Received: 9 December 2018 | Revised: 25 February 2020 | Accepted: 28 February 2020

DOI: 10.1002/sim.8532

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment

Ruth H. Keogh[1] | Pamela A. Shaw[2] | Paul Gustafson[3] | Raymond J. Carroll[4,5] | Veronika Deffner[6] | Kevin W. Dodd[7] | Helmut Küchenhoff[8] | Janet A. Tooze[9] | Michael P. Wallace[10] | Victor Kipnis[11] | Laurence S. Freedman[12,13]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

[2]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[3]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada

[4]Department of Statistics, Texas A&M University, College Station, Texas, USA

[5]School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, New South Wales, Australia

[6]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany

[7]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, USA

[8]Department of Statistics, Statistical Consulting Unit StaBLab, Ludwig-Maximilians-Universität, Munich, Germany

[9]Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

[10]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Measurement error and misclassification of variables frequently occur in epidemiology and involve variables important to public health. Their presence can impact strongly on results of statistical analyses involving such variables. However, investigators commonly fail to pay attention to biases resulting from such mismeasurement. We provide, in two parts, an overview of the types of error that occur, their impacts on analytic results, and statistical methods to mitigate the biases that they cause. In this first part, we review different types of measurement error and misclassification, emphasizing the classical, linear, and Berkson models, and on the concepts of nondifferential and differential error. We describe the impacts of these types of error in covariates and in outcome variables on various analyses, including estimation and testing in regression models and estimating distributions. We outline types of ancillary studies required to provide information about such errors and discuss the implications of covariate measurement error for study design. Methods for ascertaining sample size requirements are outlined, both for ancillary studies designed to provide information about measurement error and for main studies where the exposure of interest is measured with error. We describe two of the simpler methods, regression calibration and simulation extrapolation (SIMEX), that adjust for bias in regression coefficients caused by measurement error in continuous covariates, and illustrate their use through examples drawn from the Observing Protein and Energy (OPEN) dietary validation study. Finally, we review software available for implementing these methods. The second part of the article deals with more advanced topics.

KEYWORDS

Berkson error, classical error, differential error, measurement error, misclassification, nondifferential error, regression calibration, sample size, SIMEX, simulation extrapolation

[11]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, USA

[12]Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel

[13]Information Management Services Inc., Rockville, Maryland, USA

Correspondence
Laurence S. Freedman, Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer 52621, Israel.
Email: lsf@actcom.co.il

Funding information
Natural Sciences and Engineering Research Council of Canada (NSERC), Grant/Award Number: RGPIN-2019-03957; Patient Centered Outcomes Research Institute (PCORI), Grant/Award Number: R-1609-36207; National Institutes of Health (NIH), Grant/Award Numbers: NCI P30CA012197, U01-CA057030, R01-AI131771

1 | INTRODUCTION

Measurement error and misclassification of variables are frequently encountered in epidemiology and involve variables of considerable importance in public health such as smoking habits,[1] dietary intakes,[2] physical activity,[3] and air pollution.[4] Their presence can impact strongly on the results of statistical analyses that involve such variables. However, more often than not, investigators do not pay serious attention to the biases that can result from such mismeasurement.[5] We provide in two parts an overview of the types of error that can occur, their impacts on results from analyses, and statistical methods to mitigate the biases that they cause. Throughout, we endeavor to retain a utilitarian approach and to relate theory to practice.

Our focus throughout is on studies which are aimed at either describing the distribution of variables in a population or at understanding the relationships between variables, that is, on etiology. In studies in which the aim is instead to derive a prediction model, the considerations surrounding error-prone variables can be quite different. We also focus on relationships between a single outcome and covariates, excluding longitudinal data modeling and also survival problems with time-varying covariates.

In this first part, we provide a description of the most commonly used statistical models for measurement error and misclassification (Section 2), and the impact of such errors on estimated coefficients in regression models frequently used in epidemiology (Section 3). We describe ancillary studies needed to provide information about the measurement error model and an overview of reference measurements available for some of the most common exposures encountered in epidemiology that are measured with error (Section 4). Study design issues that are impacted by measurement error are discussed and we outline calculations for determining sample size requirements and power (Section 5). We then present two of the simpler—and more commonly used—methods used to adjust for the bias caused by measurement error in estimated regression coefficients: regression calibration (RC) and simulation extrapolation (SIMEX, Section 6). Available software for implementing such methods is listed (Section 7). The methods are illustrated by examples from a real study, somewhat simplified so as to retain clarity.

There have been a number of tutorial articles, chapters, and books on measurement error. The book of Carroll et al[6] provides a wide-ranging statistical account of issues of measurement error. The book of Buonaccorsi[7] provides background and detail on many of the topics that we discuss in this article, including misclassification and measurement error in linear models, and Gustafson's book[8] focuses on Bayesian analysis methods and considers both measurement error and misclassification. The recent book of Yi[9] focuses especially on survival analysis and longitudinal settings, which we do not cover in our two articles. The tutorial article of Keogh and White[10] discusses several methods for measurement error correction with a focus on nutritional epidemiology, Armstrong[11] provides an overview of the impact of measurement error in studies of environmental and occupational exposures, and there have been many others focusing on specific aspects. A detailed account of issues surrounding error in covariates is given in a chapter by Buzas et al,[12] covering types of error, study design considerations, and methods of analysis, with an emphasis on likelihood-based methods.

Our two articles provide a comprehensive overview of measurement error issues, from describing types of error and its impact to study design and analysis, with an emphasis on providing practical guidance. Specific topics covered in this first part that do not appear in the chapter of Buzas et al[12] include errors in the outcome variable, the effects of measurement error on estimating distributions, clarifications concerning the effects of Berkson error in multivariable models, a review of reference measurements used in different areas of epidemiology, discussion of sample size for ancillary studies, and an overview of software for RC and SIMEX. In Part 2, we additionally consider a number of advanced topics including additional methods to address covariate error, methods to estimate a distribution of an error prone covariate, methods to address outcome error, mixtures of misclassification and classical error, mixtures of Berkson and classification error, and approaches to handle imperfect or a lack of validation data.

2 | THE MAIN TYPES OF ERROR

In the statistical and epidemiological literature, one can find two separate terms for errors in variables: measurement error and misclassification. The former term is typically used for continuous variables, such as dietary intakes, and the latter term is used for categorical (including discrete) variables such as level of educational attainment. In this section, we describe the types of error that occur in continuous and categorical variables.

Suppose that we are interested in learning the regression relationship between a scalar outcome variable Y and variables X and Z. We will call the latter, X and Z, covariates. They could both be vectors, but for simplicity we start with X being scalar. We suppose that whereas the Z variables are measured exactly, X is measured with error, with the true value of X being unobserved. We denote the error-prone observed variable by X^* (although sometimes it is denoted by W in the statistical literature). To understand and measure the impact of the mismeasurement of X, we have to know the relationship of the observed X^* to the unobserved X. This relationship is specified in a statistical model.

2.1 | Measurement error in covariates (continuous variables)

There are several sources of measurement error in continuous variables. One is instrument error, arising due to limitations of the instruments used to measure the exposure. Another is error due to self-reporting. It is also common that the exposure of interest is an underlying "usual level," or average level over a defined period, of a quantity that fluctuates almost continually over time. This applies particularly when the exposure is a biological measurement (such as weight or blood pressure). In this case the true exposure may never be observed and we discuss this setting further in Section 4.2 where issues of study design, including reproducibility, are considered.

When X and X^* are continuous, their relationship is defined by a *measurement error model*. The simplest case is known as the *classical measurement error* model[13] and is defined by:

$$X^* = X + U, \tag{1}$$

where U is a random variable with mean 0, and is independent of X. Classical errors have been assumed quite frequently, although not universally, in laboratory and objective clinical measurements, for example, when modeling the relationship between serum cholesterol[14] or blood pressure and heart disease.[15] An extension of this model that is more suitable for some measurements, particularly self-reports, is the *linear measurement error* model,[13] in which

$$X^* = \alpha_0 + \alpha_X X + U. \tag{2}$$

This model describes a situation where the observed measurement includes both random error (U, with mean 0 and independent of X) and systematic error, allowing the latter to depend on the true value X. Classical error is included as a special case of this more general model (2), occurring when $\alpha_0 = 0$ and $\alpha_X = 1$. In model (2), α_0 can be said to quantify location bias—bias independent of the value of X, and α_X quantifies the scale bias—bias that depends proportionally on the value of X. The linear measurement error model has been used widely to describe the error in self-reports of dietary intake (eg, Freedman et al[16]), and in some versions the parameter α_0 has been specified as a random variable that varies across individuals.[17] Further extensions of the linear measurement error model allow X^* to depend also on other variables $\widetilde{Z}$ that may, or may not, be the same as the exactly measured variables Z in the outcome model.[18] Another extension is to allow the variance of U to depend on X (eg, Spiegelman et al[19]).

Of course, the relationship between X^* and X may, in practice, be of other forms than the linear model (2). There is a large literature on such models (see, for example, Chen et al[20]), but relatively few applications of them in epidemiology, where the most common approach has been to use transformation of the variables to recover, at least approximately, the linearity of the relationship. The most common transformation employed is the logarithmic[21] but power transformations have also been used.[22] Assuming the classical error on the log scale, that is, $\log X^* = \log X + U$, corresponds to multiplicative error.

In models (1) and (2), the measurement X^* is viewed as arising from the true value X together with a random error term U that is independent of X. Such a model is suitable for many measurements that are employed in epidemiology. However, in some circumstances, it is instead appropriate to view the true value X as arising from the measured value X^* together with an error U that is independent of X^*. This occurs, for example, when all the individuals in specific subgroups are assigned the average value of their subgroup. In that case the error model should be written:

$$X = X^* + U, \tag{3}$$

where U has mean 0 and is independent of X^*. This is known as the *Berkson error model*.[23] It occurs frequently in occupational epidemiology (eg, Oraby et al[24]) and in air pollution studies (eg, Goldman et al[25]). For example, in a study of second-hand smoke exposure, the exposure of workers to second-hand smoke in different factories may be assigned to be the average level of airborne nicotine obtained from monitoring devices placed in each factory. In air pollution studies, the exposure to certain particles of individuals living in different geographical areas may be assigned as the value from a fixed local air pollution monitor. Berkson error also arises when scores are assigned to individuals on the basis of a prediction or calibration equation, which is often not appreciated (eg, Tooze et al[26]). Suppose that the observed exposure is obtained as the expected outcome (fitted value) from a prediction model for the true exposure based on predictor variables C, $X = \theta_0 + \theta_C^\top C + \epsilon$, giving $X^* = \widehat{\theta}_0 + \widehat{\theta}_C^\top C$. In the event that the prediction model parameters are based on a large sample, it follows approximately that $X = X^* + \epsilon$, that is, the scores from the prediction model are subject to Berkson error. Tooze et al[26] give an example where the exposure of interest is basal energy expenditure, and the observed value is an estimate obtained using a previously-derived prediction equation based on an individual's age, sex, height, and weight. Variables obtained on the basis of prediction or calibration equations should therefore be handled accordingly in analyses that incorporate them. See Section 3 for a discussion in the context of outcome variables measured with error.

The Berkson error model as defined in (3) is additive. However, as for the classical error model, a multiplicative form of the Berkson error model may also be considered, meaning that the additive form holds for the log-transformed variables: $\log X = \log X^* + U$.

When $X^*$and U are normally distributed, then the Berkson error model can be re-expressed as a special case of the linear measurement error model (2), with $\alpha_X = \mathrm{var}(X^*)/(\mathrm{var}(X^*) + \mathrm{var}(U))$ and $\alpha_0 = E(X^*)(1 - \alpha_X)$. In the same vein, under these same circumstances, the linear measurement error model (2) can be re-expressed as a linear regression of X on X^*.

An important issue is whether or not the measurement error U contains any extra information about the outcome Y. When the error contains no extra information about Y, we call the error *nondifferential*. We express this concept formally, by saying the error is nondifferential with respect to outcome Y if the distribution of Y conditional on X, X^*, and Z (denoted $p(Y \mid X, X^*, Z)$) is equal to $p(Y \mid X, Z)$.[6(p36)] Otherwise the error is *differential* with respect to Y. Nondifferential error with respect to Y can also be expressed in terms of the conditional distribution of X^*, $p(X^* \mid X, Y, Z) = p(X^* \mid X, Z)$, and in terms of conditional independence between X^* and Y given X and Z, that is, $p(X^*, Y \mid X, Z) = p(X^* \mid X, Z)p(Y \mid X, Z)$. Error is often nondifferential when variables are measured at baseline in a prospective cohort study, since the measurement precedes the outcome often by a lengthy period (eg, Willett[27(p8)]). Differential error can occur in case-control studies, when measurements may take place after the outcome has occurred. One form of differential error is the well-known problem of recall bias that can occur in case-control studies (eg, Willett[27(p7)]). Here, the errors in responses to questions,

such as "how many cigarettes did you smoke per week?," are dependent on whether the person is a case or a control (the outcome), for example, a person who has or has not been diagnosed with cancer. In general, in analysis, the effects of nondifferential error are easier to correct than the effects of differential error.

Error can also be nondifferential with respect to both Y and Z, whereby X^* is known to be conditionally independent of (Y, Z) given X. This is a stronger form of nondifferentiality than the definition given in the previous paragraph. In fact, this stronger form holds for models (1)-(3), if the random term U is independent of variables Z as well as being independent of Y and X (models (1) and (2)) or X^* (model (3)).

2.2 | Misclassification in covariates (categorical variables)

As in Section 2.1, we are interested in the regression relationship between Y and covariates (X, Z), when the available data are Y, X^*, and Z. Now, however, we take up the case where X and X^* are both categorical variables. For instance, in an observational, questionnaire-based study, some participants may, wittingly or unwittingly, check the wrong box on a yes/no item. As mentioned earlier, when X is categorical the discrepancy between X and X^* is typically referred to as *misclassification* rather than measurement error.

The simplest case to start with is that in which X and X^* are both binary, and the misclassification is nondifferential. For current purposes we assume the stronger form of nondifferentiality, which is tantamount to assuming that the "noise" transforming the latent X into the observable X^* is ignorant of the values of Y and Z. Then, mimicking the linear measurement error model of Equation (2), we can view the regression relationship

$$E(X^*|X, Z, Y) = \alpha_0 + \alpha_X X, \tag{4}$$

as describing the extent of the misclassification. This serves to remind that misclassification is closely related to measurement error. However, unlike the situation with measurement error, (α_0, α_x) are necessarily constrained to ensure that (4) remains between zero and one. In fact, it is more common and intuitive to see nondifferential misclassification expressed in terms of the *sensitivity*, $\text{Sn} = \Pr(X^* = 1|X = 1, Z, Y) = \Pr(X^* = 1|X = 1) = \alpha_0 + \alpha_X$, and *specificity*, $\text{Sp} = \Pr(X^* = 0|X = 0, Z, Y) = \Pr(X^* = 0|X = 0) = 1 - \alpha_0$. See Gustafson[8(sec.3.1)] or Buonaccorsi[7(sec.2.3)] for a more detailed description of this framework.

Following from this, when choosing a model to represent differential misclassification, adding terms involving Y and/or Z to the right-hand side of (4) is unappealing, since a complicated constrained parameter space ensues. Rather, using an appropriate link function is recommended, for example the logit link function.

In principle, the idea of Berkson error also carries over to the misclassification setting. That is, there could be situations where the conditional distribution of X given X^* is the fundamental descriptor of the misclassification. This is seen less often in practical applications, but could arise in an occupational hygiene context, where *group-based exposure assessment* is used (see, for example, Tielemans et al[28]). For instance, all workers in a group, based on a common work location, are assigned the same exposure status, either $X^* = 0$ for unexposed or $X^* = 1$ for exposed; but misclassification arises, because within that group some workers' true exposure X will differ from the assigned value X^*. Berkson misclassification also occurs with pooled samples in the diagnostic setting (see, for example, Peters et al[29]).

Regardless of the type of misclassification that can reasonably be assumed, when X and X^* are both binary, the distribution of X given X^* is usually referred to in terms of *predictive values* (or *reclassification probabilities*), namely the *positive predictive value*, $PPV = \Pr(X = 1|X^* = 1)$ and the *negative predictive value* $NPV = \Pr(X = 0|X^* = 0)$. In the typical non-Berkson situation, Sn and Sp are taken as characterizing the misclassification, with PPV and NPV then being determined by Sn, Sp, and the prevalence of X, $\Pr(X = 1)$. In the atypical Berkson situation, PPV and NPV would characterize the misclassification, with Sn and Sp then being determined by PPV, NPV, and the prevalence of X^*, $\Pr(X^* = 1)$. See Buonaccorsi[7(sec.2.3)] for more details.

An interesting form of misclassification arises when the binary X^* is created by thresholding a continuous variable measured with error, in which case X would correspond to thresholding the continuous variable measured without error. That is, say $X = I\{S > c\}$ and $X^* = I\{S^* > c\}$, where S^* is a nondifferential (with respect to Y and Z) error-prone measurement of a continuous variable S. In a somewhat counterintuitive finding, Flegal et al[30] demonstrated that even if the continuous measurement S^* has nondifferential error, the binary measure X^* may have differential misclassification (see also Wacholder et al[31]). That is, conditional independence of $S^*$and (Y, Z) given S does not imply conditional independence of X^* and (Y, Z) given X.

Intermediate between a binary X and a continuous X is the case of a categorical X with more than two categories. An obvious example in epidemiology is where X represents smoking status, with three options: never, former, or current smoker. In this situation, sensitivity and specificity are subsumed by a matrix of classification probabilities, that is, $p_{ij} = \Pr(X^* = j | X = i)$, for $i, j = 1, \ldots, k$, where k is the number of categories. Work addressing this case includes that of Brenner[32] and Wang and Gustafson.[33]

While it has not commonly arisen, one need not limit misclassification to "square" situations, in which X and X^* share a common number (and labeling and interpretation) of categories. For example, 2 by 3 misclassification could arise if exposure X is truly binary ("exposed" or "unexposed"), but the exposure assessment by an expert (or consensus of experts) classifies each subject as "likely unexposed," "perhaps exposed" and "likely exposed." A 2×3 matrix of classification probabilities would then describe the extent of misclassification. Wang et al[34] considered this setting.

2.3 | Measurement error and misclassification in an outcome variable

In previous sections, we have discussed error that occurs in measuring covariates X. However, it is also possible that the outcome Y is measured with error, through the error-prone observation Y^*. As we will see in Section 3, the effects of measurement error in an outcome variable are different from those in a covariate. The effects of measurement error in an outcome variable have tended to be under-studied in the past relative to error in covariates, but there is now a growing recognition of their potential impact. As we shall see in Section 3.3, particularly important are the effects of differential error in Y, but here the definition of differential error changes. Differential error in Y occurs when Y^* is dependent on X conditional on Y (or on Y and Z), that is, $p(Y^* | Y, X) \neq p(Y^* | Y)$ or $p(Y^* | Y, X, Z) \neq p(Y^* | Y, Z)$.

Applications where a binary or categorical outcome variable is subject to misclassification have also been discussed in the literature (eg, McInturff et al,[35] Lyles et al[36]). This is somewhat more nuanced than the corresponding situation for measurement error in a continuous outcome variable, as we discuss in Section 3.3.

3 | EFFECTS OF MEASUREMENT ERROR AND MISCLASSIFICATION ON STUDY RESULTS

In this section, we focus on the impact of measurement error in or misclassification of a variable on the results of a study. We consider studies where the main analysis is based upon a model relating an outcome Y to a covariate X via a regression model. We deal first with error in a continuous X, then with misclassification in a categorical X, and finally with error in Y. In the following section on the effects of error in continuous covariates, we build up from the simplest case of a single covariate, measured with error, to the case with additional exactly measured covariates, and finally to the case of multiple error-prone covariates. The same is done in Section 3.2, which refers to misclassified covariates. It should come as no surprise that the effects of error are different for the different types of error that occur, so that this section relies on information we have already presented in Section 2. While it is a common misconception that measurement error in covariates merely leads to attenuation of effect estimates (and is thus considered less of a concern by some), we shall see that this is true only in certain special cases.

3.1 | Effects of measurement error in a continuous covariate

3.1.1 | Single covariate regression

Suppose that our analysis of the relationship between a continuous outcome Y and covariate X is based on a linear regression model

$$E(Y|X) = \beta_0 + \beta_X X. \tag{5}$$

However, because of measurement problems we use X^* instead of X and therefore explore the linear regression

$$E(Y|X^*) = \beta_0^* + \beta_X^* X^*. \tag{6}$$

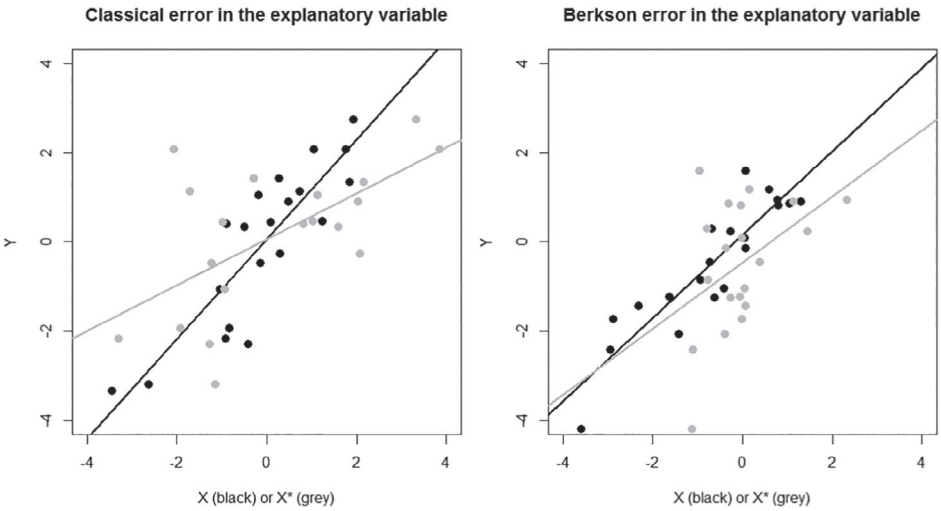

FIGURE 1 Simulated data on 20 individuals showing the effects of classical error and Berkson error in the continuous covariate X on the fitted regression line. For both plots Y was generated from a normal distribution with mean $\beta_0 + \beta_X X$ (using $\beta_0 = 0$, $\beta_X = 1$) and variance 1. *Classical error plot:* X was generated from a normal distribution with mean 0 and variance 1. X^* was generated using $X^* = X + U$. The difference in the slopes in this graph is due to attenuation from the measurement error in X^*. *Berkson error plot:* X^* was generated from a normal distribution with mean 0 and variance 1 and X was generated from the normal distribution implied by the Berkson error model $X = X^* + U$. For both error types $\text{var}(U) = 3$. The small difference in the slopes in this graph is due entirely to sampling error

In assessing the impact of the measurement error on the regression results we will be interested mainly in:

1. whether and how β_X^* is different from β_X, and
2. whether the precision with which we estimate β_X^* is different from the precision with which we estimate β_X, and
3. whether the usual statistical test of the hypothesis that $\beta_X^* = 0$ is or is not a valid test (ie, preserves the nominal significance level) of the hypothesis that $\beta_X = 0$.

When measurement error is classical and nondifferential (model (1)), then $|\beta_X^*| \leq |\beta_X|$, with equality occurring only when $\beta_X = 0$. The measurement error in X^* *attenuates* the estimated coefficient, and any relationship with Y appears less strong. More precisely we can write: $\beta_X^* = \frac{\text{cov}(Y, X^*)}{\text{var}(X^*)} = \frac{\text{cov}(Y, X+U)}{\text{var}(X+U)} = \frac{\text{cov}(Y, X)}{\text{var}(X)+\text{var}(U)} = \frac{\text{var}(X)}{\text{var}(X)+\text{var}(U)} \frac{\text{cov}(Y, X)}{\text{var}(X)} = \lambda\beta_X$, where $\lambda = \frac{\text{var}(X)}{\text{var}(X)+\text{var}(U)}$ lies between 0 and 1 ($0 < \lambda \leq 1$) and is called the *attenuation factor*,[6(p. 43)] or by some the *regression dilution factor*.[37] See the graph on the left-hand side in Figure 1. Clearly, the larger is the measurement error ($\text{var}(U)$), the smaller is the attenuation factor, and the greater is the attenuation. Besides attenuating the estimated coefficient relating Y to X, classical measurement error also makes the estimate less precise relative to its expected value. In other words, the ratio of the expected value of the estimated coefficient to its standard error (SE) is smaller than under circumstances where X is measured without error, that is, $E(\widehat{\beta}_X^*)/\text{SE}(\widehat{\beta}_X^*) < E(\widehat{\beta}_X)/\text{SE}(\widehat{\beta}_X)$, and therefore the statistical power to detect whether it is different from zero is lower. Approximately, the effective sample size is reduced by the squared correlation coefficient between X^* and X, $\rho^2_{XX^*}$, which for this model happens to be equal to the attenuation factor λ.[12,38] When measurement error is substantial (eg, $\lambda < 0.5$), its effects on the results of research studies can be profound, with key relationships being much more difficult to detect.

While measurement error in this single covariate setting results in bias and loss of power, any test of the null hypothesis that $\beta_X^* = 0$ is a valid test of the hypothesis that $\beta_X = 0$, and this is because the relationship $\beta_X^* = \lambda\beta_X$ means that β_X^* equals 0 if and only if β_X equals 0.

When the error in X^* conforms to the linear measurement error model (2), the relationship $\beta_X^* = \lambda\beta_X$ still holds[12] but λ need no longer lie between 0 and 1, since now

$$\lambda = \frac{\alpha_X \text{var}(X)}{\alpha_X^2 \text{var}(X) + \text{var}(U)}. \tag{7}$$

This means that under the linear measurement error model the effect of the measurement error is no longer necessarily an attenuation. Nevertheless, in nearly all applications α_X is positive, so that negative values of λ are virtually unknown, and in many applications, var(U) is sufficiently large to render λ less than 1, even when α_X is less than 1.[39] Regardless of the value of α_X, statistical power to detect the relationship between X and Y is reduced and the effective sample size is reduced by a factor approximately equal to $\rho^2_{XX^*}$. As with classical measurement error, the test of the null hypothesis that $\beta^*_X = 0$ is a valid test of the hypothesis that $\beta_X = 0$. Note that the expression for λ in (7) reverts to the form for classical error when $\alpha_X = 1$.

When the error in X^* has the form of the classical (1) or linear measurement error model (2), but the error is differential, then the above results no longer hold, and an additional contribution to the bias occurs in the estimated coefficient due to the covariance of outcome Y with error U. In particular, the relationship $\beta^*_X = \lambda\beta_X$ no longer holds, so that statistical tests of the null hypothesis that $\beta^*_X = 0$ generally are not valid tests of the hypothesis that $\beta_X = 0$.

When the error in X^* is Berkson (model (3)) and nondifferential, the effects on estimation are very different from those described above. In fact, there is no bias, that is $\beta^*_X = \beta_X$![11] See the graph on the right-hand side in Figure 1. However, as with the classical and linear error models, statistical power is reduced, and the effective sample size is again reduced by the factor $\rho^2_{XX^*}$.

The results in this section relating to the properties of β^*_X are based on assuming the linear model for Y given X in (5) and a linear relation between X and X^*. The linear relation between X and X^* is likely to be appropriate for many practical purposes including when X and X^* are jointly normal. If the relation between X and X^* is nonlinear, then the linear model for Y given X no longer implies a linear model for Y given X^* or vice-versa. However, transformations of X^* (and X) can often lead to approximate linearity and the above results could then be taken as good working approximations. If the linear model for Y given X is misspecified then the above expressions for the attenuation factor λ hold asymptotically. If investigators are particularly concerned that these linearity assumptions do not hold, they may try to find a transformation of the X scale on which the assumptions are more tenable, or pursue more advanced methods to accommodate nonlinearity.

3.1.2 | Regression with a single error-prone covariate and other exactly measured covariates

Most analytic epidemiological studies involve regression models with several covariates. Suppose we wish to relate outcome Y not only to X but also simultaneously to one or more exactly measured covariates, for example confounders, Z, so that

$$E(Y|X, Z) = \beta_0 + \beta_X X + \beta_Z Z, \tag{8}$$

where Z may be scalar or vector. As before, because of measurement problems we use X^* instead of X, and therefore explore the linear regression

$$E(Y|X^*, Z) = \beta^*_0 + \beta^*_X X^* + \beta^*_Z Z. \tag{9}$$

Results concerning β^*_X are similar to those in Section 3.1.1. When the error in X^* conforms to a linear measurement error model, the relationship $\beta^*_X = \lambda\beta_X$ still holds but now $\lambda = \frac{\alpha_{X|Z}\mathrm{var}(X|Z)}{\alpha^2_{X|Z}\mathrm{var}(X|Z)+\mathrm{var}(U)}$,[6] where $\alpha_{X|Z}$ is the coefficient of X in a linear measurement error model for X^* that includes the variables Z as other covariates. This expression is a simple extension of the formula given for classical measurement error by Carroll et al.[6(eq.3.10)] Because $\beta^*_X = \lambda\beta_X$, any test of the null hypothesis that $\beta^*_X = 0$ is a valid test of the hypothesis that $\beta_X = 0$. Also similar to previous results, statistical power to detect the relationship between X and Y is reduced, but now the effective sample size is reduced by a factor equal to $\rho^2_{XX^*|Z}$, where $\rho_{XX^*|Z}$ is the partial correlation of X^* with X conditional on Z.

Note also that in general the coefficients for Z, β^*_Z, are not equal to β_Z, so that estimates of these coefficients from the model with X^* substituted for X will also be biased. This bias will occur in the case of each Z-variable, unless the Z-variable is independent of X conditional on the other Z-variables or $\beta_X = 0$.[6(sec.3.3)] Moreover, due to the form of the bias, any test of the null hypothesis that $\beta^*_Z = 0$ is an invalid test of the hypothesis that $\beta_Z = 0$. These results highlight the impact of measurement error in covariates that are not the main exposure of interest on the results from an analysis, even when the main exposure is measured without error.

With Berkson error, special conditions are required for β_X^* to equal β_X and for β_Z^* to equal β_Z, namely that the error involved in X^*, U (see model (3)), is independent both of Z and of the residual error in regression model (8). Independence of the residual error is the equivalent of the nondifferential error assumption with respect to the outcome Y ($p(X^* | X, Y, Z) = p(X^* | X, Z)$). However, independence of U from Z is not guaranteed, and may even be uncommon. Thus, contrary to general perception, Berkson error in a covariate can indeed cause bias in the conventional estimate of the regression coefficients in multiple regression problems, even when the error is nondifferential. The bias caused by Berkson error that is nondifferential but correlated with Z is a multiplicative one, like the bias caused by nondifferential classical or linear measurement error. Thus the usual test of the null hypothesis remains valid. Except in some special cases, correlation of the Berkson error with Z also causes bias in the usual estimate of β_Z, but in this case the bias is additive, so the usual test of the null hypothesis is invalid. The special cases where there is no bias occur when X is independent of a given Z conditional on the other Zs, or when β_X is zero.

3.1.3 | Regression with multiple error-prone covariates

Often, particularly in nutritional epidemiology, we wish to relate the outcome Y to two or more variables that are each measured with error. For example, in the case of two such variables, the model will be

$$E(Y|X_1, X_2) = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2. \tag{10}$$

Because of measurement problems we observe X_1^* instead of X_1, and X_2^* instead of X_2, and therefore fit the linear regression model

$$E(Y|X_1^*, X_2^*) = \beta_0^* + \beta_{X1}^*X_1^* + \beta_{X2}^*X_2^*. \tag{11}$$

Results concerning the vectors of coefficients $\beta_X = (\beta_{X1}, \beta_{X2})^T$ and $\beta_X^* = (\beta_{X1}^*, \beta_{X2}^*)^T$ are different from those in the earlier sections. When the errors in the X^* variables conform to the classical measurement error model, their relationship may still be written in the form $\beta_X^* = \Lambda\beta_X$ but now $\Lambda = \text{cov}(X + U)^{-1}\text{cov}(X)$, where cov( ) is a 2-by-2 variance-covariance matrix and X and U are vectors $(X_1, X_2)^T$ and $(U_1, U_2)^T$, the latter denoting the errors in X_1^* and X_2^*, respectively. Writing out this relationship fully we obtain,

$$\beta_{X1}^* = \Lambda_{11}\beta_{X1} + \Lambda_{12}\beta_{X2} \tag{12}$$

$$\beta_{X2}^* = \Lambda_{21}\beta_{X1} + \Lambda_{22}\beta_{X2},$$

where Λ_{ij} $(i, j = 1, 2)$ denotes the (i, j)th element of the 2-by-2 matrix Λ. Thus, the simple proportional relationship between β_{X1}^* and β_{X1} (or between β_{X2}^* and β_{X2}) seen in earlier sections no longer holds. The diagonal terms of the Λ matrix, Λ_{11} and Λ_{22} are still likely to lie between 0 and 1 in most applications, so that, for example, β_{X1}^* will usually contain an attenuated contribution from the true coefficient of X_1 ($\Lambda_{11}\beta_{X1}$), but β_{X1}^* will also be affected by "residual confounding" from the mismeasured X_2 ($\Lambda_{12}\beta_{X2}$). Thus, the estimated coefficients in model (11) may be larger or smaller than the true target values in a rather unpredictable manner. Furthermore, any test of the null hypothesis that $\beta_{X1}^* = 0$ in general will no longer be a valid test of the hypothesis that $\beta_{X1} = 0$, and similarly for the test of $\beta_{X2} = 0$. As we will see in Section 6, in these circumstances, valid inference for β_{X1} (say) requires knowledge about or estimation of the parameters of the measurement error models for both X_1^* and X_2^*.

When the errors in the X^* variables conform to the linear measurement error model, the formulas are similar but slightly more complex, and the concerns over biased estimation and hypothesis testing are identical to those described above for multiple covariates having classical error.

If both covariates are subject to Berkson error then, as in Section 3.1.2, the estimated coefficients are unbiased only in special circumstances. Assuming the errors are nondifferential with respect to outcome Y, one also requires that the Berkson error in X_1^* is independent of X_2^*, and the Berkson error of X_2^* is independent of X_1^*. When bias occurs, it is accompanied by the nonvalidity of the conventional null hypothesis tests that the regression coefficients are zero, as explained in Section 3.1.2.

3.1.4 | Common nonlinear regression models

While the results described in Sections 3.1.1-3.1.3 are derived for linear regression models, they serve as good approximations in many circumstances for other regression models that specify a linear predictor function, for example, generalized linear models:

$$h(E(Y|X,Z)) = \beta_0 + \beta_X X + \beta_Z Z. \tag{13}$$

The approximation is usually good if the measurement error is small or the magnitude of β_X remains small to moderate, although the definition of "small to moderate" depends on the form of the model. Details for the commonly used logistic regression model, among others, are given by Carroll et al.[6(sec.4.8)]

For the Cox proportional hazards model, the relation between β_X^* and β_X (now log hazard ratios) is complicated by the longitudinal nature of the analysis, and by the changing set of individuals remaining at risk.[40] However, sometimes in epidemiological problems, event rates are very low and drop-outs are entirely random, so that the covariate distribution of individuals at risk remains stable throughout the follow-up. In such circumstances the results for linear regression models above would provide good approximations.[41] Carroll et al[6] discuss the impact of measurement error in nonlinear regression models and provide expressions for bias based on higher-order approximations.

3.2 | Effects of misclassification in a binary covariate

3.2.1 | Single covariate regression

Having considered continuous covariates measured with error, we now turn to the case of binary or categorical covariates that are misclassified. We start with a continuous outcome Y and a single binary covariate X that is misclassified as binary X^*, used in models described by Equations (5) and (6). In the situation of a binary covariate, the interpretation of the coefficient β_X is as a difference in the mean outcome between those with $X = 1$ and those with $X = 0$. Assuming nondifferential misclassification, as described by sensitivity Sn and specificity Sp, β_X^* is attenuated. The attenuation factor $\lambda = \beta_X^*/\beta_X$ is determined by Sn, Sp, and $\Pr(X = 1)$. For examples, see Figure 2. Given that the positive predictive value (PPV) and negative predictive value (NPV) for X^* as an error-prone measurement of X are themselves determined by Sn, Sp, and $\Pr(X = 1)$, Gustafson[8(sec.3.1)] argues that the attenuation factor is most intuitively expressed as

$$\lambda = \text{NPV} + \text{PPV} - 1. \tag{14}$$

This gives a direct expression that shows how a more error prone measurement of X yields more attenuation in estimating the coefficient for X.

Aside from the particular form of the attenuation factor, the other messages from Section 3.1.1 remain unchanged. Testing the null hypothesis that $\beta_X^* = 0$ is a valid test of the hypothesis that $\beta_X = 0$. However, the test based on (Y, X^*) data has lower power than the ideal test based on (Y, X) data, since the association between Y and X^* is necessarily weaker than the association between Y and X.[38,42]

Returning to the magnitude of attenuation, Equation (14) also permits identification of problematic situations. For example, say $\Pr(X = 1)$ is close to zero (a "rare exposure"). Then, since $\text{PPV} = \{1 + (\Pr(X = 0)/\Pr(X = 1))((1 - \text{Sp})/\text{Sn})\}^{-1}$, even a relatively high specificity can produce a very low PPV. For instance, if $\Pr(X = 1) = 0.01$, a specificity of 0.9 nevertheless leads to a PPV less than 0.1. In turn this produces massive attenuation.[8(sec.3.1)]

3.2.2 | Regression with a single misclassified covariate and other exactly measured covariates

In Section 3.1.2, we considered using, by necessity, the linear regression of Y on X^* and Z when really wanting to regress Y on X and Z. When Z is scalar (of any type), X is binary and the misclassification is nondifferential (with respect to Y and Z), an expression for the attenuation factor $\lambda = \beta_X^*/\beta_X$ is given in Section 3.2 of Gustafson's book.[8] The result does not depend on assuming that the linear model for Y given (X, Z) is correctly specified. Rather, the attenuation factor is

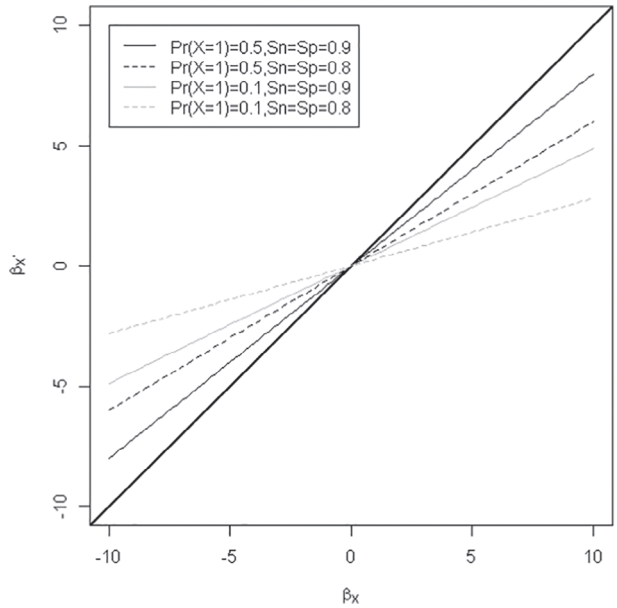

FIGURE 2 Effects of nondifferential misclassification in binary X on the regression coefficient in a linear regression of continuous Y on X. β_X is the regression coefficient in a regression of Y on X and β_X^* is the regression coefficient in a regression of Y on X^* (misclassified X). The attenuation factor λ (Equation (14)) is a function of $\Pr(X = 1)$ and the sensitivity (Sn) and specificity (Sp) of X^*. The thick line is the line $\beta_X^*=\beta_X$

defined as the ratio of large-sample limits for the estimated regression coefficients. The expression for the attenuation factor is unwieldy and not reproduced here. Some of its properties, however, are intuitively helpful. In particular, when X and Z are uncorrelated, the attenuation reduces to Equation (14). Also, with all other aspects of the problem fixed, the attenuation factor decreases as the magnitude of the correlation between X and Z increases. This permits an expansion of the list of problematic situations. We have already mentioned as problematic modest misclassification (and imperfect specificity particularly) of a rare exposure, but if that rare exposure X is strongly associated with a precisely measured covariate Z, then even stronger attenuation occurs.

Again, the main message about hypothesis testing is the same as for continuous X (Section 3.1.2). Assuming nondifferential misclassification, Y and X^* will be associated given Z if and only if Y and X are associated given Z. Hence a test for $\beta_X^* = 0$ using available data will be valid as a test for $\beta_X = 0$, but will have less power than could be achieved were (Y, X, Z) data available.

Also in line with the case of continuous X, the misclassification of X implies that the coefficients of Z estimated from regressing Y on X^* and Z are biased for the coefficients of Z in the Y given X and Z model. When Z is scalar, an explicit expression for this bias is given in Gustafson's book.[8(sec.3.2)]

3.2.3 | Regression with multiple misclassified covariates

Situations where two or more categorical covariates are subject to misclassification have not received very much attention, either theoretically or in practice. The added complexity discussed for the continuous case in Section 3.1.3 applies here as well. Even if the model of Equation (10) holds for X_1 and X_2 that are both binary, and even if the misclassification mechanism is simple, unwieldy forms for Equation (12) result. It is worth considering what "simple" or "nicely behaved" can mean in the face of two binary covariates subject to misclassification. We could assume, for example that as a pair (X_1^*, X_2^*) have nondifferential misclassification, and we could further assume "independent errors," that is, conditional independence of X_1^* and X_2^* given (X_1, X_2). Under these assumptions, and for given sensitivity and specificity of each error-prone measurement, it is easy to determine $E(Y|X_1^*, X_2^*)$ from $E(Y|X_1, X_2)$. However, we do not obtain simple and interpretable expressions. In particular, and in line with Section 3.1.3, the X_1^* coefficient will be a sum of a term involving β_{X1} and a term involving β_{X2}. A simple attenuation structure does not emerge.

3.2.4 | Other common situations

What we know about the impact of a misclassified covariate in a linear model for a continuous outcome carries over approximately, but not exactly, to generalized linear models for other types of outcomes. Closed-form expressions are

elusive here. For example, Gustafson[8(sec.3.4)] gives a numerical algorithm to determine the large-sample limits of logistic regression of a binary Y on X^* and Z, when the logistic regression of Y on X and Z is of interest. For fixed β_0 and β_Z, β_X^* is seen to vary *almost* linearly with β_X, and this relationship varies *only slightly* with β_0 and β_Z. As with so many other statistical concepts, what holds exactly in the linear model holds approximately in the generalized linear model.

However, such intuitions about attenuation do *not* extend to misclassification of a categorical covariate having more than two categories. Recall that a matrix of misclassification probabilities governs such misclassification, with entries $p_{ij} = \Pr(X^* = j | X = i)$. Even given nondifferential misclassification, it is straightforward to construct a plausible misclassification matrix for which $E(Y|X^*)$ has a different pattern than $E(Y|X)$, in which attenuation does not occur. For example, suppose X is ordinal. Then, in comparing levels $X = a$ and $X = a+1$, $E(Y|X^* = a+1) - E(Y|X^* = a)$ can be larger in magnitude than $E(Y|X = a+1) - E(Y|X = a)$. Some work that investigates the polychotomous case includes Dosemeci et al[43] and Weinburg et al,[44] but our emphasis here is really on the *irregularity* of the impact of misclassification.

3.3 | Effects of measurement error in an outcome variable

In Sections 3.1 and 3.2, we have focused on the effects of error in covariates. We consider now the effects of measurement error in an outcome variable, Y. Recall that the error prone version of Y is denoted Y^*. We assume that covariates are measured without error and, for simplicity, we focus on a single covariate X, though the results extend easily to multiple covariates.

3.3.1 | Continuous outcomes

Suppose that our analysis is based on the linear regression model

$$E(Y|X) = \beta_0 + \beta_X X. \tag{15}$$

Because of measurement error in Y, we instead use the linear regression model

$$E(Y^*|X) = \beta_0^* + \beta_X^* X. \tag{16}$$

As in the case of measurement error in a covariate, our interest is in whether and how β_X^* is different from β_X, whether the precision with which we estimate β_X^* is different from the precision with which we estimate β_X, and whether the usual statistical test of the hypothesis that $\beta_X^* = 0$ is a valid test of the hypothesis that $\beta_X = 0$.

Under the classical error model for Y^*, that is $Y^* = Y + U$ (as in Equation (1)), Y^* is an unbiased measure of Y. Hence the expectation of Y^* is equal to the expectation of Y; $E(Y^*) = E(Y)$. The same holds if we condition on any covariates X; $E(Y^*|X) = E(Y|X)$. Therefore, a regression of Y^* on X yields unbiased estimates of β_0 and β_X, in other words $\beta_0^* = \beta_0$ and $\beta_X^* = \beta_X$. See the graph on the left-hand side of Figure 3. This contrasts with the attenuation effect of classical measurement error in a single covariate X. It follows also that a test of the hypothesis that $\beta_X^* = 0$ is a valid test of the hypothesis that $\beta_X = 0$. Although the regression coefficients are not affected by replacing Y by the error prone measure Y^* when the error is classical, the fitted line from the regression of Y^* on X will have greater uncertainty than that from a regression of Y on X, and the precision with which β_X^* is estimated using Y^* is lower than that with which β_X is estimated using Y. Consequently, the power to detect an association between X and the outcome is lower when using Y^* than when using Y. One way of understanding this is to note that $\mathrm{var}(Y^*|X) = \mathrm{var}(Y|X) + \mathrm{var}(U)$. The additional variability in Y^* compared with Y is absorbed into the residual variance in the regression of Y^* on X, and the variance of the estimator $\widehat{\beta}_X^*$ is a function of the residual variance.

Suppose now, instead, that the error in Y^* takes the linear measurement error form

$$Y^* = \alpha_0 + \alpha_Y Y + U, \tag{17}$$

where U is a random variable with mean 0, constant variance, and independent of Y, as in Equation (2). Under this model we have $E(Y^*) = \alpha_0 + \alpha_Y E(Y)$. It follows, from Equation (15), that $E(Y^*|X) = (\alpha_0 + \beta_0\alpha_Y) + \alpha_Y\beta_X X$. Measurement error of this form therefore results in biased estimates of the association between X and the outcome.

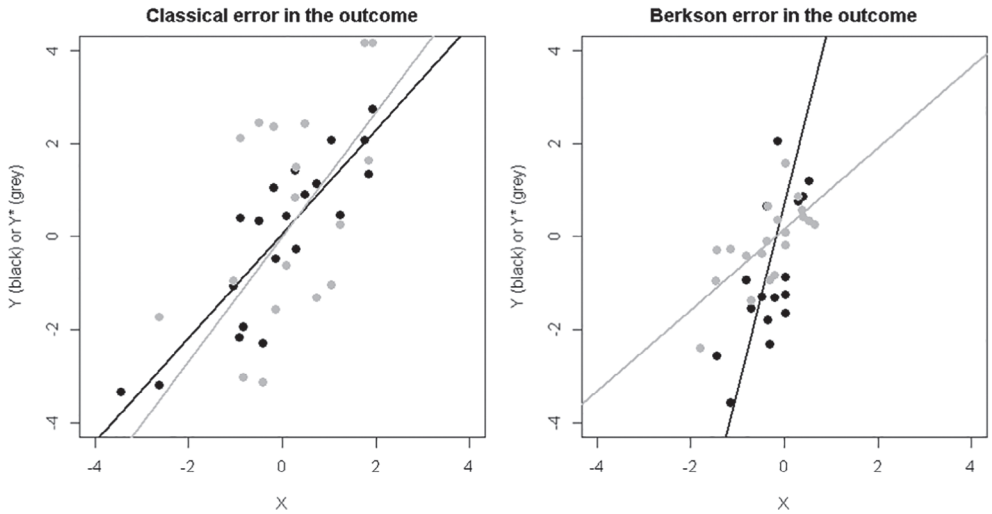

FIGURE 3 Simulated data on 20 individuals showing the effects of classical error and Berkson error in continuous Y on the fitted regression line. For both plots X was generated from a normal distribution with mean 0, variance 1 and the errors U were generated from a normal distribution with mean 0 and variance 3. *Classical error plot:* Y was generated with mean X and variance 1. Y^* was generated using $Y^* = Y + U$. The difference in slopes is due entirely to sampling error. *Berkson error plot:* Y^* was generated with mean X and variance 1. The difference in slopes is due to attenuation from the measurement error in Y. Y given X was generated from the normal distribution implied by the model for Y^* and the Berkson error model $Y = Y^* + U$

In the measurement error models for Y^* considered above, the error is nondifferential with respect to X. One example of when differential measurement error in an outcome could arise is in a randomized study of two treatments (X), in which the nature of the treatments results in differential reporting of the outcome in the two treatment groups. Differential error in Y^* may take the simple classical form, as in Equation (1), but with different error variances, var(U), in the two treatment groups. This does not result in bias in the estimate of β_X^*. However, it does result in heteroscedasticity in the residual variance. More usually, differential measurement error in an outcome could take the form of different degrees of systematic error: $Y^* = \alpha_{0X} + \alpha_{YX} Y + U$ for two groups $X = 0$ and 1. The effect of this is that an estimate of β_X^* is a biased estimate of β_X. The bias may be either towards or away from the null value 0, depending on the form of the differential error.[45]

Outcome variables may also be subject to Berkson error, though this is perhaps less common than Berkson error in covariates. We explained in Section 2.1 how Berkson error arises in variables that are derived as the result of a prediction or calibration equation. Hence Berkson error in an outcome could arise if, instead of observing Y, we observe Y^* which has been obtained as the fitted value from a prediction model for the outcome Y. The Berkson error model (Equation (3)) is $Y = Y^* + U$, where U has mean zero and is independent of Y^*. Here we focus on nondifferential Berkson error, meaning that Y^* and X are independent conditional on Y. To understand the effect of nondifferential Berkson error in an outcome variable, recall that under this error model, and when Y^* and U are normally distributed, the measured outcome follows a linear regression model $E(Y^* | Y) = \alpha_0 + \alpha_Y Y$. Using this result, we can see that the coefficient β_X^* in Equation (16) can be expressed as: $\beta_X^* = \frac{\text{cov}(X,Y^*)}{\text{var}(Y^*)} = \frac{\alpha_Y \text{cov}(X,Y)}{\text{var}(Y^*)}$. The true coefficient of interest from Equation (15) is $\beta_X = \frac{\text{cov}(X,Y)}{\text{var}(Y)}$. Also using the result that $\alpha_Y = \frac{\text{cov}(Y,Y^*)}{\text{var}(Y)}$, we have the relation $\beta_X^* = \frac{\text{cov}(Y,Y^*)}{\text{var}(Y)}\beta_X = \frac{\text{var}(Y^*)}{\text{var}(Y)}\beta_X = \frac{\text{var}(Y^*)}{\text{var}(Y^*)+\text{var}(U)}\beta_X$. Since the ratio $\frac{\text{var}(Y^*)}{\text{var}(Y^*)+\text{var}(U)}$ lies between 0 and 1, the effect of this type of error is to attenuate the estimated regression coefficient.[46] See the graph on the right-hand side in Figure 3.

Recalling that in a simple linear regression of Y on X, Berkson error in covariate X causes no bias in the estimated regression coefficient, one sees that the effects of nondifferential classical error and Berkson error in an outcome variable are the reverse of their effects in a covariate. As with differential classical error, differential Berkson error in an outcome variable may cause over-estimation or under-estimation of β_X.

3.3.2 | Binary outcomes

We have seen in Section 3.3.1 that nondifferential and unbiased measurement error (as in model (1)), yielding a surrogate Y^* for continuous Y, preserves linear model structure, that is, $E(Y^* | X) = E(Y \mid X)$, with the only impact of the

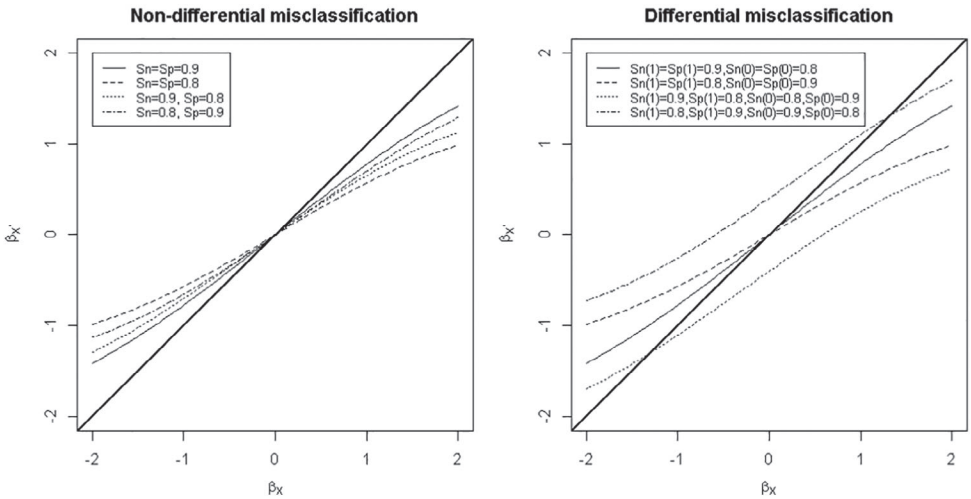

FIGURE 4 Effects of nondifferential and differential misclassification in Y on the log odds ratio. β_X^* is the log odds ratio for Y^* given X (Equation (20)) and β_X is the log odds ratio for Y given X (Equation (19)). The covariate X is binary (for simplicity) and we assume $\beta_0 = 0$ in Equation (19). Sn and Sp denote the nondifferential sensitivity and specificity for Y^*, and Sn*(X)* and Sp*(X)* denote the differential versions for $X = 0, 1$. The thick line is the line $\beta_X^* = \beta_X$

measurement error being an increase in residual variance. However, it is quickly apparent that this same relationship does not hold if Y is categorical, as we now show. Misclassification in a binary outcome (Section 2.2) can be expressed in terms of the sensitivity $\mathrm{Sn}(X) = \Pr(Y^* = 1 \mid Y = 1, X)$ and specificity $Sp(X) = \Pr(Y^* = 0 \mid Y = 0, X)$. The sensitivity and specificity may be differential, that is, dependent on X, or nondifferential, in which case $\mathrm{Sn}(X)$ and $\mathrm{SP}(X)$ do not depend on X. Noting that $E(Y|X) = \Pr(Y = 1|X)$ and $E(Y^*|X) = \Pr(Y^* = 1|X)$, these probabilities are related using the sensitivity and specificity as

$$\Pr(Y^* = 1|X) = (1 - \mathrm{Sp}(X)) + (\mathrm{Sn}(X) + \mathrm{Sp}(X) - 1)\Pr(Y = 1|X), \tag{18}$$

and are equal only when $\mathrm{Sp}(X)$ and $\mathrm{Sn}(X)$ equal 1.

The association between a binary outcome and covariate X is typically modeled using a logistic regression model, for example

$$\log \frac{\Pr(Y = 1|X)}{\Pr(Y = 0|X)} = \beta_0 + \beta_X X, \tag{19}$$

where β_X is the log odds ratio of interest. Using the measured outcome, we would instead fit the model

$$\log \frac{\Pr(Y^* = 1|X)}{\Pr(Y^* = 0|X)} = \beta_0^* + \beta_X^* X. \tag{20}$$

It can be shown that, provided the misclassification in Y is nondifferential with respect to X, meaning that $\mathrm{Sn}(X)$ and $\mathrm{Sp}(X)$ do not depend on X, the impact of the misclassification is that the log odds ratio β_X^* is attenuated relative to β_X.[47] However, if the misclassification is differential, that is if the sensitivity or specificity differ for different values of X, the effect on the log odds ratio can be a bias either away from or towards the null value 0.[7(sec.3.4)] The effects of nondifferential and differential misclassification are illustrated in Figure 4.

3.4 | Effects of measurement error on estimating the distribution of a variable

In some cases, there is interest in describing what we have termed an "outcome" variable not in relationship to other variables, but to estimate the distribution of the variable in the population. Examples include estimating the distribution of food intakes and physical activity levels for a population.[48,49] As above, we consider the true outcome, Y, to be the variable that we want to measure, and its measurement, Y^*, to be an error-prone version.

Most commonly, we are interested in a continuous measure and assume a classical error model for Y^*. As noted in Section 3.3.1, the classical error model leads to $E(Y^*) = E(Y)$, and $\mathrm{var}(Y^*) = \mathrm{var}(Y) + \mathrm{var}(U)$; in other words, the mean of the distribution of Y^* is unbiased, but the variance of Y^* overestimates the variance of Y. Under a Berkson error model, the mean of Y^* is again unbiased for the mean of Y, but $\mathrm{var}(Y^*) = \mathrm{var}(Y) - \mathrm{var}(U)$ so the variance of Y^* *underestimates*

TABLE 1 Effects of measurement error according to type of error and target of the analysis

		Nondifferential error			Differential error
Analysis	**Target**	**Classical**	**Linear**	**Berkson**	**Any**
Single error-prone covariate regression	Regression coefficient	Underestimated	Biased in either direction	Sometimes unbiased[a]	Biased in either direction
	Test of null hypothesis	Valid	Valid	Sometimes valid[a]	Invalid
	Power	Reduced	Reduced	Reduced[b]	Not applicable[b]
Regression with multiple error-prone covariates	Regression coefficients	Biased in either direction	Biased in either direction	Sometimes unbiased[a]	Biased in either direction
	Tests of null hypothesis	Invalid	Invalid	Sometimes valid[a]	Invalid
	Power	Not applicable[b]	Not applicable[b]	Reduced[b]	Not applicable[b]
Regression with error-prone outcome variable	Regression coefficients	Unbiased	Biased in either direction	Underestimated	Biased in either direction
	Tests of null hypothesis	Valid	Valid	Valid	Invalid
	Power	Reduced	Reduced	Reduced	Not applicable[b]
Distribution with an error-prone continuous variable	Mean	Unbiased	Biased in either direction	Unbiased	—
	Lower tail percentiles[c]	Underestimated	Biased in either direction	Overestimated	—
	Upper tail percentiles[c]	Overestimated	Biased in either direction	Underestimated	—

[a]Unbiased and valid only when the Berkson error is independent of the other covariates in the model.
[b]The power of a test is only meaningful when the test of the null hypothesis is valid.
[c]The percentiles affected depends on the distribution of Y.

the variance of Y.[23] For other types of error model, such as the linear measurement error model, the variance of Y^* may vary in either direction from the variance of Y. Thus, the distribution of Y^* is generally biased for important features of the distribution of Y. In Part 2, Section 3, we discuss methods to estimate the distribution of Y when only an error prone Y^* can be observed.

3.5 | Summary of results in this section

This section has dealt with the effects of measurement error and misclassification on the results of commonly used statistical procedures. To provide an overview, we summarize the main results in a table (Table 1). In Section 6, we will consider some analysis methods commonly used with continuous variables to mitigate these effects. However, because these adjustments usually require data from ancillary studies that investigate the measurement error, we will first consider such studies in Sections 4 and 5.

4 | ANCILLARY STUDIES FOR ASSESSING THE NATURE AND MAGNITUDE OF MEASUREMENT ERROR

4.1 | General principles

To adjust estimates and hypothesis tests for the effects of measurement error, one needs information on the measurement error model and its parameters. If at the time of designing the study, the measurement error model and its parameters are fully known then one may use the information to form a correct analysis method. In epidemiology, however, there is

often a severe lack of information about measurement errors and ancillary studies are sorely needed for ascertaining the nature and magnitude of the measurement error.

Ancillary studies involve the use of additional measurements alongside the error-prone measurement X^* to provide information about the measurement error. In practice, the nature of ancillary studies varies according to the type of errors of measurement expected and the availability of more accurate methods of measurement that may be used as reference values. We will see this in Section 4.3, where we present a brief survey of reference instruments available for selected exposures used in epidemiology. There is no universally accepted terminology for the different types of ancillary studies that are used. Here we describe three types, which we refer to as validation studies, calibration studies, and replicates studies.[50]

4.2 | Classification of the types of ancillary study

In a *validation study*, a measurement of the true value of the variable, X, is obtained, as well as the main error-prone measurement X^*, for some individuals. The measure of the true value X is often also referred to as the *reference measurement*. A validation study is the cleanest type of ancillary study and in this case, the model relating X^* to X can be inferred directly from the data.

Suppose X^* follows the linear measurement error model in (2). If the true value X cannot be ascertained, then a measurement that is unbiased at the individual level (ie, that is known to conform to the classical measurement error model (1)) may be used in its place. We call this a calibration study.[51] The unbiased measurement, which we denote X^{**}, is also often referred to as the *reference measurement*, and it comes with the extra requirement that its random errors are independent of the errors of the main error-prone measurement X^*.

A calibration study, as described above, can provide the data for the method of measurement error adjustment known as RC that we will present in Section 6. A single measure using the reference instrument is sufficient to enable use of RC. However, to identify all the parameters of the measurement error models for X^* and X^{**}, the reference measurement with classical error must be repeated within individuals so as to assess the magnitude of its random error. In particular, this enables the correlation coefficient between the main error-prone measurement X^* and the true value X to be estimated. A key assumption in this is that the errors in the repeated measurements obtained using the reference instrument are independent. The repeated measures should be obtained at a sufficiently distant time to ensure such independence, though not so distant that the true underlying value has changed for an individual.

When working with self-reported data, it sometimes occurs that X^* comes from a short questionnaire that is inexpensive to collect and the desire is to validate it against another more intensive self-report procedure that is used as the reference measurement. Unfortunately, this reference measurement may also be biased (although less so), and it is often found that it has errors that are correlated with those using the short questionnaire. This can be considered a type of calibration study, but one with an imperfect reference measurement. In this case, the calibration study will provide estimates of the parameters of the measurement error or misclassification model that are somewhat biased.

A special case, commonly occurring in epidemiology, is where it is assumed that the main error-prone measurement X^* has classical measurement error. Then the parameters of the measurement error model may be estimated from repeated applications of the main error-prone measurement method within individuals. No measurements of the true value of the variable are required. We refer to an ancillary study of this type as a *replicates study*; it is also sometimes known as a *reproducibility study* or a *reliability study*. Under the assumption that the errors in the repeated measurements of X^* are independent, the data from a replicates study are sufficient to estimate the parameters of the classical measurement error model. Carroll et al[6(sec.1.7)] describe how to use data from such a study to check whether the assumptions of the classical model really hold. It is sometimes found that the classical model holds only after a transformation (eg, logarithmic) of the variable.

Ancillary studies of the three types described above are best nested within the main study. For example, a subgroup of participants in a cohort study may be asked to provide not only the main error-prone measurement of exposure but also the additional measurement(s), these being the true measurement X (validation study), the reference measurement (calibration study), or the repeated measure (replicates study). In this case, the study is called an *internal* study. It may be usually desirable that the subgroup of participants are, as far as possible, a random sample (simple or stratified) of those in the main study. In settings where $E(X|X^*, Z)$ is being estimated via a regression, then sampling schemes stratified on variables in this regression could achieve better precision of the coefficients in the calibration equation compared with simple random sampling. For example, in linear regression, oversampling the

extremes and increasing the variance of X^*can be more efficient than simple random sampling in terms of decreasing the variance of the estimated regression parameters; optimal sampling schemes for multivariable regression can also be derived.[52]

Ancillary studies that are conducted on a group of individuals not participating in the main study are called *external* studies. External studies are less reliable than internal ones for determining the parameters of the measurement error model, since the estimation involves an assumption of *transportability* between the group of participants in the ancillary study and the group participating in the main study. Carroll et al[6(sec.2.2.5)] describe the dangers of transporting a model derived from an external study. However, in many circumstances, the only information available about the measurement error comes from an external study, and careful use of such information (accompanied by sensitivity analyses) can add greatly to the understanding of results (see Part 2, Section 6 of our article).

Estimation of the error variance in a Berkson model is often problematic, since in these applications reference measurements are typically difficult to obtain. In some applications, the Berkson error comes from the use of an X^* derived from a prediction equation for X, and in that case the residual error variance estimated from the source data that yielded the equation can serve as the Berkson error variance estimate. See, for example, Tooze et al.[26]

In Section 5 we will discuss the desirable size of a validation, calibration, or replicates study. For further reading on these types of study see Kaaks et al.[53]

We will see each of the types of study described above in the following survey of exposure measurements in different areas of epidemiology. Before proceeding to the survey, it should be noted that, when reporting the results of validation, calibration or replicates studies, most investigators limit themselves to presenting correlations between the measurements from their instrument and the reference instrument (sometimes adjusting for the within-person variation in the reference measurement). However, they usually do not use the information from their study to determine the measurement error model and its parameters. As a result the information required for adjusting estimates in the main study for measurement error or misclassification is not reported or used, and the study is used simply to report that the study instrument has been "validated"![5] Investigators should be encouraged to use ancillary study data to better interpret the results of their main study.

This section has implicitly focused on the situation in which error is nondifferential. However, similar principles apply when there is differential measurement error. In that case, parameters of the measurements error model depend on the outcome Y, and data in the ancillary study should be obtained in such a way that all relevant parameters can be estimated. In particular, the ancillary study requires information on the outcome. This is discussed further in Part 2, Section 2 where measurement error correction methods that address differential error are outlined.

4.3 | Reference instruments available for selected exposures used in epidemiology

4.3.1 | Nutrition

Of all the areas of epidemiology, nutritional epidemiology has probably paid the most attention to measurement errors of exposure.[54] Nearly all nutritional epidemiological studies rely on self-reported dietary intakes as their main measure of exposure. However, these are known to be subject to considerable error, especially if exposure is defined as the usual (or average) intake over a long period, the measure that is thought to be of most relevance to the epidemiology of chronic diseases. There is no known way of getting an exact value of this measure, so true validation studies do not exist.

For a few dietary components (energy, protein, potassium, and sodium) unbiased measurements of short-term intake exist—they are called recovery biomarkers—and can be used as the reference measurements in calibration studies.[53] For all other dietary components—foods (eg, vegetables, meat) and other nutrients (eg, fat, fiber)—the usual practice is to rely on a second more accurate self-report method as the reference measurement (calibration study with an imperfect reference measurement). When the main measurement X^* is a food frequency questionnaire (FFQ)—a relatively short questionnaire that asks the individual to report on average intake over the past several months (up to 12 months)—24-hour recalls or multiple-day food records in which the participant reports on intakes over a short period in the immediate past are used as the reference. This is less than ideal, since these more accurate methods are nevertheless somewhat biased and also have errors that are correlated with the errors in the FFQ report. However, they are the best method currently available.[55] Prentice and others, using data from a unique large feeding

study,[56] are currently engaged in expanding the list of dietary components for which unbiased measurements are available.

Another type of reference instrument used is a (nonrecovery) biomarker that is related to the intake of the nutrient or food consumed (eg, serum cholesterol for saturated fat intake). Such biomarkers are usually subject to a high degree of metabolic regulation that varies across individuals and consequently do not provide an unbiased measure of intake, and are not, by themselves, helpful in determining the measurement error model, although they have been used alongside other methods, to gain understanding of measurement error.[57,58]

The measurement error model for self-reported dietary intakes has been shown not to conform to the classical model,[17] so replicates studies are not a true option. However, many investigators, in the absence of anything better, have adopted the assumption that 24-hour recalls provide unbiased measurements and have based measurement error adjustment on studies of repeated measurements from this instrument (eg, Beaton et al[59]).

4.3.2 | Physical activity

As in nutritional epidemiology, in large studies physical activity has mostly been assessed by self-reports using questionnaires, of which there are many variants. A large number of smaller studies have been conducted to "validate" these questionnaires (see the National Cancer Institute (NCI) website https://epi.grants.cancer.gov/paq/validation.html) and a few studies have used a measurement error model framework to adjust estimated associations of physical activity with health outcomes, for example Spiegelman et al,[60] Ferrari et al,[3] Nusser et al,[61] Tooze et al,[26] Neuhouser et al,[62] Lim et al,[63] Matthews et al,[64] and Shaw et al.[65] The reference instruments that have been used generally fall into three main categories: doubly labeled water, accelerometers, and physical activity diaries. Doubly labeled water is a technique used to obtain an unbiased measure of total energy expenditure (TEE) and it is useful for determining the measurement error model for TEE measured by a questionnaire through a calibration study.[66]

Since many physical activity questionnaires and recalls are designed to measure physical activity level (PAL) which is defined as the ratio of TEE to basal energy expenditure (BEE), then for determining a measurement error model for PAL, a reference measure for BEE is also needed, and may be provided by direct or indirect calorimetry. BEE has also been estimated by using a prediction equation; however, this measure of BEE exhibits Berkson error (Equation (3)). In this case, it may be necessary to have a calorimetry measure of BEE on at least a subset of participants to form a suitable reference measurement for PAL.[26]

Unbiased reference measures of other physical activity measurements that can be derived from questionnaires, such as hours of moderate or vigorous activity, are currently lacking. Accelerometers do provide information on such measurements and, although not completely unbiased, they may be used as the reference in a calibration study with an imperfect reference measurement. Physical activity diaries are more accurate than questionnaires,[3] but still rely on self-report. They may therefore be regarded in a similar manner to 24-hour recalls or multiple-day records of food intake, not ideal as references but usable in circumstances where other references, such as accelerometers or doubly labeled water, are infeasible or unsuitable (eg, the activity measure of interest is something that cannot be measured by either of them, such as the amount of time spent in anaerobic exercise).

4.3.3 | Smoking

Since smoking is a causal factor in a range of chronic diseases, it is often collected as a potential confounding variable in chronic disease epidemiology studies. The usual mode of collection is through self-report questionnaires. The most common method of "validation" of self-reported smoking status is biochemical. Three different metabolites may be measured: thiocyanate in the blood, urine or saliva; cotinine in the saliva, blood, urine or hair; and exhaled carbon monoxide.[67] Although these measurements are made on a continuous scale, they have been used mostly as binary indicators (smoker or nonsmoker) using a predetermined cut-off point (that has varied among investigators). Thus, these studies report misclassification rates (sensitivity and specificity), rather than measurement error model parameters or correlations. When calculating these rates, the investigators have usually assumed that the biochemical measurement yields the true smoking status, and, thus, that the study is a validation study. Biochemical validation has been used most frequently in smoking cessation trials, where it has become the standard method of assessing the outcome.[68] Its use in observational studies is less widespread, but nevertheless many validation studies have been conducted in this setting. For a review of cotinine-based validation studies, see Rebagliato.[69]

4.3.4 | Air pollution

Research studies into links between air pollution exposure and health outcomes have often taken the form of longitudinal studies, where time series of air pollution levels in different geographical areas are compared with levels of a disease, such as asthma, at the same location and time (with a possible lag effect). Exposures are often assessed using a mathematical model that is applied to serial measurements of the concentration of certain particles in the air at fixed locations and to other information such as temperature, wind strength and direction, and topology, so as to provide an estimate of the ambient pollution at a given time and location. Zeger et al[4] discuss the type of measurement error inherent in such estimated exposures when used as measures of exposure at the individual level, and conclude that it is a mixture of Berkson and classical errors (see Part 2, Section 5.1 of our article). The accepted gold-standard for measuring personal exposure is through a personal monitoring device, which is assumed to provide unbiased measures of true exposure. For a review, see Koehler and Peters.[70] For example, exposure at the individual level was recorded in the PTEAM study on a personal monitor for measuring inhalable (PM10) particles or fine (PM2.5) particles[71]; this can then be used in a calibration study. In the Augsburger Environmental Study repeated measures from a personal monitor measuring ultrafine particle concentrations were available from an external sample, giving an external replicates study.[72] For the Augsburger Environmental Study, methods for handling the mixture of classical and Berkson error were developed in Deffner et al[73] and applied to the data. A valuable resource for investigation of measurement error modeling methods are the data from the Nine City Validation Study[74] on personal daily exposure to PM2.5 particles compared to PM2.5 of ambient origin based on the nearest EPA monitor and spatio-temporal smoothed exposure estimates. These data may be requested at the website https://www.hsph.harvard.edu/pm2-5-validation-dataset/.

4.3.5 | Other exposures

There is, of course, no end of other exposures that are relevant to questions of public health and it can be assumed that many of them cannot be measured without substantial error. In some cases, such as blood pressure measurement, a gold standard measurement (intra-arterial blood pressure) exists and validation or calibration studies of more approximate measurement procedures (eg, sphygmomanometer) can be conducted to determine the measurement error model. In other cases, it is thought that the measurement, if not exact, is at least unbiased (eg, serum cholesterol) and replicates studies are sufficient to estimate the statistical magnitude of the error and make corrections for its impact. However, often neither of these situations exists and the best that can be done is to compare the main measurement method used with a method that although imperfect is thought to be better (calibration study with an imperfect reference measurement).[75] For example, body mass index (BMI), as a measure of body fat, may be compared with percent body fat measured by bioelectrical impedance analysis.[76] In Part 2, Section 6 of our article, we discuss these many cases where it is known that the measured exposure is subject to considerable measurement error, but the measurement error model is in some sense unknown.

5 | DESIGN OF STUDIES WHERE ONE OR MORE OF THE MAJOR COVARIATES IS MEASURED WITH ERROR

In view of the impacts that measurement error or misclassification have on study results, it is advisable to take account of the error at the design stage. To do that, aside from defining the main aim of the study and its target estimate, one needs information on the measurement error model and its parameters. Only then can one make the appropriate adjustment to the design in the form of a change in sample size, or more fundamentally a change in the measurement of the error-prone variable in question. In addition, as we already showed in Section 4, the information about the measurement error model plays a central role in making adjustments for measurement error in the analysis of the main study. A number of authors have discussed design issues in the presence of measurement error.[12,77-79]

In Section 4, we described ancillary studies that are conducted to obtain information about the measurement error model. We will now deal with determining how large such studies should be and then proceed to methods for calculating statistical power in the presence of measurement error as an aid to deciding on the design of the main study.

Our focus in this section is on the size of the ancillary study. An alternative scenario is that a total cost is assigned and that sample sizes for the main study and ancillary study are derived according to requirements for meeting a specified objective. More work is needed on sample size calculations for this situation.

5.1 | Size of ancillary studies

Relatively little attention has been paid to the appropriate size of an ancillary substudy, and we provide here a guideline. A first principle is that the size of the ancillary study is decided in relation to its contribution to the main study's goal. Therefore, we must specify, among other things, what is the target of interest in the main study, and what is the desired statistical power for testing the exposure-outcome association.

As in Section 3.1.1, we limit ourselves to the situation of a single exposure X that is measured by X^* with nondifferential linear measurement error, and focus on the aim of estimating the slope in a simple linear regression model of a health outcome Y on X. To recap, we consider the model:

$$E(Y|X) = \beta_0 + \beta_X X. \tag{21}$$

If we were to use X^* in place of X, then we would obtain instead a different regression model:

$$E(Y|X^*) = \beta_0^* + \beta_X^* X^*. \tag{22}$$

In the main study, Y and X^* are observed in all participants. In the ancillary substudy, a "reference" measurement, R , equal to X or an unbiased measurement of X, is observed in addition to X^*. In the terminology of Section 4 we are therefore in the situation of a validation or calibration study. As discussed in Section 3.1.1, the linear model in (22) does not always hold, but is often a good approximation.

Recall from Section 3.1 that, when X^* is measured with nondifferential linear error, the relation $\beta_X^* = \lambda\beta_X$ holds, where λ is the attenuation coefficient. Therefore, a simple "adjusted estimate" of β_X is obtained by dividing the estimate of β_X^* (that comes from the main study) by an estimate of λ that is obtained from the ancillary study (see Section 6). In simple cases, λ is estimated by the slope of the linear regression of R on X^*. The variance of the adjusted estimate of β_X can then be expressed by the approximation[80]

$$\mathrm{var}(\widehat{\beta}_X) = \frac{\mathrm{var}(\widehat{\beta}_X^*)}{\lambda^2} + \frac{\beta_X^{*2}\mathrm{var}(\widehat{\lambda})}{\lambda^4}. \tag{23}$$

The second term on the right-hand side of (23) represents the extra uncertainty introduced into the estimate of β_X by the uncertainty in the value of λ. We may choose the size of the validation/calibration study to minimize the impact of this extra uncertainty, specifically so that the second term of the right-hand side of (23) will be a small fraction, f, of the first term. For example, if the size of the main study will provide 50% power for a test of the null hypothesis that $\beta_X^* = 0$ at the 5% significance level when the true value is β_X^*, that is, that approximately $\mathrm{var}(\widehat{\beta}_X^*) = \beta_X^{*2}/4$, we obtain $\mathrm{var}(\widehat{\lambda}) = f\lambda^2/4$. If the investigator is planning to test the association with exposure at a two-sided level α (maybe different from 0.05) with power $1-\omega$ (maybe different from 0.5), then the factor 4 may be replaced by $[\Phi^{-1}(1-\alpha/2)+\Phi^{-1}(1-\omega)]^2$, where Φ is the standard normal cumulative distribution function. In the validation study case where the reference measurement R equals X, we can apply the formula for the variance of a regression slope and, simplifying, we obtain the formula for the sample size of the validation study n_v, as:

$$n_v = \frac{\left\{\Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \Phi^{-1}(1-\omega)\right\}^2 (1-\rho_{XX^*}^2)}{f\rho_{XX^*}^2}, \tag{24}$$

where $\rho_{XX^*}^2$ is the squared correlation coefficient between X and X^*.

For example, if $f = 0.1$, $\alpha = 0.05$, $\omega = 0.50$ and $\rho_{XX^*} = 0.4$, then a sample size $n_v = 210$ will ensure that the variance of the adjusted estimate of β_X will not increase by more than about 10% ($f = 0.1$) as a result of the uncertainty in estimating λ. Clearly, the larger the degree of measurement error, the larger the validation/calibration study that will be needed. If

instead of a study with 50% power, the investigator plans a study with 90% power ($\omega = 0.10$), then $n_v = 550$. Thus, the higher the power desired in the main study, the larger the required validation/calibration study.

Note that the quantity ρ_{XX^*} will not be known precisely before the validation study is conducted—indeed it is one of the quantities that we hope to estimate from the validation study data. Consequently, a plausible value will need to be specified in order to use formula (24). This is not unlike needing to specify an exposure's hitherto unknown effect on the outcome in order to calculate the sample size of a main study.

If reference measurement R is an unbiased, but not exact, measure of exposure with random errors, as in a calibration study, then a different formula based on the approximation $\text{var}(\widehat{\lambda}) = f\lambda^2/[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\omega)]^2$, may be derived for n_v. Kaaks et al[53] supply details for the setting where there is a log-linear relation between the incidence rate of disease and error-prone exposure of interest, albeit with very different notation. For problems with two or more exposures measured with error, although the same principles may be used, the sample size formulas have not been derived.

5.2 | Calculating statistical power and sample size in the presence of measurement error

5.2.1 | Continuous X and X^*

In what follows, we assume that the investigator is planning the sample size by specifying the regression coefficient associated with X. In this situation, adjustment for the measurement error needs to be made, as specified below. If, however, the investigator prefers to specify the regression coefficient associated with X^* (thus taking into account within this specification the measurement error) then no further adjustment is required and the material that follows in this subsection is irrelevant.

For simplicity, we assume, as in Section 2.1, that the study is designed in order to elucidate the association between a continuous exposure X and health outcome Y, and that X is univariate. We assume that in the main study we observe X^*, and not X, where X^* measures X with nondifferential linear measurement error. There may or may not be other exactly measured covariates Z that need to be in the model relating Y to X.

In Section 3.1.1, just after Equation (7), we noted that when Y is continuous and is linked to exposure X through a linear regression, then when there are no covariates Z, the measurement error effectively lowers the sample size by $\rho^2_{XX^*}$, the squared correlation between X and X^*. If, however, there are covariates Z, the effect of measurement error effectively lowers the sample size by the factor $\rho^2_{XX^*|Z}$ the squared partial correlation between X and X^* given Z. Thus, to achieve comparable power to what would happen if X were observed, using X^* requires increasing the sample size, sometimes dramatically.

Indeed, the sample size needs to be increased by the factor $1/\rho^2_{XX^*}$, or, where covariates Z are included, by the factor $1/\rho^2_{XX^*|Z}$. For example, if the correlation between X and X^* is 0.40, measurement error means that the required sample size will be $6.25 = 1/(0.40)^2$ times larger than it would be if X were observable.

Consider now logistic and Cox regression. As described in Section 3.1.4, the results given above for linear regression serve as good approximations for logistic, Poisson and Cox regression, under the proviso that β_X remains "small to moderate"—see also Devine and Smith,[81] McKeown-Eyssen and Tibshirani,[82] and White et al.[83] In Tosteson et al,[84] power and sample size issues are described for logistic regression but without the "small to moderate" assumption. These authors show in their Figure 2 an example wherein the sample size actually needs to be increased by another 2%-5% for logistic regression above the increase discussed in the previous paragraph.

This discussion suggests the following strategies for calculating sample size and power in studies where the main exposure X is measured with error:

1. Use any information available about the distribution of X and its anticipated association with Y to set a sample size, n_X, in the ideal event that X could have been observed. Also, for this sample size, n_X, compute the resulting power function, say $\text{pow}_X(\beta_X)$. See Self and Mauritsen[85] for such calculations for generalized linear models. Then, when using X^*, inflate the sample size as described above to $n_{X^*} = n_X/\rho^2_{XX^*|Z}$, and evaluate the power function as $\text{pow}_{X^*}(\beta_X) = \text{pow}_X(\lambda\beta_X)$, where λ is the attenuation factor defined in Section 3.1.
2. For a binary outcome and logistic regression, unless using the method in Tosteson et al,[84] inflate the sample size by an extra 2%-5%.

The following is an example of an adjustment to the sample size calculation required for a hypothetical cohort study of the association between the sodium-potassium (Na-K) intake ratio of an individual with all-cause mortality. Previous

information indicates that the variance of the log of the true usual (ie, long-term average) sodium-potassium intake ratio in the USA has a population standard deviation of 0.35. Suppose also that the expected hazard ratio for all-cause mortality per change of 0.35 in log usual Na-K intake ratio is 1.17, consistent with a study based on a NHANES cohort.[86] Then the number of deaths, D, required to provide $1-\omega$ power to detect such an effect using a two-sided Wald test at level α is given by $D = \left(z_{1-\frac{\alpha}{2}} + z_{1-\omega}\right)^2 / (\sigma_X^2 \beta_X^2)$,[87] where β_X is the log hazard ratio for a change of 1 unit in continuous variable X, and σ_X^2 is the variance of X. Setting $\alpha = 0.05$, $\omega = 0.1$, $\sigma_X^2 = 0.35^2$, and $\beta_X = \frac{\log(1.17)}{0.35} = 0.45$, we obtain $D = 423$. However, the dietary report instrument we wish to use in our cohort is a FFQ, which measures dietary intake with error. From the OPEN validation study of dietary report instruments in which unbiased measures of sodium and potassium intakes were measured,[88] the correlation between true FFQ Na-K intake ratio and the true ratio conditional on age and gender can be estimated as 0.45. Then the study should be designed to observe $423/(0.45^2) = 2{,}089$ deaths. This approximately 5-fold increase in the required observed number of deaths could be achieved by increasing sample size, increasing length of follow-up, including a larger proportion of elderly persons in the cohort, or some combination of these.

5.2.2 | Binary X and X^*

In Section 3.2.1, we described attenuation for a continuous outcome Y when both X and X^* are binary. The guidelines in Section 5.2.1 will be useful also in this case. While formulae for power and sample size determination can be obtained for the binary X and X^* case, to the best of our knowledge there are no publicly available programs to do this. It would be useful to have such programs.

6 | ANALYSIS OF STUDIES WHERE ONE OR MORE OF THE MAJOR COVARIATES IS MEASURED WITH ERROR

In Section 3 we described the impact of measurement error in covariates on study results in the event that no statistical adjustments are made for the presence of that measurement error. In Sections 4 and 5 we outlined what additional information is needed to learn about the form and magnitude of measurement error and how to accommodate the need to address the impact of measurement error at the design stage of a study. We now describe in this section two of the simpler methods for adjusting the statistical analysis so that the resulting estimates of key parameters will be free (or approximately free) of the bias induced by measurement error—RC and SIMEX. We focus on regression problems with one or more covariates that are measured with error that conforms to the classical or linear measurement error model. More complex methods for this problem and methods for other types of problem, including misclassification, are described in Part 2 of this article.

It should be noted that, when the measurement error causes attenuation, the estimates from the two methods described will have variance that is larger than the estimates based on applying the standard analysis as if all variables were measured exactly—such estimates are sometimes referred to as naïve estimates and their bias is described in Section 3.1. How much larger the variance of the adjusted estimates will be will depend on the level of attenuation and the size of the validation, calibration, or replicates study (see Section 4.1). Thus practitioners are often confronted with a bias-variance trade-off in their choice of estimation method. In the analyses below, the estimated standard errors of the estimates derived from the correction methods are made to account for the extra uncertainty introduced by the measurement error.

We will illustrate RC with an example from the OPEN study.[88] This was a dietary intake calibration study (Section 4.2) using unbiased reference measurements, conducted in 484 adult volunteers aged 40-69 years, resident in Maryland, USA in 1999-2000. Participants reported on their dietary intake using a FFQ, provided two 24-hour urines for measuring sodium and potassium intake, and provided samples for measuring total energy intake through doubly labeled water. The target dietary measures are considered to be average daily potassium intake density and sodium intake density, where the density is the ratio of the intake to total energy intake. The issue to be addressed will be the association of these intakes with a person's BMI. The questionnaire responses are considered to have linear measurement error and the urinary data are considered to have classical measurement error (since they measure only a single day's intake with unbiased, independent assay error). The errors in the questionnaire and urinary measurements are assumed to be uncorrelated. Furthermore, the errors in both questionnaire and urinary measurements of potassium and sodium intake density are assumed to be nondifferential with respect to BMI. The dataset is referenced as "Selected OPEN data."[89]

6.1 | Regression calibration

RC is one of the most popular methods of adjusting for nondifferential covariate measurement error. The method is intuitive and relatively easy to implement. For the setting of linear regression of a continuous outcome on covariates, one or more of which is measured with error, RC yields an unbiased estimate of the target regression coefficient if the calibration model (see below) is exactly known, or the parameters of the calibration model can be estimated consistently. There is an extensive literature on this method, with early descriptions including those by Prentice,[41] Carroll and Stefanski,[90] Armstrong,[91] and Rosner et al.[80,92,93]

6.1.1 | Regression calibration with continuous outcome and one covariate measured with error

Suppose that our analysis of the relationship between a continuous outcome Y and variable X is based on the linear regression model $E(Y|X,Z) = \beta_0 + \beta_X X + \beta_Z Z$, where X is scalar and Z may be scalar or vector. Suppose further that X^* measures X with nondifferential linear measurement error. The RC estimator is obtained by performing the regression of outcome Y with the unobserved X replaced, not by X^*, but by the predicted value of X given the observed data X^* and Z, namely $E(X|X^*,Z)$. By iterated expectation, one has:

$$E(Y|X^*,Z) = E_{X|X^*,Z}\{E(Y \mid X^*,Z,X)\} = E_{X|X^*,Z}\{E(Y \mid Z,X)\} = E_{X|X^*,Z}\{\beta_0 + \beta_X X + \beta_Z Z\} = \beta_0 + \beta_X E(X|X^*,Z) + \beta_Z Z.$$

The second equality is due to the crucial assumption that conditioned on X and Z, X^* contains no additional information about Y, that is, that the error in X^* is nondifferential. From this equation we see that by regressing Y on $E(X|X^*,Z)$ and Z, we obtain consistent estimates of the coefficients for both X and Z. Note, this holds true regardless of whether X is continuous, binary, or ordinal so long as the model for $E(X|X^*,Z)$ is correct.

The expression for $E(X|X^*,Z)$ is called the calibration equation. If the calibration equation is linear, namely, $E(X|X^*,Z) = \lambda_0 + \lambda_{X^*}X^* + \lambda_Z Z$, then parameter λ_{X^*} is identical to the attenuation factor discussed in Section 3. In the case of the classical error model (1) of Section 2, when X and measurement error U are normally distributed, one has that $E(X|X^*) = \lambda X^* + (1-\lambda)E(X^*)$, where λ is the attenuation factor discussed in Section 3.1.1. This formulation of the calibration equation helps illustrate the effect of measurement error. When the error variance is small relative to the variance of X, λ is close to 1 and the estimated $E(X|X^*)$ makes only a small adjustment to X^*; when the error variance is very large, $E(X|X^*)$ will be close to $E(X^*) = E(X)$.

As discussed in Section 4, estimation of the parameters in the model for $E(X|X^*,Z)$ requires an ancillary study. In a validation study in which the true measures of X are observed in some individuals (either internal or external to the main study), the calibration equation can be estimated directly through a regression of X on X^* and Z. In a calibration study, such as that in the OPEN study, a reference measurement X^{**} can be obtained on a subset of individuals, and this is thought to follow the classical measurement error model $X^{**} = X + V$, where V is random error with mean zero. In this case $E(X^{**}|X^*,Z) = E(X+V|X^*,Z) = E(X|X^*,Z)$. Thus, one can obtain the predicted value for $E(X|X^*,Z)$ simply by regressing the reference measurement X^{**} value from a similar time period on the observed X^* and Z. In a replicates study in which repeat measurements X_1^* and X_2^* are available on at least a subset of individuals, and where both are assumed to be subject to the classical measurement error model (1), the predicted value for $E(X|X^*,Z)$ can be obtained by regressing X_2^* on X_1^* and Z. Carroll et al[6(sec.4.4.2)] also provides a formula for the best linear approximation of $E(X|X^*,Z)$ using replicate data.

A few authors have considered ways to improve upon the RC estimator, that is to identify methods that give more precise estimates, in some circumstances by making better use of the data. Efficient maximum likelihood methods for the linear and logistic outcome models with a replicates study were described by Bartlett et al.[94] Spiegelman et al[95] described efficient approaches for the logistic outcome model with either validation and calibration substudies; this approach considers a weighted estimator that combines separate estimates of the target parameter from the main study and from the calibration/validation substudy.

It is worth noting that RC can also be used for a misclassified binary covariate when there is a validation study which can be used to estimate $E(X|X^*,Z)$. When X is binary the expectation needed for RC can be estimated from a replicates study with three or more repeated measures of X^* in a subset of individuals. Two replicates can be sufficient in some settings, but this generally relies on either simplifying assumptions about the error model (such as one of sensitivity or specificity are 100% or they are equal) or the two replicates are observed simultaneously with the binary outcome Y. In the latter case, the information regarding the relationship between X and Y can be used

to identify the necessary parameters, but this method has not been observed to perform well when this relationship is weak.[96]

Sometimes, to obtain more precise estimation of X, one may wish to include further covariates $Z^{\#}$ in the calibration equation in addition to the covariates Z originally planned to be in the model for Y.[22] The $Z^{\#}$ variables that may be used for this purpose are those that are predictive of X , but independent of Y given X and Z, and therefore are not included in the outcome model.

The RC method is closely connected to even simpler methods for measurement error correction, which are based on the relation $\beta_X = \beta_X^*/\lambda$ shown in Section 3.1.2, where λ is the attenuation factor and also the slope in the regression of X on X^* and Z in the calibration equation used in RC.[80,91,97] An estimate of β_X can be obtained from the above relation using an estimate of β_X^* from the naïve analysis unadjusted for measurement error together with an estimate of λ. An estimate of λ may be available from an external study or it can be estimated from an ancillary study as the slope in the calibration regression model. In the latter case, this simpler approach is identical to RC as described above, which instead makes use of the predicted values of $E(X|X^*, Z)$ from the calibration equation.

When the outcome model is a linear regression, the attenuation factor λ can also be estimated using the method of moments. For example, in the situation of a single covariate measured with classical error and no other covariates, $\lambda = \text{var}(X)/\text{var}(X^*)$. The variance of the error-prone measurement X^* can be estimated directly from the main study. If a validation study is available, var(X) can be estimated directly within it. In a replicates study, the covariance between the replicate measurements provides an estimate of var(X), using the assumption of classical error and uncorrelated error in the replicate measurements. In this simple setting of classical measurement error the method of moments estimator is equivalent to the RC estimator, because the parameters of the regression model for $E(X|X^*)$ used in RC can be defined in terms of the first and second moments. When X^* follows the linear measurement error model (2) the expression for λ is $\lambda = \frac{\alpha_X \text{var}(X)}{\alpha_X^2 \text{var}(X)+\text{var}(U)}$, as given in (7). The individual components of this expression can be estimated using the method of moments from a validation study. However, in the case of a calibration study estimating var(X) and var(U) requires two or more measures of the unbiased reference measurement, whereas RC can be performed using only one such measurement.[17,58] The method of moments approach is conceptually straightforward and, like RC, it extends to situations with covariates Z or with multiple error-prone measurements (Section 6.1.3). As in RC, use of approximations or bootstrapping is still needed to obtain standard errors of the correct estimates (see Section 6.1.2). Unlike RC, the method of moments does not extend to nonlinear models, though it does extend to the situation of differential error (see Part 2). The method of moments is covered in several of the key texts on measurement error methods.[6,7,9,97]

We now illustrate RC using an example from the OPEN study, introduced earlier in Section 5, in which we examine the association between BMI and usual (average daily) log potassium density intake. Our dependent variable (Y) is BMI, X is usual log potassium density intake, X^* is the log FFQ potassium density reported which is considered to have linear measurement error, and we control for age and sex (Z). We first perform an analysis unadjusted for the measurement error in FFQ-reported potassium density intake, and regress BMI on log FFQ potassium density intake, age and sex for the 483 participants who have observations for the variables. The result is shown in Table 2(1), with the coefficient for log potassium density estimated as −1.69 with standard error 0.93 and a z-value of −1.81. This indicates that there may be a negative association between potassium density intake and BMI, with a decrease of approximately 1 BMI unit for every doubling of potassium density intake, not a very large effect.

To perform RC, we develop a prediction equation relating true log potassium density intake to FFQ log potassium density, age, and sex. We assume here that the appropriate form of the equation is linear. The equation is estimated by regressing an unbiased measure of log potassium density intake, the average of the two log urinary potassium values minus the log doubly labeled water value (the reference instrument, X^{**}), on FFQ log potassium density (X^*), age, and sex (Z). In most applications the urinary measures, which create a much higher participant burden, and the doubly labeled water exam, which is expensive, are performed in a smaller random subset of participants. Although in our dataset we have these measures in nearly all the participants of this study, we have here sampled only 250 participants to develop the prediction equation. The result is shown in Table 2(2), and the prediction equation for log potassium density intake is: $-0.48 + 0.45 \times \log(\text{FFQ potassium density}) + 0.046 \times \text{sex} + 0.0059 \times \text{age}$. The predicted values obtained using this equation are now used for the log potassium density variable in the regression of BMI on log potassium density, age, and sex.

TABLE 2 Results from analysis of data from 483 participants in the OPEN study: single covariate measured with error

(1) Analyses of the association of log potassium density intake with BMI, using an unadjusted analysis and regression calibration

	Unadjusted analysis			Regression calibration		
Variable	**Est.**	**SE**	***z*-Value**	**Est.**	**SE**	***z*-Value**
log FFQ potassium density	−1.69	0.93	−1.81	−3.76	2.43[a]	−1.55
Sex (F vs M)	−0.38	0.49	−0.77	−0.20	0.58[a]	−0.36
Age (years)	0.039	0.029	1.32	0.061	0.034[a]	1.81

(2) Prediction model used by regression calibration, based on a random subsample of 250 participants

Variable	**Est.**	**SE**	***z*-Value**
Intercept	−0.48	0.16	−2.96
log FFQ potassium density	0.45	0.089	5.02
Sex (F vs M)	0.046	0.044	1.02
Age (years)	0.0059	0.0027	2.23

Note: Estimated coefficients (Est.), standard errors (SE) and *z*-values.
[a]From 5000 bootstrap samples.

The result is shown in Table 2(1), with the estimated coefficient for log potassium now equal to −3.76. The adjusted estimate of the coefficient indicates a stronger association with BMI than the unadjusted estimate, with a 30% increase in potassium density now associated with a decrease of approximately 1.0 ($3.76 \times \log 1.33$) BMI unit. It is worth noting that the adjusted coefficient for log potassium density, −3.76, is equal to the unadjusted estimate, −1.69, divided by the coefficient for log FFQ potassium density in the prediction equation, 0.45; this follows the multiplicative relationship between the unadjusted estimate of the coefficient and its true value that was described in Section 3.1.1.

There are many examples of RC, particularly in the nutritional epidemiology literature. A situation that is challenging for RC arises when Z is a strong predictor of both X and the outcome Y, resulting in $E(X|X^*, Z)$ and Z being highly correlated. Detailed discussion of this issue is provided by Prentice et al[98] and Freedman et al.[99]

6.1.2 | Standard error estimation for regression calibration

Due to the extra uncertainty in the parameters estimated in the calibration model, one cannot use the usual model standard errors from the outcome regression model when performing RC, as these will be too small, resulting in confidence intervals that are too narrow. Instead, standard errors for the estimated coefficients in the outcome model may be obtained either from a bootstrap variance estimator[100] or a sandwich estimator obtained by stacking the calibration and outcome model estimating equations[6(appendixB3),99]. Due to its ease of implementation, the bootstrap is commonly used in practice. Commonly, a nonparametric bootstrap is used, and is applied separately to the ancillary study sample and the main study sample; that is, for internal calibration, validation, or replicates studies the bootstrap is stratified by membership to the ancillary study. In some studies, notably when the ancillary study is large, the uncertainty in the estimation of the calibration equation parameters is small relative to that in the estimation of the outcome model parameters, and in this case, ignoring uncertainty in the estimation of the calibration equation parameters may not have a large impact.[101] However, we demonstrated in Section 5.1 that uncertainty in the estimation of the calibration equation parameters can have a large impact. It is good practice to make corrections to the model SEs and doing so is straightforward in practice using bootstrapping.

For the example given in Section 6.1.1, we bootstrapped the entire RC procedure, starting with the estimation of the prediction equation for log potassium density and then incorporated the re-estimated predicted values for $E(X|X^*, Z)$ into the regression model for BMI. The bootstrap standard errors of the estimated coefficients in that model are shown in Table 2(1). Note that, just as the adjusted coefficient for log potassium density is much larger than the unadjusted value

(−3.76 vs −1.69), so is its standard error (2.43 vs 0.93). In fact, the relative increase in the standard error is larger than the relative increase in the coefficient, so that the adjusted Wald z-value (−1.55) is smaller than the unadjusted value (−1.81). Actually, as seen in Table 1, for a single covariate measured with nondifferential error, the test based on the unadjusted Wald z-value is valid, so there is no need to test the null hypothesis of no association using the adjusted analysis. Moreover, a test based on the adjusted Wald z-value would be less powerful—however, this is because such a test is based on an incorrect assumption of normality for the z-value. A correct test based on the adjusted analysis can be performed using percentile-based bootstrap confidence intervals. Frost and Thompson[37] described alternative methods for constructing confidence intervals with correct coverage in this setting, using Fieller's Theorem.

6.1.3 | Regression calibration when more than one covariate is measured with error

It is not uncommon that multiple error-prone covariates $(X_1, \dots, X_p)$ are of interest (such as smoking, alcohol intake, and physical activity as predictors of blood pressure). As seen in Sections 3.1.3 and 3.2.3, coefficients estimated in a regression that ignores the error in multiple covariates can be biased in either direction. RC can be applied to provide consistent estimates of the coefficients in the outcome model, so long as appropriate data are available to fit the series of calibration equations $E(X_i|X^*, Z)$, $i = 1, \dots, p$. Note that in this case, there is one equation for each covariate that is measured with error, and X^* typically includes all the error-prone covariates. When there is an internal validation or calibration study with simultaneous observation of the vector of error-prone covariates X^*, exact covariates Z and exact or unbiased reference measurements for vector X, then a multivariate regression model can be fit for $E(X\,|X^*, Z)$. Rosner et al[92] develop this approach for the setting of logistic regression with a rare disease outcome and one or more continuous error-prone covariates. When classical measurement error occurs, it is sufficient that independent replicates of a vector X^* are measured in a replicates study (see Section 4.2). Carroll et al[6(sec.4.4.2)] present a RC-based approach for this case, allowing varying number of replicates across individuals. A key assumption for this approach is that the replicate observations are independent given the true X. Methods to estimate interactions between two error-prone variables have also been developed.[102,103]

We illustrate RC with more than one error-prone covariate using an extension of the example in Section 6.1.1. We consider the regression of BMI on log potassium density and log sodium density intakes, while controlling for age and sex. As noted above, the error-prone measurements of potassium density and sodium density are presumed to have nondifferential error with respect to BMI. Hence RC was appropriate for application to this example. In Part 2 we consider instead using absolute average daily sodium intake as the main covariate; this absolute measure *is* subject to differential error, making RC an unsuitable correction method. The estimated regression coefficients of the unadjusted model, using FFQ reported intakes, are presented in left-hand part of Table 3(1). The estimated coefficients for log potassium density and log sodium density are −1.93 and 1.51, with z-values of −2.03 and 1.26, respectively. Unlike in the case of a single error-prone covariate model (Table 2(1)) hypothesis tests based on these z-values are invalid as highlighted in Table 1. Table 3(2) shows the results of the calibration models for log potassium density and log sodium density, and the right-hand part of Table 3(1) shows the adjusted estimated coefficients for these variables. As in Section 6.1.2, the standard errors are obtained by bootstrapping. One can see that both estimated coefficients are larger in magnitude than their unadjusted versions, the log potassium density coefficient by a factor of 1.6 and the log sodium density coefficient by a factor of 2.4. However, neither coefficient reaches the conventional statistical significance level of 0.05. It is interesting that the similar magnitude but opposite signs of the log potassium density and log sodium density coefficients do suggest that the association with BMI is through the sodium/potassium ratio, a dietary intake measure that has been found to be related to cardiovascular disease.[104]

6.1.4 | Regression calibration in nonlinear models

When the outcome model is nonlinear, RC provides estimates that are only approximately unbiased; however, it has been observed to work remarkably well in generalized linear models logistic, and Poisson regression.[6(sec.4.8),80,105] As discussed in Section 3.1.4, RC usually works well if the magnitude of β_X remains "small to moderate" or the measurement error variance is "small".[6(sec.4.8)] For Cox regression, Prentice[41] demonstrated that RC gives a reasonable approximation in cases of rare disease and either a log hazard ratio of moderate size or small measurement error variance. Xie et al[40]

TABLE 3 Results from analysis of data from 483 participants in the OPEN study: two covariates measured with error

(1) Analyses of the association of log potassium density intake and log sodium density intake with BMI, using an unadjusted analysis and regression calibration						
	Unadjusted analysis			Regression calibration		
Variable	**Est.**	**SE**	***z*-Value**	**Est.**	**SE**	***z*-Value**
log FFQ potassium density	−1.93	0.95	−2.03	−3.17	2.34[a]	−1.35
log FFQ sodium density	1.51	1.20	1.26	3.56	3.82[a]	0.93
Sex (F vs M)	−0.33	0.44	−0.67	−0.09	0.60[a]	−0.16
Age (years)	0.038	0.029	1.30	0.044	0.038[a]	1.14

(2) Prediction models used by regression calibration, based on a random subsample of 250 participants			
Variable	**Est.**	**SE**	***z*-Value**
Prediction model for log potassium density			
Intercept	−0.48	0.17	−2.84
log FFQ potassium density	0.45	0.091	4.94
log FFQ sodium density	−0.006	0.107	−0.06
Sex (F vs M)	0.046	0.046	1.00
Age (years)	0.0059	0.0027	2.26
Prediction model for log sodium density			
Intercept	0.20	0.16	1.27
log FFQ potassium density	−0.14	0.09	−1.66
log FFQ sodium density	0.42	0.10	4.11
Sex (F vs M)	−0.025	0.043	−0.59
Age (years)	0.0038	0.0025	1.49

Note: Estimated coefficients (Est.), standard errors (SE) and *z*-values.
[a]From 5000 bootstrap samples.

demonstrated that the performance of RC in Cox regression for larger log hazard ratios can be improved by re-estimating $E(X|X^*, Z)$ on each risk set. This approach is called "risk set regression calibration" and works well when recalibration is done periodically (say at every fifth percentile) rather than at each failure time.[106] Liao et al[107] extended the risk set approach for time-dependent covariates. For highly nonlinear models, RC can be problematic, and Carroll et al[6(chap.4)] give methods to improve its performance.

6.2 | Simulation extrapolation (SIMEX)

Another method which, like RC, enjoys considerable use in practice, is SIMEX. SIMEX is a very general method for measurement error correction in estimating complex regression models in the presence of classical measurement error (see model (1)) in the covariates. It was proposed by Cook and Stefanski[108] and enhanced by Carroll et al.[109] The method has been extended to a correction for misclassification (MC-SIMEX) by Küchenhoff et al[110,111]—see Part 2, Section 2.5 for more details. The basic idea of SIMEX is to add more error to the error prone covariate X^*, to see the impact this has on the parameter estimates from the outcome regression model, and to then extrapolate back to the situation with no measurement error. This is equivalent to estimating the relationship between the measurement error variance var(U) and the parameter estimates in the regression ignoring the measurement error, and to extrapolate back to the situation in which the error variance is 0. As a simple example, for single covariate regression (see Section 3.1.1) the theoretical relationship between the biased regression slope β_X^*, based on the regression of Y on X^*, is displayed in Figure 5 as a function of var(U), $\beta_X^*(\text{var}(U))$. In practice, this function is estimated from simulated datasets obtained by adding further

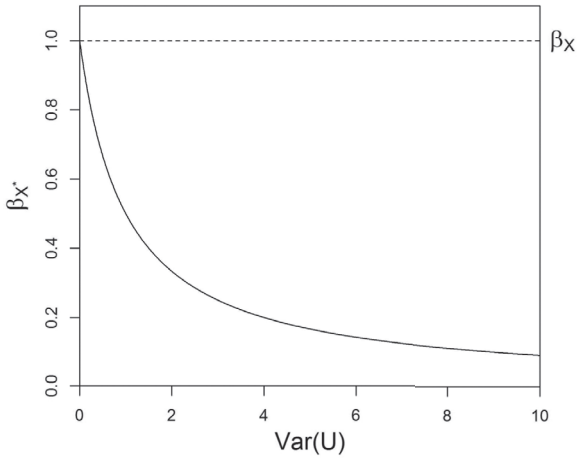

FIGURE 5 SIMEX: Relationship between the measurement error variance var(U) and the regression coefficient β_X^*. β_X is the regression coefficient in a regression of Y on X and β_X^* is the regression coefficient in a regression of Y on X^*. X^* is assumed to follow the classical error model $X^* = X + U$, and $\beta_X^* = \frac{\text{var}(X)}{\text{var}(X)+\text{var}(U)}\beta_X$. Here, var($X$) = 1 and $\beta_X = 1$

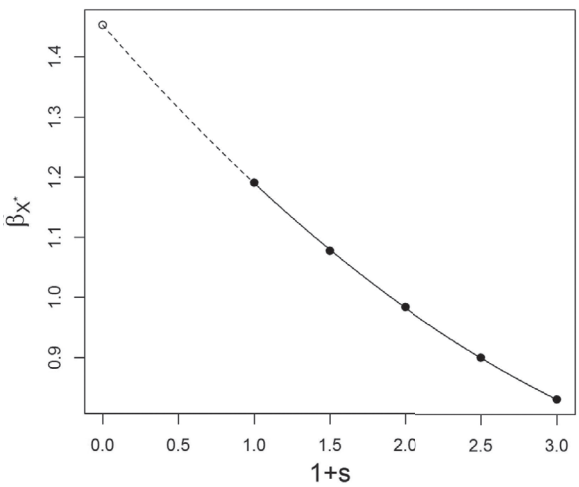

FIGURE 6 SIMEX estimation for the association between individual heart rate as outcome variable and log-transformed individual particle number concentration (a measure of air pollution exposure) measured in number per cm^3. The SIMEX estimator is assessed assuming a measurement error variance of 0.03, which is determined through comparison measurements, a quadratic extrapolation function, number of simulations $B = 100$ and $\boldsymbol{s} = (1, 1.5, 2, 2.5, 3)$. The analysis is based on longitudinal data from an observational study described in Peters et al.[72] The model accounts for temperature, relative humidity, time trend, and time of the day. The solid curve is obtained from the fit of the extrapolation model to the pseudo-datasets, and the dotted line represents the extrapolated part

measurement error to the error-prone covariate and estimating the regression coefficient β_X in each dataset, and then using a model to relate the estimated regression coefficients to var(U). If the estimator for this functional relationship between the error variance and $\beta_X^*$is consistent, then in the case of an error-free covariate, $\beta_X^*(0) = \beta_X$. So the function $\beta_X^*(\text{var}(U))$ is extrapolated back to var(U) = 0 so as to estimate β_X.

More broadly, we assume a general regression model with a covariate X measured with classical measurement error U, that is, we observe $X^* = X + U$. Further covariates Z without measurement error may be included in the model (see Section 3.1.2). The parameters of interest are denoted by the vector β_X. Given data $(Y_i, X_i^*, Z_i)_{i=1}^n$, the unadjusted estimator ignoring measurement error is denoted by $\widehat{\beta}_X^*[(Y_i, X_i^*, Z_i)_{i=1}^n]$.

The simulation and extrapolation steps then proceed as follows:

Simulation step: For each value of a fixed grid of values $s_1, \ldots, s_m$ (≥ 0), B new pseudo-datasets are simulated by $X_{ib}^*(s_k) = X_i^* + \sqrt{s_k \text{var}(U)} U_{ibk}, i = 1, \ldots, n; b = 1, \ldots, B; k = 1, \ldots, m$,where U_{ibk} are independent identically distributed standard normal variables. Note that the measurement error variance of $X_{ib}^*(s_k)$ is $(1 + s_k)\text{var}(U)$. For each pseudo-dataset the unadjusted estimator is given by $\widehat{\beta}_X^*[(Y_i, X_{ib}^*(s_k), Z_i)_{i=1}^n]$. We denote the average over the B repetitions by $\widehat{\beta}_{Xs_k}^*$, which we calculate by $\widehat{\beta}_{Xs_k}^* = B^{-1}\sum_{b=1}^B \widehat{\beta}_X^*[(Y_i, X_{ib}^*(s_k), Z_i)_{i=1}^n], k = 1, \ldots, m$. Furthermore, we set $s_0 = 0$ and $\widehat{\beta}_{Xs_0}^* = \widehat{\beta}_X^*[(Y_i, X_i^*, Z_i)_{i=1}^n]$.

Extrapolation step: To make the necessary extrapolation, we use a parametric approximation for $\beta_X^*((1 + s)\text{var}(U))$, setting it equal to a function of s denoted by $G(s, \Gamma)$, where Γ denotes the parameters of the model. We estimate the parameters Γ by least squares with $[\widehat{\beta}_{Xs_k}^*]_{k=0}^m$ as the independent variable data and functions of $[1 + s_k]_{k=0}^m$ as the dependent variables, yielding an estimator $\widehat{\Gamma}$. The estimated parametric function is then extrapolated to $s = -1$, the case of no measurement error. The SIMEX estimator is then defined by

$$\widehat{\beta}_{\text{SIMEX}} = G(-1, \widehat{\Gamma}). \tag{25}$$

TABLE 4 Overview of available software for performing regression calibration

Package/procedure	Website location or information	Notes	References
Procedure rcal within the Stata package merror	http://www.stata.com/merror/ See also: http://www.stat.tamu.edu/~carroll/eiv.SecondEdition/statacode.php	For generalized linear models with X^* having classical measurement error and (a) the error variance is known; (b) replicate measurements of X^* are available or (c) when there is also available an "instrumental variable" that is correlated with X and whose errors are independent of the errors in X^*.	Hardin et al[116]
Procedure eivreg within Stata	http://www.stata.com/manuals13/reivreg.pdf	For linear regression where X^* has classical measurement error and the error variance (or ratio of the error variance to total variance) is known.	Hardin et al[116]
NCI SAS macros	https://epi.grants.cancer.gov/diet/usualintakes/macros.html	(a) For X^* measured in all individuals that, after suitable transformation, satisfies a linear measurement error model, together with an X^{**} measured in a subsample that, after suitable transformation, has classical measurement error. (b) For X^* measured in all individuals that, after suitable transformation, has classical measurement error. A substantial subsample should have at least one repeat value of X^*. In these options X^* and X^{**} may be univariate, bivariate or multivariate. X^* and X^{**} may have excess zeros.	Kipnis et al[22]
Spiegelman SAS macro %blinplus	https://www.hsph.harvard.edu/donna-spiegelman/software/	For univariate or multivariate X^* measured in all individuals in the main study and a validation study where both X and X^* are measured in all individuals. X^* satisfies the linear measurement error model (2).	Rosner et al[92]
Spiegelman SAS macro %relibpls8	https://www.hsph.harvard.edu/donna-spiegelman/software/	For univariate or multivariate X^* measured in all individuals and repeat measurements of X^*. X^* satisfies the classical measurement error model (1).	Rosner et al[93]
Spiegelman SAS macro %rrc	https://www.hsph.harvard.edu/donna-spiegelman/software/	For a time-varying covariate X in a Cox regression model. X^* satisfies the linear measurement error model (2). The method uses risk-set regression calibration for estimating the risk parameters (see Section 5.1.4).	Liao et al[107]

When β_X is a vector, the SIMEX estimator can be applied separately for each component.

The estimator $\widehat{\beta}_{\text{SIMEX}}$ is consistent when the extrapolation function is correctly specified. In complex regression models, the extrapolation function is unknown and usually does not have a suitable parametrization. However, a quadratic function, defined as $G_Q(s, \Gamma) = \gamma_0 + \gamma_1 s + \gamma_2 s^2$, is a good approximation in many practical applications. For the estimation of the variance of $\widehat{\beta}_{\text{SIMEX}}$, three methods are available: (i) an asymptotic approach using the delta method, see Carroll et al[109]; (ii) a simulation-based method proposed by Stefanski and Cook[112]; and (iii) the nonparametric bootstrap performing the whole SIMEX procedure in each bootstrap sample. When the measurement error variance is estimated from an ancillary study, the bootstrap procedure includes a sample from the ancillary study and simultaneously a sample from the main study in each repetition. Then, the uncertainty about the measurement error variance is also addressed. The SIMEX method is illustrated in an example using the R-package SIMEX by Lederer and Küchenhoff.[113] Figure 6 displays the SIMEX extrapolation curve for the regression coefficient of heart rate on log particle number concentration (a measure of air pollution exposure) in individuals with impaired glucose metabolism or diabetes.[72]

The SIMEX method is attractive due to its very general applicability. It requires only a well-defined estimation procedure leading to a consistent estimator in the case of error-free covariates. It also requires knowledge about the

TABLE 5 Overview of available software for performing simulation extrapolation (SIMEX)

Package/procedure	Website location or information	Notes	References
Package simex within the R language	https://cran.r-project.org/web/packages/simex/simex.pdf	(a) Performs SIMEX for a wide range of models with X^* having classical measurement error and known error variance. (b) Performs MC-SIMEX for a wide range of models with categorical X^* having misclassification with a known matrix of misclassification probabilities. The jackknife method (Stefanski and Cook[112]) and an asymptotic approach (Carroll et al[109]; Küchenhoff et al[110]) are available for estimating the standard errors of the estimated coefficients.	Cook and Stefanski,[108] Küchenhoff et al,[111] Lederer and Küchenhoff[113]
Procedures simex and simexplot within the Stata package merror	http://www.stata.com/merror/	Performs SIMEX for generalized linear models with covariate X^* having classical measurement error and known error variance or repeat measurements.	Hardin et al[117]
Package simexaft within the R language	https://cran.r-project.org/web/packages/simexaft/simexaft.pdf	Performs SIMEX for the accelerated failure time model with covariate X^* having classical measurement error and known error variance or repeat measurements. X^* may be multivariate.	He et al[118]

measurement error model parameters, for example through an ancillary study of one of the types described in Section 4, but this is necessary for most procedures of measurement error correction. This makes the procedure attractive for complex models where other methods are not feasible. Furthermore, the plot of the SIMEX procedure yields a convenient description of the effect of measurement error on parameter estimation. The Achilles heel of the procedure is the usually unknown extrapolation function. Especially for large measurement error, when extrapolation is to a more distant point, the procedure should be checked by simulation studies. Note that SIMEX just needs an error structure that can be described by $X^* = X + U$. Here, U can be correlated with Y and Z, which makes SIMEX applicable for special types of nondifferential measurement error. SIMEX has also been extended for more complex measurement error structures; for a recent example with spatial data see Alexeff et al.[114] In another example, SIMEX has been applied to adjust for measurement error in a survival outcome, see Oh et al.[115]

7 | SOFTWARE FOR ANALYSIS

One of the main barriers in the past to the use of the analysis methods described in Section 6 was the lack of specific software for their implementation, though the situation is gradually improving. There are now available several programs or macros for performing RC, including the *rcal* command for generalized linear models in Stata.[116] They are summarized in Table 4. There are also now available three packages for performing SIMEX. They are summarized in Table 5. One performs SIMEX for a wide range of regression models (*simex* in *R*),[113] one performs SIMEX for generalized linear models (procedures *simex* and *simexplot* within the *merror* package in Stata),[117] and one performs SIMEX for accelerated lifetime models (*simexaft* in *R*).[118]

In supplemental materials available at https://github.com/PamelaShaw/STRATOS-TG4-Guidance-Paper, we provide code which demonstrates the implementation of the RC and SIMEX analyses provided in each of the data examples.

8 | CONCLUSION

We have presented here the background and key results required for understanding the impact of measurement error on the results of epidemiological research studies, and some relatively simple methods available to adjust for such measurement error in continuous covariates used in regression models. In most cases, the limiting factor for performing these adjusted analyses will be the availability of quantitative information regarding the measurement error. However,

our impression is that even when such quantitative information is available, the adjusted analyses are not being performed.[5,119]

In Part 2 we will describe a variety of additional topics, many of which build on what we have covered here. In particular, we will present other methods of adjusting for measurement error, such as a likelihood-based approach, multiple imputation and moment reconstruction, and Bayesian methods. We will also describe methods of adjustment for misclassified discrete variables and methods for adjusting estimates of distributions. Finally, we will touch on a number of more advanced topics, such as mixtures of Berkson and classical error and variable selection, concluding with a discussion of how to proceed when the information on measurement error is incomplete.

ACKNOWLEDGEMENTS

This research is supported in part by the National Institutes of Health (NIH) grants R01-AI131771 (P.A.S.), U01-CA057030 (R.J.C.), NCI P30CA012197 (J.A.T.); Patient Centered Outcomes Research Institute (PCORI) Award R-1609-36207 (P.A.S.); and Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2019-03957 (P.G.). The statements in this manuscript are solely the responsibility of the authors and do not necessarily represent the views of NIH, PCORI, or NSERC.

DATA AVAILABILITY STATEMENT

The OPEN Study data that illustrate the methods presented in this paper are available upon request to RFAB@mail.nih.gov. The request should specify the dataset used in analyses presented in the papers by Keogh et al (2020) and Shaw et al (2020). More information about these data can be obtained at https://epi.grants.cancer.gov/past-initiatives/open/. The software code files for implementing our examples based on the OPEN study may be found online at https://github.com/PamelaShaw/STRATOS-TG4-Guidance-Paper.

ORCID

Ruth H. Keogh https://orcid.org/0000-0001-6504-3253
Pamela A. Shaw https://orcid.org/0000-0003-1883-8410
Paul Gustafson https://orcid.org/0000-0002-2375-5006
Laurence S. Freedman https://orcid.org/0000-0002-6767-7900

REFERENCES

1. Murray RP, Connett JE, Lauger GG, Voelker HT. Error in smoking measures: effects of intervention on relations of cotinine and carbon monoxide to self-reported smoking. The Lung Health Study Research Group. *Am J Public Health*. 1993;83(9):1251-1257. https://doi.org/10.2105/ajph.83.9.1251.
2. Thiébaut ACM, Freedman LS, Carroll RJ, Kipnis V. Is it necessary to correct for measurement error in nutritional epidemiology? *Ann Intern Med*. 2007;146(1):65. https://doi.org/10.7326/0003-4819-146-1-200701020-00012.
3. Ferrari P, Friedenreich C, Matthews CE. The role of measurement error in estimating levels of physical activity. *Am J Epidemiol*. 2007;166(7):832-840. https://doi.org/10.1093/aje/kwm148.
4. Zeger SL, Thomas D, Dominici F, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect*. 2000;108(5):419-426. https://doi.org/10.1289/ehp.00108419.
5. Shaw PA, Deffner V, Keogh RH, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Ann Epidemiol*. 2018;28(11):821-828. https://doi.org/10.1016/j.annepidem.2018.09.001.
6. Carroll R, Ruppert D, Stefanski L, Crainiceanu C. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman and Hall; 2006.
7. Buonaccorsi J. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman and Hall; 2010.
8. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton, FL: Chapman and Hall; 2004.
9. Yi G. *Statistical Analysis with Measurement Error or Misclassification*. New York, NY: Springer; 2017.
10. Keogh R, White I. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Stat Med*. 2014;33(12):2137-2155. https://doi.org/10.1002/sim.6095.
11. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998;55(10):651-656. https://doi.org/10.1136/oem.55.10.651.
12. Buzas J, Stefanski L, Tosteson T. Measurement error. In: Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*. New York, NY: Springer; 2014:1241-1282.
13. Cochran W. Errors of measurement in statistics. *Dent Tech*. 1968;10(4):637-666.

14. Chen Z, Peto R, Collins R, MacMahon S, Lu J, Li W. Serum cholesterol concentration and coronary heart disease in population with low cholesterol concentrations. *BMJ*. 1991;303(6797):276-282. https://doi.org/10.1136/bmj.303.6797.276.
15. MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*. 1990;335(8692):765-774. https://doi.org/10.1016/0140-6736(90)90878-9.
16. Freedman LS, Carroll RJ, Wax Y. Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *Am J Epidemiol*. 1991;134(3):310-320. https://doi.org/10.1093/oxfordjournals.aje.a116086.
17. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol*. 2003;158(1):14-21. https://doi.org/10.1093/aje/kwg091.
18. Prentice RL, Sugar E, Wang C, Neuhouser M, Patterson R. Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. *Public Health Nutr*. 2002;5(6a):977-984. https://doi.org/10.1079/PHN2002382.
19. Spiegelman D, Logan R, Grove D. Regression calibration with heteroscedastic error variance. *Int J Biostat*. 2011;7(1):4-34. https://doi.org/10.2202/1557-4679.1259.
20. Chen X, Hong H, Nekipelov D. Nonlinear models of measurement errors. *J Econ Lit*. 2011;49:901-937. https://doi.org/10.2307/23071661.
21. Lyles RH, Kupper LL. A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics*. 1997;53(3):1008-1025. https://doi.org/10.2307/2533560.
22. Kipnis V, Midthune D, Buckman DW, et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*. 2009;65(4):1003-1010. https://doi.org/10.1111/j.1541-0420.2009.01223.x.
23. Berkson J. Are there two regressions? *J Am Stat Assoc*. 1950;45(250):164. https://doi.org/10.2307/2280676.
24. Oraby T, Sivaganesan S, Bowman JD, et al. Berkson error adjustment and other exposure surrogates in occupational case-control studies, with application to the Canadian INTEROCC study. *J Expo Sci Environ Epidemiol*. 2018;28(3):251-258. https://doi.org/10.1038/jes.2017.2.
25. Goldman GT, Mulholland JA, Russell AG, et al. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health*. 2011;10:61. https://doi.org/10.1186/1476-069X-10-61.
26. Tooze JA, Troiano RP, Carroll RJ, Moshfegh AJ, Freedman LS. A measurement error model for physical activity level as measured by a questionnaire with application to the 1999-2006 NHANES questionnaire. *Am J Epidemiol*. 2013;177(11):1199-1208. https://doi.org/10.1093/aje/kws379.
27. Willett W. *Nutritional Epidemiology*. 3rd ed. New York, NY: Oxford University Press; 2013.
28. Tielemans E, Kupper LL, Kromhout H, Heederik D, Houba R. Individual-based and group-based occupational exposure assessment: some equations to evaluate different strategies. *Ann Occup Hyg*. 1998;42(2):115-119. https://doi.org/10.1016/s0003-4878(97)00051-3.
29. Peters PJ, Westheimer E, Cohen S, et al. Screening yield of HIV antigen/antibody combination and pooled HIV RNA testing for acute HIV infection in a high-prevalence population. *JAMA*. 2016;315(7):682-690. https://doi.org/10.1001/jama.2016.0286.
30. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol*. 1991;134(10):1233-1246. https://doi.org/10.1093/oxfordjournals.aje.a116026.
31. Wacholder S, Dosemeci M, Lubin JH. Blind assignment of exposure does not always prevent differential misclassification. *Am J Epidemiol*. 1991;134(4):433-437. https://doi.org/10.1093/oxfordjournals.aje.a116105.
32. Brenner H. Notes on the assessment of trend in the presence of nondifferential exposure misclassification. *Epidemiology*. 1992;3(5):420-427.
33. Wang D, Gustafson P. On the impact of misclassification in an ordinal exposure variable. *Epidemiol Methods*. 2014;3(1):97-106. https://doi.org/10.1515/em-2013-0017.
34. Wang D, Shen T, Gustafson P. Partial identification arising from nondifferential exposure misclassification: how informative are data on the unlikely, maybe, and likely exposed? *Int J Biostat*. 2012;8(1). https://doi.org/10.1515/1557-4679.1397.
35. McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. *Stat Med*. 2004;23(7):1095-1109. https://doi.org/10.1002/sim.1656.
36. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology*. 2011;22(4):589-597. https://doi.org/10.1097/EDE.0b013e3182117c85.
37. Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *J R Stat Soc Ser A*. 2000;163:173-189. https://doi.org/10.2307/2680496.
38. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Stat Med*. 1988;7(1–2):257-274. https://doi.org/10.1002/sim.4780070126.
39. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. *J Natl Cancer Inst*. 2011;103(14):1086-1092. https://doi.org/10.1093/jnci/djr189.
40. Xie SX, Wang CY, Prentice RL. A risk set calibration method for failure time regression by using a covariate reliability sample. *J R Stat Soc Ser B*. 2001;63(4):855-870. https://doi.org/10.1111/1467-9868.00317.
41. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1982;69(2):331-342. https://doi.org/10.1093/biomet/69.2.331.
42. Bross I. Misclassification in 2 X 2 tables. *Biometrics*. 1954;10(4):478. https://doi.org/10.2307/3001619.
43. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol*. 1990;132(4):746-748. https://doi.org/10.1093/oxfordjournals.aje.a115716.

44. Weinberg CA, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? *Am J Epidemiol*. 1994;140(6):565-571. https://doi.org/10.1093/oxfordjournals.aje.a117283.
45. Keogh RH, Carroll RJ, Tooze JA, Kirkpatrick SI, Freedman LS. Statistical issues related to dietary intake as the response variable in intervention trials. *Stat Med*. 2016;35(25):4493-4508. https://doi.org/10.1002/sim.7011.
46. Hyslop DR, Imbens GW. Bias from classical and other forms of measurement error. *J Bus Econ Stat*. 2001;19(4):475-481. https://doi.org/10.1198/07350010152596727.
47. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol*. 1997;146(2):195-203. https://doi.org/10.1093/oxfordjournals.aje.a009251.
48. Tooze JA, Kipnis V, Buckman DW, et al. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Stat Med*. 2010;29(27):2857-2868. https://doi.org/10.1002/sim.4063.
49. Mielgo-Ayuso J, Aparicio-Ugarriza R, Castillo A, et al. Physical activity patterns of the spanish population are mostly determined by sex and age: findings in the ANIBES Study. *PLoS One*. 2016;11(2):e0149969. https://doi.org/10.1371/journal.pone.0149969.
50. Kaaks R, Ferrari P, Ciampi A, Plummer M, Riboli E. Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments. *Public Health Nutr*. 2002;5(6A):969-976. https://doi.org/10.1079/PHN2002380.
51. Kaaks R, Riboli E. Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol*. 1997;26(suppl 1):S15-S25. https://doi.org/10.1093/ije/26.suppl_1.s15.
52. O'Brien TE, Funk GM. A gentle introduction to optimal design for regression models. *Am Stat*. 2003;57(4):265-267. https://doi.org/10.2307/30037294.
53. Kaaks R, Riboli E, van Staveren W. Calibration of dietary intake measurements in prospective cohort studies. *Am J Epidemiol*. 1995;142(5):548-556. https://doi.org/10.1093/oxfordjournals.aje.a117673.
54. Subar AF, Freedman LS, Tooze JA, et al. Addressing current criticism regarding the value of self-report dietary data. *J Nutr*. 2015;145(12):2639-2645. https://doi.org/10.3945/jn.115.219634.
55. Freedman LS, Commins JM, Willett W, et al. Evaluation of the 24-hour recall as a reference instrument for calibrating other self-report instruments in nutritional cohort studies: evidence from the validation studies pooling projecT. *Am J Epidemiol*. 2017;186(1):73-82. https://doi.org/10.1093/aje/kwx039.
56. Lampe JW, Huang Y, Neuhouser ML, et al. Dietary biomarker evaluation in a controlled feeding study in women from the Women's Health Initiative cohort. *Am J Clin Nutr*. 2017;105(2):466-475. https://doi.org/10.3945/ajcn.116.144840.
57. Michels KB, Welch AA, Luben R, Bingham SA, Day NE. Measurement of fruit and vegetable consumption with diet questionnaires and implications for analyses and interpretation. *Am J Epidemiol*. 2005;161(10):987-994. https://doi.org/10.1093/aje/kwi115.
58. Keogh RH, White IR, Rodwell SA. Using surrogate biomarkers to improve measurement error models in nutritional epidemiology. *Stat Med*. 2013;32(22):3838-3861. https://doi.org/10.1002/sim.5803.
59. Beaton GH, Milner J, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *Am J Clin Nutr*. 1979;32(12):2546-2559. https://doi.org/10.1093/ajcn/32.12.2546.
60. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "Alloyed Gold Standard". *Am J Epidemiol*. 1997;145(2):184-196. https://doi.org/10.1093/oxfordjournals.aje.a009089.
61. Nusser SM, Beyler NK, Welk GJ, Carriquiry AL, Fuller WA, King BMN. Modeling errors in physical activity recall data. *J Phys Act Health*. 2012;9(suppl 1):S56-S67.
62. Neuhouser ML, Di C, Tinker LF, et al. Physical activity assessment: biomarkers and self-report of activity-related energy expenditure in the WHI. *Am J Epidemiol*. 2013;177(6):576-585. https://doi.org/10.1093/aje/kws269.
63. Lim S, Wyker B, Bartley K, Eisenhower D. Measurement error of self-reported physical activity levels in New York City: assessment and correction. *Am J Epidemiol*. 2015;181(9):648-655. https://doi.org/10.1093/aje/kwu470.
64. Matthews CE, Kozey Keadle S, Moore SC, et al. Measurement of active and sedentary behavior in context of large epidemiologic studies. *Med Sci Sports Exerc*. 2018;50(2):266-276. https://doi.org/10.1249/MSS.0000000000001428.
65. Shaw PA, McMurray R, Butte N, et al. Calibration of activity-related energy expenditure in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *J Sci Med Sport*. 2019;22(3):300-306. https://doi.org/10.1016/J.JSAMS.2018.07.021.
66. Schoeller DA. Measurement of energy expenditure in free-living humans by using doubly labeled water. *J Nutr*. 1988;118(11):1278-1289. https://doi.org/10.1093/jn/118.11.1278.
67. Tolonen H, Wolf H, Jakovljevic D, Kuulasmaa K, Project. EHRM. *Review of Surveys for Risk Factors of Major Chronic Diseases and Comparability of the Results. Section 7: Smoking*; 2002. www.thi.fi/publications/ehrm/product1/section7.htm.
68. West R, Hajek P, Stead L, Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. *Addiction*. 2005;100(3):299-303. https://doi.org/10.1111/j.1360-0443.2004.00995.x.
69. Rebagliato M. Validation of self reported smoking. *J Epidemiol Commun Health*. 2002;56(3):163-164. https://doi.org/10.1136/jech.56.3.163.
70. Koehler KA, Peters TM. New methods for personal exposure monitoring for airborne particles. *Curr Environ Heal Reports*. 2015;2(4):399-411. https://doi.org/10.1007/s40572-015-0070-z.
71. Ozkaynak H, Xue J, Spengler J, Wallace L, Pellizzari E, Jenkins P. Personal exposure to airborne particles and metals: results from the Particle TEAM study in Riverside, California. *J Expo Anal Environ Epidemiol*. 1996;6(1):57-78.

72. Peters A, Hampel R, Cyrys J, et al. Elevated particle number concentrations induce immediate changes in heart rate variability: a panel study in individuals with impaired glucose metabolism or diabetes. *Part Fibre Toxicol*. 2015;12:7. https://doi.org/10.1186/s12989-015-0083-7.
73. Deffner V, Küchenhoff H, Breitner S, Schneider A, Cyrys J, Peters A. Mixtures of Berkson and classical covariate measurement error in the linear mixed model: bias analysis and application to a study on ultrafine particles. *Biom J*. 2018;60(3):480-497. https://doi.org/10.1002/bimj.201600188.
74. Kioumourtzoglou M-A, Spiegelman D, Szpiro AA, et al. Exposure measurement error in PM2.5 health effects studies: a pooled analysis of eight personal exposure validation studies. *Environ Health*. 2014;13(1):2. https://doi.org/10.1186/1476-069X-13-2.
75. Coggon D, Rose G, Barker D. *Epidemiology for the Uninitiated*. 5th ed. London, UK: BMJ Publishing Group; 2003. http://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated.
76. Burkhauser RV, Cawley J. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. *J Health Econ*. 2008;27(2):519-529. https://doi.org/10.1016/J.JHEALECO.2007.05.005.
77. Spiegelman D. Cost-efficient study designs for relative risk modeling with covariate measurement error. *J Stat Plan Inference*. 1994;42(1–2):187-208. https://doi.org/10.1016/0378-3758(94)90196-1.
78. Reilly M. Optimal sampling strategies for two-stage studies. *Am J Epidemiol*. 1996;143(1):92-100. https://doi.org/10.1093/oxfordjournals.aje.a008662.
79. Holford TR, Stack C. Study design for epidemiologic studies with measurement error. *Stat Methods Med Res*. 1995;4(4):339-358. https://doi.org/10.1177/096228029500400405.
80. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*. 1989;8(9):1051-1069. https://doi.org/10.1002/sim.4780080905.
81. Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty. *Stat Med*. 1998;17(12):1375-1389. https://doi.org/10.1002/(SICI)1097-0258(19980630)17:12%3C1375::AID-SIM857%3E3.0.CO;2-D.
82. McKeown-Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *Am J Epidemiol*. 1994;139(4):415-421. https://doi.org/10.1093/oxfordjournals.aje.a117014.
83. White E, Kushi LH, Pepe MS. The effect of exposure variance and exposure measurement error on study sample size: implications for the design of epidemiologic studies. *J Clin Epidemiol*. 1994;47(8):873-880. https://doi.org/10.1016/0895-4356(94)90190-2.
84. Tosteson TD, Buzas JS, Demidenko E, Karagas M. Power and sample size calculations for generalized regression models with covariate measurement error. *Stat Med*. 2003;22(7):1069-1082. https://doi.org/10.1002/sim.1388.
85. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. *Biometrics*. 1988;44(1):79. https://doi.org/10.2307/2531897.
86. Yang Q, Liu T, Kuklina EV, et al. Sodium and potassium intake and mortality among US adults. *Arch Intern Med*. 2011;171(13):1183-1191. https://doi.org/10.1001/archinternmed.2011.257.
87. Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control Clin Trials*. 2000;21(6):552-560. https://doi.org/10.1016/S0197-2456(00)00104-5.
88. Subar AF, Kipnis V, Troiano RP, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. *Am J Epidemiol*. 2003;158(1):1-13. https://doi.org/10.1093/aje/kwg092.
89. Selected OPEN Data. https://epi.grants.cancer.gov/past-initiatives/open/. Accessed September 5, 2019.
90. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *J Am Stat Assoc*. 1990;85(411):652-663. https://doi.org/10.1080/01621459.1990.10474925.
91. Armstrong B. Measurement error in the generalised linear model. *Commun Stat Simul Comput*. 1985;14(3):529-544. https://doi.org/10.1080/03610918508812457.
92. 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(4):734-745. https://doi.org/10.1093/oxfordjournals.aje.a115715.
93. 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(11):1400-1413. https://doi.org/10.1093/oxfordjournals.aje.a116453.
94. Bartlett JW, De Stavola BL, Frost C. Linear mixed models for replication data to efficiently allow for covariate measurement error. *Stat Med*. 2009;28(25):3158-3178. https://doi.org/10.1002/sim.3713.
95. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat Med*. 2001;20(1):139-160. https://doi.org/10.1002/1097-0258(20010115)20:1<139::AID-SIM644>3.0.CO;2-K.
96. White I, Frost C, Tokunaga S. Correcting for measurement error in binary and continuous variables using replicates. *Stat Med*. 2001;20(22):3441-3457. https://doi.org/10.1002/sim.908.
97. Fuller W. *Measurement Error Models*. New York: John Wiley & Sons, Inc; 1987.
98. Prentice RL, Huang Y, Kuller LH, et al. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. *Epidemiology*. 2011;22(2):170-179. https://doi.org/10.1097/EDE.0b013e31820839bc.
99. Freedman LS, Midthune D, Carroll RJ, et al. Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. *Am J Epidemiol*. 2011;174(11):1238-1245. https://doi.org/10.1093/aje/kwr248.
100. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall; 1994.

101. Collaboration FS. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Stat Med*. 2009;28(7):1067-1092. https://doi.org/10.1002/sim.3530.
102. Murad H, Freedman LS. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Stat Med*. 2007;26(23):4293-4310. https://doi.org/10.1002/sim.2849.
103. Murad H, Kipnis V, Freedman LS. Estimating and testing interactions when explanatory variables are subject to non-classical measurement error. *Stat Methods Med Res*. 2016;25(5):1991-2013. https://doi.org/10.1177/0962280213509720.
104. Cook NR, Obarzanek E, Cutler JA, et al. Joint effects of sodium and potassium intake on subsequent cardiovascular disease: the Trials of Hypertension Prevention follow-up study. *Arch Intern Med*. 2009;169(1):32-40. https://doi.org/10.1001/archinternmed.2008.523.
105. Whittemore AS. Errors-in-variables regression using Stein estimates. *Am Stat*. 1989;43(4):226-228. https://doi.org/10.1080/00031305.1989.10475663.
106. Shaw PA, Prentice RL. Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics*. 2012;68(2):397-407. https://doi.org/10.1111/j.1541-0420.2011.01690.x.
107. Liao X, Zucker DM, Li Y, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics*. 2011;67(1):50-58. https://doi.org/10.1111/j.1541-0420.2010.01423.x.
108. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Assoc*. 1994;89(428):1314-1328. https://doi.org/10.1080/01621459.1994.10476871.
109. Carroll RJ, Kuchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J Am Stat Assoc*. 1996;91(433):242. https://doi.org/10.2307/2291401.
110. Küchenhoff H, Lederer W, Lesaffre E. Asymptotic variance estimation for the misclassification SIMEX. *Comput Stat Data Anal*. 2007;51(12):6197-6211. https://doi.org/10.1016/j.csda.2006.12.045.
111. Küchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*. 2006;62(1):85-96. https://doi.org/10.1111/j.1541-0420.2005.00396.x.
112. Stefanski LA, Cook JR. Simulation-extrapolation: the measurement error jackknife. *J Am Stat Assoc*. 1995;90(432):1247-1256. https://doi.org/10.2307/2291515.
113. Lederer W, Küchenhoff H. Simex: SIMEX- and MCSIMEX-algorithm for measurement error models; 2013. http://cran.r-project.org/package=simex.
114. Alexeeff SE, Carroll RJ, Coull B. Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics*. 2016;17(2):377-389. https://doi.org/10.1093/biostatistics/kxv048.
115. Oh EJ, Shepherd BE, Lumley T, Shaw PA. Considerations for analysis of time-to-event outcomes measured with error: bias and correction with SIMEX. *Stat Med*. 2018;37(8):1276-1289. https://doi.org/10.1002/sim.7554.
116. Hardin JW, Schmeidiche H, Carroll RJ. The regression-calibration method for fitting generalized linear models with additive measurement error. *Stata J*. 2003;3(4):361-372. https://doi.org/10.1177/1536867X0400300406.
117. Hardin JW, Schmiediche H, Carroll RJ. The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata J*. 2003;3(4):373-385. https://doi.org/10.1177/1536867X0400300407.
118. He W, Yi GY, Xiong J. Accelerated failure time models with covariates subject to measurement error. *Stat Med*. 2007;26(26):4817-4832. https://doi.org/10.1002/sim.2892.
119. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. 2018;98:89-97. https://doi.org/10.1016/j.jclinepi.2018.02.023.

How to cite this article: Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment. *Statistics in Medicine*. 2020;1–35. https://doi.org/10.1002/sim.8532