of red meat intake is on colorectal cancer risk. Since information on food intake is collected through bi-annual questionnaires, the resulting measurements are error-prone. In Section 5, we will analyze these data using the various methods studied in this paper to learn more about the scientific questions at hand.

Our paper contains the following sections. In Section 2, we introduce the notation and the basic model setup. In Section 3, we develop a number of methods to correct for bias induced by the misclassified exposure variable. Because all methods considered in the paper produce consistent estimators, the discussion focuses on efficiency comparisons of the estimators of the model parameters, where information about misclassification is assumed known. In Section 4, we develop inference methods for the situations where the misclassification probabilities are unknown, but are estimated from a validation study. In Section 5, we analyze the three real data using the proposed methods. Numerical assessment of the proposed methods is conducted in Section 6. Concluding remarks are given in the last section.

2 Notation and framework

2.1 Response model

Let Y be the outcome variable, X be the binary exposure variable, and Z be a vector of covariates. We consider the case where these variables are linked by a generalized linear model. That is, given $\{X, Z\}$, outcome variable Y follows an exponential family distribution

$$f_{Y|X,Z}(y|x,z;\xi) = \exp \left\{ \frac{y\xi - b(\xi)}{a(\phi)} + c(y;\xi) \right\}, \quad (1)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, ϕ is the dispersion parameter, and ξ is the canonical parameter. Here the lower case letters y, x and z stand for realizations of Y , X and Z , respectively, and the dependence on x and z of the right-hand side of (1) is suppressed in the notation as shown below.

This distribution form gives that the conditional mean and variance of Y , given $\{X, Z\}$, are, respectively, the first and second derivatives of $b(\xi)$ with

$$E(Y|X, Z) = b'(\xi) \quad \text{and} \quad \text{var}(Y|X, Z) = a(\phi)b''(\xi),$$

where $b'(\cdot)$ and $b''(\cdot)$ represent the first and second derivatives of $b(\cdot)$, respectively.

To explicitly reflect the influence of the covariates on the outcome variable, we further postulate the conditional mean $E(Y|X, Z)$ via the regression model

$$g\{E(Y|X, Z)\} = \beta_0 + \beta_x X + \beta_z^T Z, \quad (2)$$

where $g(\cdot)$ is assumed to be known and monotone, and $\beta = (\beta_0, \beta_x, \beta_z^T)^T$ is the vector of regression parameters. The objective is to estimate the parameter β and to highlight this, in the following development we often

explicitly show β when presenting an estimating function or a score function but suppress other parameters in the notation.

If the variables are all correctly measured, then inference about β may be based on the likelihood function:

$$L(\beta) = f_{Y|X,Z}(y|X, Z), \quad (3)$$

where the conditional probability density or mass function $f_{Y|X,Z}(y|X, Z)$ is determined by (1) and (2).

Let $S_{Y|X,Z}(\beta; y, X, Z) = (\partial/\partial\beta) \log f_{Y|X,Z}(y|X, Z)$ and $I(\beta) = E\{-(\partial/\partial\beta^T) S_{Y|X,Z}(\beta; Y, X, Z)\}$ denote the score function and the expected information matrix, respectively.

2.2 Misclassification probabilities

Suppose that covariates Z are perfectly measured, but X is subject to misclassification with a surrogate measurement X^* . Let $p_{00} = P(X^* = 0|X = 0, Z)$ be the specificity and let $p_{11} = P(X^* = 1|X = 1, Z)$ be the sensitivity. Let $\tilde{\pi} = P(X = 1|Z)$ be the prevalence of the exposure variable. Here for simplicity, the dependence on Z is suppressed in the symbols p_{00} , p_{11} , and $\tilde{\pi}$.

Then the conditional probability of X^* given Z is given by

$$P(X^* = 1|Z) = p_{11}\tilde{\pi} + (1 - p_{00})(1 - \tilde{\pi}).$$

Applying the Bayesian Theorem gives the conditional probabilities

$$P(X = 1|X^* = 1, Z) = \frac{p_{11}\tilde{\pi}}{p_{11}\tilde{\pi} + (1 - p_{00})(1 - \tilde{\pi})}$$

and

$$P(X = 0|X^* = 0, Z) = \frac{p_{00}(1 - \tilde{\pi})}{(1 - p_{11})\tilde{\pi} + p_{00}(1 - \tilde{\pi})}; \quad (4)$$

these probabilities are called the *positive predictive value* (PPV) and *negative predictive value* (NPV), respectively. For ease of notation, we let p_{11}^* and p_{00}^* denote the PPV and NPV, respectively.

In subsequent development, we assume that misclassification is nondifferential with

$$f_{Y|X,X^*,Z}(y|x, x^*, z) = f_{Y|X,Z}(y|x, z), \quad (5)$$

where $f_{Y|X,X^*,Z}(y|x, x^*, z)$ represents the conditional probability density or mass function of Y given $\{X, X^*, Z\}$. This assumption basically says that conditional on the true covariates X and Z , the surrogate measurement X^* is not predictive for the value of the response variable Y .

3 Analysis methods with known misclassification probabilities

In some applications, estimation of the parameter β may be conducted using the estimating equations approaches. It is imperative to require that estimating functions be both computable and sensible, in addition to allowing for identifiability of the parameter β . To be computable, estimating functions must be expressed in terms of the observed variables together with the parameters of interest. By sensible estimating functions, we mean to find those functions that produce estimators having nice statistical properties, such as consistency and asymptotic normality. While various strategies can be mobilized for this purpose, it is important to understand how these methods intrinsically differ in efficiency.

In this section, we elaborate on several estimation procedures and compare their relative efficiency to the likelihood method. Without loss of generality and to highlight the ideas, we assume that the misclassification probabilities and the prevalence of exposure are known for now. This assumption could be feasible for situations where misclassification arises by design in order to save cost or time. In circumstances where misclassification of the exposure variable is believed but there is no knowledge of the misclassification degree, this assumption can be imposed to conduct sensitivity analyses. Moreover, this assumption allows a direct comparison among the following methods to be described, as discussed in Section 3.5.

3.1 Induced likelihood method

The induced probability density or mass function of Y given $\{X^*, Z\}$ is given by

$$\begin{aligned} f_{Y|X^*, Z}(y|X^*, Z) &= E\{f_{Y|X, Z}(y|X, Z)|X^*, Z\} \\ &= \sum_{x=0,1} f_{Y|X, Z}(y|X=x, Z)P(X=x|X^*, Z), \end{aligned} \quad (6)$$

where the probability density or mass function $f_{Y|X, Z}(y|X=x, Z)$ is the same as that appears in (3), and $P(X=x|X^*, Z)$ is determined by (4).

Let $S_{Y|X^*, Z}^*(\beta; y, X^*, Z) = (\partial/\partial\beta) \log f_{Y|X^*, Z}(y|X^*, Z)$ be the score function. Then the expected information matrix is given by

$$I^*(\beta) = E\{-(\partial/\partial\beta^T)S_{Y|X^*, Z}^*(\beta; Y, X^*, Z)\}, \quad (7)$$

where the expectation is taken with respect to the joint probability density or mass function $f(y, x^*, z)$ of Y and $\{X^*, Z\}$.

This function is determined by

$$f(y, x^*, z) = \sum_{x=0,1} f_{Y|X, Z}(y|x, z)f_{X^*|X, Z}(x^*|x, z)f_{X|Z}(x|z)f_Z(z) \quad (8)$$

where $f_{Y|X, Z}(y|x, z)$ is given by (3), $f_{X^*|X, Z}(x^*|x, z)$ and $f_{X|Z}(x|z)$ are respectively determined by (12) and (13) in Section 3.3, and $f_Z(z)$ is the marginal probability density or mass function of Z .

The induced likelihood method was explored by various authors for different models. For instance, [32] employed this approach for the Cox proportional hazards models where continuous covariates are subject to measurement error. Spiegelman et al. [11] investigated this strategy for logistic regression with covariate misclassification and measurement error.

3.2 Subtraction correction method

It is often tempting to ignore the difference between X^* and X , and conducting the naive analysis is unfortunately far too common in practice. That is, the naive analysis uses the score function $S_{Y|X, Z}(\beta; y, X, Z)$ under the true model and replaces X with X^* to form the estimating function for β . Although this substitution gives a computable estimating function $S_{Y|X, Z}(\beta; y, X^*, Z)$, the substitution destroys the unbiasedness of the original score function $S_{Y|X, Z}(\beta; y, X, Z)$. That is, the expectation of $S_{Y|X, Z}(\beta; y, X^*, Z)$ is no longer zero, yielding that the resulting estimator for β is not necessarily consistent.

To obtain a consistent estimator of β , we modify function $S_{Y|X, Z}(\beta; y, X^*, Z)$ by subtracting its expectation. Define

$$U_{sub}^*(\beta; y, X^*, Z) = S_{Y|X, Z}(\beta; y, X^*, Z) - E\{S_{Y|X, Z}(\beta; Y, X^*, Z)\},$$

then the estimating function $U_{sub}^*(\beta; y, X^*, Z)$ is both unbiased and computable. As a result, estimation of β can be carried out using $U_{sub}^*(\beta; y, X^*, Z)$.

Given the setup in Section 2, $U_{sub}^*(\beta; y, X^*, Z)$ is specifically given by

$$U_{sub}^*(\beta; y, X^*, Z) = S_{Y|X, Z}(\beta; y, X^*, Z) - \int S_{Y|X, Z}(\beta; y, X^*, Z)f_{Y|X^*, Z}(y|X^*, Z)d\eta(y) \quad (9)$$

where $f_{Y|X^*, Z}(y|X^*, Z)$ is determined by (6), and $d\eta(y)$ is used to indicate an integral or a summation for all the possible values of Y , depending on whether Y is continuous or discrete.

Let $I_{sub}^*(\beta) = E\{U_{sub}^*(\beta; Y, X^*, Z)U_{sub}^{*T}(\beta; Y, X^*, Z)\}$ and $J_{sub}^*(\beta) = E\{(\partial/\partial\beta^T)U_{sub}^*(\beta; Y, X^*, Z)\}$. Then the Godambe information matrix (e.g. [3], Section 1.3) for the estimating function is given by

$$\mathcal{J}_{sub} = J_{sub}^*(\beta)I_{sub}^{*-1}(\beta)J_{sub}^{*T}(\beta). \quad (10)$$

While the idea of the subtraction correction method is conceptually simple, this scheme has not been widely employed. As discussed by [33], the subtraction correction method requires evaluation of the expectation $E\{S_{Y|X, Z}(\beta; Y, X^*, Z)\}$, which is often difficult to express in an analytic form for many settings. Under the additive hazards model with continuous covariates subject to measurement error for survival data, [34] implemented this strategy and obtained a tractable estimating function form. Here, we leverage the unique properties of discrete variables and successfully implement the subtraction correction scheme to handle misclassification-prone covariates.

3.3 Expectation correction method

Define

$$U_{exp}^*(\beta; Y, X^*, Z) = E\{S_{Y|X,Z}(\beta; Y, X, Z)|Y, X^*, Z\}, \quad (11)$$

where the expectation is taken with respect to the conditional distribution of X given $\{Y, X^*, Z\}$.

Function $U_{exp}^*(\beta; Y, X^*, Z)$ can be invoked to estimate β because it is both computable and unbiased; the unbiasedness of $U_{exp}^*(\beta; Y, X^*, Z)$ follows immediately from using the law of total expectation and the unbiasedness of the score function $S_{Y|X,Z}(\beta; Y, X, Z)$:

$$E_{Y,X,X^*,Z}[E\{S_{Y|X,Z}(\beta; Y, X, Z)|Y, X^*, Z\}] = E_{Y,X,X^*,Z}\{S_{Y|X,Z}(\beta; Y, X, Z)\} = 0,$$

where the expectations are evaluated with respect to the models for the corresponding distributions indicated by the associated random variables.

Given the model setup in Section 2, the conditional probability mass function of X^* given $\{X, Z\}$ is

$$f_{X^*|X,Z}(x^*|x, z) = (1 - p_{00})^{x^*(1-x)} p_{00}^{(1-x^*)(1-x)} p_{11}^{x^*x} (1 - p_{11})^{(1-x^*)x} \quad (12)$$

and the conditional probability mass function of X given Z is

$$f_{X|Z}(x|z) = \tilde{\pi}^x (1 - \tilde{\pi})^{(1-x)}. \quad (13)$$

Then the conditional probability mass function of X given $\{Y, X^*, Z\}$ is given by

$$f_{X|Y,X^*,Z}(x|y, x^*, z) = \frac{f_{Y|X,Z}(y|x, z) f_{X^*|X,Z}(x^*|x, z) f_{X|Z}(x|z)}{\sum_{x=0,1} f_{Y|X,Z}(y|x, z) f_{X^*|X,Z}(x^*|x, z) f_{X|Z}(x|z)},$$

where $f_{Y|X,Z}(y|x, z)$ is given by (3). As a result, $U_{exp}^*(\beta; y, X^*, Z)$ is given by

$$U_{exp}^*(\beta; y, X^*, Z) = \sum_{x=0,1} S_{Y|X,Z}(\beta; y, x, Z) f_{X|Y,X^*,Z}(x|y, X^*, Z). \quad (14)$$

Let $I_{exp}^*(\beta) = E\{U_{exp}^*(\beta; Y, X^*, Z) U_{exp}^{*T}(\beta; Y, X^*, Z)\}$ and $J_{exp}^*(\beta) = E\{(\partial/\partial\beta^T) U_{exp}^*(\beta; Y, X^*, Z)\}$. Then the Godambe information matrix of such an estimating function is given by

$$\mathcal{J}_{exp} = J_{exp}^*(\beta) I_{exp}^{*-1}(\beta) J_{exp}^{*T}(\beta). \quad (15)$$

When error-prone covariates are continuous, [35] developed the expectation correction method to accommodate measurement error when replicate measurements are available. Here, we explore this strategy to account for bias due to misclassification-prone discrete covariates.

3.4 Corrected score method

In contrast to the expectation correction method, we describe the corrected score method. The idea is to find a computable estimating function, denoted by $U_{cor}^*(\beta; Y, X^*, Z)$, so that a conditional expectation of this function can recover the score function $S_{Y|X,Z}(\beta; Y, X, Z)$. That is, as long as

$$E\{U_{cor}^*(\beta; Y, X^*, Z)|X, Y, Z\} = S_{Y|X,Z}(\beta; Y, X, Z), \quad (16)$$

then using $U_{cor}^*(\beta; y, X^*, Z)$ would produce a consistent estimator for β under certain regularity conditions.

Given the model setup of the response and misclassification processes, it is easily seen that setting

$$U_{cor}^*(\beta; Y, X^*, Z) = \frac{S(X=1) - S(X=0)}{p_{00} + p_{11} - 1} \cdot X^* + \frac{(p_{00} - 1)S(X=1) + p_{11}S(X=0)}{p_{00} + p_{11} - 1} \quad (17)$$

would satisfy (16). Here, $S(X=k)$ is the value of $S_{Y|X,Z}(\beta; Y, X, Z)$ evaluated at $X=k$ for $k=0,1$.

Let $I_{cor}^*(\beta) = E\{U_{cor}^*(\beta; Y, X^*, Z) U_{cor}^{*T}(\beta; Y, X^*, Z)\}$ and $J_{cor}^*(\beta) = E\{(\partial/\partial\beta^T) U_{cor}^*(\beta; Y, X^*, Z)\}$. Then the Godambe information matrix of such an estimating function is given by

$$\mathcal{J}_{cor} = J_{cor}^*(\beta) I_{cor}^{*-1}(\beta) J_{cor}^{*T}(\beta). \quad (18)$$

The corrected score method was proposed by [36] for generalized linear measurement error models with mis-measured continuous covariates. It has been successfully used for various regression models, such as Normal, Poisson and Gamma regression models ([2], Ch. 7). Extensions of this strategy to a range of settings have been discussed by several authors, including [12, 20, 37, 38]. Here, we use the corrected score scheme to develop unbiased estimating functions for estimation pertinent to regression models with misclassified discrete covariates.

3.5 Efficiency comparison

Standard estimating function theory shows that each of the preceding estimation methods results in a consistent estimator for regression parameter β , provided regularity conditions. The consistency of the estimators is ensured by the unbiasedness of the induced score function in Section 3.1 and the estimating functions in Sections 3.2–3.4. It is noted that the corrected score approach is valid under fewer modeling assumptions than the other three methods – in general, the corrected score approach does not require the distributional specification of the true covariate process but the other methods do need this. So in this sense, the corrected score approach is more robust than the other methods, as discussed by Carroll et al. [2] and Yi [3], among others.

To make a fair comparison among the four methods in Sections 3.1–3.4 in terms of the relative performance, here we assume that the parameters associated with the misclassification process are known. This assumption enables the four methods to share the same model assumptions; their differences are solely reflected from the different formulation perspectives. A general development is provided in section 4 that deals with the case where such an assumption is invalid. In Appendix A, we show the following result.

Theorem:

Suppose that the parameters $(\tilde{\pi}, p_{00}, p_{11})$ under the model setup of Section 2 are known. Then under mild regularity conditions, the induced likelihood method and the expectation correction method yield the same estimator of β .

It should be noted that the equivalence of the induced likelihood and the expectation correction methods holds only when the parameters $(\tilde{\pi}, p_{00}, p_{11})$ are known. This relationship no longer holds when the parameters $(\tilde{\pi}, p_{00}, p_{11})$ are estimated from a validation study, as demonstrated in the simulations reported in Section 6.

Asymptotically, the induced likelihood method and the expectation correction method are the most efficient. In terms of relative performance of the subtraction correction and the corrected score methods, compared to the induced likelihood method, we calculate the relative efficiency using the information carried with each type of estimating functions. For each component β_j of β , we define

$$\mathcal{R}_{sub,j} = \mathcal{J}_{sub,j}(\beta) / I_j^*(\beta), \text{ and } \mathcal{R}_{cor,j} = \mathcal{J}_{cor,j}(\beta) / I_j^*(\beta),$$

where $\mathcal{J}_{sub,j}(\beta)$, $\mathcal{J}_{cor,j}(\beta)$ and $I_j^*(\beta)$ represent the j th diagonal element, respectively, corresponding to $\mathcal{J}_{sub}(\beta)$ in (10), $\mathcal{J}_{cor}(\beta)$ in (18) and $I^*(\beta)$ in (7). Then, the ratios of $\mathcal{R}_{sub,j}$ and $\mathcal{R}_{cor,j}$ can be used to compare the efficiency of the subtraction correction and the corrected score methods, relative to the induced likelihood method, for estimators of the parameter component β_j .

To gain a direct understanding of the efficiency differences among the preceding estimation methods, we consider a logistic regression response model, where the outcome variable Y is binary and function $g(\cdot)$ in (2) is the logit function $g(t) = \log\{t/(1-t)\}$, given by

$$P(Y = y|X, Z) = \frac{\exp\{(\beta_0 + \beta_x X + \beta_z^T Z)y\}}{1 + \exp(\beta_0 + \beta_x X + \beta_z^T Z)} \quad (19)$$

for $y = 0, 1$. In Appendix B, we provide detailed expressions for the following efficiency comparisons.

Now, we numerically investigate efficiency differences between the preceding estimation methods under a number of scenarios with various degrees of misclassification probabilities and the prevalence of the outcome as well as different magnitudes of covariate effects. We consider the case where Z is a binary variable with probability 0.2 for assuming value 1, sensitivity $p_{11} = 0.6$ or 0.8 , and specificity $p_{00} = 0.6$ or 0.8 . In all these settings, we take $\beta_0 = \log 1.2$, leading to $P(Y = 1) \approx 60\%$, and $\beta_z = \log 1.2$. Let β_x be $-\log 20$ or $\log 2.0$.

Regarding the prevalence $\tilde{\pi} = P(X = 1|Z)$, we consider two cases. In the first case, we assume that X and Z are independent, and consider two values for $\tilde{\pi}$: 0.3 and 0.7. The results are displayed in Table 1. In the second case, we assume that X and Z are correlated, and the prevalence $\tilde{\pi}$ is given by

$$\text{logit } \tilde{\pi} = \alpha_0 + \alpha_z Z$$

with $(\alpha_0, \alpha_z) = (1, 0.5)$, leading to the correlation between X and Z being about 15%. The results are summarized in Table 2.

Table 1: Efficiency comparison for $\hat{\beta}_x$ and $\hat{\beta}_z$.

π_x	(p_{00}, p_{11})	$\beta_x = \log 0.05$		$\beta_x = \log 2.0$		$\beta_x = \log 2.0$		$\beta_x = \log 2.0$	
0.3	(0.6,0.6)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.998	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.150	$\mathcal{R}_{cor,z}$	0.142	$\mathcal{R}_{cor,x}$	0.619	$\mathcal{R}_{cor,z}$	0.634
		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.990	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.415	$\mathcal{R}_{cor,z}$	0.445	$\mathcal{R}_{cor,x}$	0.877	$\mathcal{R}_{cor,z}$	0.896
	(0.8,0.6)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.993	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.497	$\mathcal{R}_{cor,z}$	0.482	$\mathcal{R}_{cor,x}$	0.907	$\mathcal{R}_{cor,z}$	0.911
		$\mathcal{R}_{sub,x}$	0.999	$\mathcal{R}_{sub,z}$	0.986	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.691	$\mathcal{R}_{cor,z}$	0.737	$\mathcal{R}_{cor,x}$	0.963	$\mathcal{R}_{cor,z}$	0.971
	(0.8,0.8)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.998	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.150	$\mathcal{R}_{cor,z}$	0.142	$\mathcal{R}_{cor,x}$	0.619	$\mathcal{R}_{cor,z}$	0.634
		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.990	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.415	$\mathcal{R}_{cor,z}$	0.445	$\mathcal{R}_{cor,x}$	0.877	$\mathcal{R}_{cor,z}$	0.896
0.7	(0.6,0.6)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.168	$\mathcal{R}_{cor,z}$	0.098	$\mathcal{R}_{cor,x}$	0.629	$\mathcal{R}_{cor,z}$	0.612
		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.509	$\mathcal{R}_{cor,z}$	0.383	$\mathcal{R}_{cor,x}$	0.920	$\mathcal{R}_{cor,z}$	0.904
	(0.8,0.6)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	0.999	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.563	$\mathcal{R}_{cor,z}$	0.344	$\mathcal{R}_{cor,x}$	0.912	$\mathcal{R}_{cor,z}$	0.887
		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.812	$\mathcal{R}_{cor,z}$	0.653	$\mathcal{R}_{cor,x}$	0.981	$\mathcal{R}_{cor,z}$	0.968
	(0.8,0.8)	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.168	$\mathcal{R}_{cor,z}$	0.098	$\mathcal{R}_{cor,x}$	0.629	$\mathcal{R}_{cor,z}$	0.612
		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.509	$\mathcal{R}_{cor,z}$	0.383	$\mathcal{R}_{cor,x}$	0.920	$\mathcal{R}_{cor,z}$	0.904

Those entries with value 1.000 are not exactly identical to 1 but are nearly 1.

Table 2: Efficiency comparison for $\hat{\beta}_x$ and $\hat{\beta}_z$ with correlated covariates.

(p_{00}, p_{11})										
(0.6,0.6)	$\beta_x = \log 0.05$	$\mathcal{R}_{sub,x}$	0.999	$\mathcal{R}_{sub,z}$	0.999	$\beta_x = \log 2.0$	$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.167	$\mathcal{R}_{cor,z}$	0.081		$\mathcal{R}_{cor,x}$	0.627	$\mathcal{R}_{cor,z}$	0.603
(0.6,0.8)		$\mathcal{R}_{sub,x}$	0.998	$\mathcal{R}_{sub,z}$	0.997		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.513	$\mathcal{R}_{cor,z}$	0.337		$\mathcal{R}_{cor,x}$	0.921	$\mathcal{R}_{cor,z}$	0.902
(0.8,0.6)		$\mathcal{R}_{sub,x}$	0.999	$\mathcal{R}_{sub,z}$	0.993		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.558	$\mathcal{R}_{cor,z}$	0.286		$\mathcal{R}_{cor,x}$	0.908	$\mathcal{R}_{cor,z}$	0.874
(0.8,0.8)		$\mathcal{R}_{sub,x}$	0.998	$\mathcal{R}_{sub,z}$	0.991		$\mathcal{R}_{sub,x}$	1.000	$\mathcal{R}_{sub,z}$	1.000
		$\mathcal{R}_{cor,x}$	0.816	$\mathcal{R}_{cor,z}$	0.585		$\mathcal{R}_{cor,x}$	0.980	$\mathcal{R}_{cor,z}$	0.964

Those entries with value 1.000 are not exactly identical to 1 but are nearly 1.

It is seen that the subtraction correction method is more efficient than the corrected score method, and is almost as efficient as the induced likelihood method; the efficiency of this method does not seem to change with the misclassification degree or the values of β_x we consider. For a given value of β_x , the efficiency of the corrected score method on estimation of both β_x and β_z decreases as the degree of misclassification increases, regardless of the correlation strength between X and Z ; such a misclassification effect on reducing the efficiency of the corrected score method is exacerbated when β_x assumes a large absolute value.

As commented earlier, the corrected score method is more robust than other methods in that its results are not affected by the distributional specification of the covariate processes. The trade-off between the robustness and efficiency for these methods would be a key to decide which method is more suitable for a specific application.

4 Methods for main study/validation study designs

In applications, the misclassification probabilities and the prevalence of exposure are usually unknown and must be estimated from a validation study. In this section, we develop estimation procedures for the response model parameters to incorporate this feature. Our development covers two types of validation studies, *internal* or *external*, together with the main study.

In the main study/internal validation design (e.g. [39]), the data available are $\{(y_i, x_i^*, z_i) : i \in \mathcal{M}\}$ and $\{(y_i, x_i^*, x_i, z_i) : i \in \mathcal{V}\}$, respectively, where $\mathcal{M}$ and $\mathcal{V}$ contain n and m subjects, respectively, and $\mathcal{V}$ is a subset of $\mathcal{M}$. On the other hand, in the main study/external validation design (e.g. [40]), the available data are $\{(y_i, x_i^*, z_i) : i \in \mathcal{M}\}$ and $\{(x_i^*, x_i, z_i) : i \in \mathcal{V}\}$, respectively, where $\mathcal{V}$ and $\mathcal{M}$ do not overlap, and there are no response measurements for subjects in $\mathcal{V}$. With the main study/external validation design, we assume that given Z_i , the conditional distribution of (X_i, X_i^*) for $i \in \mathcal{V}$ is the same as that of (X_i, X_i^*) for $i \in \mathcal{M}$ so that the information carried by the study $\mathcal{V}$ can be transported to the main study $\mathcal{M}$ when carrying out inferences where X_i , X_i^* and Z_i represent the i th copy of X , X^* and Z , respectively. The feasibility of this assumption is justified by subject matter considerations. This assumption is typically reasonable for scenarios where both main and external validation studies are carried out to the same population using the same data collection procedures, and may often be reasonable even when this is not the case [41, 42].

We are interested in the relationship between the response variable Y_i and the true covariates (X_i, Z_i) , which are modeled by (1) and (2). Let $\mu\theta = (\beta^T, \vartheta^T)^T$ be the vector of model parameters, where β is of prime interest, and ϑ is the vector of nuisance parameters, including the misclassification probabilities possibly together with the parameters associated with the conditional distribution of X_i given Z_i , depending on the method to be used. As noted in Section 3.5, the validity of the four methods relies on different model assumptions. The induced likelihood, the subtraction correction and expectation correction methods all require modeling the response process and the misclassification process as well as modeling the conditional distribution of X given Z ; but the corrected score method only needs a model for the response and misclassification processes, with the conditional distribution of X given Z left unspecified.

Let $f_{X^*|X,Z}(x^*, x|z)$ denote the joint probability mass function of X^* and X given Z . When using the induced likelihood method, the subtraction correction method, or the expectation correction method, the nuisance parameter vector is $\vartheta_c = (p_{11}, p_{00}, \tilde{\pi})^T$. In this instance, for $i \in \mathcal{V}$, we define

$$L_i^\vartheta = f_{X^*|X,Z}(x_i^*, x_i|z_i), \quad (20)$$

which can be equivalently written as $f_{X^*|X,Z}(x_i^*|x_i, z_i)f_{X|Z}(x_i|z_i)$, and thus expressed in terms of the model setup in Section 2.2.

In contrast, when working with the corrected score method, the nuisance parameter vector is $\vartheta_c = (p_{11}, p_{00})^T$ and for $i \in \mathcal{V}$, L_i^ϑ is defined to be $f_{X^*|X,Z}(x_i^*|x_i, z_i)$ instead.

4.1 Induced likelihood method with validation data

With a main study/internal validation study design, the subjects in the main study contribute through the conditional distribution of $\{Y_i, X_i^*\}$ given Z_i while the subjects in the validation study contribute through the conditional distribution of $\{Y_i, X_i, X_i^*\}$ given Z_i . Therefore, under the nondifferential misclassification mechanism, the likelihood function for the entire data is given by

$$L_I(\theta) = \prod_{i \in \mathcal{M} \setminus \mathcal{V}} L_i^* \cdot \prod_{i \in \mathcal{V}} (L_i \cdot L_i^\vartheta)$$

where $L_i^* = f_{Y, X^*|Z}(y_i, x_i^*|z_i)$ is proportional to (6), L_i is given by (3), and L_i^ϑ is calculated from (20).

Let $S_I(\theta) = \log L_I(\theta)/\partial\theta$ be the score function of the induced likelihood for the observed data. Then solving

$$S_I(\theta) = 0 \quad (21)$$

gives an estimator of θ , $\hat{\theta}_I$, where 0 represents the zero vector of the same dimension of θ . In the following development, the symbol 0 may represent the number zero as well as a zero vector or a zero matrix whose dimension is clear from the context.

Let $S_{i\beta} = \partial \log L_i / \partial \beta$ and $S_{i\vartheta}^\vartheta = \partial \log L_i^\vartheta / \partial \vartheta$. Define $S_{i\vartheta}^* = (S_{i\beta}^{*T}, S_{i\vartheta}^{*T})^T$, where $S_{i\beta}^* = \partial \log L_i^* / \partial \beta$, and $S_{i\vartheta}^* = \partial \log L_i^* / \partial \vartheta$. Under regularity conditions and when the ratio m/n approaches a positive constant ρ as $n \rightarrow \infty$, $\hat{\theta}_I$ is a consistent estimator of θ , and $\sqrt{n}(\hat{\theta}_I - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix Σ_I given by $\Sigma_I = A_I^{-1}$, where

$$A_I = -(1-\rho)E\left(\frac{\partial S_{i\vartheta}^*}{\partial \theta^T}\right) - \rho \begin{pmatrix} E\left(\frac{\partial S_{i\beta}^*}{\partial \beta^T}\right) & 0 \\ 0 & E\left(\frac{\partial S_{i\vartheta}^\vartheta}{\partial \vartheta^T}\right) \end{pmatrix}.$$

This result can be proved by modifying standard likelihood theory; a sketch is given in Appendix C.

On the other hand, under the main study/external validation study design, the likelihood function is given by

$$L_E(\theta) = \prod_{i \in \mathcal{M}} L_i^* \cdot \prod_{i \in \mathcal{V}'} L_i^\vartheta$$

where L_i^* is determined by (6), and L_i^ϑ is given by (20). Then solving

$$\sum_{i \in \mathcal{M}} S_{i\vartheta}^*(\theta) + \sum_{i \in \mathcal{V}'} \begin{pmatrix} 0 \\ S_{i\vartheta}^\vartheta(\vartheta) \end{pmatrix} = 0 \quad (22)$$

gives an estimator of θ , $\hat{\theta}_E$, where the parameters are explicitly expressed in the score functions $S_{i\vartheta}^*$ and $S_{i\vartheta}^\vartheta$.

Under regularity conditions and when the ratio m/n approaches a positive constant ρ as $n \rightarrow \infty$, $\hat{\theta}_E$ is a consistent estimator of θ , and $\sqrt{n}(\hat{\theta}_E - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix Σ_E given by $\Sigma_E = \frac{1}{1+\rho}A_E^{-1}$, where

$$A_E = -\frac{1}{(1+\rho)}E\left(\frac{\partial S_{i\vartheta}^*}{\partial \theta^T}\right) - \frac{\rho}{1+\rho} \begin{pmatrix} 0 & 0 \\ 0 & E\left(\frac{\partial S_{i\vartheta}^\vartheta}{\partial \vartheta^T}\right) \end{pmatrix}.$$

This result can be proved by modifying standard likelihood theory; a sketch is given in Appendix D, which also shows why the asymptotic variances of $\hat{\theta}_E$ and $\hat{\theta}_I$ are different.

4.2 Unbiased estimating equations for main study/validation study designs

In this subsection, we explore three unbiased estimating equations methods for correcting for bias due to misclassification when validation data are available. Specifically, we consider the main study/internal validation

study design and the main study/external validation study design, and develop inference procedures based on the *subtraction correction*, *expectation correction* and *corrected score* methods, as in Section 3.

First, we describe inference procedures using the subtraction correction method. With the main study/internal validation study design, the estimating function is given by

$$U_{sub,I}(\theta) = \sum_{i \in \mathcal{M} \setminus \mathcal{V}} \begin{pmatrix} U_{sub,i}^*(\beta, \vartheta) \\ 0 \end{pmatrix} + \sum_{i \in \mathcal{V}} \begin{pmatrix} S_i(\beta) \\ S_{i\vartheta}^\vartheta(\vartheta) \end{pmatrix}, \quad (23)$$

where $U_{sub,i}^*$ is given by (9), and $S_{i\beta}(\beta)$ is the score function $S_{i\beta}$ with the dependence on the parameter β explicitly spelled out. Estimation of θ is then obtained by solving

$$U_{sub,I}(\theta) = 0$$

for θ ; let $\hat{\theta}_{sub,I}$ denote the resulting estimator.

The formulation of (23) reflects different uses of the information from the main study and the validation study. The first term of (23) includes the contributions from the main study; its zero subvector shows that this term contributes no information on estimation of ϑ , the parameter which governs the misclassification process. The second term of (23) involves the measurements of (Y_i, X_i^*, X_i, Z_i) from the validation sample; this sample contributes information for the estimation of both β and ϑ , as suggested by the dependence on β of $S_{i\beta}(\beta)$ and the dependence on ϑ of $S_{i\vartheta}^\vartheta(\vartheta)$.

Since the main study subjects contribute no data for estimating ϑ , this estimation procedure is identical to the two-stage procedure: (1) use the likelihood score function $S_{i\vartheta}^\vartheta(\vartheta)$ based on the validation data $\{(x_i^*, x_i, z_i) : i \in \mathcal{V}\}$ to obtain an estimator $\hat{\vartheta}_{sub,I}$ of the misclassification parameters $\vartheta_{sub,I}$, and (2) solve

$$\sum_{i \in \mathcal{M} \setminus \mathcal{V}} U_{sub,i}^*(\beta, \hat{\vartheta}_{sub,I}) + \sum_{i \in \mathcal{V}} S_{i\beta}(\beta) = 0$$

to obtain the estimator $\hat{\beta}_{sub,I}$ of $\beta_{sub,I}$.

Under regularity conditions and that the ratio m/n approaches a positive constant ρ as $n \rightarrow \infty$, $\hat{\theta}_{sub,I}$ is a consistent estimator of θ , and $\sqrt{n}(\hat{\theta}_{sub,I} - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix $\Sigma_{sub,I}$ given by $\Sigma_{sub,I} = A_{sub,I}^{-1} B_{sub,I} (A_{sub,I}^{-1})^T$, where

$$\begin{aligned} A_{sub,I} &= -(1-\rho) \begin{pmatrix} E\left(\frac{\partial U_{sub,i}^*}{\partial \beta^T}\right) & E\left(\frac{\partial U_{sub,i}^*}{\partial \vartheta^T}\right) \\ 0 & 0 \end{pmatrix} - \rho \begin{pmatrix} E\left(\frac{\partial S_{i\beta}}{\partial \beta^T}\right) & 0 \\ 0 & E\left(\frac{\partial S_{i\vartheta}^\vartheta}{\partial \vartheta^T}\right) \end{pmatrix}, \text{ and} \\ B_{sub,I} &= (1-\rho) \begin{pmatrix} E(U_{sub,i}^* U_{sub,i}^{*T}) & 0 \\ 0 & 0 \end{pmatrix} + \rho \begin{pmatrix} E(S_{i\beta} S_{i\beta}^T) & 0 \\ 0 & E(S_{i\vartheta}^\vartheta S_{i\vartheta}^{\vartheta T}) \end{pmatrix}. \end{aligned}$$

On the other hand, under the main study/external validation study design, the estimating function is given by

$$U_{sub,E}(\theta) = \sum_{i \in \mathcal{M}} \begin{pmatrix} U_{sub,i}^*(\beta, \vartheta) \\ 0 \end{pmatrix} + \sum_{i \in \mathcal{V}} \begin{pmatrix} 0 \\ S_{i\vartheta}^\vartheta(\vartheta) \end{pmatrix},$$

where $U_{sub,i}^*(\beta, \vartheta)$ is given by (9) with the dependence on the parameters spelled out. Then solving

$$U_{sub,E}(\theta) = 0$$

gives an estimator of $\theta_{sub,E}$. Let $\hat{\theta}_{sub,E}$ denote this estimator.

Under regularity conditions and that the ratio m/n approaches a positive constant ρ as $n \rightarrow \infty$, $\hat{\theta}_{sub,E}$ is a consistent estimator of $\theta_{sub,E}$, and $\sqrt{n}(\hat{\theta}_{sub,E} - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix $\Sigma_{sub,E}$ given by $\Sigma_{sub,E} = \frac{1}{1+\rho} A_{sub,E}^{-1} B_{sub,E} (A_{sub,E}^{-1})^T$, where

$$\begin{aligned} A_{sub,E} &= -\frac{1}{1+\rho} \begin{pmatrix} E\left(\frac{\partial U_{sub,i}^*}{\partial \beta^T}\right) & E\left(\frac{\partial U_{sub,i}^*}{\partial \vartheta^T}\right) \\ 0 & 0 \end{pmatrix} - \frac{\rho}{1+\rho} \begin{pmatrix} 0 & 0 \\ 0 & E\left(\frac{\partial S_{i\vartheta}^\vartheta}{\partial \vartheta^T}\right) \end{pmatrix}, \text{ and} \\ B_{sub,E} &= \frac{1}{1+\rho} \begin{pmatrix} E(U_{sub,i}^* U_{sub,i}^{*T}) & 0 \\ 0 & 0 \end{pmatrix} + \frac{\rho}{1+\rho} \begin{pmatrix} 0 & 0 \\ 0 & E(S_{i\vartheta}^\vartheta S_{i\vartheta}^{\vartheta T}) \end{pmatrix}. \end{aligned}$$

Analogously, inference procedures may be developed using the expectation correction method and the corrected score method. For the expectation correction method, we replace $U_{sub,i}^*(\beta, \vartheta)$ with $U_{exp,i}^*(\beta, \vartheta)$ where $U_{exp,i}^*$ is given by (14). For the corrected score method, we replace $U_{sub,i}^*(\beta, \vartheta)$ with $U_{cor,i}^*(\beta, \vartheta)$ where $U_{cor,i}^*$ is given by (17). Other relevant quantities are modified accordingly. We let $\hat{\theta}_{exp,I}$, $\hat{\theta}_{exp,E}$, $\hat{\theta}_{cor,I}$, and $\hat{\theta}_{cor,E}$ denote the corresponding estimators.

4.3 Discussion and extension

Sections 4.1 and 4.2 describe four methods of accommodating the misclassification effects into estimation procedures, which end up solving estimating equations. Since those equations have no analytic solutions, we must use an iterative algorithm to find the solutions. In our following numerical studies, we employ the Newton-Raphson algorithm to obtain the estimates of the parameters.

In the foregoing development, the prevalence $\tilde{\pi} = P(X = 1|Z)$ is a constant, i.e. $P(X = 1|Z) = P(X = 1) = \tilde{\pi}$, implying that X and Z are uncorrelated. When X and Z are correlated, $\tilde{\pi}$ will be a function of Z and we let $\tilde{\pi}(Z) = P(X = 1|Z) = h(Z; \alpha)$, where $h(\cdot)$ is a specified link function, and α is a vector of parameters. A common choice of this sort of parametric regression model is the logistic regression model. Then, the conditional distribution of X given Z is given by

$$f_{X|Z}(x|z) = \tilde{\pi}(z)^x \{1 - \tilde{\pi}(z)\}^{(1-x)} = \{h(z; \alpha)\}^x \{1 - h(z; \alpha)\}^{(1-x)},$$

and in the preceding derivations, $\tilde{\pi}$ is replaced by α in the parameter vector ϑ and relevant quantities are modified accordingly.

5 Illustrative examples

In this section, we apply the proposed methods to analyze three epidemiologic studies. There are two objectives of these analyses: (1) using different analytic methods to analyze the same data may help us understand the nature of the studies more comprehensively than applying a single method; and (2) such analyses illustrate the usage of the proposed methods.

5.1 Calcium intake in relation to Distal Colon Cancer in the Nurses' Health Study

First, we analyze data from the Nurses' Health Study (NHS) [30]. One of the objectives of the study is to understand how calcium intake is related to distal colon cancer incidence. The study consisted of 72011 female nurses who were free of cancer in 1984, and were followed up to 2012 for distal colon cancer occurrence, during which time 345 individuals developed distal colon cancer. Let $Y = 1$ if a participant has distal colon cancer, and 0 otherwise.

At the entry, calcium intake was assessed through a food frequency questionnaire (FFQ). The FFQ measures dietary intake with some error and more accurate information can be obtained from a diet record (DR) ([30], Ch. 6). Let X represent the accurate measurement of calcium intake obtained from the DR method and let X^* be the surrogate measurement of calcium intake obtained from the FFQ. In epidemiology and clinical research, it is standard practice to dichotomize or otherwise categorize continuous variables. Measures of association expressed in terms of regression slopes and related quantities are difficult for non-technical audiences to interpret, and do not lend themselves to clinical decision making, where patients need to be classified as either normal or abnormal. Thus, as in the original publication by Wu et al. ([43], Table 4), we group calcium intake as a binary variable defined as 'high Ca' if calcium intake was greater than 700 mg/day and 0 otherwise. Specifically, X and X^* take value 1 if the total calcium intake was greater than 700 mg/day, and 0 otherwise. The prevalence of high calcium intake was measured by FFQ 44%, i.e. $P(X^* = 1) \approx 0.44$. In this study, X and X^* were measured in 191 subjects, and among the rest only X^* was measured. In this validation subsample, the specificity and sensitivity were estimated as $\hat{p}_{00} = 0.60$ and $\hat{p}_{11} = 0.88$, respectively ([30], pp. 122–126).

In addition to including the error-prone covariate, X , we also adjusted for the effects of baseline body mass index (BMI) (in kg/m²) and baseline aspirin use, covariates which are assumed to be error-free. BMI was expressed in terms of the following categories: missing, BMI < 22, 22 < BMI < 25, 25 < BMI < 30, and 30 < BMI; and four dummy variables, Z_1, Z_2, Z_3 and Z_4 , were used to express BMI. Aspirin use was coded as yes (1) or no (0), and is denoted as Z_5 .

We assumed a logistic regression model for the relationship between the outcome and covariates:

$$P(Y = 1|X, Z_1, \dots, Z_5) = \frac{\exp(\beta_0 + \beta_x X + \sum_{i=1}^5 \beta_{z_i} Z_i)}{1 + \exp(\beta_0 + \beta_x X + \sum_{i=1}^5 \beta_{z_i} Z_i)},$$

where $\beta_0, \beta_x, \beta_{z_1}, \dots, \beta_{z_5}$ are the regression parameters.

We applied the four methods discussed in Section 4 to correct for the bias due to misclassification in the estimated effect of total calcium intake. As a comparison, we also reported the results obtained from the naive

method that disregards the difference between X^* and X . We report the estimate of odds ratio (OR) together with a 95% confidence interval, the estimate of β_x and its associated standard error, and the p-value. These results are displayed in the top panel of Table 3.

Table 3: Analysis results for three studies of diet on health using the proposed methods and the naive method.

Study	Outcome (# of cases)	Prevalence		Specificity (%)	$\hat{\beta}_x$ (S.E.) OR (95% C.I.) p-value				
		(%)	(%)		Naive Method	Induced Likelihood	Subtraction Correction	Expectation Correction	Corrected Score Method
NHS									
Total calcium intake (≤ 700 mg/day v.s. > 700 mg/day)	Distal colon cancer (345)	44	88	60	-0.265 (0.117) 0.77 (0.61,0.96) 0.02	-0.507 (0.234) 0.60 (0.38,0.95) 0.03	-0.522 (0.239) 0.59 (0.37,0.95) 0.03	-0.507 (0.231) 0.60 (0.38,0.94) 0.03	NA ^a
HPFS									
Low-carbohydrate diet score (≤ 15 v.s. > 15)	Type II diabetes (2666)	52	68	69	0.174 (0.043) 1.19 (1.09,1.29) < 0.001	NA ^a	0.539 (0.194) 1.71 (1.17,2.51) 0.005	0.624 (0.241) 1.87 (1.16,2.99) 0.01	0.762 (0.535) 2.14 (0.75,6.11) 0.15
HPFS									
Total red meat intake (≤ 2 servings/week v.s. > 2 servings/week)	Colorectal cancer (1281)	93	84	85	0.273 (0.133) 1.31 (1.01,1.71) 0.04	1.014 (0.857) 2.76 (0.51,14.81) 0.24	1.047 (0.887) 2.85 (0.50,16.20) 0.24	0.998 (0.836) 2.71 (0.53,13.96) 0.23	-0.116 (0.064) 0.89 (0.78,1.01) 0.07

^aDid not converge.

The naive method yielded a smaller estimate of β_x and a higher OR than those obtained by the induced likelihood method, the subtraction correction method, and the expectation correction method. As expected, the standard error for the naive estimate is smaller than those for the estimates from the correction methods. The results obtained from the three correction methods are fairly similar. All these methods suggested that low calcium intake is associated with a significantly higher risk of developing distal colon cancer. Interestingly, the corrected score method did not yield meaningful estimates due to non-convergence, although different initial values were tried for the numerical optimization routine used. This phenomenon has been further investigated in Section 6 via simulation studies, which showed that the non-convergence rate of the corrected score method was relatively high if the misclassification probabilities were high.

5.2 Carbohydrate intake in relation to Type II diabetes (T2D) in the health professionals follow-up study

In this subsection, we investigate the effect of carbohydrate quality on the risk of type 2 diabetes (T2D) [31]. 40507 male health professionals, who were free of T2D, cardiovascular disease, and cancer in 1986, were followed to 2006 for T2D incidence. During the follow-up period, 2666 individuals developed T2D. Let $Y = 1$ if a participant developed T2D and 0 otherwise.

Baseline diet was assessed through the FFQ and the low-carbohydrate diet score was developed according to [44]. Consistent with [31], we defined a binary low-carbohydrate diet score, which equaled 1 if the score was greater than 15 and 0 otherwise, where the cutoff, 15, is the median of the low-carbohydrate diet score; [31] reported their primary findings as a dichotomous variable also, contrasting the extreme quintiles of the distribution. Let X denote the binary low-carbohydrate diet score measured by DR in the validation study, and let X^* be the score measured by the FFQ method.

We estimated the misclassification probabilities among the 127 participants in the HPFS dietary validation study [45]. The specificity was estimated as $\hat{p}_{00} = 0.69$, and the sensitivity was estimated as $\hat{p}_{11} = 0.68$. The prevalence of high low-carbohydrate diet score was estimated about 52% in the main study.

Since BMI was found to be the most important confounder in the analysis by [31], we set Z to be BMI (kg/m^2). We fit the logistic model to the data:

$$P(Y = 1|X, Z) = \frac{\exp(\beta_0 + \beta_x X + \beta_z Z)}{1 + \exp(\beta_0 + \beta_x X + \beta_z Z)},$$

where β_0 , β_x and β_z are regression parameters.

We applied the four methods discussed in Section 4 to correct for the bias due to misclassification in the estimated effect of total calcium intake, and compared it to the results obtained from the naive method which disregards the difference between X^* and X .

These results are displayed in the middle panel of Table 3. The naive method yielded smaller estimates of both β_x and the odds ratio than those produced from the subtraction correction method, the expectation correction method, and the correct score method. As expected, the standard error produced by the naive method is smaller than those produced from the three correction methods. The three bias correction methods produced relatively similar estimates, but quite different standard errors. In particular, the standard errors obtained from the corrected score method were substantially larger than those from the subtraction correction method and the expectation correction method. The naive method, the subtraction correction method and the expectation correction method all suggested that the low-carbohydrate diet score is significantly associated with increased risk of T2D, but the evidence varies. In this study, the induced likelihood method was unable to yield meaningful estimates due to non-convergence, although different initial values were tried. This phenomenon has been further investigated in Section 6 by simulation, where it was found that the non-convergence rate of the induced likelihood method was the highest (about 44%) of the four methods when misclassification probabilities were high, as in this example.

5.3 Red meat intake in relation to colorectal cancer incidence in the health professionals follow-up study

In this example, we applied the methods to a study of red meat intake in relation to colorectal cancer risk in the Health Professionals Follow-up study (HPFS) [46]. 49980 male health professionals who were free of cancer in 1986 were followed up to 2010 for colorectal cancer incidence. During this study period, 1281 individuals developed colorectal cancer. Let $Y = 1$ if a participant has colorectal cancer and 0 otherwise.

Red meat intake at the baseline was assessed by the FFQ. We defined a binary red meat intake indicator, which was 1 if red meat intake was greater than 2 servings/week and 0 otherwise, where 2 servings/week represented a moderate consumption of red meat among HPFS participants [46]. Let X denote the binary red meat intake indicator obtained from the DR, and let X^* denote the measurements obtained from the FFQ. The prevalence of high red meat intake was around 93% in HPFS. As in the T2D example, we estimated the misclassification probabilities among the 127 subjects from the HPFS validation study [45]. The estimated specificity was $\hat{p}_{00} = 0.85$ and the estimated sensitivity was $\hat{p}_{11} = 0.84$.

As previously, we set Z to be BMI, a potential confounder in this association [46]. The logistic model was fitted to the data:

$$P(Y = 1|X, Z) = \frac{\exp(\beta_0 + \beta_x X + \beta_z Z)}{1 + \exp(\beta_0 + \beta_x X + \beta_z Z)},$$

where β_0 , β_x and β_z are the regression parameters.

We applied the four methods discussed in Section 4 to correct for misclassification in red meat intake. We also reported the results obtained from the naive method which disregarded the difference between X^* and X .

The results are displayed in the bottom panel of Table 3. The naive method yielded smaller estimates of the effect than those produced from the induced likelihood method, the subtraction correction method, and the expectation correction method. As expected, the standard error produced by the naive method was smaller than those produced by the three correction methods. The three correction methods produce relatively similar estimates and standard errors. After correction for bias due to exposure misclassification, no significant effects are evidenced. However, the naive method indicated moderate evidence for an increase of colorectal cancer risk in relation to red meat intake. In this study, results from the corrected score method were quite different than those from the other misclassification correction methods, likely due to instability in the performance of this method.

5.4 Summary and remarks

In this section, we applied the four correction methods to adjust for the bias in estimation and inference due to binary exposure misclassification. The misclassification probabilities were estimated from a validation study. It was demonstrated in these three studies that ignoring misclassification in the analysis would yield results that are considerably different from those obtained from the methods which account for misclassification effects.

Regarding the performance of the correction methods, the subtraction correction and expectation correction methods perform stably under situations with different degrees of misclassification, and they yielded very similar results. On the other hand, the performance of the induced likelihood and corrected score methods was unstable and affected by the magnitude of misclassification. Sometimes, the induced likelihood and corrected score methods failed due to non-convergence; and even when they did provide estimates and standard errors, those results may not have been reliable due to the apparent sensitivity of their performance to the severity of misclassification.

The four misclassification correction methods are valid from a theoretical viewpoint. They all yield consistent estimators with asymptotic normal distributions under standard regularity conditions as the sample size approaches infinity. However, their *finite* sample performance differs greatly. The three studies here offer us important insight into the choice of suitable methods for correcting for misclassification effects. To further assess the finite sample performance of the four misclassification correction methods, in the next section we carry out simulation studies for various scenarios.

6 Simulation studies

To further assess the finite sample performance of the various methods described in Section 4, in this section we report the results of simulation studies for a range of relevant scenarios for main study/validation study designs. Specifically, we assess the performance of the induced likelihood, the subtraction correction, the expectation correction, and the corrected score methods for three cases with known $(\tilde{\pi}, p_{00}, p_{11})$, a main study/internal validation, and a main study/external validation. For comparison purposes, we also apply the likelihood method to the simulated main study data where the true values of X_i are used.

We set the sample size to $n = 600$, and considered two scenarios where the validation sample size m takes $\rho = 20\%$ or 40% of the main study size. One thousand simulations were run for each parameter configuration.

6.1 Simulation with independent covariates

For subject i , covariate Z_i was generated independently from the standard normal distribution $N(0,1)$, and the true covariate X_i was independently generated from a Binomial distribution $BIN(1, \tilde{\pi})$ with $\tilde{\pi} = 0.5$. The surrogate measurement, X_i^* , was generated from the Bernoulli distribution with the (mis)classification probabilities $(p_{11}, p_{00}) = (0.9, 0.9)$, $(0.8, 0.8)$, or $(0.7, 0.7)$, given the generated value of X_i . The response measurement, Y_i , was simulated from the logistic regression model

$$\text{logit } P(Y_i = 1|X_i, Z_i) = \beta_0 + \beta_x X_i + \beta_z Z_i,$$

where $\beta_0 = \log 1.2$, $\beta_x = \log 2.0$, and $\beta_z = \log 1.2$.

We report the average of the finite sample biases (BIAS), the average of model-based standard errors (SE), the empirical standard error (ESE), the coverage rate (CR) of 95% confidence intervals for the parameters β_x and β_z , and the non-convergence rates of the Newton-Raphson algorithm used to solve the equations.

Table 4 displays the results for the case with known $(\tilde{\pi}, p_{00}, p_{11})$. The result in Theorem of Section 3.5 is confirmed by the identical numerical results for the estimators $\hat{\beta}_{ind}$ and $\hat{\beta}_{exp}$.

Concerning the estimates for β_x , the methods all performed very well, especially when misclassification probabilities were not substantial. As misclassification probabilities increased, the corrected score method tended to incur higher variability; but it still yielded reasonably reliable results. As expected, compared to the likelihood method in the absence of error in X , all the error correction methods produced more variable estimates, and the variability became more substantial as the misclassification probabilities increased. On the other hand, all the methods gave rise to comparable results for the estimation of β_z , regardless of the degree of misclassification in X .

Table 4: Simulation studies for the proposed estimators: the parameters (π, p_{00}, p_{11}) are Known.

(p_{00}, p_{11})	Estimator †	Estimating β_x				Estimating β_z				Noncov(%)‡
		Bias	ESE	SE	MCP(%)	Bias	ESE	SE	MCP(%)	
	$\hat{\beta}_{true}$	−0.001	0.176	0.173	95.0	−0.005	0.084	0.089	95.8	0.0
(0.9, 0.9)	$\hat{\beta}_{ind}$	−0.004	0.219	0.219	95.3	−0.004	0.084	0.089	95.8	0.0
	$\hat{\beta}_{sub}$	−0.004	0.219	0.219	95.3	−0.004	0.084	0.089	95.8	0.0
	$\hat{\beta}_{exp}$	−0.004	0.219	0.219	95.3	−0.004	0.084	0.089	95.8	0.0
	$\hat{\beta}_{cor}$	−0.004	0.219	0.219	95.3	−0.004	0.084	0.089	95.6	0.0
(0.8, 0.8)	$\hat{\beta}_{ind}$	−0.005	0.292	0.293	95.2	−0.004	0.089	0.089	95.7	0.0
	$\hat{\beta}_{sub}$	−0.005	0.292	0.293	95.2	−0.004	0.089	0.089	95.7	0.0
	$\hat{\beta}_{exp}$	−0.005	0.292	0.293	95.2	−0.004	0.089	0.089	95.7	0.0
	$\hat{\beta}_{cor}$	0.000	0.295	0.295	95.3	−0.003	0.089	0.089	96.3	0.0
(0.7, 0.7)	$\hat{\beta}_{ind}$	−0.010	0.445	0.446	96.2	−0.003	0.089	0.089	95.8	0.0
	$\hat{\beta}_{sub}$	−0.010	0.446	0.446	95.8	−0.003	0.089	0.089	95.8	0.0
	$\hat{\beta}_{exp}$	−0.010	0.445	0.446	96.2	−0.003	0.089	0.089	95.8	0.0
	$\hat{\beta}_{cor}$	−0.003	0.456	0.457	96.5	−0.003	0.095	0.100	96.7	0.0

†: $\hat{\beta}_{true}$: true likelihood estimator; $\hat{\beta}_{ind}$: induced likelihood estimator; $\hat{\beta}_{sub}$: subtraction correction estimator; $\hat{\beta}_{cor}$: corrected score estimator.

‡: nonconvergence proportion

Table 5 and Table 6 present the results for main study/internal validation and main study/external validation scenarios, respectively. It is clear that in both designs, all the proposed methods successfully corrected for the bias due to misclassification, and produced consistent estimators. As expected, the performance of those methods deteriorated as misclassification became more serious or the validation study size decreased. Consistent with the patterns observed in the numerical studies in Section 3.5, the estimator $\hat{\theta}_{cor}$ was the least efficient among all the estimators. In the main study/internal validation case, the subtraction correction method was outperformed by the induced likelihood and by the expectation correction methods. But for the main study/external validation scenario, these three methods performed similarly, although sometimes, the subtraction correction method produced slightly better results than the induced likelihood and the expectation correction methods.

Finally, we observed that numerically, all the methods were reasonably stable except for the induced likelihood method. The induced likelihood method most frequently failed to converge, although the frequency of this was practically acceptable. This method was sensitive to the choice of initial values, while other methods were relatively insensitive. If the initial value is not wisely set, the induced likelihood method may not converge. We found that the induced likelihood method can output reliable results if an initial value is set to the estimate obtained from the naive method.

Table 5: Simulation studies for the proposed estimators: a main/internal validation study design.

(p_{00}, p_{11})	ρ	Estimator	Estimating β_x				Estimating β_z				Noncov(%)
			Bias	ESE	SE	MCP(%)	Bias	ESE	SE	MCP(%)	
		$\hat{\beta}_{true}$	-0.001	0.176	0.173	95.0	-0.005	0.084	0.089	95.8	0.0
(0.9,0.9)	0.2	$\hat{\beta}_{ind}$	-0.010	0.212	0.210	95.3	-0.003	0.084	0.089	95.9	3.1
		$\hat{\beta}_{sub}$	-0.002	0.230	0.221	95.5	-0.004	0.084	0.089	95.6	0.2
		$\hat{\beta}_{exp}$	-0.009	0.214	0.212	95.2	-0.004	0.084	0.089	95.6	0.1
		$\hat{\beta}_{cor}$	-0.004	0.224	0.217	95.4	-0.004	0.084	0.089	95.6	0.0
	0.4	$\hat{\beta}_{ind}$	-0.004	0.202	0.197	95.1	-0.004	0.084	0.089	95.6	0.0
		$\hat{\beta}_{sub}$	-0.005	0.202	0.197	94.8	-0.004	0.084	0.089	95.5	0.0
		$\hat{\beta}_{exp}$	-0.005	0.202	0.197	95.1	-0.004	0.084	0.089	95.3	0.0
		$\hat{\beta}_{cor}$	-0.004	0.205	0.202	94.8	-0.004	0.084	0.089	95.6	0.0
(0.8,0.8)	0.2	$\hat{\beta}_{ind}$	-0.009	0.255	0.257	95.2	-0.003	0.084	0.089	95.9	2.8
		$\hat{\beta}_{sub}$	-0.006	0.265	0.268	95.1	-0.004	0.084	0.089	95.8	0.0
		$\hat{\beta}_{exp}$	-0.014	0.253	0.257	95.3	-0.004	0.084	0.089	96.0	0.4
		$\hat{\beta}_{cor}$	0.009	0.292	0.302	95.3	-0.003	0.089	0.089	95.9	0.4
	0.4	$\hat{\beta}_{ind}$	-0.001	0.228	0.224	94.7	-0.004	0.084	0.089	95.9	0.0
		$\hat{\beta}_{sub}$	-0.002	0.235	0.232	94.6	-0.004	0.084	0.089	95.8	0.0
		$\hat{\beta}_{exp}$	-0.003	0.226	0.224	94.7	-0.004	0.084	0.089	95.6	0.0
		$\hat{\beta}_{cor}$	0.004	0.257	0.261	94.5	-0.003	0.089	0.089	95.6	0.0
(0.7,0.7)	0.2	$\hat{\beta}_{ind}$	-0.016	0.322	0.313	94.5	-0.004	0.084	0.089	95.7	2.4
		$\hat{\beta}_{sub}$	-0.007	0.368	0.368	95.5	-0.004	0.084	0.089	95.4	0.0
		$\hat{\beta}_{exp}$	-0.026	0.316	0.313	94.7	-0.004	0.084	0.089	95.5	0.7
		$\hat{\beta}_{cor}$	0.020	0.455	0.542	97.0	-0.002	0.095	0.105	96.4	1.6
	0.4	$\hat{\beta}_{ind}$	-0.001	0.255	0.249	95.4	-0.004	0.084	0.089	95.7	0.0
		$\hat{\beta}_{sub}$	-0.003	0.283	0.279	95.1	-0.004	0.084	0.089	95.7	0.0
		$\hat{\beta}_{exp}$	-0.004	0.255	0.249	95.5	-0.004	0.084	0.089	95.6	0.0
		$\hat{\beta}_{cor}$	0.017	0.395	0.407	96.5	-0.003	0.089	0.095	96.2	0.0

The data were generated in the same manner as that for Table 4, except that the main study/internal validation study design is considered here where the size of the main study is 600, and the size of the validation study is set as 120 for $\rho = 20\%$ or 240 for $\rho = 40\%$.

Table 6: Simulation studies for the proposed estimators: a main/external validation study design.

(p_{00}, p_{11})	ρ	Estimator	Estimating β_x				Estimating β_z				Noncov(%)
			Bias	ESE	SE	MCP(%)	Bias	ESE	SE	MCP(%)	
		$\hat{\beta}_{true}$	-0.001	0.176	0.173	95.0	-0.005	0.084	0.089	95.8	0.0
(0.9,0.9)	0.2	$\hat{\beta}_{ind}$	-0.004	0.230	0.226	94.9	-0.002	0.089	0.089	95.8	0.5
		$\hat{\beta}_{sub}$	-0.004	0.230	0.226	94.9	-0.002	0.089	0.089	95.3	0.5
		$\hat{\beta}_{exp}$	-0.004	0.230	0.226	94.9	-0.002	0.089	0.089	95.3	0.5
		$\hat{\beta}_{cor}$	-0.002	0.237	0.230	94.5	-0.001	0.089	0.089	95.6	0.0
	0.4	$\hat{\beta}_{ind}$	-0.001	0.224	0.221	95.8	-0.006	0.089	0.089	95.4	0.0
		$\hat{\beta}_{sub}$	-0.001	0.224	0.221	95.9	-0.006	0.089	0.089	95.2	0.0
		$\hat{\beta}_{exp}$	-0.001	0.224	0.221	95.9	-0.006	0.089	0.089	95.2	0.0
		$\hat{\beta}_{cor}$	0.002	0.226	0.224	95.9	-0.005	0.089	0.089	95.2	0.0

(0.8,0.8)	0.2	$\hat{\beta}_{ind}$	0.005	0.321	0.324	95.1	-0.001	0.089	0.089	95.8	0.0
		$\hat{\beta}_{sub}$	0.005	0.321	0.322	95.1	-0.001	0.089	0.089	95.6	0.0
		$\hat{\beta}_{exp}$	0.005	0.321	0.324	95.1	-0.001	0.089	0.089	95.6	0.0
		$\hat{\beta}_{cor}$	0.026	0.341	0.358	95.7	0.000	0.089	0.095	95.6	0.0
	0.4	$\hat{\beta}_{ind}$	0.003	0.300	0.307	95.9	-0.005	0.089	0.089	95.2	0.0
		$\hat{\beta}_{sub}$	0.003	0.300	0.307	95.9	-0.005	0.089	0.089	95.2	0.0
		$\hat{\beta}_{exp}$	0.003	0.300	0.307	95.9	-0.005	0.089	0.089	95.2	0.0
		$\hat{\beta}_{cor}$	0.015	0.308	0.316	96.1	-0.004	0.089	0.089	95.8	0.0
(0.7,0.7)	0.2	$\hat{\beta}_{ind}$	0.012	0.503	0.555	96.6	0.001	0.089	0.089	95.5	1.2
		$\hat{\beta}_{sub}$	0.019	0.512	0.562	96.6	0.001	0.089	0.089	95.4	0.6
		$\hat{\beta}_{exp}$	0.022	0.516	0.568	96.6	0.001	0.089	0.089	95.5	0.5
		$\hat{\beta}_{cor}$	0.042	0.521	0.654	97.0	0.002	0.100	0.118	97.3	4.0
	0.4	$\hat{\beta}_{ind}$	0.010	0.488	0.499	96.9	-0.004	0.089	0.089	95.7	0.0
		$\hat{\beta}_{sub}$	0.010	0.487	0.499	97.1	-0.004	0.089	0.089	95.7	0.0
		$\hat{\beta}_{exp}$	0.010	0.487	0.498	96.9	-0.004	0.089	0.089	95.7	0.0
		$\hat{\beta}_{cor}$	0.037	0.506	0.548	97.5	0.000	0.100	0.110	97.0	1.0

The data were generated in the same manner as that for Table 4, except that the main study/external validation study design is considered here where the size of the main study is 600, and the size of the validation study is set as 120 for $\rho = 20\%$ or 240 for $\rho = 40\%$.

6.2 Simulation with dependent covariates

To further evaluate the finite sample performance of the proposed methods, we considered more extreme scenarios where the misclassification probabilities were increased and the ratio of the external validation sample to the main study was reduced to $\rho = 10\%$. Here, the sample size of the main study was set to 1000. In addition, we incorporated an association between the covariates using the model

$$\text{logit } P(X_i = 1|Z_i) = \alpha_0 + \alpha_z Z_i$$

with $(\alpha_0, \alpha_z) = (1, 0.5)$. All other data generation procedures were the same as for Table 6. We considered more extreme values for p_{00} and p_{11} , with the specificity and sensitivity equal to (0.6,0.8) or (0.8,0.6), resembling the data in Section 5.

The results of those simulations are given in Table 7. Similar patterns to those observed in Section 6.1 were found. However, in terms of the numerical stability of the proposed methods, both the induced likelihood method and the correct score method were found to be unstable with higher frequency of failure to converge, especially when the sensitivity and specificity were low. The expectation correction and the subtraction correction methods were more stable, although unsurprisingly, their performance deteriorated as misclassification became more substantial.

Table 7: Simulation studies for the proposed estimators: a main/external validation study design with correlated covariates.

(p_{00}, p_{11})	$c\rho$	Estimator	Estimating β_x				Estimating β_z				Noncov(%)
			Bias	ESE	SE	MCP(%)	Bias	ESE	SE	MCP(%)	
		$\hat{\beta}_{true}$	0.007	0.146	0.149	95.4	-0.002	0.070	0.070	95.1	0.0
(0.8,0.8)	0.1	$\hat{\beta}_{ind}$	0.015	0.301	0.316	96.2	-0.004	0.080	0.082	96.0	2.3
		$\hat{\beta}_{sub}$	0.027	0.312	0.326	96.5	-0.006	0.082	0.084	96.0	0.7
		$\hat{\beta}_{exp}$	0.020	0.309	0.320	96.3	-0.005	0.081	0.083	96.0	0.7
		$\hat{\beta}_{cor}$	0.041	0.346	0.432	95.2	-0.010	0.087	0.097	96.1	1.6
(0.6,0.8)		$\hat{\beta}_{ind}$	-0.041	0.423	0.497	96.4	0.003	0.087	0.095	96.6	9.4
		$\hat{\beta}_{sub}$	0.038	0.476	0.569	96.8	-0.006	0.095	0.104	96.8	1.7
		$\hat{\beta}_{exp}$	0.031	0.480	0.556	96.6	-0.005	0.094	0.101	96.9	1.2
		$\hat{\beta}_{cor}$	0.018	0.455	0.692	96.1	-0.008	0.099	0.148	96.7	10.8

(0.8,0.6)	$\hat{\beta}_{ind}$	-0.146	0.398	0.472	95.9	0.012	0.087	0.091	94.6	23.7
	$\hat{\beta}_{sub}$	0.055	0.517	0.593	97.2	-0.008	0.099	0.108	96.5	1.7
	$\hat{\beta}_{exp}$	0.003	0.467	0.527	96.6	-0.002	0.095	0.098	95.8	4.6
	$\hat{\beta}_{cor}$	-0.025	0.473	0.716	96.1	0.001	0.095	0.136	96.1	17.7
(0.6,0.6)	$\hat{\beta}_{ind}$	-0.441	0.620	0.896	96.9	0.040	0.110	0.128	95.1	44.5
	$\hat{\beta}_{sub}$	-0.168	0.688	1.052	99.1	0.018	0.104	0.143	97.2	24.3
	$\hat{\beta}_{exp}$	-0.284	0.729	0.990	97.8	0.026	0.113	0.136	96.3	26.2
	$\hat{\beta}_{cor}$	-0.333	0.582	1.022	95.3	0.038	0.103	0.221	96.0	40.2

Simulation setting: $\text{logit } P(X_i = 1|Z_i) = \alpha_0 + \alpha_z Z_i$ with $(\alpha_0, \alpha_z) = (1, 0.5)$. All other data generation procedures were the same as for Table 6. The sample size of the main study and the external study is 1000 and 100, respectively.

More simulation studies were conducted to compare the performance of the methods, and similar patterns are observed. Those results are included in the Supplementary Material.

7 Discussion

Measurement error and misclassification in covariates commonly arises in practice. Ignoring this feature in data analysis would typically yield biased point and interval estimates. Many methods have been proposed to correct for measurement error and misclassification effects for various settings. In particular, with continuous covariates subject to measurement error, methods, such as the induced likelihood, the expectation correction method, the corrected score method, and their extensions, have been developed by various authors for different models and data designs. For instance, the induced likelihood method was explored by [32] for the Cox proportional hazards model with a mis-measured continuous covariate, X . [32] considered main study/internal validation and main study/external validation designs as well as the case with replicate measurement data, where the distribution of the true covariate X is required to be modeled. The corrected score method was proposed by [36] for conducting valid inference for generalized linear measurement error models, where the measurement error models are assumed known. The expectation correction method was developed by [35] to accommodate measurement error effects in continuous covariates, where replicate surrogate measurements were available.

Although these methods have been generalized and applied to handle a variety of problems with error-contaminated continuous covariates, there has been limited work on adapting these strategies to validly analyze data with misclassification in discrete covariates.

In this paper, we have systematically explored valid inferential procedures to account for misclassification for settings where: (1) the misclassification parameters are known, (2) the misclassification parameters are unknown and a main study/internal validation design is available, and (3) the misclassification parameters are unknown and a main study/external validation design is available. The development takes advantage of the unique features of discrete variables, which enables us to address misclassification effects under a unified model setup, and to evaluate the relative performance of various methods, especially their *finite sample performance*, so that we are able to understand the differences between commonly discussed methods.

Our investigations suggest that the expectation correction method performs best under various settings, from both asymptotic and finite sample perspectives. This method was the most efficient and stable. When the misclassification probabilities are known and mild regularity conditions hold, this method is identical to the induced likelihood method. However, when the misclassification probabilities are estimated as required in main study/internal validation and the main study/external validation designs, this method outperformed the induced likelihood method. The induced likelihood method was, on the contrary, sensitive to initial values, and its finite sample performance was sometimes unstable. On the other hand, the subtraction correction method performed nearly as well as the expectation correction method. These three methods require the same model assumptions and they are structural methods that require estimates of the conditional probabilities of X given Z . In contrast, the corrected score method is a functional strategy that requires no model assumption for the conditional distribution of X given Z . However, when relevant models are correctly specified, the corrected score method had the least ideal performance. In the case where the parameters for the misclassification model are known, this method suffers substantial loss of efficiency. When the parameters for the misclassification model are unknown and estimated from the main study/internal validation or the main study/external validation design, the corrected score method does not always yield reliable results, and its finite sample performance had the highest frequency of failure to converge among all the methods considered.

Our methods have been applied to analyze three nutritional epidemiologic studies, where food frequency questionnaires were used to measure food intake in the main study and validation data were available. The results suggested that ignoring misclassification gave rise, in some cases, to conclusions quite different from those obtained from methods that adjust for bias due to misclassification.

Acknowledgements

The authors thank the referees for their comments on the initial version. The research of Yi and Yan was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). Spiegelman's research was supported by grants from NIH/NIEHS (R01 ES 09411) and NIH/NCI (R01 CA050597).

Appendix

A

As in the main text, let $f(y, x, x^*, z)$ denote the joint probability density or mass function for the complete set of random variables $\{Y, X, X^*, Z\}$, and $f(y, x^*, z)$ denote the joint probability density or mass function for the "observed" set of random variables $\{Y, X^*, Z\}$. Let $f_{Y|X,Z}(y|x, z)$ and $f_{Y|X^*,Z}(y|x^*, z)$ denote the conditional probability density or mass function of Y given $\{X, Z\}$ and of Y given $\{X^*, Z\}$, respectively. Let β be the parameters associated with $f_{Y|X,Z}(y|x, z)$ which is of prime interest.

Assume that the nondifferential misclassification mechanism (5) is true and that the parameters associated with $f_{Y|X,Z}(y|x, z)$ are distinct from the parameters governing other models. Then we have

$$(\partial/\partial\beta) \log f_{Y|X^*,Z}(y|x^*, z) = (\partial/\partial\beta) \log f(y, x^*, z) \quad (24)$$

and

$$(\partial/\partial\beta) \log f_{Y|X,Z}(y|x, z) = (\partial/\partial\beta) \log f(y, x, x^*, z), \quad (25)$$

which are, respectively, denoted as $S_{Y|X^*,Z}^*(\beta; y, x^*, z)$ and $S_{Y|X,Z}(\beta; y, x, z)$. That is, $S_{Y|X^*,Z}^*(\beta; y, x^*, z)$ represents the (conditional) score function based on the observed data $\{Y, X^*, Z\}$, and $S_{Y|X,Z}(\beta; y, x, z)$ is the (conditional) score function based on the full data $\{Y, X, X^*, Z\}$.

We now show that

$$S_{Y|X^*,Z}^*(\beta; Y, X^*, Z) = E\{S_{Y|X,Z}(\beta; Y, X, Z) | Y, X^*, Z\},$$

which is justified as follows:

$$\begin{aligned} & E\{S_{Y|X,Z}(\beta; Y, X, Z) | Y, X^*, Z\} \\ &= \int S_{Y|X,Z}(\beta; Y, x, Z) f_{X|Y,X^*,Z}(x | Y, X^*, Z) d\eta(x) \\ &= \int (\partial/\partial\beta) \log f(Y, x, X^*, Z) \cdot \frac{f(Y, x, X^*, Z)}{f(Y, X^*, Z)} d\eta(x) \\ &= \int \frac{\partial f(Y, x, X^*, Z)}{\partial\beta} \cdot \frac{1}{f(Y, X^*, Z)} d\eta(x) \\ &= \frac{1}{f(Y, X^*, Z)} \int \frac{\partial f(Y, x, X^*, Z)}{\partial\beta} d\eta(x) \\ &= \frac{1}{f(Y, X^*, Z)} \cdot \left\{ \frac{\partial}{\partial\beta} \int f(Y, x, X^*, Z) d\eta(x) \right\} \\ &= \frac{1}{f(Y, X^*, Z)} \cdot \left\{ \frac{\partial}{\partial\beta} f(Y, X^*, Z) \right\} \\ &= S_{Y|X^*,Z}^*(\beta; Y, X^*, Z), \end{aligned}$$

where the second identity is due to (25), the fifth identity assumes that the order of integration and differentiation can be exchanged, and the last step is due to (24). Here $d\eta(x)$ stands for the Lebesgue measure if X is continuous or the counting measure if X is discrete; in the setting we consider, $d\eta(x)$ is the counting measure.

By (11), we obtain that

$$U_{exp}^*(\beta; Y, X^*, Z) = S_{Y|X^*, Z}^*(\beta; Y, X^*, Z).$$

That is, under certain regularity conditions, $U_{exp}^*(\beta; Y, X^*, Z)$ is the score function of the induced likelihood method. Therefore, when the nuisance parameter ϑ is known, the induced likelihood and the expectation correction methods produce identical results for inference about the β parameter.

However, if the value of the nuisance parameter ϑ is unknown but is estimated from a validation sample, the estimates obtained from induced likelihood and the expectation correction methods may differ. This can be clearly seen by comparing the corresponding estimating functions for the main study/internal validation and the main study/external validation cases discussed in Section 4.

B

Under the logistic model (19), the induced likelihood (6) becomes

$$\begin{aligned} f_{Y|X^*, Z}(y|X^*, Z) &= \frac{\exp\{(\beta_0 + \beta_z^T Z)y\}}{1 + \exp(\beta_0 + \beta_z^T Z)} \cdot p_{00}^{*(1-X^*)} (1 - p_{11}^*)^{X^*} \\ &\quad + \frac{\exp\{(\beta_0 + \beta_x + \beta_z^T Z)y\}}{1 + \exp(\beta_0 + \beta_x + \beta_z^T Z)} \cdot (1 - p_{00}^*)^{1-X^*} p_{11}^{*X^*}. \end{aligned}$$

Then the expected information matrix (7) is given by

$$I^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} S_{Y|X^*, Z}^*(\beta; y, x^*, z) S_{Y|X^*, Z}^{*T}(\beta; y, x^*, z) \cdot f(y, x^*, z) d\eta(z),$$

where $S_{Y|X^*, Z}^*(\beta; y, x^*, z) = (\partial/\partial\beta) \log f_{Y|X^*, Z}(y|x^*, z)$.

To calculate $\mathcal{J}_{sub}$, we first calculate $U_{sub}^*(\beta; y, x^*, z)$ using (9):

$$U_{sub}^*(\beta; y, x^*, z) = S_{Y|X^*, Z}(\beta; y, x^*, z) - \sum_{y=0,1} S_{Y|X^*, Z}(\beta; y, x^*, z) f_{Y|X^*, Z}(y|x^*, z).$$

Then we compute $I_{sub}^*(\beta)$ and $J_{sub}^*(\beta)$, which are given by

$$I_{sub}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} U_{sub}^*(\beta; y, x^*, z) U_{sub}^{*T}(\beta; y, x^*, z) \cdot f(y, x^*, z) d\eta(z)$$

and

$$J_{sub}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} (\partial/\partial\beta) U_{sub}^*(\beta; y, X^*, Z) \cdot f(y, x^*, z) d\eta(z),$$

respectively. Therefore, $\mathcal{J}_{sub}$ is obtained from (10).

To calculate $\mathcal{J}_{exp}$ and $\mathcal{J}_{cor}$, first we note that the score function under the true model (19) is given by

$$S_{Y|X, Z}(\beta; y, X, Z) = \begin{pmatrix} 1 \\ X \\ Z \end{pmatrix} \cdot \left\{ y - \frac{\exp(\beta_0 + \beta_x X + \beta_z^T Z)}{1 + \exp(\beta_0 + \beta_x X + \beta_z^T Z)} \right\}.$$

Then

$$I_{exp}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} U_{exp}^*(\beta; y, x^*, z) U_{exp}^{*T}(\beta; y, x^*, z) \cdot f(y, x^*, z) d\eta(z),$$

$$J_{exp}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} (\partial/\partial\beta) U_{exp}^*(\beta; y, x^*, z) \cdot f(y, x^*, z) d\eta(z),$$

$$I_{cor}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} U_{cor}^*(\beta; y, X^*, Z) U_{cor}^{*T}(\beta; y, X^*, Z) \cdot f(y, x^*, z) d\eta(z),$$

and

$$J_{cor}^*(\beta) = \int \sum_{x^*=0,1} \sum_{y=0,1} (\partial/\partial\beta) U_{cor}^*(\beta; y, x^*, z) \cdot f(y, x^*, z) d\eta(z),$$

where $U_{exp}^*(\beta; y, x^*, z)$ is given by (14), and $U_{cor}^*(\beta; y, x^*, z)$ is determined by (17). Consequently, $\mathcal{J}_{exp}$ is obtained from (15), and $\mathcal{J}_{cor}$ is calculated from (18).

C

Let $\delta_i = I(i \in \mathcal{V})$. Solving (21) is equivalent to solving

$$\sum_{i \in \mathcal{M} \setminus \mathcal{V}} \begin{pmatrix} \partial \log L_i^* / \partial \beta \\ \partial \log L_i^* / \partial \theta \end{pmatrix} + \sum_{i \in \mathcal{V}} \begin{pmatrix} \partial \log L_i / \partial \beta \\ \partial \log L_i^\theta / \partial \theta \end{pmatrix} = 0,$$

which can be written as

$$\sum_{i \in \mathcal{M}} \begin{pmatrix} (1 - \delta_i) S_{i\beta}^* + \delta_i S_{i\beta} \\ (1 - \delta_i) S_{i\theta}^* + \delta_i S_{i\theta}^\theta \end{pmatrix} = 0. \quad (26)$$

Let $H_{i\beta} = (1 - \delta_i) S_{i\beta}^* + \delta_i S_{i\beta}$, $H_{i\theta} = (1 - \delta_i) S_{i\theta}^* + \delta_i S_{i\theta}^\theta$ and $H_{i\theta} = (H_{i\beta}^T, H_{i\theta}^T)^T$. Then solving (26) is equivalent to solving $\sum_{i \in \mathcal{M}} H_{i\theta} = 0$, which yields the estimator $\hat{\theta}_I$.

Applying the Taylor series expansion to $\sum_{i \in \mathcal{M}} H_{i\theta}(\hat{\theta}_I) = 0$ gives

$$\sum_{i \in \mathcal{M}} H_{i\theta} + \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^T} (\hat{\theta}_I - \theta) + o_p(1) = 0,$$

leading to

$$\sqrt{n}(\hat{\theta}_I - \theta) = \left(-\frac{1}{n} \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} H_{i\theta} + o_p(1). \quad (27)$$

Let

$$A_I = \lim_{n \rightarrow \infty} E \left(-\frac{1}{n} \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right) \quad \text{and} \quad B_I = \lim_{n \rightarrow \infty} \text{var} \left(\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} H_{i\theta} \right).$$

After calculations, A_I and B_I can be written as

$$A_I = -(1 - \rho) E \left(\frac{\partial S_{i\theta}^*}{\partial \theta^T} \right) - \rho \begin{pmatrix} E \left(\frac{\partial S_{i\beta}}{\partial \beta^T} \right) & 0 \\ 0 & E \left(\frac{\partial S_{i\theta}^\theta}{\partial \theta^T} \right) \end{pmatrix}$$

and

$$B_I = (1 - \rho) E(S_{i\theta}^* S_{i\theta}^{*T}) + \rho \begin{pmatrix} E(S_{i\beta} S_{i\beta}^T) & 0 \\ 0 & E(S_{i\theta}^\theta S_{i\theta}^{\theta T}) \end{pmatrix},$$

which implies, by the Bartlett identities, that $A_I = B_I$. Then applying the Central Limit Theorem to (27) gives that

$$\sqrt{n}(\hat{\theta}_I - \theta) \xrightarrow{d} N(0, A_I^{-1}) \quad \text{as } n \rightarrow \infty.$$

D

Solving (22) is equivalent to solving

$$\sum_{i \in \mathcal{M}} \begin{pmatrix} \partial \log L_i^* / \partial \beta \\ \partial \log L_i^* / \partial \theta \end{pmatrix} + \sum_{i \in \mathcal{V}} \begin{pmatrix} 0 \\ \partial \log L_i^\theta / \partial \theta \end{pmatrix} = 0,$$

which can be written as

$$\sum_{i \in \mathcal{M} \cup \mathcal{V}} \begin{pmatrix} (1 - \delta_i) S_{i\beta}^* \\ (1 - \delta_i) S_{i\theta}^* + \delta_i S_{i\theta}^\theta \end{pmatrix} = 0.$$

Using the same notation and arguments for (27), we now obtain

$$\sqrt{n+m}(\hat{\theta}_E - \theta) = \left(-\frac{1}{n+m} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right)^{-1} \frac{1}{\sqrt{n+m}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} H_{i\theta} + o_p(1).$$

Rewrite this as

$$\sqrt{\frac{m}{n} + 1} \sqrt{n}(\hat{\theta}_E - \theta) = \left(-\frac{1}{n+m} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right)^{-1} \frac{1}{\sqrt{n+m}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} H_{i\theta} + o_p(1), \quad (28)$$

and let

$$A_E = \lim_{n \rightarrow \infty} E \left(-\frac{1}{n+m} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right) \quad \text{and} \quad B_E = \lim_{n \rightarrow \infty} \text{var} \left(\frac{1}{\sqrt{n+m}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} H_{i\theta} \right).$$

Using the similar calculations to A_I and B_I give

$$A_E = -\frac{1}{(1+\rho)} E \left(\frac{\partial S_{i\theta}^*}{\partial \theta^T} \right) - \frac{\rho}{1+\rho} \begin{pmatrix} 0 & 0 \\ 0 & E \left(\frac{\partial S_{i\theta}^\theta}{\partial \theta^T} \right) \end{pmatrix},$$

and

$$B_E = \frac{1}{(1+\rho)} E(S_{i\theta}^* S_{i\theta}^{*T}) + \frac{\rho}{1+\rho} \begin{pmatrix} 0 & 0 \\ 0 & E(S_{i\theta}^\theta S_{i\theta}^{\theta T}) \end{pmatrix},$$

implying that $A_E = B_E$. Then applying the Central Limit Theorem to (28) gives that

$$\sqrt{n}(\hat{\theta}_E - \theta) \xrightarrow{d} N \left(0, \frac{1}{1+\rho} A_E^{-1} \right) \quad \text{as } n \rightarrow \infty.$$

References

- [1] Fuller WA. Measurement error models. New York: Wiley, 1987.
- [2] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models, 2nd ed. Boca Raton: Chapman & Hall/CRC, 2006.
- [3] Yi CY. Statistical analysis with measurement error or misclassification: strategy, method and application. New York: Springer Science+Business Media, LLC., 2017.
- [4] Stefanski LA, Carroll RJ. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*. 1987;74:703–16.
- [5] Rosner BA, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*. 1989;8:1051–70.
- [6] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol*. 1990;132:734–45.
- [7] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Am J Epidemiol*. 1992;136:1400–13.
- [8] Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Assoc*. 1994;89:1314–28.
- [9] Wang N, Davidian M. A note on covariate measurement error in nonlinear mixed models. *Biometrika*. 1996;83:801–12.
- [10] Lin X, Carroll RJ. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J Am Stat Assoc*. 2000;95:520–34.
- [11] Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J Am Stat Assoc*. 2000;95:51–61.
- [12] Huang Y, Wang CY. Consistent functional methods for logistic regression with error in covariates. *J Am Stat Assoc*. 2001;96:1469–82.
- [13] Liang H, Wang N. Partially linear single-index measurement error models. *Stat Sin*. 2005;15:99–116.
- [14] Spiegelman D, Zhao B, Kim J. Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Stat Med*. 2005;24:1657–82.
- [15] Zucker DM, Spiegelman D. Inference for the proportional hazards model with misclassified discrete-valued covariates. *Biometrics*. 2004;60:324–34.
- [16] Zucker DM, Spiegelman D. Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Stat Med*. 2008;27:1911–1933.
- [17] Sugar EA, Wang C-Y, Prentice RL. Logistic regression with exposure biomarkers and flexible measurement error. *Biometrics*. 2007;63:143–51.

- [18] Yi GY. A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*. 2008;9:501–12.
- [19] 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.
- [20] Yi GY, Ma Y, Carroll RJ. A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*. 2012;99:151–65.
- [21] Yi GY, Ma Y, Spiegelman D, Carroll RJ. Functional and structural methods with mixed measurement error and misclassification in covariates. *J Am Stat Assoc*. 2015;110:681–96.
- [22] Akazawa K, Kinukawa K, Nakamura T. A note on the corrected score function adjusting for misclassification. *Jpn Stat Soc*. 1998;28:115–23.
- [23] Buonaccorsi JP, Laake P, Veierød M. A note on the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*. 2005;61:831–6.
- [24] Dalen I, Buonaccorsi JP, Sexton J, Laake P, Thoresen M. Correcting for misclassification of a categorized exposure in binary regression using replication data. *Stat Med*. 2009;28:3368–410.
- [25] Gustafson P. Measurement error and misclassification in statistics and epidemiology. Chapman & Hall/CRC, 2004.
- [26] 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.
- [27] Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V. Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Stat Med*. 2008;27:2756–83.
- [28] Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. *Clin Epidemiol*. 1988;41:1167–74.
- [29] Spiegelman D, Gray R. Cost-efficient study designs for binary response data with generalized Gaussian measurement error in the covariate. *Biometrics*. 1991;47:851–70.
- [30] Willett WC. Nutritional epidemiology, 2nd ed. New York: Oxford University Press, 1998.
- [31] de Koning L, Fung TT, Liao X, Chiuve SE, Rimm EB, Willett WC, Spiegelman D, Hu FB. Low-carbohydrate diet scores and risk of type 2 diabetes in men. *Am J Clin Nutr*. 2011;93:844–50.
- [32] Zucker DM. A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *J Am Stat Assoc*. 2005;100:1264–77.
- [33] Yi GY, Reid N. A note on mis-specified estimating functions. *Stat Sin*. 2010;20:1749–69.
- [34] Yan Y, Yi GY. A Class of functional methods for error-contaminated survival data under additive hazards models with replicate measurements. *J Am Stat Assoc*. 2016;111:684–95.
- [35] Wang CY, Pepe MS. Expected estimating equations to accommodate covariate measurement error. *J Royal Stat Soc Ser B*. 2000;62:509–24.
- [36] Nakamura T. Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*. 1990;77:127–37.
- [37] Yi GY. Robust methods for incomplete longitudinal data with mismeasured covariates. *Far East J Theor Statist* 2005;16:205–34.
- [38] Yi GY, Lawless JF. Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error. *Can J Stat*. 2012;40:530–49.
- [39] Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs. *Stat Med*. 2001;29:139–60.
- [40] Thurston SW, Spiegelman D, Ruppert D. Equivalence of regression calibration methods for main study/external validation study designs. *J Stat Plann Inference*. 2003;113:527–39.
- [41] Hart JE, Liao X, Hong B, Puett RC, Yanosky JD, Suh H, Kioumourtoglou MA, Spiegelman D, Laden F. The association of long-term exposure to PM_{2.5} on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health* 2015;14:38.
- [42] Kioumourtoglou MA, Spiegelman D, Szpiro AA, Sheppard L, Kaufman JD, Yanosky JD, Williams R, Laden F, Hong B, Suh H. Exposure measurement error in PM_{2.5} health effects studies: a pooled analysis of eight personal exposure validation studies. *Environ Health*. 2014;13:2.
- [43] Wu K, Willett WC, Fuchs CS, Colditz GA, Giovannucci EL. Calcium intake and risk of colon cancer in women and men. *J National Cancer Inst*. 2002;94:437–46.
- [44] Halton TL, Willett WC, Liu S, Manson JE, Albert CM, Rexrode K, Hu FB. Low-carbohydrate-diet score and the risk of coronary heart disease in women. *N Engl J Med*. 2006;355:1991–2002.
- [45] Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am J Epidemiol*. 1992;135:1114–26, discussion 1127–366.
- [46] Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC. Intake of fat, meat, and fiber in relation to risk of colon cancer in men. *Cancer Res*. 1994;54:2390–7.

Supplementary Material: The online version of this article offers supplementary material (DOI:<https://doi.org/10.1515/ijb-2017-0002>).