Epidemiologic Methods

Volume 1, *Issue* 1 2012 *Article* 9

Correcting for Bias due to Misclassification when Error-Prone Continuous Exposures Are Misclassified

Ruth H. Keogh, *MRC Biostatistics Unit, Cambridge, UK*
Alexander D. Strawbridge, *MRC Biostatistics Unit, Cambridge, UK*
Ian White, *MRC Biostatistics Unit, Cambridge, UK*

Recommended Citation:

Keogh, Ruth H.; Strawbridge, Alexander D.; and White, Ian (2012) "Correcting for bias due to misclassification when error-prone continuous exposures are misclassified," *Epidemiologic Methods*: Vol. 1: Iss. 1, Article 9.
DOI: 10.1515/2161-962X.1011

Correcting for Bias due to Misclassification when Error-Prone Continuous Exposures Are Misclassified

Ruth H. Keogh, Alexander D. Strawbridge, and Ian White

Abstract

To investigate the association between a continuous exposure and an outcome it is common to categorize the exposure and estimate the relative associations between categories. Error in measurement of the continuous exposure results in misclassification when the exposure is categorized. In this paper we investigate methods for correcting for this misclassification. We consider applications of methods for continuous exposures and for fundamentally categorical exposures. A particular challenge is that even nondifferential error in the underlying continuous exposure can result in differential misclassification in the categorized exposure, i.e. misclassification dependent on the outcome. For continuous exposures, there exist a range of methods for correcting for the effects of exposure measurement error on the exposure-outcome association, including regression calibration (RC), multiple imputation (MI), moment reconstruction (MR) and simulation extrapolation (SIMEX). There are also correction methods for use with genuinely categorical exposures, using estimated misclassification probabilities. Alongside simple methods using estimated misclassification probabilities, we also consider two RC based methods, MI and MR of the continuous exposure followed by categorization, and a new SIMEX method. Simulation studies are used to compare the methods when the true exposure is available in a validation study and the more common situation in which replicate or additional error-prone exposure measurements are available in a subsample. We restrict attention to the case where the underlying association between the continuous exposure and the outcome is linear on the appropriate scale. RC and SIMEX methods fail to correct adequately for bias. However, MI and MR perform well. Methods using estimated misclassification probabilities also perform well, provided differential misclassification is assumed, however these methods are restricted to estimation of odds ratios and have other practical drawbacks. MI and MR have the benefit of being flexible for use with different analysis models, with quantile-based cutpoints, and more easily accommodate covariate adjustment. In summary, we found that MI and MR can be applied to correct exposure-outcome associations for the effects of misclassification error when the association is linear. Extending MI and MR for use with categorized continuous exposures under nonlinear exposure-outcome associations is now an important area for further research.

KEYWORDS: measurement error, categorized exposures, regression calibration, multiple imputation, moment reconstruction, SIMEX

Author Notes: Ruth Keogh and Alexander Strawbridge should be considered the joint first authors.
This work was supported by the Medical Research Council [Unit Programme number U105260558]. Ruth Keogh is now at the Department of Medical Statistics, London School of Hygiene and Tropical Medicine (ruth.keogh@lshtm.ac.uk).

1 Introduction

Many exposures in epidemiology studies are subject to measurement error. Error arises, for example, due to fluctuation in exposure over time when the exposure of interest is usual level and inferences are based on a single measurement; due to the measurement processing stage (e.g. laboratory error); due to self-reporting; and due to limitations of measurement instruments. We focus on exposures measured on a continuous scale. Examples include biological exposures, such as blood pressure (MacMahon et al., 1990) and plasma fibrinogen (The Fibrinogen Studies Collaboration, 2006), and self-reported measurements of dietary intake in nutritional epidemiology (Willett, 1998). Error in exposure measurements results in biased estimates of exposure-outcome associations (Carroll et al., 2006)

In epidemiological analyses it is common to categorize continuous exposures and to investigate the exposure-outcome association in exposure categories relative to a reference category. We refer to this as a categorized exposure analysis. The cutpoints may be fixed, i.e. based on predefined values, or may be based on quantiles of the exposure distribution in the study population. While a categorized exposure analysis has drawbacks (Greenland, 1995a,b) it is widely used because it provides a simple method of investigating nonlinearity in the exposure-outcome association and a convenient way of presenting results. Measurement error in the continuous exposure results in misclassification of individuals when the exposure is categorized, leading to biased estimates of the exposure-outcome association in a categorized exposure analysis. In this paper we investigate methods for correcting for the effects of misclassification when error-prone continuous exposures are categorized. At present there exists no single clear correction method for this situation and the contribution of this paper is to investigate potential candidates for such a method.

When a continuous exposure is categorized, it is clear that individuals with a continuous exposure value close to one of the cutpoints which define the exposure categories are more likely to be misclassified than individuals with continuous exposure values further from a cutpoint. The result is that knowing both the categorized true exposure and the categorized mismeasured exposure could give more information about the outcome than the categorized true exposure alone, because the probability of misclassification depends on the proximity of an exposure value to a cutpoint, which in turn is related to the outcome. Therefore, even nondifferential error in a continuous exposure measurement - that is, error which is independent of the outcome - can give rise to differential misclassification in the categorized exposure. This phenomenon is discussed in detail by Flegal et al. (1991) and Gustafson and Le (2002).

Methods for correcting for the effects of measurement error in continuous exposures on the associations found in categorized exposure analyses have received little attention in the literature. Perhaps the most intuitive approach is to employ methods for correcting for misclassification of fundamentally categorical exposures, that is categorical exposures which are not derived from an underlying continuous measure. These methods are based on estimated misclassification probabilities (Barron, 1977, Morrissey and Spiegelman, 1999, Kosinski and Flanders, 1999, Chu et al., 2008). A disadvantage of these methods is that they are focused on estimation of odds ratios, and therefore not applicable in a more general context. There also exist a number of methods for correcting for the effects of error in a continuous exposure on the continuous exposure-outcome association. These include regression calibration (RC) (Rosner et al., 1989, 1992, Spiegelman et al., 1997, Carroll et al., 1999, 2006), which is widely used, multiple imputation (MI) (Cole et al., 2006, Freedman et al., 2008), moment reconstruction (MR) (Freedman et al., 2004, 2008), and simulation extrapolation (SIMEX) (Cook and Stefanski, 1994, Carroll et al., 2006, Staudenmayer and Ruppert, 2004). In this paper we investigate whether RC, MI, MR and SIMEX can be adapted for use in correcting for the effects of misclassification when the mismeasured observed continuous exposure is categorized. A method based on RC for use in categorized exposure analyses was recently proposed by Natarajan (2009), though we show later that this method is flawed.

In this paper we restrict our attention to the case in which the underlying association between the continuous exposure and the outcome is linear on the appropriate scale, e.g. in a logistic model. Possible extensions to the case of nonlinear associations are noted in the discussion. In the first part of the paper we focus on nondifferential error in the continuous measurement, that is error not associated with the outcome. In this situation the continuous mismeasured exposure gives no additional information about the outcome beyond that given by the true exposure. We consider both classical (random) measurement error in the continuous exposure, where the observed exposure W can be written as $W = X + U$ where X is the true exposure and U is random error with constant variance, and systematic error depending on the true level of exposure, where $W = \alpha_0 + \alpha_1 X + U$. In a later section we extend our investigations to differential error in the continuous exposure measurement. Heteroscedastic error, in which the variance of U depends on X, is considered in the discussion.

Any correction method requires information about the relationship between the true and mismeasured exposures. This could be from information external to the current study, but we focus on the case where additional information is available within the study population. This is ideally in the form of a validation study in which the true exposure is observed in a subsample. However, in many situations

it is not possible to obtain a measure of true exposure. In the case of classical measurement error, correction methods can be applied using repeated measurements in a subsample. In the case of systematic measurement error in the main exposure measurement, correction methods can be applied if a different measure of exposure is available in a subsample which is superior to the main measure and which provides an unbiased measure of the true exposure.

The plan of the paper is as follows. In Section 2.1 we set out the measurement error problem in more detail and outline three scenarios under which we will consider using correction methods, which depend on the additional information available in the study. In Section 2.2 we briefly outline simple correction methods for fundamentally categorical exposures, based on estimated misclassification probabilities. In Section 2.3 we describe the adaptation of correction methods for continuous exposures to the categorized exposure situation. The performance of the methods is investigated using simulation studies covering a range of scenarios, as described in Section 3. In Section 4 we extend the simulation study to the case of differential error in the continuous exposure and we conclude with a discussion in Section 5.

2 Methods

2.1 Measurement error

Let X denote the true but unobserved continuous exposure measurement and Y denote the outcome of interest, e.g. disease status (0,1). For simplicity we focus on a dichotomized continuous exposure $X_C = I(X > C)$ where C is a predefined cutpoint. X_C therefore takes value 1 if $X > C$ and value 0 otherwise. In a categorized exposure analysis the exposure-outcome model of interest is

$$g(E(Y|X_C)) = b_{0X} + b_{1X}X_C \tag{1}$$

where for example $g(x) = x$ for a linear regression and $g(x) = \{1 + \exp(-x)\}^{-1}$ for a logistic regression. The focus of this paper is on estimating b_{1X}, which is the regression coefficient in a linear regression, and the log odds ratio in a logistic regression.

The observed continuous exposure measurement is denoted W_1 where

$$W_1 = \alpha_0 + \alpha_1 X + u_1 \tag{2}$$

such that $E(u_1) = 0$ and $\text{var}(u_1) = \sigma^2_{u_1}$. The errors u_1 are assumed independent of X and Y, that is error in W_1 is nondifferential ($E(Y|X,W_1) = E(Y|X)$). When $\alpha_0 = 0$

and $\alpha_1 = 1$, (2) is the classical measurement error model. Other values for (α_0, α_1) represent systematic error, with $\alpha_1 \neq 1$ indicating systematic error depending on the true exposure. The naive approach to estimating b_{1X} is to replace X with the observed exposure W_1 using

$$g(E(Y|W_{1C})) = b_{0W} + b_{1W}W_{1C} \tag{3}$$

where $W_{1C} = I(W_1 > C)$. The estimate of b_{1W} is not an unbiased estimate of b_{1X} because of the misclassification in W_{1C}. Below we investigate methods for obtaining unbiased estimates of b_{1X}. As noted in the introduction, additional measurements are required to make corrections for measurement error. We consider three scenarios:

Scenario (a) True exposure measurements X are available for a subset of the study population. This situation may arise when the true exposure measurement is expensive to obtain and an error-prone but cheaper instrument is used in studies involving a large number of individuals. For example, in a large prospective cohort, participants may self-report anthropometric variables such as height and weight, which may be measured by a nurse in a subsample for use in a validation study. In this scenario W_1 can follow any error model of the form in (2).

Scenario (b) W_1 follows the classical measurement error model ($\alpha_0 = 0, \alpha_1 = 1$ in model (2)) and a repeated measurement W_2 is available in a subset of the study population. In this scenario we have $W_j = X + u_j, j = 1,2$, where the errors u_1 and u_2 are independent and have the same variance $\sigma^2_{u_1}$. An example of this is when repeated measures of blood pressure are available, where the exposure of interest is usual level.

Scenario (c) W_1 is subject to systematic error depending on true exposure ($\alpha_1 \neq 1$ and α_0 takes any value in model (2)) and *two* additional exposure measurements of a *different* type, W_2 and W_3, are available and are subject to classical measurement error. Here $W_j = X + u_j, j = 2,3$, where the errors u_2 and u_3 are independent of each other, of X and Y, and of u_1, and have the same variance $\sigma^2_{u_2}$. Examples of this scenario arise in nutritional epidemiology where food frequency questionnaires, known to be subject to systematic error, are used as the main method of measuring dietary intake in large studies, and repeated food record measurements or biological measurements are available in a subsample and assumed to follow the classical measurement error model.

In scenarios (b) and (c) the true exposure X is completely unobserved. We do not specifically consider the situation in which the observed exposure measurements are subject to a constant shift relative to the true exposure such that $W_1 = \alpha_0 + X + u_1, \quad \alpha_0 \neq 0$. In that situation methods for scenario (c) would be appropriate for a fixed cutpoint C. However, if the cutpoint were a quantile of the exposure distribution then methods for scenarios (a) and (b) would apply.

In Section 4 we consider the more complex case in which measurement error in W_1 is differential, that is depends in some way on the outcome Y.

2.2 Methods using estimated misclassification probabilities

We begin by outlining correction methods using estimated misclassification probabilities. These are perhaps the first obvious candidates to consider, due to their simplicity and their common use with fundamentally categorical exposures. The methods in this section refer to a binary outcome $Y = 0, 1$ and estimation specifically of an odds ratio. For the dichotomized true exposure, X_C, the odds ratio of interest is

$$OR = \frac{\Pr(Y = 1|X_C = 1)\Pr(Y = 0|X_C = 0)}{\Pr(Y = 0|X_C = 1)\Pr(Y = 1|X_C = 0)}. \tag{4}$$

A crucial result for the methods in this section is that the odds ratio can also be written in the form

$$OR = \frac{\Pr(X_C = 1|Y = 1)\Pr(X_C = 0|Y = 0)}{\Pr(X_C = 0|Y = 1)\Pr(X_C = 1|Y = 0)}. \tag{5}$$

There exist matrix-based methods for estimating the odds ratio in (5) using estimated misclassification probabilities (Barron, 1977, Morrissey and Spiegelman, 1999). Let $\eta^X_{xy} = \Pr(X_C = x|Y = y)$ and $\eta^W_{wy} = \Pr(W_C = w|Y = y)$. We also define the positive and negative predictive probabilities $\text{ppv}_y = \Pr(X_C = 1|W_{1C} = 1, Y = y)$ and $\text{npv}_y = \Pr(X_C = 0|W_{1C} = 0, Y = y)$. Estimation of ppv_y and npv_y is discussed below. An alternative approach would be to formulate the misclassification using sensitivities and specificities (Barron, 1977). It can be shown that the probabilities η^X_{xy} are estimated using $\hat{\eta}^X_{1y} = \widehat{\text{ppv}}_y\hat{\eta}^W_{1y} + (1 - \widehat{\text{npv}}_y)\hat{\eta}^W_{0y}$, $\hat{\eta}^X_{0y} = (1 - \widehat{\text{ppv}}_y)\hat{\eta}^W_{0y} + \widehat{\text{npv}}_1\hat{\eta}^W_{01}$ $(y = 0, 1)$. Hence, the odds ratio of interest is estimated by

$$\widehat{OR} = \frac{\hat{\eta}^X_{11}\hat{\eta}^X_{00}}{\hat{\eta}^X_{10}\hat{\eta}^X_{01}}. \tag{6}$$

The probabilities $\text{ppv}_y, \text{npv}_y (y = 0, 1)$ are written here to allow differential misclassification by conditioning on Y. However, typically for fundamentally categorical exposures the most common approach would be to assume nondifferential misclassification. In our later simulation studies we investigate allowing differential or non-differential misclassification.

In the situation of interest in this paper, in which the exposure categories are formed by categorizing an underlying continuous exposure, there are two ways of estimating the probabilities ppv_y and npv_y, which we outline below.

2.2.1 Misclassification probability method 1 (MP1)

The simplest way to estimate the misclassification probabilities is to treat the the observed X_C and W_C as purely categorical, that is to ignore the underlying continuous values. In scenario (a), where the true exposure X is observed in a subset of the study population, ppv_y and $\text{npv}_y (y = 0, 1)$ can be estimated directly in the subset. When X is not observed, in general at least three observations of the misclassified exposure are required to estimate misclassification probabilities if they are treated as fundamentally categorical (Hui and Walter, 1980). We do not pursue this here.

2.2.2 Misclassification probability method 2 (MP2)

The second way of estimating the misclassification probabilities is to use the underlying continuous values. Using this method, the positive predictive probability (ppv_y) can be written

$$\begin{aligned} \Pr(X_C = 1 | W_{1C} = 1, Y = y) &= \Pr(X > C | W_1 < C, Y = y) \\ &= \int_C^\infty \frac{\Pr(X > C | w, Y = y)}{\Pr(W_1 > C | Y = y)} f_{W|Y}(w|y) \mathrm{d}w \end{aligned} \quad (7)$$

where $f_{W|Y}$ denotes the conditional distribution of W_1 given Y. To estimate the misclassification probabilities this way therefore requires assumptions about the distribution of W_1 given Y and about the distribution of X given (W_1, Y). Here we outline the case in which both conditional distributions are assumed normal. In this case we let $\mu_{W|Y} = E(W_1|Y)$, $\sigma^2_{W|Y} = \text{var}(W_1|Y)$, $\mu_{X|WY} = E(X|W_1, Y)$ and $\sigma^2_{X|WY} = \text{var}(X|W_1, Y)$. The probability in (7) can then be written as

$$\left\{1 - \Phi\left(\frac{C - \mu_{W|Y}}{\sigma_{W|Y}}\right)\right\}^{-1} \int_C^\infty \left\{1 - \Phi\left(\frac{C - \mu_{X|WY}}{\sigma_{X|WY}}\right)\right\} \phi\left(\frac{w - \mu_{W|Y}}{\sigma_{W|Y}}\right) \mathrm{d}w \quad (8)$$

where $\Phi(.)$ denotes the cumulative density function for the standard normal distribution and $\phi(.)$ denotes the corresponding probability density function. The negative predictive probability can be estimated in a similar way.

In scenario (a), where X is observed, $\mu_{X|WY}$ and $\sigma^2_{X|WY}$ can be estimated in the validation subset. Unlike method MP1, this method for estimating misclassification probabilities can be used in scenario (b) when only two observations of the mismeasured exposure are available. In this case $\mu_{X|WY}$ and $\sigma^2_{X|WY}$ can be estimated using the joint distribution of W_1, W_2 given Y. In scenario (c) we similarly use W_2 and W_3 to estimate $\mu_{X|WY}$ and $\sigma^2_{X|WY}$. Further details are given in Appendices A and B.

2.3 Use of correction methods for continuous exposures

In this section we discuss applications of correction methods for continuous mismeasured exposures in the categorized exposure situation. These are discussed in the context of a more general form for the exposure-outcome association, which includes the logistic model.

2.3.1 Regression calibration (RC)

Consider a linear association, on some scale, between the continuous exposure and outcome,

$$g(E(Y|X)) = \beta_{0X} + \beta_{1X}X.$$

When W_1 is observed instead of X as the main measurement, it can be shown that unbiased estimates of (β_{0X}, β_{1X}) can be obtained by replacing X with $E(X|W_1)$. This is called RC (Rosner et al., 1989, Carroll et al., 2006) and relies on the assumption that error in W_1 is nondifferential. This method is exact under a linear exposure-outcome model ($g(x) = x$), and an approximation under a logistic model (Rosner et al., 1989). The expectation $E(X|W_1)$ can be found by assuming a linear association between X and W_1,

$$X = \lambda_0 + \lambda_1 W_1 + e. \tag{9}$$

The model in (9) can be fitted easily if X is observed alongside W_1 for a subset of individuals (scenario (a)). In scenarios (b) and (c) λ_0 and λ_1 can be estimated by a regression of W_2 on W_1 no matter what the values of α_0 and α_1 in the error model for W_1. W_2 must, however, be subject only to classical measurement error. This is a special case in which a third measurement W_3 is not required under scenario (c).

Measurement error correction using RC does not extend to the situation in which mismeasured continuous exposures are categorized because of its reliance on the assumption of nondifferential error. In the simulation study we investigate the effect of falsely making this assumption on the resulting estimates of b_{1X} using two RC based methods outlined below.

Natarajan's method (RC1) Natarajan (2009) proposed a RC based method for correcting for the misclassification which arises when continuous exposures are categorized. In this, first the fitted values for the continuous exposure are found using (9), giving $\tilde{X} = \lambda_0 + \lambda_1 W_1$. Next, the fitted values $\tilde{X}$ are dichotomized using

the cutpoint C to obtain the binary variable $\tilde{X}_C = I(\tilde{X} > C)$, which is then used in the exposure-outcome model:

$$g(E(Y|\tilde{X}_C)) = b_0^{rc1} + b_1^{rc1}\tilde{X}_C. \tag{10}$$

Although this seems an obvious approach it is flawed because the transformation of W_1 to $\tilde{X}$ results only in a shift of the exposure measurements and not in a change in the ordering of individuals with respect to their exposure measurements. Suppose that instead of having a fixed cutpoint, the cutpoint was defined by the median, then $W_{1C} = \tilde{X}_C$ for all individuals and this method would give identical results to the naive method. By using $\tilde{X}$ to find $\tilde{X}_C$ as above it is implicitly assumed that the misclassification of individuals is nondifferential. In scenario (c), although only one of W_2 and W_3 is required to fit model (4), in the simulation we fitted the model by regressing the mean of W_2 and W_3 on W_1.

Alternative RC based approach (RC2) An alternative RC approach is somewhat related to the misclassification probabilities approach in Section 2.2. In this method we fit the model

$$g(E(Y|W_{1C})) = b_0^{rc2} + b_1^{rc2}E(X_C|W_{1C}). \tag{11}$$

This is a direct extension of RC to categorical exposures and as such requires the assumption that misclassification in W_{1C} is nondifferential, though this is known not to be the case. Suppose we were to proceed with this approach. In scenario (a) $E(X_C|W_{1C})$ can be estimated in the validation sample, e.g. using logistic regression. In scenarios (b) and (c) we assume that X and W_1 have a joint normal distribution and estimate $E(X_C|W_{1C} = 1) = \Pr(X > C|W_1 > C)$ using a similar method to that described in Section 2.2.2, but removing the conditioning on Y.

2.3.2 Multiple imputation (MI)

MI was introduced by Rubin (1987) and is becoming widely used to handle missing data in studies of different types. The key idea in MI for missing exposure measurements is that missing values are imputed by drawing a random value from the distribution of the exposure conditional on all observed data, including the outcome. To account for the uncertainty in the imputed values a number of imputed values are obtained for each missing data point, creating M imputed data sets for the full cohort. The resulting data sets are analysed separately but identically and the resulting estimates are combined (Rubin, 1987). There is now a large literature on MI for missing data - see for example White and Carlin (2010) and White et al. (2011) for summaries.

Cole et al. (2006) and Freedman et al. (2008) proposed using MI to correct for measurement error in continuous exposures, by treating the true continuous exposure values as missing data. In this method X is replaced in the exposure-outcome model by imputed values from the model

$$X = \gamma_0 + \gamma_1 W_1 + \gamma_2 Y + \varepsilon. \tag{12}$$

The imputed measurement is given by

$$X_{MI}(W_1, Y) = E(X|W_1, Y) + \varepsilon^*$$

where $E(X|W_1, Y)$ is obtained using model (12), and ε^* is a random draw from the distribution of the residuals from the regression of X on W_1 and Y. The procedure is repeated to give M multiply imputed data sets. We propose that to estimate the association between dichotomized X and outcome Y the imputed values $X^{MI(m)}$ are dichotomized to give the binary variable $X_C^{MI(m)} = I(X^{MI(m)} > C)$, $(m = 1, \ldots, M)$. In each imputed data set we fit the model:

$$g(E(Y|X_C^{(m)})) = b_0^{MI(m)} + b_1^{MI(m)} X_C^{MI(m)}$$

and estimate the parameter of interest b_{1X} as

$$\hat{b}_1^{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{b}_1^{MI(m)}.$$

The imputation model can be fitted directly in scenario (a). To fit the model under scenarios (b) and (c) we require assumptions about the joint distribution of (X, W_1, W_2) and (X, W_1, W_2, W_3) respectively. This is outlined in Appendices A and B, assuming joint normal distributions. To make the most of the available data, we allow a different imputation model for individuals in the subsample with additional exposure measurements.

2.3.3 Moment reconstruction (MR)

MR is an alternative correction method for continuous exposures proposed by Freedman et al. (2004). The idea of MR is to find values $X_{MR}(W_1, Y)$ such that the first two joint moments of $X_{MR}(W_1, Y)$ with Y are the same as the first two joint moments of X with Y. It has been shown (Freedman et al., 2004, 2008) that $X_{MR}(W_1, Y)$ is

$$X_{MR}(W_1, Y) = E(X|Y) + (W_1 - E(W_1|Y)) \sqrt{\frac{\text{var}(X|Y)}{\text{var}(W_1|Y)}}. \tag{13}$$

When the distributions of $X|Y$ and $W_1|Y$ are normal, it follows the joint distribution of $X_{MR}(W_1, Y)$ and Y is the same as that of X and Y (Freedman et al., 2004). Having

obtained the moment reconstructed values we define $X_C^{MR} = I(X_{MR}(W_1, Y) > C)$ and fit the categorized exposure model

$$g(E(Y|X_C^{MR})) = b_0^{MR} + b_1^{MR} X_C^{MR}.$$

Calculation of the moment reconstructed values requires estimation of $E(X|Y)$ and $\text{var}(X|Y)$. Under scenario (a) $E(X|Y)$ and $\text{var}(X|Y)$ can be estimated in the validation subset. Estimation of these quantities under scenarios (b) and (c) can be performed by maximum likelihood, as outlined in Appendices A and B.

2.3.4 Simulation extrapolation (SIMEX)

In the SIMEX procedure for continuous exposures, the change in the estimated exposure-outcome association parameter under different degrees of exposure measurement error is modelled using pseudo-datasets formed by adding additional measurement error to the observed measurements W_1 (Cook and Stefanski, 1994, Carroll et al., 2006). The model is then used to extrapolate back to the case of no measurement error. Here we adapt the SIMEX procedure to accommodate categorized continuous variables. We refer to this as group-SIMEX.

As in standard SIMEX for continuous exposures, J simulated sets of values for W_1 are generated using

$$W_j^*(\zeta) = W_1 + \sqrt{\zeta}\sigma_{u_1} Z_j, \quad j = 1, \ldots, J \tag{14}$$

for several values of ζ, where Z_j is a vector of standard normal random deviates. J is typically chosen to be between 100 and 200. Typical values considered for ζ are $\{0.5, 1, 1.5, 2\}$, where $\zeta = 0$ represents the original observed measurement W_1. In group-SIMEX the categorized values $W_{Cj}^*(\zeta) = I(W_j^*(\zeta) > C)$ are obtained. The model of interest is then fitted using each simulated set of values $W_j^*(\zeta)$, under each value of ζ:

$$g(E(Y|W_{Cj}^*(\zeta))) = b_0^{SIM(j)} + b_1^{SIM(j)} W_{Cj}^*(\zeta). \tag{15}$$

For a particular value of ζ we let $\hat{b}_0^{SIM}(\zeta)$ and $\hat{b}_1^{SIM}(\zeta)$ denote the mean of the parameter estimates across the J simulated data sets. A regression model is fitted which relates $\hat{b}_0^{SIM}(\zeta)$ and $\hat{b}_1^{SIM}(\zeta)$ to ζ, and the model is used to extrapolate back to the error-free situation where $\zeta = -1$. Examples of suitable extrapolation models include quadratic and linear-rational extrapolant models (Cook and Stefanski, 1994). When W_1 is subject to systematic error (scenario (c)) Alpizar-Jara et al. (1998) suggest performing an initial calibration step, and performing SIMEX on $W_1^* = (W_1 - \alpha_0)/\alpha_1$ using measurement error variance $\sigma_{u_1}^{*2} = \sigma_{u_1}^2/\alpha_1^2$. See Appendices A and B for some further details.

Another SIMEX method has been proposed for fundamentally categorical exposures subject to misclassification, called misclassification-SIMEX (Kuchenhoff et al., 2006, Carroll et al., 2006). In this, a matrix of misclassification probabilities is assumed for the observed exposures and misclassification is increased using ζ and then extrapolated to the case with no misclassification. However, this method assumes that the probability of misclassification is uniform within categories, and therefore does not accommodate differential misclassification.

3 Simulation study

3.1 Description

For a sample of 5000 individuals the true continuous exposure X was generated according to a normal distribution with mean 0 and variance 1. The observed exposure W_1 was generated using (2). We considered both a classical measurement error model for W_1 using $(\alpha_0 = 0, \alpha_1 = 1)$, and systematic error in W_1 using $(\alpha_0 = 0, \alpha_1 = 0.5)$. The errors u_1 were generated from a normal distribution with mean 0 and variance $\sigma^2_{u_1}$, using values $\sigma^2_{u_1} = 0.25, 1$. Dichotomized true and observed exposures X_C and W_{1C} were derived using a fixed cutpoint at $C = 0$ or $C = 1$.

We consider a binary outcome Y, generated according to a logistic regression model $\Pr(Y = 1|X) = \{1 + e^{-(\beta_{0X} + \beta_{1X} X)}\}^{-1}$ using values $\beta_{1X} = \log(1.5), \log(2)$ and $\beta_{0X} = -2.5$, resulting in approximately 8% and 9% of individuals having the outcome $Y = 1$ for each value of β_{1X} respectively.

In scenario (a) the true exposure, X, was assumed to be observed in a random subset of 10% or 50% of the population, and W_1 can be subject to either classical or systematic error. Scenario (b) refers to classical measurement error in W_1 and scenario (c) to systematic error, and in both cases X is assumed unobserved. We generated two additional observed exposures W_2, W_3 according to the classical measurement error model $W_j = X + u_j (j = 2, 3)$, where the errors u_2, u_3 are independently distributed normal with means 0 and variances $\sigma^2_{u_2} = \sigma^2_{u_1}$ and independent of u_1. The additional exposure measurements W_2, W_3 were assumed to be available in a randomly selected subset of 10% or 50% of the study population. In scenario (b) only W_2 is needed in the correction methods while in scenario (c) both W_2 and W_3 are required.

The simulation was repeated 1000 times. In each simulated dataset we fitted the analysis model using the true exposure, X, and using the mismeasured exposure W_1, which is referred to as the naive method. The methods outlined in Section 2 were then applied, as detailed further below. In scenario (a) only we applied the method of Section 2.2.1, using misclassification probabilities estimated

by treating the categorized exposure as fundamentally categorical (MP1). This was performed first assuming nondifferential misclassification and then allowing differential misclassification. We are interested in the results using this method under nondifferential misclassification because it is that which is commonly used for fundamentally categorical exposures and as such may be the first analysis to consider in our situation. The method of Section 2.2.2 in which misclassification probabilities are estimated using the underlying continuous measurements (MP2) was applied in scenarios (a), (b) and (c). For this we restrict ourselves to the case in which differential misclassification is assumed. We then applied the methods Section 2.3: RC using Natarajan's method (RC1), RC using misclassification probabilities (RC2), MI, MR, and group-SIMEX (gSIMEX). In all methods, the aim was to estimate the log odds ratio parameter b_{1X} in the logistic regression model $\Pr(Y=1|X_C)=\{1+e^{-(b_{0X}+b_{1X}X_C)}\}^{-1}$. The focus is on assessing bias in estimates of b_{1X} under the different methods.

When X was observed in a subsample of the study population (scenario (a)) the true dichotomized values X_C were used in place of fitted or imputed values. This does not apply for the methods in Section 2.2 based on misclassification probabilities. Using MI, the number of multiply imputed data sets equaled the percentage of missing data, and for the group-SIMEX procedure we used $J=100$ pseudo-datasets, values $\zeta=\{0,0.5,1,1.5,2\}$, and a quadratic extrapolant. For consistency, all methods were implemented using maximum likelihood as described in the appendices, with the exception of RC1 which was performed using the commonly used method of moments approach. The simulations were performed using R.

3.2 Results

Table 1 shows the estimated log odds ratio, $\hat{b}_{1X}$, in scenario (a) in which true exposure X was observed in a subset of individuals (10% or 50%) for each of the methods described above. Tables 2 and 3 show the corresponding results in scenarios (b) and (c), where X was not observed for any individuals, but where repeated or additional error-prone measurements were available instead.

As was expected, the naive method results in attenuated log odds ratio estimates, with the attenuation being more severe as the variability of the errors increases ($\sigma^2_{u_1}$) and when W_1 is subject to systematic error (Table 1).

In scenario (a) method MP1 resulted in badly attenuated odds ratio estimates when nondifferential misclassification was assumed, with the attenuation being much worse than under the naive approach. However, the extension to allow differential misclassification under this method resulted in almost unbiased estimates, with some fairly negligible attenuation when only 10% of individuals were

in the validation subsample. The extension to this method in which misclassification probabilities were estimated using the underlying continuous measurements (MP2) was successful when we allowed differential misclassification. Notice that the empirical standard deviations of the estimates were substantially lowered when we used the continuous exposure measurements to estimate misclassification probabilities. Method MP2, allowing for differential misclassification, also performed well in scenarios (b) and (c) (Tables 2 and 3).

The RC method proposed by Natarajan (2009) (RC1) did not work well in general. In scenario (a), where the true value of X was used in the analysis where it was observed, the estimated log odds ratio for those above versus below the cutpoint at $C = 0$ is a weighted average of the estimate obtained using X_C and that obtained under the naive method using W_{1C}. Hence in scenario (a) the attenuation under RC1 was less severe when a greater proportion of the population have the true exposure observed (Table 1). Note that in scenarios (b) and (c), where X was completely unobserved, method RC1 gave almost the same attenuated estimate of the log odds ratio as the naive method when the cutpoint was at 0 (Tables 2 and 3). This was because the median of the exposure distribution happened to be at the cutpoint 0, which as discussed in Section 2.3.1, leads to individuals being categorized in the same way using W_1 and using the calibrated values. Method RC2 performed better than RC1 when X was observed in a subsample, giving approximately unbiased estimates when 50% of individuals had X observed. However, the estimates became biased when only 10% had X observed, with the bias depending on the choice of cutpoint. In scenarios (b) and (c) method RC2 resulted in severe upward bias in the log odds ratio estimates.

MI and MR performed well under all scenarios, giving unbiased or almost-unbiased estimates of the log odds ratio.

Group-SIMEX gave attenuated log odds ratio estimates. Estimates became more biased as the error variability increased, though the attenuation was less severe than when using the naive method or RC1. The bias was more severe when W_1 was subject to systematic error (scenario (a) (Table 1), and scenario (c) (Table 3)). The degree of extrapolation required under the group-SIMEX method increases as measurement error increases and so any error in the extrapolation model is amplified.

The simulation study was repeated for a situation in which the proportion of cases $(Y = 1)$ was higher (approximately 25%). The results were not materially different from those reported above and are not shown here.

4 Extension to differential error

So far we have assumed an error model of the form $W_1 = \alpha_0 + \alpha_1 X + u_1$ where the errors u_1 have mean zero and constant variance, and are independent of both X and Y. Classical measurement error is a special case ($\alpha_0 = 0, \alpha_1 = 1$). In this section we extend the simulation study to the situation in which the observed continuous exposure measurement W_1 has differential error, that is error which depends on the outcome Y. Differential error is believed to occur especially in case-control studies in which exposure information is obtained retrospectively. For example, cases may report their exposure more accurately.

We describe different types of differential error in the context of the previously considered error model, $W_1 = \alpha_0 + \alpha_1 X + u_1$, and we focus on a binary outcome $Y = 0, 1$. One possibility is that the variability of the errors u_1 depends on Y, given by $\sigma^2_{u_1 y}(y = 0, 1)$. A second possibility is that the mean of W_1 given X differs in the two outcome groups such that there are different slopes, α_{1y} $(y = 0, 1)$, in the error model. A third simple possibility is that there are different intercept terms, α_{0y} $(y = 0, 1)$, in the error model. We do not consider this case further here.

For a continuous exposure with differential error of one of the three types described above, Freedman et al. (2008) found in simulation studies that MI and MR gave almost unbiased log odds ratio estimates in a logistic regression using the continuous exposure. In the simulation study we restrict our attentions to the misclassification probability methods of Section 2.2 (allowing differential misclassification) and to MI and MR, which were the only methods to perform well in the simulations study for the types of exposure measurement error considered earlier.

For the simulation study the true exposure X and outcome Y were generated as described in Section 3.1. We focus on two types of differential error in the continuous observed exposure measurements:

(i) $W_1 = X + u_1$, where the u_1 were generated from a normal distribution with mean 0 and variance depending on y, $\sigma^2_{u_1 y}$, with values $\sigma^2_{u_1 0} = 2, \sigma^2_{u_1 1} = 0.5$.
(ii) $W_1 = \alpha_{1y} X + u_1$, where $\alpha_{10} = 0.5, \alpha_{11} = 1$ and the u_1 were generated from a normal distribution with mean 0 and constant variance $\sigma^2_{u_1} = 1$.

The above parameter values mimic those used by Freedman et al. (2008). As before, the exposure is dichotomized using a fixed cutpoint $C = 0$ or $C = 1$. In situation (i) we consider scenario (a) where X is observed in a subsample of the study population, and scenario (b) where a repeated measurement of the same type W_2 is available in a subsample. W_2 was generated in the same way as W_1. For situation (ii) we consider scenarios (a) and (c). In scenario (c) we must assume that two additional measurements of a *different* type are available in a subsample, and that

these measurements do not suffer differential error. Therefore, W_2 and W_3 were generated independently from a classical measurement error model with constant error variance 1.

For both differential error models considered, the parameters required to perform both MI and MR were estimated separately within the subgroups with $Y = 0$ and $Y = 1$. The imputed values were dichotomized as before.

The results from the simulation are shown in Tables 4 and 5. All methods (MP1, MP2, MI, MR) were found to continue to perform well in the presence of differential error in the continuous exposure.

5 Discussion

In this paper we investigated methods for correcting for the effects of misclassification which results when continuous exposures are categorized for use in categorized exposure analyses. While there exist correction methods for use with fundamentally categorical exposures and also methods for use with continuous exposures, the situation in which continuous exposures are categorized in the exposure-outcome analysis has received little attention. We summarized methods for fundamentally categorical exposures, which are based on correction of an odds ratio using estimated misclassification probabilities, and outlined their use when continuous exposures are categorized. We also described possible adaptations to four methods used to correct for error in continuous exposures: regression calibration (RC), multiple imputation (MI), moment reconstruction (MR) and simulation-extrapolation (SIMEX). The methods were compared using simulation studies for the case a dichotomized continuous exposure and a logistic regression model with underlying linear exposure-outcome association. As expected, the naive method in which the error-prone categorized exposure is used in place of the true exposure results in attenuated log odds ratio estimates within categories.

A particular challenge in this situation is that even nondifferential error in continuous exposure measurements can translate into differential error when the exposure measurements are categorized. Methods for fundamentally categorical exposures can accommodate this by estimating misclassification probabilities separately by the binary outcome, and this approach was found to work well in simulation studies. Using the underlying continuous exposures to estimate the misclassification probabilities resulted in more precise estimates, under some distributional assumptions. The commonly used method for categorical exposures which assumes nondifferential misclassification performed badly in the simulations and should not be used for categorized continuous exposures.

There does not appear to be an obvious extension of RC to accommodate categorization of continuous exposures and RC based methods performed badly in the simulation studies. We focused for simplicity on fixed cutpoints, but we also note that the RC method proposed by Natarajan (2009) can perform no better than the naive method when the cutpoint is defined by a quantile of the exposure distribution and true exposure is completely unobserved (scenarios (b) and (c)). An extension to SIMEX, referred to as group-SIMEX, was found to give attenuated effect estimates, though the attenuation was less severe than under the naive method or RC1. We also investigated misclassification-SIMEX (Kuchenhoff et al., 2006, Carroll et al., 2006), which suffers the same problem of assuming nondifferential misclassification and did not perform well (results not shown).

We proposed a simple adaptation of the MI method proposed by Cole et al. (2006) and Freedman et al. (2008), in which multiply imputed continuous exposure measurements are categorized. Similarly we proposed categorization of imputed continuous exposure measurements otained using MR (Freedman et al., 2004). Both MI and MR performed well in all situations considered, giving unbiased or nearly unbiased estimates of the log odds ratio. In further simulations we applied the methods for linear exposure-outcome models, as opposed to logistic models, which again led to similar results across methods (results not shown). We also assessed the use of MI and MR with more than two exposure categories, finding that these methods continue to work well when the number of exposure categories increases (results not shown). In contrast to RC, MI and MR give imputed values which have approximately the same distribution as the true exposure, hence the subsequent categorization allows for differential misclassification. It is this feature of MI and MR that makes these methods suitable for extension to categorized continuous covariates.

Correction methods based on estimated misclassification probabilities are restricted to estimation of odds ratios (see Section 2.2) and also do not extend directly to incorporate additional adjustment for covariates. However, the approach described in Section 2.2 can be found to be equivalent under differential misclassification to a likelihood-based approach (Lyles, 2002), which can be extended to allow covariate adjustment. Kosinski and Flanders (1999) outlined a likelihood-based method for estimating the odds ratio under misclassified exposures when the true exposure X is not observed, where the probability of true exposure is modelled as dependent on the outcome Y plus an additional covariate. The presence of the covariate is crucial for identifiability. Dalen et al. (2009) proposed a correction method for use with categorized continuous exposures for use specifically in a logistic regression, in which the probabilities $Pr(X_C = 1|Y = y)$ are estimated by assuming a distribution for continuous X conditional on Y, or by treating the conditional distribution nonparametrically.

A drawback of methods based on estimation of misclassification probabilities is that they are restricted to binary outcomes and estimation of odds ratios. Hence they do not apply in linear regressions or proportional hazards regression, for example. An advantage of using MI or MR is that they are not restricted to a specific analysis model. We have not considered survival outcomes in this paper, but it has been shown that MI can be used to impute missing covariates in a proportional hazards regression (White and Royston, 2009). MR has not, to our knowledge, been extended to a survival analysis setting. A particularly attractive feature of the MI and MR methods is that they involve imputing a value (or values) for the true exposure, categorizing, and implementing the original analysis model.

In this paper we considered dichotomized continuous exposures created using a fixed cutpoint. For exposures for which there are no commonly used predefined cutpoints, quantile cutpoints are often used. We have already made some references to this. A further drawback of the methods based on estimated misclassification probabilities is that they do not extend in an obvious way to accommodate non-fixed cutpoints. For MI and MR, quantile-based cutpoints can be derived from the distribution of the multiply imputed or moment reconstructed exposure values respectively. Further simulation studies using cutpoints based on the quantiles of the distribution yielded similar results to those shown.

To perform MI and MR requires assumptions about conditional or joint conditional distributions of observed continuous exposure measurements, and we assumed normal distributions. Methods using misclassification probabilities in which the probabilities are estimated using the underlying continuous measurements require similar assumptions. However, the methods can also be applied by making these assumptions about measurements on a transformed scale. Inferences can then be made about exposure categories on the original scale provided the transformation is monotonic.

One type of error which have not considered in this paper is heteroscedastic error in W_1 such that variability of the error depends on X; typically error in W_1 increases as X increases. This type of error at the continuous level will result in individuals in the upper end of exposure being more prone to misclassification when the continuous exposure is categorized. It can be investigated graphically whether there is heteroscedastic error, both when the true exposure is available in a validation subset, or using replicate measurements (Carroll et al., 2006). In many situations it may be possible to apply a transformation, $h(\cdot)$ say, to X and W_1, such that $h(W_1) = \theta + h(X) + u_1$ is the classical measurement error model with a constant shift θ (Carroll et al., 2006). In some cases a log transformation may be suitable, and in general Box-Cox transformations may be considered. Provided the transformation used is monotonic, the classification of individuals will be the same whether categorized on the basis of X or $h(X)$ and whether on the basis of W_1 or $h(W_1)$. It

follows that the correction methods could be applied using the transformed-scale continuous measurements, as noted above. Guo and Little (2011) suggested a MI based correction method specifically for use with continuous exposures subject to heteroscadastic error, as described here, and Spiegelman et al. (2011) described a RC based method. An adaptation of the MI method of Guo and Little (2011) to the categorized exposure situation may be possible, however we believe our suggested method using transformations could work well in many situations. This is an area for further work.

In this paper we focused on a linear association between the continuous exposure and the outcome, on the appropriate scale. An important reason for using a categorized exposure analysis is to avoid assuming a particular form for this association, and the method is often used to assess nonlinearity. Development of methods that apply for categorized exposure analyses when the exposure-outcome association is nonlinear on the appropriate scale is an area for future work. The success of MI and MR for linear exposure-outcome associations on the appropriate scale suggests that, with modifications, these methods may be good candidates for extension to a more general situation. Use of MI to correct for measurement error in dichotomized exposures relies on the assumptions that true and observed exposure measurements have a joint normal distribution conditional on the outcome Y. Under nonlinear associations the variance of X will depend on Y and the distribution of $X|Y$ may not be normal. Possible adaptations to MI therefore include allowing separate imputation models by outcome groups and drawing residuals from a non-normal distribution. MR using the first two moments may suffer similarly under nonlinear associations. Thomas et al. (2011) recently suggested an extension to the MR method proposed by Freedman et al. (2004), called moment-adjusted imputation, which uses more than the first two moments to impute 'corrected' exposure values therefore allowing greater flexibility. This method is also a target for extension to the nonlinear situation.

The aim of this paper was to investigate the use of measurement error correction methods in exposure-outcome analyses involving categorized continuous exposures. In summary, methods for categorical exposures using misclassification probabilities which allow differential misclassification performed well. MI and MR also worked well. Use of MI or MR is attractive and has a number of advantages over methods based on misclassification probabilities. MI and MR are now targets for further extension to the situation in which the underlying continuous exposure-outcome association is nonlinear.

					MP1: Non-Diff		MP1: Diff		MP2: Diff		RC1		RC2		MI		MR		gSIMEX	
σ_u^2	C		Using X_C	Naive	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\beta_{1X} = \log(1.5)$, $(\alpha_0 = 0, \alpha_1 = 1)$																				
0.25	0	Mean	0.650	0.580	0.405	0.404	0.647	0.651	0.645	0.651	0.587	0.615	0.662	0.648	0.647	0.648	0.647	0.649	0.622	0.622
		SD	0.109	0.105	0.074	0.072	0.252	0.135	0.154	0.103	0.107	0.109	0.356	0.157	0.156	0.109	0.162	0.118	0.157	0.157
	1	Mean	0.707	0.614	0.447	0.446	0.696	0.710	0.696	0.703	0.655	0.676	0.685	0.699	0.703	0.700	0.701	0.701	0.643	0.645
		SD	0.120	0.119	0.091	0.089	0.300	0.152	0.164	0.117	0.126	0.124	0.390	0.170	0.168	0.120	0.178	0.129	0.161	0.165
1	0	Mean	0.650	0.457	0.226	0.226	0.647	0.650	0.642	0.650	0.474	0.552	0.662	0.648	0.646	0.645	0.647	0.647	0.488	0.487
		SD	0.109	0.106	0.056	0.052	0.312	0.146	0.210	0.110	0.107	0.109	0.356	0.157	0.213	0.119	0.228	0.131	0.156	0.160
	1	Mean	0.707	0.473	0.253	0.254	0.698	0.710	0.696	0.706	0.570	0.631	0.685	0.699	0.700	0.699	0.698	0.698	0.488	0.487
		SD	0.120	0.110	0.065	0.062	0.363	0.164	0.221	0.119	0.150	0.137	0.390	0.170	0.223	0.128	0.243	0.138	0.154	0.153
$\beta_{1X} = \log(1.5)$, $(\alpha_0 = 0, \alpha_1 = 0.5)$																				
0.25	0	Mean	0.650	0.457	0.226	0.226	0.647	0.650	0.642	0.650	0.474	0.552	0.662	0.648	0.646	0.645	0.647	0.647	0.566	0.564
		SD	0.109	0.106	0.056	0.052	0.312	0.146	0.210	0.110	0.107	0.109	0.356	0.157	0.213	0.119	0.228	0.131	0.188	0.192
	1	Mean	0.707	0.548	0.188	0.187	0.691*	0.708	0.695	0.706	0.570	0.631	0.685	0.699	0.700	0.699	0.698	0.698	0.615	0.613
		SD	0.120	0.164	0.070	0.064	0.380*	0.165	0.222	0.119	0.150	0.137	0.390	0.170	0.223	0.128	0.243	0.138	0.221	0.223
1	0	Mean	0.650	0.287	0.085	0.084	0.649	0.650	0.641	0.649	0.321	0.464	0.662	0.648	0.647	0.644	0.645	0.645	0.319	0.317
		SD	0.109	0.104	0.033	0.031	0.345	0.151	0.251	0.118	0.105	0.108	0.356	0.157	0.256	0.128	0.296	0.143	0.167	0.166
	1	Mean	0.707	0.304	0.086	0.086	0.690	0.709	0.694	0.706	0.520	0.616	0.685	0.699	0.697	0.698	0.693	0.694	0.324	0.321
		SD	0.120	0.118	0.041	0.036	0.394	0.169	0.264	0.124	0.272	0.155	0.390	0.170	0.266	0.136	0.308	0.152	0.170	0.171
$\beta_{1X} = \log(2)$, $(\alpha_0 = 0, \alpha_1 = 1)$																				
0.25	0	Mean	1.106	0.982	0.671	0.671	1.108	1.107	1.098	1.102	0.994	1.043	1.133	1.105	1.100	1.103	1.100	1.104	1.054	1.054
		SD	0.112	0.107	0.076	0.070	0.266	0.141	0.155	0.106	0.110	0.111	0.363	0.160	0.155	0.112	0.165	0.120	0.160	0.161
	1	Mean	1.169	1.019	0.744	0.744	1.161	1.170	1.158	1.163	1.080	1.118	1.167	1.165	1.163	1.163	1.160	1.163	1.070	1.071
		SD	0.110	0.107	0.088	0.081	0.263	0.136	0.154	0.107	0.112	0.112	0.336	0.152	0.153	0.110	0.161	0.117	0.150	0.154
1	0	Mean	1.106	0.766	0.374	0.372	1.110	1.106	1.095	1.099	0.798	0.930	1.133	1.105	1.098	1.098	1.096	1.100	0.819	0.818
		SD	0.112	0.105	0.058	0.051	0.329	0.151	0.208	0.110	0.107	0.111	0.363	0.160	0.205	0.120	0.226	0.132	0.157	0.159
	1	Mean	1.169	0.785	0.421	0.422	1.161	1.170	1.159	1.165	0.941	1.044	1.167	1.165	1.162	1.161	1.155	1.161	0.816	0.815
		SD	0.110	0.102	0.068	0.059	0.316	0.146	0.207	0.111	0.131	0.120	0.336	0.152	0.204	0.116	0.219	0.127	0.147	0.146
$\beta_{1X} = \log(2)$, $(\alpha_0 = 0, \alpha_1 = 0.5)$																				
0.25	0	Mean	1.106	0.766	0.374	0.372	1.110	1.106	1.095	1.099	0.798	0.930	1.133	1.105	1.098	1.098	1.096	1.100	0.949	0.948
		SD	0.112	0.105	0.058	0.051	0.329	0.151	0.208	0.110	0.107	0.111	0.363	0.160	0.205	0.120	0.226	0.132	0.191	0.193
	1	Mean	1.169	0.910	0.332	0.330	1.161**	1.168	1.160	1.166	0.941	1.044	1.167	1.165	1.162	1.161	1.155	1.161	1.023	1.021
		SD	0.110	0.139	0.077	0.063	0.328**	0.148	0.208	0.110	0.131	0.120	0.336	0.152	0.204	0.116	0.219	0.127	0.198	0.201
1	0	Mean	1.106	0.479	0.140	0.139	1.111	1.105	1.096	1.098	0.536	0.777	1.133	1.105	1.100	1.097	1.096	1.099	0.531	0.530
		SD	0.112	0.101	0.036	0.031	0.360	0.157	0.245	0.115	0.101	0.107	0.363	0.160	0.246	0.128	0.289	0.144	0.161	0.162
	1	Mean	1.169	0.509	0.147	0.146	1.160	1.169	1.161	1.165	0.856	1.006	1.167	1.165	1.161	1.161	1.152	1.158	0.539	0.534
		SD	0.110	0.109	0.047	0.038	0.340	0.152	0.244	0.115	0.224	0.134	0.336	0.152	0.243	0.124	0.281	0.140	0.158	0.159

Table 1: Scenario (a): Mean and empirical standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome Y on X_C across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when X is observed in 10% or 50% of the study population.*6 simulations omitted because no cases above the cutpoint were in the validation sample,** 1 simulation omitted for the same reason.

					MP2: Diff		RC1		RC2		MI		MR		gSIMEX	
σ_u^2	C		Using X_C	Naive	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\beta_{1X} = \log(1.5)$																
0.25	0	Mean	0.650	0.580	0.650	0.649	0.580	0.580	0.824	0.824	0.649	0.645	0.649	0.649	0.621	0.622
		SD	0.109	0.105	0.116	0.104	0.107	0.106	0.150	0.150	0.112	0.100	0.117	0.115	0.158	0.157
	1	Mean	0.707	0.614	0.703	0.702	0.648	0.646	0.982	0.980	0.703	0.698	0.704	0.703	0.644	0.644
		SD	0.120	0.119	0.125	0.116	0.131	0.130	0.192	0.191	0.121	0.111	0.131	0.129	0.163	0.162
1	0	Mean	0.650	0.457	0.652	0.649	0.456	0.457	0.918	0.916	0.651	0.641	0.651	0.651	0.487	0.487
		SD	0.109	0.106	0.197	0.123	0.108	0.106	0.217	0.214	0.194	0.119	0.161	0.134	0.156	0.160
	1	Mean	0.707	0.473	0.707	0.705	0.552	0.547	1.251	1.239	0.702	0.691	0.702	0.700	0.488	0.488
		SD	0.120	0.110	0.208	0.130	0.166	0.163	0.316	0.295	0.205	0.126	0.174	0.144	0.153	0.152
$\beta_{1X} = \log(2)$																
0.25	0	Mean	1.106	0.982	1.105	1.102	0.981	0.982	1.394	1.394	1.104	1.096	1.105	1.104	1.052	1.054
		SD	0.112	0.107	0.117	0.105	0.108	0.107	0.154	0.152	0.113	0.103	0.122	0.119	0.162	0.161
	1	Mean	1.169	1.019	1.166	1.164	1.069	1.068	1.630	1.628	1.166	1.157	1.166	1.165	1.070	1.070
		SD	0.110	0.107	0.117	0.106	0.117	0.115	0.177	0.173	0.114	0.101	0.120	0.117	0.152	0.151
1	0	Mean	1.106	0.766	1.109	1.101	0.767	0.767	1.540	1.537	1.109	1.088	1.109	1.103	0.816	0.817
		SD	0.112	0.105	0.199	0.123	0.106	0.105	0.221	0.212	0.197	0.119	0.164	0.137	0.157	0.160
	1	Mean	1.169	0.785	1.175	1.168	0.914	0.911	2.074	2.056	1.171	1.148	1.170	1.165	0.816	0.815
		SD	0.110	0.102	0.200	0.122	0.146	0.140	0.329	0.283	0.197	0.118	0.165	0.132	0.148	0.145

Table 2: Scenario (b) $(\alpha_0 = 0, \alpha_1 = 1)$: Mean and empirical standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome Y on X_C across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when W_2 is assumed to be observed in 10% or 50% of the study population.

					MP2: Diff		RC1		RC2		MI		MR		gSIMEX	
σ_u^2	C		Using X_C	Naive	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\beta_{1X} = \log(1.5)$																
0.25	0	Mean	0.650	0.457	0.646	0.643	0.456	0.457	0.913	0.914	0.647	0.642	0.650	0.648	0.583	0.576
		SD	0.109	0.106	0.226	0.120	0.107	0.106	0.215	0.213	0.226	0.120	0.149	0.127	0.192	0.191
	1	Mean	0.707	0.548	0.700	0.699	0.550	0.546	1.009	1.008	0.698	0.694	0.702	0.698	0.726	0.734
		SD	0.120	0.164	0.239	0.128	0.163	0.162	0.308	0.303	0.237	0.130	0.157	0.138	0.342	0.330
1	0	Mean	0.650	0.287	0.647	0.641	0.287	0.287	0.982	0.975	0.648	0.636	0.651	0.648	0.397	0.395
		SD	0.109	0.104	0.318	0.148	0.105	0.103	0.378	0.357	0.318	0.147	0.228	0.150	0.207	0.204
	1	Mean	0.707	0.304	0.698	0.696	-0.057*	0.333	1.331	1.313	0.696	0.686	0.702	0.697	0.401	0.404
		SD	0.120	0.118	0.332	0.156	2.507*	0.963	0.564	0.517	0.332	0.156	0.240	0.160	0.206	0.208
$\beta_{1X} = \log(2)$																
0.25	0	Mean	1.106	0.766	1.096	1.093	0.766	0.767	1.533	1.533	1.098	1.092	1.104	1.101	0.974	0.970
		SD	0.112	0.105	0.226	0.117	0.106	0.105	0.218	0.211	0.225	0.121	0.152	0.131	0.195	0.194
	1	Mean	1.169	0.910	1.161	1.160	0.911	0.910	1.676	1.675	1.159	1.153	1.166	1.162	1.218	1.223
		SD	0.110	0.139	0.224	0.117	0.141	0.140	0.273	0.259	0.222	0.118	0.149	0.125	0.290	0.278
1	0	Mean	1.106	0.479	1.100	1.091	0.480	0.479	1.639	1.627	1.101	1.083	1.109	1.100	0.662	0.661
		SD	0.112	0.101	0.318	0.142	0.101	0.101	0.397	0.349	0.317	0.143	0.234	0.152	0.199	0.198
	1	Mean	1.169	0.509	1.163	1.158	0.394*	0.700	2.224	2.196	1.160	1.142	1.170	1.160	0.666	0.669
		SD	0.110	0.109	0.316	0.142	2.221*	0.385	0.585	0.484	0.315	0.142	0.231	0.148	0.194	0.193

Table 3: Scenario (c) ($\alpha_0 = 0, \alpha_1 = 0.5$): Mean and empirical standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome Y on X_C across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when W_2, W_3 are assumed to be observed in 10% or 50% of the study population. * Based on 995 datasets since $\max(X_C) < C$ for five datasets.

				MP1: Diff		MP2: Diff		MI		MR	
C		Using X_C	Naive	10%	50%	10%	50%	10%	50%	10%	50%
Scenario (a): $\beta_{1X} = \log(1.5)$											
0	Mean	0.650	0.517	0.668	0.651	0.658	0.651	0.655	0.649	0.678	0.654
	SD	0.109	0.106	0.302	0.140	0.193	0.110	0.198	0.115	0.366	0.154
1	Mean	0.707	0.128	0.713	0.709	0.704	0.705	0.707	0.709	0.709	0.711
	SD	0.120	0.112	0.351	0.159	0.236	0.126	0.248	0.133	0.424	0.176
Scenario (a): $\beta_{1X} = \log(2)$											
0	Mean	1.106	0.870	1.137	1.107	1.120	1.107	1.114	1.106	1.151	1.110
	SD	0.112	0.107	0.319	0.145	0.206	0.117	0.212	0.121	0.384	0.158
1	Mean	1.169	0.474	1.182	1.171	1.165	1.162	1.168	1.167	1.181	1.172
	SD	0.110	0.102	0.310	0.144	0.205	0.113	0.209	0.121	0.368	0.158
Scenario (b): $\beta_{1X} = \log(1.5)$											
0	Mean	0.650	0.517	-	-	0.673	0.654	0.672	0.647	0.664	0.653
	SD	0.109	0.106	-	-	0.195	0.117	0.193	0.111	0.155	0.125
1	Mean	0.707	0.128	-	-	0.686	0.705	0.682	0.687	0.695	0.703
	SD	0.120	0.112			0.284	0.151	0.280	0.145	0.238	0.158
Scenario (b): $\beta_{1X} = \log(2)$											
0	Mean	1.106	0.870	-	-	1.149	1.113	1.147	1.101	1.133	1.111
	SD	0.112	0.107	-	-	0.225	0.126	0.224	0.120	0.179	0.134
1	Mean	1.169	0.474	-	-	1.166	1.167	1.161	1.144	1.169	1.167
	SD	0.110	0.102	-	-	0.252	0.136	0.249	0.132	0.217	0.142

Table 4: Differential error, type (i), scenario (a): Mean and empirical standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome Y on X_C across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when (a) X is observed in 10% or 50% of the study population, (b) W_2 is observed in 10% or 50% of the study population

				MP1: Diff		MP2: Diff		MI		MR	
C		Using X_C	Naive	10%	50%	10%	50%	10%	50%	10%	50%
Scenario (a): $\beta_{1X} = \log(1.5)$											
0	Mean	0.650	0.444	0.658	0.649	0.651	0.644	0.646	0.644	0.644	0.647
	SD	0.109	0.106	0.320	0.151	0.221	0.114	0.229	0.123	0.258	0.128
1	Mean	0.707	0.782	0.690	0.699	0.690	0.699	0.698	0.700	0.698	0.698
	SD	0.120	0.110	0.353	0.163	0.259	0.130	0.262	0.137	0.270	0.134
Scenario (a): $\beta_{1X} = \log(2)$											
0	Mean	1.106	0.743	1.127	1.105	1.111	1.101	1.104	1.102	1.100	1.101
	SD	0.112	0.105	0.326	0.154	0.229	0.120	0.235	0.127	0.253	0.129
1	Mean	1.169	1.079	1.165	1.165	1.111	1.101	1.159	1.161	1.166	1.163
	SD	0.110	0.102	0.305	0.144	0.229	0.120	0.228	0.120	0.250	0.122
Scenario (c): $\beta_{1X} = \log(1.5)$											
0	Mean	0.650	0.444	-	-	0.688	0.650	0.687	0.647	0.650	0.644
	SD	0.109	0.106	-	-	0.345	0.154	0.344	0.152	0.347	0.163
1	Mean	0.707	0.782	-	-	0.629	0.687	0.624	0.672	0.700	0.697
	SD	0.120	0.110	-	-	0.460	0.177	0.463	0.175	0.368	0.172
Scenario (c): $\beta_{1X} = \log(2)$											
0	Mean	1.106	0.743	-	-	1.166	1.110	1.166	1.103	1.104	1.093
	SD	0.112	0.105	-	-	0.402	0.167	0.405	0.165	0.343	0.157
1	Mean	1.169	1.079	-	-	1.120	1.148	1.116	1.130	1.169	1.159
	SD	0.110	0.102	-	-	0.376	0.154	0.374	0.151	0.345	0.159

Table 5: Differential error, type (ii), scenarios (a) and (c): Mean and empirical standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome Y on X_C across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when (a) a repeat X is observed in 10% or 50% of the study population, (c) W_2, W_3 are assumed to be observed in 10% or 50% of the study population.

References

Alpizar-Jara, R., Stefanski, L., Pollock, K., and Laake, J. (1998). Assessing the effects of measurement errors in line transect sampling. *North Carolina State University, Institute of Statistics Mimeograph Series, Number 2508*.

Barron, B. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, 33: 414–418.

Carroll, R., Maca, J., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, 86: 541–554.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman & Hall/CRC, 2nd edition.

Chu, R., Gustafson, P., and Le, N. (2008). Bayesian adjustment for exposure misclassification in case-control studies. *Statistics in Medicine*, 29: 994–1003.

Cole, S., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement error correction. *International Journal of Epidemiology*, 35: 1074–1081.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, 89: 1314–1328.

Dalen, I., Buonaccorsi, J., Sexton, J., Laake, P., and Thoresen, M. (2009). Correction for misclassification of a categorized exposure in binary regression using replication data. *Statistics in Medicine*, 28: 3386–3410.

Flegal, K., Keyl, P., and Neito, F. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134: 1233–1244.

Freedman, L., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60: 172–181.

Freedman, L., Midthune, D., Carroll, R., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27: 5195–5216.

Greenland, S. (1995a). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*, 6: 450–454.

Greenland, S. (1995b). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, 6: 356–365.

Guo, Y. and Little, R. (2011). Regression analysis with covariates that have heteroscedastic measurement error. *Statistics in Medicine*, 30: 2278–2294.

Gustafson, P. and Le, N. (2002). Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*, 58: 878–887.

Hui, S. and Walter, S. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36: 167–171.

Kosinski, A. and Flanders, W. (1999). Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: A regression approach. *Statistics in Medicine*, 18: 2795–2808.

Kuchenhoff, H., Mwalili, S., and Lesffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62: 85–96.

Lyles, R. (2002). A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*, 58: 1034–1037.

MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A., and Stamler, J. (1990). Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335: 765–774.

Morrissey, M. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons. *Biometrics*, 55: 338–344.

Natarajan, L. (2009). Regression calibration for dichotomized mismeasured predictors. *International Journal of Biostatistics*, 5(1): DOI:10.2202/1557–4679.1143.

Rosner, B., Willett, W., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8: 1051–1069.

Rosner, B., Willett, W. C., and Spiegelman, D. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136: 1400–1413.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*, New York: Wiley.

Spiegelman, D., Logan, R., and Grove, D. (2011). Regression calibration with heteroscedastic error variance. *International Journal of Biostatistics*, 7: doi:10.2202/1557–4679.1259.

Spiegelman, D., Schneeweiss, S., and McDermott, A. (1997). Measurement error correction for logistic regression models with an "alloyed gold standard". *American Journal of Epidemiology*, 145: 184–196.

Staudenmayer, J. and Ruppert, D. (2004). Local polynomial regression and simulation-extrapolation. *Journal of the Royal Statistical Society (Series B)*, 66: 17–30.

The Fibrinogen Studies Collaboration (2006). Regression dilution methods for meta-analysis: assessing long-term variability in plasma fibrinogen among 27 247 adults in 15 prospective studies. *International Journal of Epidemiology*, 35: 1570–1578.

Thomas, L., Stefanski, L., and Davidian, M. (2011). A Moment-Adjusted Imputation Method for Measurement Error Models. *Biometrics*: DOI: 10.1111/j.1541–0420.2011.01569.x.

White, I. and Carlin, J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29: 2920–2931.

White, I. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28: 1982–1998.

White, I., Royston, P., and Wood, A. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30: 377–399.

Willett, W. (1998). *Nutritional Epidemiology*, Oxford University Press, 2nd edition.

Appendices

A Estimation of required parameters in scenario (b) under different correction methods

We assume a joint normal distribution for $W_1, W_2|Y$ to estimate $\mu_{X|Y}$ and $\sigma^2_{X|Y}$ required to estimate misclassification probabilities (Section 2.2.2), to fit the MI model in (12) (Section 2.3.2), and to obtain moment reconstructed values (Section 2.3.3):

$$W_1, W_2|Y \sim N\left\{\begin{pmatrix} E(X|Y) \\ E(X|Y) \end{pmatrix}, \begin{pmatrix} \text{var}(X|Y)+\sigma^2_{u_1} & \text{var}(X|Y) \\ \text{var}(X|Y) & \text{var}(X|Y)+\sigma^2_{u_1} \end{pmatrix}\right\}, \quad (16)$$

$$X_1, W_1|Y \sim N\left\{\begin{pmatrix} E(X|Y) \\ E(X|Y) \end{pmatrix}, \begin{pmatrix} \text{var}(X|Y) & \text{var}(X|Y) \\ \text{var}(X|Y) & \text{var}(X|Y)+\sigma^2_{u_1} \end{pmatrix}\right\}. \quad (17)$$

This gives

$$E(X|W_1,Y) = \frac{W_1\text{var}(X|Y)+E(X|Y)\sigma^2_{u_1}}{\text{var}(X|Y)+\sigma^2_{u_1}}, \text{var}(X|W_1,Y) = \frac{\text{var}(X|Y)\sigma^2_{u_1}}{\text{var}(X|Y)+\sigma^2_{u_1}}$$

where $E(X|Y)$, $\text{var}(X|Y)$, and $\sigma^2_{u_1}$ are estimated by maximum likelihood using (16). Note that a method of moments approach could alternatively have been be used.

To calculate $X_{IM}(W_1, W_2, Y)$ we think of $X, W_2|W_1, Y$ or $X, W_1|W_2, Y$ as having a bivariate normal distribution, giving

$$\begin{aligned} E(X|W_1, W_2, Y) &= \frac{\bar{W}_{12}\text{var}(X|W_1,Y)+E(X|W_1,Y)\sigma^2_{u_1}}{\text{var}(X|W_1,Y)+\sigma^2_{u_1}}, \\ \text{var}(X|W_1, W_2, Y) &= \frac{\text{var}(X|W_1,Y)\sigma^2_{u_1}}{\text{var}(X|W_1,Y)+\sigma^2_{u_1}}, \end{aligned} \tag{18}$$

where $\bar{W}_{12}$ is the mean of W_1 and W_2.

For group-SIMEX (Section 2.3.4) in scenarios (a) and (b), $\sigma^2_{u_1}$ was estimated respectively by assuming bivariate normal distributions for (X, W_1) and (W_1, W_2). This is as above, but removing the conditioning on Y. Method RC2 (Section 2.3.1) is also performed using the above joint distribution, but again without the conditioning on Y.

B Estimation of required parameters in scenario (c) under different correction methods

We assume a joint normal distribution for $W_2, W_3|Y$ to estimate $\mu_{X|Y}$ and $\sigma^2_{X|Y}$, which are required to estimate misclassification probabilities (Section 2.2.2), and to obtain moment reconstructed values (Section 2.3.3). As outlined in Appendix A, this enables estimation of $E(X|Y)$, $\text{var}(X|Y)$ and $\sigma^2_{u_1}$ by maximum likelihood.

To fit the MI model in (12) in scenario (c) (Section 2.3.2) we assume a bivariate normal distribution for W_2, W_3 conditional on W_1, Y:

$$W_2, W_3|W_1, Y \sim N\left\{ \begin{pmatrix} E(X|W_1,Y) \\ E(X|W_1,Y) \end{pmatrix}, \begin{pmatrix} \text{var}(X|W_1,Y)+\sigma^2_{u_2} & \text{var}(X|W_1,Y) \\ \text{var}(X|W_1,Y) & \text{var}(X|W_1,Y)+\sigma^2_{u_2} \end{pmatrix} \right\}$$

$E(X|W_1, Y)$ and $\text{var}(X|W_1, Y)$ can be estimated by maximum likelihood, letting $E(X|W_1, Y) = \alpha_0 + \alpha_1 W_1 + \alpha_2 Y$, $\text{var}(X|W_1, Y) + \sigma^2_{u_2} = \exp(\theta_1)$, and $\text{var}(X|W_1, Y) = \exp(\theta_1 + \theta_2)/(1 + \exp(\theta_2))$ (Freedman et al., 2008).

To find $X_{IM}(W_1, W_2, W_3, Y)$ we think of $X, W_2|W_1, W_3, Y$ or $X, W_3|W_1, W_2, Y$ as having a bivariate normal distribution and use results similar to those given in (18).

For group-SIMEX in scenario (c) $\sigma^2_{u_1}$ can be estimated by assuming a multivariate normal distribution for W_1, W_2, W_3.