HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2021 July 20.

Published in final edited form as:

Stat Med. 2020 July 20; 39(16): 2232–2263. doi:10.1002/sim.8531.

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2 –more complex methods of adjustment and advanced topics

Pamela A Shaw,
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

Paul Gustafson,
Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Raymond J Carroll,
Department of Statistics, Texas A&M University, College Station TX, 77843 USA and School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway NSW 2007, Australia

Veronika Deffner,
Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität, Munich 80539, Germany

Kevin W Dodd,
Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892, USA

Ruth H Keogh,
Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK

Victor Kipnis,
Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892, USA

Janet A Tooze,
Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem NC 27157, USA

Michael P Wallace,

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/.

Supporting Information

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.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

Helmut Küchenhoff,
Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität, Munich 80539, Germany

Laurence S Freedman
Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer 52621, Israel and Information Management Services Inc., Rockville, MD 20850, USA

Abstract

We continue our review of issues related to measurement error and misclassification in epidemiology. We further describe methods of adjusting for biased estimation caused by measurement error in continuous covariates, covering likelihood methods, Bayesian methods, moment reconstruction, moment-adjusted imputation and multiple imputation. We then describe which methods can also be used with misclassification of categorical covariates. Methods of adjusting estimation of distributions of continuous variables for measurement error are then reviewed. Illustrative examples are provided throughout these sections. We provide lists of available software for implementing these methods and also provide the code for implementing our examples in the Supporting Information. Next, we present several advanced topics, including data subject to both classical and Berkson error, modeling continuous exposures with measurement error and categorical exposures with misclassification in the same model, variable selection when some of the variables are measured with error, adjusting analyses or design for error in an outcome variable, and categorizing continuous variables measured with error. Finally, we provide some advice for the often met situations where variables are known to be measured with substantial error, but there is only an external reference standard or partial (or no) information about the type or magnitude of the error.

Keywords

Bayesian methods; Bias analysis; Distribution estimates; Likelihood methods; Moment Reconstruction; Multiple imputation

1. Introduction

In the first part of this paper we presented the basic concepts underlying the effects of measurement error and misclassification of variables, described validation and other types of studies that provide information regarding the statistical properties of the error involved in measurement, discussed study design and impact of measurement error on sample size, and presented some methods of adjusting inference for measurement error in simple but commonly occurring situations in epidemiology. In this second part, we present some more complex methods of adjusting estimates or inference for measurement error and misclassification (Section 2), discuss methods to estimate a distribution of an outcome subject to error (Section 3), review the software available for performing such analyses

(Section 4), and describe some recent developments regarding more advanced problems (Section 5). The methods described in both Parts 1 and 2 of our tutorial are based on knowledge of the type and magnitude of the measurement error. In Section 6, we provide advice on how to deal with the all-too-common situations in which such information is imperfect or not available for study participants.

2. Analysis of studies where one or more of the major covariates is measured with error – more complex methods of adjustment

In Section 6 of Part 1, we described two methods of adjusting estimates of association between exposure and outcome when a continuous exposure is measured with error – regression calibration and simulation-extrapolation (SIMEX). Each of these methods is conceptually simple. For regression calibration, the exposure measured with error is replaced by a predicted value of the true exposure and the main analysis proceeds as usual, albeit with adjustment for the standard errors of the estimated association parameters. With SIMEX, one repeatedly introduces more measurement error to approximate a curve for the relationship between the measurement error variance and the regression coefficient in order to estimate the value of that parameter in the absence of measurement error. Sections 2.1–2.4 deal with some more complex but general methods for continuous variables that have measurement error. Some of these methods, such as the Bayesian approach or multiple imputation, can also handle covariate misclassification. Section 2.5 discusses approaches for categorical variables subject to misclassification.

2.1 Likelihood methods

Likelihood methods are pervasive in statistics. This section considers maximum likelihood estimation in measurement error problems. However, likelihood is also a building block for Bayesian inference, which will be discussed in Section 2.2. In the measurement error literature, discussion of maximum likelihood methods is given in the books by Carroll et al,[1] Buonaccorsi[2] and Yi.[3]

Figure 1 illustrates the steps in obtaining the likelihood function in order to carry out measurement error adjustment and perform the likelihood analysis. For non-Berkson error (i.e. classical or linear measurement error model), these steps are as follows:

Step 1: Perform a likelihood analysis. One must specify a parametric model for every component of the data. Any likelihood analysis begins with the model one would use if X were observable. We denote the likelihood of this model as $f_{Y|X,Z}(Y|X,Z,\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ denotes the parameters of the model. For example, in logistic regression, with $H(s) = \exp(s)/\{1 + \exp(s)\}$, the likelihood function is: $\{H(\beta_0 + X^T\beta_X + Z^T\beta_Z)\}^Y \{1 - H(\beta_0 + X^T\beta_X + Z^T\beta_Z)\}^{1-Y}$.

Step 2: Choose the error model. This could be a classical error model, a linear measurement error model, a Berkson model, etc. Presuming non-Berkson error, the likelihood of the model for X* given (X, Y, Z) can be denoted by $f_{X^*|X,Y,Z}(X^*|X,Y,Z,\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denotes the parameters of the model. In the case of non-differential classical measurement error, for

example, if the measurement error is normally distributed with constant variance σ_U^2, then $f_{X^*|X,Y,Z}(X^*|X,Y,Z,\sigma_U^2) = (2\pi\sigma_U^2)^{-\frac{1}{2}}\exp\left\{-(X^*-X)^2/(2\sigma_U^2)\right\}$. Note that this is a slightly stronger version of non-differential error, which in general only requires X* to be conditionally independent of Y given X. Here, X* is conditionally independent of Y and Z given X.

Step 3: If one has a classical or linear measurement error model, specify a distribution for the unobserved X given the observable covariates Z, which we call $f_{X|Z}(X|Z,\boldsymbol{\gamma})$. The need to estimate the distribution of the unobserved X given Z is described in detail in Chapter 8 of Carroll et al.[1] For example, one might assume that X is normally distributed with mean $\gamma_0 + Z^T\boldsymbol{\gamma}_Z$ and variance σ_X^2. In this example,

$$f_{X|Z}(X|Z,\boldsymbol{\gamma}) = (2\pi\sigma_X^2)^{-\frac{1}{2}}\exp\left\{-(X-\gamma_0-Z^T\gamma_Z)^2/(2\sigma_X^2)\right\}.$$

Step 4: Form the likelihood. When X is not observed and is continuous, the likelihood function of the observed (Y,X*) given Z is $\int f_{Y|X,Z}(Y|X,Z,\beta)f_{X^*|X,Z}(X^*|X,Z,\alpha)f_{X|Z}(X|Z,\gamma)dX$. If X is discrete, the likelihood is $\sum f_{Y|X,Z}(Y|X,Z,\beta)f_{X^*|X,Z}(X^*|X,Z,\alpha)f_{X|Z}(X|Z,\gamma)$.

Step 5: Find the values of the parameters $(\beta,\alpha,\boldsymbol{\gamma})$ that maximize the likelihood.

As a brief aside, note that Steps 2 through 4 are specific to non-Berkson error and the analogous procedures for Berkson error would be rather different. Typically, a non-differentiality assumption would be needed, so that the Step 1 specification is in fact for (Y|X, X*, Z). Then the other required specification is the Berkson model for (X|X*,Z), and the product of the two specified densities describes (Y,X|X*,Z). This is then integrated to yield a likelihood function based on (Y|X*,Z).

Steps 4 and 5 (or their counterpart in the case of Berkson error) involve the sometimes hard work of computing and maximizing the likelihood function to obtain parameter estimates. Because X is latent, that is, unobservable, these steps can be difficult or time-consuming, because one must integrate out the possibly high dimensional latent variable. Below, we provide a few details about computing and maximizing the likelihood function.

The overall likelihood based on a sample of n individuals is the product of each individual's likelihood function. Typically, one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways to maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques. The second general approach is to view the problem as a missing-data problem, and then use missing-data techniques; see for example Little and Rubin,[4] Tanner,[5] and Geyer and Thompson.[6]

Computing the likelihoods analytically is usually easier if X is discrete, as the conditional likelihoods are simply sums of terms. For likelihoods in which X is continuous, standard numerical methods for integration, such as Gaussian quadrature, can be applied. When

sufficient computing resources are available, the likelihood can be computed using Monte Carlo techniques.

There are many computer routines for minimizing functions. Since we want to maximize the log likelihood, it is typical to multiply the log likelihood by −1 and then minimize it: the inverse of the Hessian matrix in such a computation serves as an estimate of the joint covariance matrix of all the parameters. See Section 4.4 for further comments on software for performing likelihood-based analyses.

The above description covers cases where X is not observed. In cases where X is observed for a subset of individuals in an internal validation study, the likelihood of the observed (Y, X, X*) conditional on Z must be computed for those individuals separately from the remainder of the participants and then the two sets of likelihoods combined. Similarly, if the internal validation study involves measurement of, not X, but an unbiased measurement X** of X, then an additional measurement error model must be specified for X** and the likelihood of the observed (Y, X*, X**), conditional on Z, computed separately for the individuals having measurements of X**.

To illustrate the likelihood approach, we use an example already introduced in Part 1 of this paper (Section 6). To recap briefly, the Observing Protein and Energy Intake (OPEN) study[7] was a dietary intake validation study using unbiased reference measurements, conducted in 484 adult volunteers. Participants reported on their dietary intake using a food frequency questionnaire (FFQ), provided two 24-hour urine samples for measuring potassium intake, and provided samples for measuring total energy intake through a technique known as doubly labeled water. The target dietary measure is considered to be average daily potassium density intake, i.e. the ratio of potassium intake to total energy intake. The questionnaire responses are considered to have linear measurement error and the urinary biomarker data are considered to have classical measurement error (since they measure only a single day's intake). The issue to be addressed is the association of log potassium density intake with a person's body mass index (BMI). The dataset is referenced as "Selected OPEN data" [2018].[8]

In this analysis, although each participant provided urine samples to measure potassium intake, we assume, as in Part 1, that these were available in only the first 250 participants, so that the "reference" measure is available in only a subsample of the 484 participants. The analysis was performed using the CALIS procedure in SAS. The models required in Steps 1–3, namely the BMI outcome model (model for Y), the FFQ log potassium density intake measurement error model (model for X*) and the log potassium density intake exposure model (model for X) are specified in the upper part of Table 1. Note that in this example X itself is not observed, even in the validation subset, and instead the biomarker log potassium density X**, an unbiased measure of X, is observed in a subset. The measurement error model for X** must also be included (see Table 1). Estimates of the regression coefficients in the outcome model are presented in the middle part of Table 1. These results are later compared with those of Bayesian methods (Section 2.2), moment reconstruction (Section 2.3), multiple imputation (Section 2.4) and regression calibration (Part 1, Table 2) – see the discussions at the end of Sections 2.2, 2.3 and 2.4.

2.2 Bayesian Methods

Arguably there are advantages and disadvantages to taking a Bayesian standpoint when addressing measurement error problems. Perhaps the biggest advantage of a Bayesian approach to measurement error correction is an inherent logical and conceptual simplicity. After specifying appropriate sub-models and prior distributions for the unknown parameters therein, the remaining steps are then a matter of computation. A joint posterior distribution of the unknown parameters ensues, and all inferences stem from this in a logical manner. However, this does require committing to specific prior distributions for all unknown parameters, and not all users will be comfortable with this.

As soon as it was recognized that Markov Chain Monte Carlo (MCMC) computational techniques greatly expand the "domain of applicability" for Bayesian methods,[9] applications to measurement error adjustment quickly followed.[10–12] An overview of the Bayesian approach to adjusting for measurement error is provided by Gustafson.[13] Bartlett and Keogh[14] make specific comparisons between Bayesian adjustments for measurement error and regression calibration (Part 1, Section 6.1), maximum likelihood (Part 2, Section 2.1) and multiple imputation (Part 2, Section 2.4).

To give a specific example, suppose one has a parametric exposure model $f_{X|Z}(X|Z,\boldsymbol{\gamma})$ for (X | Z) parameterized by $\boldsymbol{\gamma}$, a parametric outcome model $f_{Y|X,Z}(Y|X,Z,\boldsymbol{\beta})$ for (Y | X, Z) parameterized by β, and a parametric measurement error model $f_{X^*|X,Y,Z}(X^*|X,Y,Z,\boldsymbol{\alpha})$ for (X* | Y, X, Z) parameterized by α. Note that these are the same specifications as required for a likelihood analysis, as described in Section 2.1. Further, assume there is a validation subsample, such that additionally the actual exposure X is observed for the first n of the N study subjects. Then the joint posterior density of all parameters (in this instance $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$) and latent variables (in this instance $X_{(n+1):N}$) can be expressed as

$$f_{post}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, X_{(n+1):N} \mid X_{1:n}, (X^*, Y, Z)_{1:N}\right) \propto f_{X|Z}(X_{1:N} \mid Z_{1:N}, \boldsymbol{\gamma}) \times f_{Y|X,Z}(Y_{1:N} \mid (X,Z)_{1:N}, \boldsymbol{\beta}) \times f_{X^*|Y,X,Z}(X^*_{1:N} \mid (Y, X, Z)_{1:N}, \boldsymbol{\alpha}) \times f_{prior}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where the four terms on the right-hand side are, in order, the exposure model density, the outcome model density, the measurement error model density, and the prior density of all the parameters. MCMC methods can be applied to draw simulated samples from this joint posterior density, hence the drawn (α, β, γ) values (upon ignoring the drawn $X_{(n+1):N}$ values) are representative of the posterior distribution of the unknown parameters given the observed data. Bayesian point and interval estimates are thereby computed as appropriate summaries of this MCMC output. Note that this approach frees the user from having to explicitly join together two likelihood functions, one for the unvalidated observations and another for the validated observations as described in Section 2.1. Arguably this is a simplifying feature of proceeding in a Bayesian fashion.

As was seen to be the case with likelihood methods, using the Bayesian paradigm to "glue together" three sub-models for exposure, outcome, and measurement has the appealing feature that uncertainty propagates across these sub-models in a manner that is both principled and automatic. For instance, reported uncertainties (say posterior standard

deviations or credible intervals directly computed from the MCMC output) about estimated outcome model parameters fully acknowledge the uncertainty about measurement error model parameters and exposure model parameters. So the data analyst is less burdened by issues of whether uncertainties are correctly propagated than is the case for, say, regression calibration (see Part 1, Section 6.1) or SIMEX (see Part 1, Section 6.2) approaches.

Against this coherence and logical simplicity, there are some challenges. General-purpose MCMC software for Bayesian analysis is available, including packages such as WinBUGS, JAGS, and STAN. However, some very specialized sub-model specifications may not be supported by some packages. There is also the more pervasive issue that MCMC works better for some models and datasets than others, such that there is a need to examine the sampled draws to rule out problems with convergence and/or mixing of the MCMC algorithm. (MCMC methods draw realizations of a Markov chain specially constructed to have the posterior distribution of parameters and latent variables as its stationary distribution, relying on the fact that a Markov chain converges to its stationary distribution under weak assumptions.) Bayesian computing is not yet at the level of an “automated black box”. Additionally, it can be more challenging to relax modeling assumptions when working under the Bayesian paradigm. That is, going from parametric to semi-parametric or nonparametric analysis becomes quite intricate, even though there has been much research on Bayesian nonparametric methods over the last decade. Papers by Sarkar et al[15,16] and Sinha and Wang[17] are recent examples that bring nonparametric Bayes technology into measurement error adjustment problems. Finally, as with all Bayesian analyses, some see the requisite specification of a prior distribution of the unknown parameters as a blessing, while others perceive it as a curse.

Much of the measurement error literature presumes a “hard” source of information about the measurement error magnitude, via observed replicates of a measurement X* that has classical measurement error, or data from a validation subsample. Implicit here is the notion that if the amount of data increases in the right way, then the values of *all* the parameters, including those describing the measurement error process, would be revealed, i.e., estimated consistently. However, Bayesian methods also offer the alternative possibility of using “soft” information. For instance, in the absence of replicates or a validation study, subject-area experts could assert a range of plausible measurement error magnitudes; a prior distribution that puts the vast majority of its mass on this plausible range could then be chosen. Of course, this sort of uncertainty would not diminish as more data are collected; so one must bear in mind that the final answer incorporates the usual statistical uncertainty arising because the sample size is finite, as well as the uncertainty in the experts’ opinions about the measurement error magnitude. Related, there is no “free lunch”. If one places a very diffuse (or even “improper”) prior distribution on the measurement error magnitude, a correspondingly diffuse (or even “improper”) posterior distribution will result. No useful measurement error correction can arise without either data or expert opinion to inform the magnitude of the measurement error.

We illustrate the Bayesian approach through the same example presented for the maximum likelihood analysis in Section 2.1. The analysis was performed in RJAGS. The exposure model, outcome model and measurement error models are the same models for the

likelihood approach in Section 2.1 and are specified in the upper part of Table 1. Prior distributions with minimal information were adopted for the parameters of the exposure model, outcome model and measurement error models; all regression coefficients were given normal priors (with mean zero and variance 1000) and precision (reciprocal of variance) parameters were given gamma priors (with shape and rate both set to 0.01). Estimates of the regression coefficients in the outcome model are presented at the bottom of Table 1. The results are rather similar to those of maximum likelihood presented in the middle of Table 1, and we will see they are also similar to those of the moment reconstruction and multiple imputation approaches to be discussed in the next section. The posterior standard deviation for the target parameter (describing the relationship between potassium intake and BMI) (1.43) is comparable to the standard error obtained for the maximum likelihood estimate (1.25).

2.3 Moment reconstruction and moment-adjusted imputation

Moment reconstruction (MR) and moment-adjusted imputation (MAI) are methods for handling covariate measurement error in which the goal is to construct a quantity $X_M(X^*, Y)$ that has the same distribution as X, and such that (X_M,Y) has the same joint distribution as (X,Y). If covariates Z are also to be included in the regression of Y on X, then the above distributions are conditional on Z. The quantity $X_M(X^*, Y)$ is generally constructed by estimating moments of the joint distribution of (X, Y) from validation data, and then is substituted for X into the desired outcome regression model to produce an estimate β_X. Standard errors that account for the extra variability in the resulting estimate of β_X, which comes from the uncertainty in the parameter estimates used to construct $X_M(X^*, Y)$, are necessary and can be obtained using the bootstrap. The bootstrap sample in this case is stratified on membership in the validation subset.

In MR[18], $X_M(X^*, Y, Z)$ is constructed by matching only the first two moments of the joint distribution for (X, Y). In the case of classical measurement error this is achieved by defining $X_M(X^*, Y, Z)$ as:

$$X_M(X^*, Y, Z) = E(X^*Y, Z) + G\{X^* - E(X^*Y, Z)\} \quad (1)$$

where $G = \text{var}(X \mid Y, Z)^{1/2}\{\text{var}(X^* \mid Y, Z)\}^{-1/2}$. This expression can be extended to the linear measurement error model by replacing the first $E(X^* \mid Y, Z)$ on the right hand side with $E(X \mid Y, Z)$, keeping G as before.[19] When the measurement error parameters are assumed known and the error is non-differential, MR is equivalent to regression calibration in linear regression, and is therefore consistent. When the measurement error is non-differential and the error model parameters are estimated in an ancillary study, MR is not equivalent to regression calibration, but both are consistent. MR is also consistent for logistic regression with normally distributed covariates, unlike regression calibration which is only approximately consistent [Carroll et al,[1] p.91]. Under conditions of differential measurement error, MR is still consistent as the necessary moments are estimated conditional on Y, and in this it can prove advantageous compared to regression calibration, which is biased.

MAI is an extension of MR, in which the moments of $(X_M(X^*, Y, Z),Y \mid Z)$ match more than the first two moments of $(X, Y \mid Z)$. Thomas et al[20] recommended matching the first 4

moments and observed that when X is normal, its performance is similar to regression calibration in linear and non-linear regression models. However, MAI has been shown superior to regression calibration for logistic regression where the distribution of X is far from normal. MR and MAI may also be used for several covariates measured with error.[18,21]

We illustrate MR with an example from the OPEN study, similar to the one in the previous two sections. We consider the regression of BMI on log sodium intake, while controlling for age and sex. We choose this example as there is evidence in the data that the measurement error in self-reported FFQ sodium intake is differential with respect to the BMI outcome variable, so that regression calibration is an inappropriate method of adjustment. The estimated regression coefficients of the unadjusted model, using FFQ-reported sodium intake, are presented in the second column of part A of Table 2. The estimated coefficient for log sodium intake is 1.13, and its z-value 1.97. However, the z-value (unlike for a single covariate with non-differential error) is invalid because the error is differential (see Part 1, Table 1). Parts B1 and B2 of Table 2 show the results of the models for E(X|Y,Z) and E(X*|Y,Z), which are both needed for the construction of $X_M(X^*,Y,Z) = E(X|Y,Z) + G\{X^* - E(X^*|Y,Z)\}$. The third column of Part A of Table 2 shows the MR adjusted estimated coefficients for log sodium intake, age and sex. The standard errors are obtained by bootstrap. One can see that the estimated coefficient for log sodium intake is 12.21, about 10 times larger than the unadjusted estimate, with a z-value of 4.53. This means that a 30% increase in sodium intake (change in log sodium intake of 0.26) is associated with an increase of 3.2 BMI units (95% CI: 1.8–4.5), rather than 0.3 units obtained from the unadjusted model. (Note that this very large effect cannot be directly causal since sodium is a micronutrient, supplying no calories. However, it indicates that high sodium intake is associated with higher BMI, probably because sodium intake is strongly correlated with energy intake.) There is also a notable change in the coefficient for sex from a non-significant negative association ($z=-0.46$) in the unadjusted analysis to a significant positive association ($z=3.16$), i.e. higher BMI in women than men for a given age and sodium intake (but see also the result for multiple imputation, given in Section 2.4).

MR can also be used as an alternative to regression calibration when measurement error is non-differential. In that case, it can be more efficient or less efficient than regression calibration depending on the type of data at hand. If we apply MR to the same problem as used in Part 1, Section 6.1.1, namely the regression of BMI on log potassium density, we obtain a remarkably improved result over that from regression calibration, in terms of the variance of the adjusted estimate. One can directly compare the estimates for regression calibration presented in Part 1, Table 2 with those obtained for MR here in the third column of Table 3. The estimated regression coefficient for log potassium density is −8.13 (compared to −3.76 for regression calibration) with a bootstrap standard error of 1.77 (compared to 2.49 for regression calibration). For the full result of the final model, see the third column of Table 3. In this case, MR is more efficient than regression calibration and reveals a significant negative association of potassium density with BMI. The circumstances that cause the greater precision of MR over regression calibration are (i) the quite strong association between outcome and exposure, and (ii) the availability of the biomarker in more than 50% of the participants. In many studies, the association between the outcome and exposure variable is much weaker (for example in studies of disease incidence) and the

validation data are available in a much smaller proportion of participants, causing regression calibration to be more efficient than MR (see Freedman et al[19]). We will further consider how MR compares to multiple imputation for this example in the next section.

2.4 Multiple imputation

When there is an internal validation subset, in which (X, X*, Z, Y) are all observed, then the case of measurement error is really just a problem of missing data.[22] If individuals in the validation sample are a random sample of the main study population then the unobserved X is missing completely at random (MCAR); if the sample is dependent by design on covariates, then X is missing at random (MAR) [Little and Rubin,[4] Chapter 1]. In either case, the distribution of X | X*, Z, Y is the same for those in the validation sample as those who are not; thus, the values of X can be imputed from a model for X | X*, Z, Y. Multiple imputation (MI), in which the unobserved values X are imputed m times, allows estimation of the coefficients in the outcome model and their standard errors [Little and Rubin,[4] Chapter 10]. Under a correctly specified model for X | X*, Z, Y, MI will produce consistent estimates for β_X and consistent standard errors. Like MR and MAI, and unlike regression calibration, MI can handle differential measurement error, since Y is used for imputing the unknown X. The same procedure can be used if the error is assumed to be non-differential, and a more efficient version of MI may also be constructed under this assumption. Freedman et al[19] found that in circumstances where regression calibration outperformed default MI, the "non-differential" MI method performed similarly to regression calibration. Here, we focus entirely on the default version, which accommodates differential error.

As with any setting, the success of MI relies on having sufficient data to build a reliable imputation model and on correct specification of that model. For this reason, MI is generally not recommended when only an external validation study is available, and coincident measures of Y are not available.[19,23]

It is also possible to use MI when there is an internal validation subset in which, instead of X, a measure of X that has classical measurement error is obtained, as well as X*, Y and Z (a calibration study – see Part 1, Section 4.2). However, in this case, implementation of the method is a little more involved than usual MI. Details are provided in Section A2 of Appendix A in Freedman et al.[19] This method is the one used in the examples that follow. Note, Keogh and White[24] described an MI approach for use in the setting of a replicates study, assuming availability of repeated measure of the error-prone covariate in some individuals, and assuming classical error. More recently, another approach for the setting of a validation or replicates study had been described[25] based on a modification of the substantive model compatible imputation approach for missing data described by Bartlett et al (2015),[26] and accompanying software is available in R.[27]

To illustrate MI, we use the same two examples as given for MR in Section 2.3. First, we consider the regression of BMI on log sodium intake, while controlling for age and sex. The results are presented in Table 2. Part C of the table shows the model that is used as a basis for imputing the unknown values of true log sodium intake. Note that this is based on a regression of the biomarker log sodium intake on the FFQ log sodium intake, BMI, age and sex. One can see that the main variables influencing the imputation are BMI and sex. Note

the strong effect of BMI in the imputation model does not imply differential error. The fourth column of Part A shows the estimated model of BMI on log sodium intake, age and sex, based on 500 multiple imputations. Generally, using a relatively large number of multiple imputations is recommended, since one is imputing 100% of the values for the unknown true log sodium intake, although no formal recommendations have been established on the number of imputations required in this context. The results are similar, although not identical, to those obtained from MR (third column, part A of Table 2). The coefficient for log sodium intake is again about 10 times the unadjusted estimate and is highly statistically significant. The coefficient for sex is large and positive but, unlike for MR, does not attain conventional statistical significance. Discrepancies between the results of the two methods are not very common, but a careful analyst would perform both methods, where possible, to check on the stability of results. What is clear from the results of both methods is that the association of BMI with sodium intake appears far stronger than that indicated in the unadjusted analysis.

As mentioned with MR, one may use MI also in cases of non-differential measurement error. We applied MI to the example of Part 1, Section 6.1.1, which was the analysis of the association of BMI with log potassium density. The results are shown in the final column of Table 3. The results are in accord with those of MR, indicating a strong negative association of BMI with potassium density intake. The same remarks made in Section 2.3, about the relative efficiency of MR compared to regression calibration, apply also to MI. The results of MR and MI presented in Table 3, showing a strong negative association between BMI and log potassium density, are rather similar to those for the likelihood method presented in Table 1. However, the standard error (1.25) of the regression coefficient for log potassium density intake is considerably smaller for the likelihood method than for the MR and MI analyses (1.77 and 2.03, respectively). A possible explanation for this is that the maximum likelihood analysis included all the data, whereas the MR and MI analyses, for simplicity, omitted 13 participants who provided one urine sample only (instead of two) for the measurement of potassium density. When the datasets are identical one would expect MI and maximum likelihood to yield very similar estimates and standard errors.

The performance of MI has been compared with other methods, including regression calibration and MR in settings of linear and logistic regression[19,23,28] and Cox regression.[22,29] These authors found that the optimal method depends on the size of the validation subset and degree of measurement error. Shepherd et al[28] noted MI worked well in the setting of correlated covariate and outcome measurement error in the linear model. With censored survival data, implementing MI can be especially challenging; Bang et al[29] recommended implementing multiple methods to compare sensitivity of results to assumptions since in reality one rarely knows the true model for the error structure. In the case of a linear outcome and linear non-differential measurement error model, a method of moments approach (MOM) can also be applied.[2] In this case we expect the performance of MOM to be similar to that of regression calibration, as seen by Shaw et al. 2018.[30]

2.5 Analysis of studies where one or more categorical covariates are subject to misclassification

Section 2 of this paper and Section 6 of Part 1 have addressed various methods that can be applied when a continuous covariate X is measured with error. The "methods menu" one can choose from when faced with a categorical covariate subject to misclassification is similar, but not identical.

Likelihood (Section 2.1) and Bayesian methods (Section 2.2) transfer directly and simply from the continuous covariate case to the discrete case. In fact, these methods are arguably more attractive in the discrete case, since concern about possible model misspecification (for the distribution of the unobservable X) is typically reduced. If X is binary and there are no precisely measured covariates, then X must follow a Bernoulli distribution, so that there is no concern about misspecification. If there are precisely measured covariates Z, then a model for X|Z is required, and misspecification could arise. For instance, a logistic regression relationship between X and Z might be posited, and might be wrong. But more fundamental concerns about the *shape* of the X distribution do not apply when X is categorical. So likelihood and Bayesian methods, as per Sections 2.1 and 2.2, can be applied as described. The variable type for X is not particularly consequential for these methods.

In some simple situations, notably when Y and Z (if applicable) are also categorical, closed-form estimation of parameters is sometimes possible. This literature dates back to at least Barron[31] who proposed the closed-form "matrix method" applicable when there are main study data in the form of a 2 by 2 table for binary (X*, Y) and validation data in the form of a 2 by 2 by 2 table for binary (X, X*, Y). Subsequently, Marshall[32] proposed an alternative closed-form estimator, known as the "inverse matrix method", by framing the classification in terms of predictive values rather than specificity and sensitivity. Later work, notably that of Morrissey and Spiegelman,[33] Lyles[34] and Greenland[35] served to both (i), quantify the efficiency of these closed-form estimators relative to iteratively computed maximum-likelihood estimators, and (ii), understand nuances of how the various methods work under differential and non-differential misclassification assumptions.

In more involved contexts, the expectation-maximization (EM) algorithm is quite straightforwardly applied to compute maximum likelihood estimates of parameters in the (Y | X, Z) model.[36] Also, just as a categorical X variable is particularly amenable to the EM algorithm for likelihood estimation, it is also particularly amenable to MCMC methods (and Gibbs sampling specifically) for computing Bayesian estimates. See Joseph et al,[37] Gustafson et al,[38] Johnson et al[39] and Prescott and Garthwaite[40] for examples.

Both regression calibration and moment reconstruction are less obvious strategies to pursue explicitly when X is categorical. However, some of the estimators discussed in Morrissey and Spiegelman,[33] Lyles[34] and Greenland[35] indeed end up having a regression calibration spirit. That is, they can be viewed as replacing X with an estimate of E(X | X*, Z). Also, multiple imputation can certainly be applied to problems involving a categorical X. In fact, this approach will be rather similar to a Bayesian analysis using MCMC computation.

The SIMEX method, described in Part 1, Section 6.2, has been extended to handle a categorical X variable that is subject to misclassification, using a method termed MC-SIMEX. Suppose we have a regression model with a discrete covariate X which is subject to misclassification.[41] The misclassification process is described by the matrix Π, which is defined by its components

$$\pi_{ij} = \Pr(X^* = i \mid X = j), i = 1, ..., r; j = 1, ..., r. \quad (2)$$

Π is a r x r matrix, where r is the number of possible outcomes for X. MC-SIMEX employs the function ($s \geq 0$) defined by:

$$\beta_{X*}(s) = \beta_{X*}(\Pi^s), \quad (3)$$

where $\beta_{X*}(\Pi^s)$ denotes the value of the coefficient β* when X* is subject to misclassification by Π^s, defined as $E\Lambda^s E^{-1}$, with Λ being the diagonal matrix of eigenvalues and E the corresponding matrix of eigenvectors. For integer values of s, $\Pi^{1+s} = \Pi^s * \Pi$, where * denotes matrix multiplication, and for s = 0, $\Pi^0 = I_{rxr}$. The central idea of the MC-SIMEX method is to add extra misclassification to X*. Namely, if X* has misclassification probabilities Π in relation to variable X, and X*(s) is related to X* by the misclassification matrix Π^s, then X*(s) is related to X by the misclassification matrix Π^{1+s}, when these two misclassification mechanisms are independent. Thus, the SIMEX algorithm can be applied to misclassification in the same manner as the original SIMEX. For details, including the variance estimation, see e.g. Küchenhoff et al.[42]

As an example of adjusting for misclassification in a binary explanatory variable X, we consider the study reported by Kraus et al[43] on risk factors for sudden infant death syndrome (SIDS). Here X is defined as an indicator of maternal use of antibiotics during pregnancy, as ascertained from medical record review, while X* indicates the mother's self-report of antibiotic use on a questionnaire. The study employed a case-control design (though the same analysis would apply for a cross-sectional or prospective study), recruiting 797 controls (Y=0) and 775 cases (Y=1) of SIDS. Since medical record review was only conducted for a subset of 217 of the controls and 211 of the cases, we are presented with a misclassified data problem with an internal validation study. The data are presented in Table 4. If we ignore the X measurements available for the validated study subjects and simply focus on the (X*,Y) association, we estimate a log odds ratio (OR) of 0.35 (corresponding to OR=1.42), with a standard error of 0.13, indicating a positive association.

A number of authors have illustrated misclassification adjustment methods using data from this study, including Greenland[44] and Chu et al,[45] who contrast multiple methods, including maximum likelihood methods, Bayes methods, and SIMEX. Suspecting the possibility of differential misreporting, they applied a likelihood ratio test for the null hypothesis of conditional independence of X* and Y given X to the data on validated subjects, obtaining some evidence against the null (P = 0.096). Thus, we focus on an adjustment method that allows for differential misclassification. While Section 2.1 alludes to general implementation challenges for likelihood methods arising because X is latent, obtaining maximum likelihood estimates in the present context is actually quite straightforward. As detailed by Lyles,[34] we

can re-parameterize the problem in terms of the (X*,Y) and (X|X*,Y) distributions, rather than the (X,Y) and (X*|X,Y) distributions, with all the study units contributing to estimation of (X*,Y) but only the validated study units contributing to the estimation of (X|X*,Y). Lyles[34] shows that one can then work back to obtain closed-form estimates and standard errors of parameters in the original parameterization. Applying this method to the data at hand results in an estimated (X,Y) log odds-ratio of 0.19 (corresponding to OR=1.21), with a standard error of 0.22. Note that this adjustment for misclassification pushes the point estimate *toward* the null, as can arise when misclassification is differential. For comparison, if we presume non-differential misclassification, then the maximum likelihood estimate of the (X,Y) log odds-ratio, as determined by numerical maximization, is 0.40 (corresponding to OR=1.49), with a standard error of 0.19. As must arise when non-differential misclassification is presumed, relative to the naïve estimate, the adjustment moves the estimate away from the null, and increases the corresponding measure of uncertainty. Generally, this example is another where the medical conclusions to be drawn from the analysis are changed substantially by taking reporting error into account.

3. Analysis methods for estimating distributions

In Part 1, Section 3.4, we briefly discussed the impact of measurement error on estimating the distribution of a random variable Y. In this section, we consider methods for estimating the distribution of Y using error-prone observations Y*. In contrast to prior sections, here the error-prone variable measures the outcome of interest (Y) rather than a covariate (X). We focus on the case where Y is continuous. For example, it might be of interest to know selected percentiles of the distribution of Y, or related quantities such as the interquartile range. Alternatively, one might wish to know what proportion of the distribution falls above or below specific cut-points. In Section 3.4 of Part 1, we briefly considered how such distributional quantities can be biased if estimated by an error-prone Y*. Most of the methods presented here focus on the case where Y* follows the classical measurement error (model (1) in Section 2.1 of Part 1). Development of these methods has been most pronounced in the field of nutrition, where the desired random variable Y is the long-term average daily consumption (“usual intake”) of a food or nutrient, and Y* is typically the reported intake from a 24-hour dietary recall, which queries everything eaten or drunk during the previous 24 hours. Under this scenario, a substantial amount of the measurement error U is assumed to be due to day-to-day variation in diet that makes a single day an imprecise proxy for a long-term average.[46,47] Other sources of systematic error are routinely assumed to be zero in these methods, although there are exceptions.[48]

3.1 A simple case

Consider the simplest case of model (1) (Part I), where $E(Y) = \mu_Y$, $E(U) = 0$, $var(Y) = \sigma^2_Y$ and $var(U) = \sigma^2_U$. Then $E(Y^*) = \mu_Y$ and $var(Y^*) = \sigma^2_Y + \sigma^2_U$. It follows that the distribution of

$$\widetilde{Y} = E(Y^*) + \sqrt{\frac{\sigma^2_Y}{\sigma^2_Y + \sigma^2_U}}(Y^* - E(Y^*)) \quad (4)$$

has the same first two moments (μ_Y, σ^2_Y) as the distribution of Y. This approach, applied to interval data, was used by the National Research Council (NRC),[49] where replicated observations Y* were available to permit separate estimation of the required variance components in equation (4). In general, for this classical measurement error setting, one can estimate var(U) by the within-person variance when there are replicate measurements; var(Y) can then be estimated by subtracting the estimate of var(U) from var(Y*). If Y and U are normally distributed, then Y* is as well, and matching the first two moments of Y is equivalent to fully characterizing the distribution. Thus, the empirical distribution of $\widetilde{Y}$ may be used as an estimate of the distribution of Y. This approach of constructing a set of representative observations to be used as a basis for an empirical distribution estimator can be extended to the more complex cases discussed below.

3.2 Use of normality transformations

The NRC report[49] highlighted the fact that 24-hour recall data (Y*) tend to be skewed, suggesting that the normality assumption is not tenable in the original scale. Therefore, transformations are routinely applied to observed data as a first step in distribution estimation. This is a complication that requires a careful choice of assumption about how Y and Y* are related. In the NRC analysis, formula (4) was applied to log transformed data, and each estimated percentile of the distribution of $\widetilde{Y}$ was exponentiated to obtain the corresponding percentile in the original scale. This approach is consistent with the model

$$g(Y^*) = g(Y) + U \tag{5}$$

where g(·) is an invertible transformation. That is, Y* is unbiased for Y on the transformed scale (and therefore biased for Y on the original scale). The transformation is also presumed to result in well-behaved (e.g., normally distributed) errors U.

We illustrate this approach with an example taken from data obtained in the OPEN study (for a short description of the study, see the example given in Section 2.1). Here we consider estimating the distribution of usual sodium intake in a population typical of those participating in the study. Besides the self-reported intakes (which have some bias), two measurements of 24-hour urinary sodium were available, a biomarker for sodium intake that is thought to be unbiased, but subject to random day-to-day variation and random assay error. Due to this random variation, these measurements (our Y*) have error when the target is to measure usual (i.e. long-term average) intake, Y. In addition, sodium intakes tend to have a skew distribution that is approximately log-normal in shape. We therefore assume model (5) where g is the logarithmic function. The value of the shrinkage factor $\sqrt{\frac{\sigma^2_Y}{\sigma^2_Y + \sigma^2_U}}$ shown in equation (4) (but applied on the logarithmic scale) was 0.72, indicating a relatively large day-to-day variation in sodium intake. Table 5 shows the percentiles of the distribution estimated from a single biomarker measurement Y* assuming it has no random error (the incorrect assumption) versus that based on model (5). The latter is calculated from applying the NRC method to the first measurement of log(Y*), estimating σ^2_Y and σ^2_U from the repeat measurement, and followed by back-transformation. Figure 2 presents the density functions

after smoothing. Both Table 5 and Figure 2 show the substantial shrinkage of the distribution obtained when using the adjustment for measurement error.

Later authors, beginning with Nusser et al[50] assumed the model:

$$g(Y^*) = \mu + U \tag{6}$$

where μ is the individual's long-term average of the transformed Y*, and

$$Y = E_U(Y^*|\mu) = E_U(g^{-1}(\mu + U)|\mu),$$

which assumes that Y* is unbiased for Y on the original scale. Because $g(\cdot)$ is typically nonlinear, estimating the distribution of Y now requires integration over the distribution of U. It is often impossible to decide from the available data which model, (5) or (6), is the more appropriate.

3.3 Model-assisted versus model-based approaches

The NRC method uses the empirical distribution of $\widetilde{Y}$ as the basis for estimating the distribution of Y. However, the routine use of normality transformations in later approaches[50–53] permits the use of exact percentiles from a normal distribution as the basis for estimation. In small samples, the empirical distribution of $\widetilde{Y}$ can be quite granular, leading to a granular approximation of the distribution of Y. Using normal distribution percentiles allows smooth estimated distributions of Y, but at the expense of relying on the normality assumption (after appropriate transformation). The NRC approach, and extensions such as the Multiple Source Method (MSM) method[54] have been characterized as "model assisted", in contrast to the other, "model-based" approaches.[55]

3.4 Extensions for inclusion of covariates, semi-continuous data, and multivariate estimation

A more general version of model (6) is given by the mixed-effects model

$$g(Y^*) = \mu(Z) + r + U \tag{7}$$

where μ(Z) is a function of observed covariates Z of an individual, r is a random effect that is constant across replicate observations on the same subject, and U is random within-subject error. Then,

$$Y = E(Y^* \,|\, r, Z) = E(g^{-1}(\mu(Z) + r + U) \,|\, r, Z)\,. \tag{8}$$

Under (7–8), the variation in the distribution of Y comes in part from the variation explained by covariates Z and in part from residual deviations r. Model (7) can be fit using standard software for (nonlinear) mixed models. The software typically requires assuming normality of the random effects, which favors the use of model-based approaches. To maintain the proper distribution of covariates, a Monte Carlo simulation approach is often used to generate a dataset for calculating the empirical distribution mentioned earlier. In this

approach, predictions of μ(Z) from sampled individuals are added to multiple randomly generated r values before integrating over the (presumed normal) errors U.

Many dietary components are “episodically consumed”, where the single-day reported intakes (Y*) can be zero, even if long-term intake (Y) is positive. The observed data Y* can therefore have a large proportion of zeros, as well as a skewed distribution of positive values. In these situations, simple transformations applied to Y* will not even approximate a normal distribution. Models for such semi-continuous data are motivated by writing the average value of Y* as a conditional expectation:

$$\begin{aligned} E(Y^*) &= E(Y^* \mid Y^* > 0)\Pr(Y^* > 0) + E(Y^* \mid Y^* = 0)\Pr(Y^* = 0) \\ &= E(Y^* \mid Y^* > 0)\Pr(Y^* > 0). \end{aligned} \tag{9}$$

This formulation expresses the average daily consumption as the product of the average consumption on consumption days E(Y* | Y* > 0) and the probability of consuming on a specific day Pr(Y* > 0). This approach led several authors[51,53,54,56,57] to consider models that used binary indicators of zero vs. nonzero consumption Y* to inform estimation of the probability part of the model and used the transformed nonzero values of Y* to inform estimation of the amount part of the model. These methods were further extended [Freedman et al,[58] Zhang et al,[59]] to allow flexible joint modeling of multiple components, which permits analysis of ratios and high-dimensional indices. A detailed description of these extended models is beyond the scope of this work.

3.5 Nonparametric Estimation of Distribution Functions

There is a very large literature on nonparametric estimation of distribution functions. The papers concentrate on estimating the density function, and this is often called *density deconvolution*. This literature has two major themes. The first uses kernel density functions, while the second usually uses infinite mixtures of normal and/or Bsplines. For the first, see Carroll and Hall,[60] Stefanski and Carroll,[61] and Fan,[62] while for the case of heteroscedastic measurement error, see Delaigle and Meister.[63] For the second, see Staudenmayer et al,[64] and Sarkar, et al.[15] For multivariate density deconvolution, see Masry[65] and Sarkar, et al.[66] The articles by Sarkar et al.[15,66] are very general, allowing heteroscedastic measurement error with unknown distributions for that measurement error, as well as of course unknown distributions for the latent variable. There is substantial software available for these estimation methods. References to online sources of the available code, which cover methods for kernel-based deconvolution and Bayesian semiparametric density deconvolution, are provided in Table 6. Non-parametric maximum likelihood is another approach considered by several authors.[67,68]

4. Software for analysis

One of the main barriers in the past to the use of the analysis methods described in Sections 2 and 3 was the lack of specific software for implementing them. The situation is now gradually improving. Here, we describe software programs, macros or packages that are now available for performing some of the methods described in those sections. Note that software for performing regression calibration and SIMEX was described in Part 1, Section 7. We

also provide the code that conducted our analyses of the OPEN data at https://github.com/PamelaShaw/STRATOS-TG4-Guidance-Paper.

4.1 Software for Bayesian methods

Over the last several decades, the most common software used for applied Bayesian work has been *BUGS*.[69–71] Fitting a Bayesian model to data using *BUGS* involves specifying model and prior distributions within the *BUGS* language, and then using both a *BUGS* interface and a *BUGS* engine to get the work done. In particular, the engine carries out MCMC sampling of the posterior distribution of parameters and latent variables given observed data. The interface serves to deliver the model and prior specifications and the data to the engine, and to then process the Monte Carlo output from the engine into inferential quantities. While the *BUGS* language is unique, a common point of confusion is that there are multiple possible engines and interfaces. Commonly used engine/interface combinations include *WinBUGS*/*R2WinBUGS*,[72] and *JAGS*/*rjags*.[73] A worked example of implementation using R and JAGS is provided in supplementary materials to the paper of Bartlett and Keogh 2018.[74]

Generally, and for measurement error modeling specifically, using *BUGS* is intermediate between a completely "do-it-yourself" workflow and a fully automated macro. The user need not have in-depth knowledge of MCMC algorithms, i.e., the user is not required to select, code, and tune a particular algorithm. However, the user must express the chosen model and prior distributions in the *BUGS* language. On balance this seems a plus, as these specifications are then much more customizable than would be the case with a fully automated macro having a "hard-wired" model specification. Three specific examples of measurement error models expressed in the *BUGS* language appear in Section 9.3 of Lunn et al.[71] Many problems could be approached by extending one of these examples.

Regardless of how the workflow is implemented, and as alluded to in Section 2.2, using MCMC to compute point and interval parameter estimates requires somewhat more scrutiny and oversight than with other statistical methods. Issues of sampler "burn-in" and "mixing" arise, so that some level of human judgment is needed to assess whether the amount of Monte Carlo sampling utilized is indeed sufficient to numerically approximate posterior quantities well. This process is streamlined, but not automated, with the *BUGS* interfaces mentioned above. Simple summaries and diagnostic plots are readily provided to attest to the trustworthiness of the computational output.

Recently, the *Stan* probabilistic programming language has been developed, along with an *R* interface.[75] Compared to *BUGS* engines, *Stan* makes use of rather different MCMC algorithms, with excellent performance reported in many contexts. Chapter 11 of the *Stan* Reference Manual[76] illustrates the coding of a measurement error model in Stan.

One package that is specifically written for Bayesian analysis of measurement error problems in R is *BayesME*, available at http://www.stat.tamu.edu/~carroll/matlab_programs/software.php. This package is based on Sarkar et al[15] and Sarkar et al,[16] and deals with nonparametric density and regression estimation when the measurement error is heteroscedastic, unknown and may depend on X.

4.2 Software for moment reconstruction, moment-adjusted imputation and multiple imputation

When a validation substudy is available (Section 4.2), i.e. in which the true X is observed, MI may be implemented using standard multiple imputation packages. Available packages include *mice*,[77] and *smcfcs* in R,[27] *mi impute* and *smcfcs* in Stata,[78] and PROC MI in SAS. The *smcfcs* package in R has been extended to accommodate measurement error correction in the settings of a validation substudy or a replicates substudy, and allows measurement error and missing data to be addressed simultaneously.[27]

No packages are available for the implementation of MR or moment-adjusted imputation (MAI). Thus, MR and MAI require a program to construct the "predicted" values of X. However, from thereon those predicted values may be used in standard regression programs to yield the measurement error adjusted estimates of the regression coefficients. Valid standard errors of these estimates may then be obtained by bootstrap methods.

4.3 Software for estimating distributions

As mentioned in Section 3, several packages or macros are available for estimating the distribution of a variable Y, using measurements Y* that have classical measurement error. Most, but not all, of these packages have been developed for nutritional data but may be used for other types of data. All packages require that all individuals have at least one measurement of Y* and that a substantial number have one or more repeat measurements. The user will note that some of these programs deal not only with continuous Y* variables, but also with semi-continuous Y* that have a positive probability of a zero value. Nevertheless, most of them assume that Y is continuous even when Y* is semi-continuous, and they yield an estimated continuous distribution. Standard errors are typically obtained via replication methods such as the bootstrap (or under specific survey designs, balanced repeated replication). The packages and macros are summarized in Table 6. The resources listed in Table 6 include software for methods relying on a variety of necessary assumptions, including the semiparametric and nonparametric methods described in Section 3.5.

4.4 Software for other methods

Section 2.1 includes some comments on what is needed to program likelihood methods. These methods are often implemented though custom-built programs. Rabe-Hesketh et al[79] describe how to conduct maximum likelihood estimation in *Stata* when X is normally distributed. In many problems of measurement error, the *SAS* procedures *MIXED*, *NLMIXED* and *CALIS* can be used, and were used for the NCI Method *SAS* macros (see Table 6). The *SAS* manual has a nice introduction illustrating how *CALIS* can be used for measurement error modeling (see: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_calis_sect001.htm), and the procedure can be used for nonlinear modeling as well.

The R software package SIMEX, introduced in Part 1 for error in a continuous exposure, also includes the *mcsimex* function, which implements the MC-SIMEX adaption for misclassified categorical exposures described in Section 2.5.[80,81] There is also a web site that has *R* and *Matlab* programs (http://www.stat.tamu.edu/~carroll/matlab_programs/

software.php), to deal with measurement error that is a mixture of classical error and Berkson error, as often occurs in radiation research and in other fields (see Section 5.1 for this topic).

5. Special topics

In Part 1 and in previous sections of this second part we have provided information regarding the effects of measurement error and misclassification on estimates obtained from some standard analyses, how to adjust for these effects and the software available to implement such adjustments. However, there is much that can be added to this basic information. In this section, we present a few selected more advanced topics. The first two topics concern situations where the data are affected by a mixture of types of errors. The third topic involves model building and variable selection in the presence of measurement error, the fourth topic involves the design and analysis of studies whose main outcome variable is measured with error, and the fifth topic involves categorization of continuous exposures that are measured with error.

5.1 Analysis of data subject to both Berkson and classical measurement error

We focus on epidemiologic projects involving an exposure that is measured by two or more methods, some of which involve Berkson errors and some classical-type measurement errors, which are then combined into a single measure. This occurs for example in radon studies,[82] and in radiation studies, such as at Hiroshima,[83] the Nevada Test Site Thyroid Disease Study,[84,85] the Hanford Thyroid Disease Study[86–88] and studies of the Chernobyl nuclear accident.[89–91] There is a similar literature in occupational epidemiology, where direct measurements of exposure are taken on individuals, but other measurements of the same exposure are “grouped”, for example, the time spent in the location of the specific exposure (e.g. in a uranium mine), with an overall estimate of exposure at that location being derived from the combined information. In environmental epidemiology, exposure to pollutants might be based partly on a spatial model of pollution in that region and the amounts of time spent by the individual in different locations within the region, and partly on some direct measurements taken from an individual.

In all these types of problems, there is then a *calculated dose* X*, a true dose X that is unobserved, and exactly measured covariates Z that are included in the outcome model and that are potentially related to X. We assume that, other than X, the covariates in the outcome model are measured exactly.

Analysis of such data is based on the statistical idea of linking the direct and indirect measurements via some version of a latent variable. Specifically, a *latent intermediate variable*, L links the true and calculated doses through a model such as

$$X = L + U_{Berk};$$
$$X^* = L + U_{Clas};$$
$$L = f(Z, \theta) + \varepsilon$$

L can be hard to interpret; in some settings, it might be useful to think of L as an average underlying dose for a given set of covariates. Here, U_{Berk} is the component of Berkson error with mean zero and variance σ^2_{Berk}, U_{Clas} is classical error with mean zero and variance σ^2_{Clas}, $f(Z,\theta)$ describes the relationship of L to the covariates Z, and ϵ is the remaining variability not explained by the covariates or the measurement errors, with mean zero and variance σ^2_{ϵ}. In the Nevada Test Site example, X= the true radiation dose, X* = the derived radiation dose, which relied on participant characteristics such as age, sex and self-reported milk consumption; and we assume X and X* are related via a latent variable L, where L may be a function of other precisely measured covariates Z determining exposure such as age, sex, distance from test site, etc. If U_{Berk} has zero variance, then the above model is a purely classical measurement error model; if U_{Clas} has zero variance then the above model is a purely Berkson model. Carroll et al[1] [pp. 193–6] provide more details on the analysis of such joint models, giving examples of the use of regression calibration (see also Reeves et al;[82] Mallick et al[84]) and maximum likelihood. In practice, knowledge of the sizes of the measurement error variances σ^2_{Berk} and σ^2_{Clas} is critical to analysis. This can be particularly difficult for Berkson errors (see Part 1, Section 4.2). In case of such difficulty, sensitivity analyses can be conducted.

The impact of a mixture of Berkson and classical errors depends critically on the ratio of σ^2_{Berk} to σ^2_{Clas}. When this ratio is very large and Berkson error dominates, the impact is close to that expected from purely Berkson error. When the ratio is small, the impact is close to that expected from classical error; and when the ratio is near one, and the Berkson error is non-differential, then the impact is an average of the impacts of both – in other words estimated regression coefficients are attenuated, although to a lesser degree than with purely classical error, and loss of power is similar to that found with either classical error or Berkson error.

The literature given above includes a host of variations on the model given above, often specific to the application. For example, Li et al[85] consider the same model as above, but give reasons to allow the Berkson errors to be correlated among groups of individuals. For the Chernobyl accident, Masiuk et al[91] argue that a better model has classical additive heteroscedastic measurement errors as well as Berkson multiplicative measurement errors.

5.2. Analysis of exposure variables subject to both measurement error and misclassification

In previous sections, we have described methods for dealing with two distinct cases: (a) a continuous variable measured with error; and (b) a categorical variable subject to misclassification. What can be done, though, if data have a combination of both?

The only work we are aware of that combines issues of measurement error and misclassification in a single analysis is that of Spiegelman et al,[92] White et al[93] and Yi et al.[94] Each of these papers considers a main study/validation study design, where the validation study includes both the true and the error-prone observations of the variables subject to measurement error and misclassification, while the much larger main study has only the

error-prone versions. Additionally, White et al[93] considers a main study/replicates study design.

White et al[93] proposed a regression calibration approach for a continuous outcome when there is both a continuous covariate and binary covariate subject to measurement error, and discuss the necessary supportive data for an identifiable model depending on whether a validation study or replicate data are available. Spiegelman et al[92] consider a binary outcome, and use logistic regression with maximum likelihood to obtain estimates and inference. Yi et al[94] consider methods applicable for all generalized linear models, with a binary covariate subject to misclassification and a continuous covariate measured with error. They discuss methods based on (i) full maximum likelihood, as in Spiegelman et al;[92] (ii) an estimating function method based on ideas of semiparametric methods such as in Tsiatis and Ma[95] and Ma and Tsiatis;[96] (iii) an augmented regression calibration method; and (iv) an augmented SIMEX method. Methods (ii)-(iv) aim at providing robustness to distributional assumptions about the error-prone continuous variable.

Of methods (ii)-(iv), the augmented SIMEX method is easiest to describe. We denote by X_{cont} and X_{cat} the continuous and categorical predictors subject to measurement error and misclassification, respectively. Their mismeasured versions are X^*_{cont} and X^*_{cat}. The procedure is based on the idea that if X_{cont} were observed, then one has a simple misclassification problem, with X_{cat} misclassified, which can be solved by misclassification methods described in Section 2.5, for example by positing a model for the misclassification distribution of X^*_{cat} given Z, X_{cont}, and X_{cat}. Then one applies ordinary SIMEX to the method that would have been used if X_{cat} had been observed.[91]

The augmented SIMEX and regression calibration methods have the advantage that they are easily implemented using regular software for SIMEX or regression calibration, once one has computer code also for solving a misclassification problem.

5.3 Variable selection when some covariates are measured with error

In many biomedical settings, one seeks to develop a parsimonious regression model from a set of candidate predictors. We saw in Part 1, Section 3, that when there is at least one covariate in a multivariable regression model that is measured with error, the estimated coefficients for that X and any other covariates can be subject to bias because of the underlying correlation structures. Furthermore, as discussed in Part 1, Sections 3.1.3 and 3.2.3, when there are multiple such error-prone or misclassified covariates in a regression model, the direction of the bias can be in either direction (towards or away from the null). One way of viewing the cause of this bias in general linear regression is that measurement error induces bias in the least squares estimating equation so that its expectation at the true parameter vector β is no longer zero. Thus, many of the statistics used in variable selection procedures, such as the deviance or p-values associated with regression coefficients, will also be biased. Consequently, measurement error in one or more covariates puts any model selection procedure at added risk of selecting an incorrect set of variables and estimating regression coefficients with bias. This is not a concern for prognostic modeling, where only the risk prediction is of interest; however, it is a concern when one wishes to interpret the

model coefficients or infer biological importance of the variables selected. Zhang et al[97] also remark that even when risk prediction is the sole interest, problems in model selection can occur when the measurement error structure in the data used to develop the prediction model is different from that in the data used for prediction.

There are many modeling procedures used to perform variable selection. Some methods have been developed to accommodate measurement error in variable selection and we highlight a few here. These methods have focused on penalized regression approaches, which use a penalty function to effectively control the model dimension. Whereas conventional stepwise procedures have been shown to be subject to instability and overfitting,[98–101] penalized regression procedures are becoming increasingly popular because of their better operating characteristics, particularly when there is a large number of candidate predictors relative to the sample size.[99,101,102] Penalized regression methods typically add a penalty to the usual parameter estimating equation (e.g. the score), which then addresses both dimension reduction and parameter estimation in a single step.

For linear and partially linear regression models, Liang and Li[103] develop a corrected score type approach in which a term, proportional to $\beta^T \mathrm{var}(U)\beta$, which offsets the bias caused by classical measurement error in the covariate vector X*, is subtracted from the estimating equation to then achieve consistent estimation. Here var(U) is the measurement error covariance matrix. To this end the authors propose minimizing the following adjusted least squares expression:

$$\frac{1}{2}\sum_{i=1}^{n}\left\{Y_i - X_i^{*T}\beta - \nu(Z_i)\right\}^2 - \frac{n}{2}\beta^T \mathrm{var}(U)\beta, \tag{10}$$

where $\nu(Z_i)$ is a general function of a precisely observed covariate Z, which is estimated with local splines. This adjusted least squares expression is incorporated into a general penalized regression framework and the authors discuss choices of the penalty, such as the L_0, L_1 or the smoothly clipped absolute deviation (SCAD) penalty, that provide a variable selection framework. The authors show that under certain conditions, asymptotically, this procedure can perform as well as if the true model were known. The method assumes that var(U) is either known or can be estimated from repeat measurements of X*. These authors also develop a similar penalized quantile regression procedure, building on the work of He and Liang[104] who had developed a quantile regression to handle covariate measurement error.

Ma and Li[105] develop a more general penalized estimating variable selection approach that can be applied to both parametric and semi-parametric measurement error models. Their method is applicable to any consistent estimating equation, including generalized linear models, and can be applied to a large class of regression models.

Currently, we are not aware of any available software to implement the methods described above; however, software for penalized regression methods is now widely available and could be used as a base for building the required software for selection of error-prone variables.

In this section, we have highlighted only a few approaches in detail. Other approaches to address variable selection in high dimensional data include Datta and Zou,[106] Loh and Wainwright,[107] Sorensen et al,[108] Yang and Xia,[109] Tian and Xue,[110] and Wang et al.[111] Zhang et al[97] discuss generally model selection in the setting linear regression models with measurement error.

5.4 Design and analysis when the outcome variable is measured with error

The effects of measurement error in outcome variables were outlined in Part 1, Section 3.3. Here, we discuss implications for study design and analysis and summarize methods for correcting for the effects of measurement error and misclassification of the type that produces biased estimates.

5.4.1 Classical error in a continuous outcome—Classical measurement error in a continuous outcome variable in a linear regression does not result in biased estimates of regression coefficients (Part 1, Section 3.3.1). Therefore, a standard linear regression analysis can be used without any alterations. However, classical error in a continuous outcome results in lower precision of estimated regression coefficients and this should be accounted for in the study design. Consider the linear regression model $Y = \beta_0 + \beta_X X + \varepsilon$, where Y denotes the error-free outcome, and the corresponding linear regression model $Y^* = \beta_0^* + \beta_X^* X + \varepsilon^*$ using the error-prone outcome Y*. The variance of the estimate is $\text{var}(\hat{\beta}_X^*) = \text{var}(\varepsilon^*)(X^T X)^{-1}$ and the relationship between variances of estimates from models using Y and Y* is $\frac{\text{var}(\hat{\beta}_X^*)}{\text{var}(\hat{\beta}_X)} = \frac{\text{var}(\varepsilon^*)}{\text{var}(\varepsilon)}$. Let n_Y denote the sample size required to achieve a desired standard error for $\hat{\beta}_X$. To achieve the same standard error for $\hat{\beta}_X^*$ therefore requires a sample size of $n_Y \frac{\text{var}(\varepsilon^*)}{\text{var}(\varepsilon)}$, where $\text{var}(\varepsilon^*) > \text{var}(\varepsilon)$.

5.4.2 Systematic error in a continuous outcome—When the outcome variable has linear measurement error (see model (2) in Part 1, Section 2.1), i.e. $Y^* = \alpha_0 + \alpha_Y Y + U$, where U has mean zero and is independent of Y, the unadjusted regression coefficients estimated using $Y^* = \beta_0^* + \beta_X^* X + \varepsilon^*$ will be biased. Consistent estimates could be obtained by using $\frac{Y^* - \alpha_0}{\alpha_Y}$ in place of Y*. This of course requires values for α_0 and α_Y, which may be known from previous studies or may need to be estimated. Buonaccorsi[112,113] and Buonaccorsi and Tosteson[114] devised methods for obtaining unbiased intervention effect estimates in this setting, when there is available either a validation study or replicates of an unbiased measure (e.g. a biomarker) in a sub-study. These methods were summarized by Carroll et al[1] (Section 15). First, consider the setting with a validation study available, in which Y (as well as X) is observed. An estimate of β_X can be obtained from the data in the validation subset, and we denote this by $\hat{\beta}_X^{(1)}$. This estimate is consistent but clearly inefficient because it is based on only a subset of the data. A second estimate, $\hat{\beta}_X^{(2)}$, can be obtained from a regression of $\frac{Y^* - \hat{\alpha}_0}{\hat{\alpha}_Y}$ on X, where $\hat{\alpha}_0$ and $\hat{\alpha}_Y$ are estimates obtained from a

regression of Y* on Y. The variance-covariance matrix for $\hat{\beta}_X^{(1)}$ and $\hat{\beta}_X^{(2)}$, denoted Σ, can be obtained using a stacked estimating approach (Carroll et al[1], Appendix A; Keogh et al[115]). Note that the variance for $\hat{\beta}_X^{(1)}$ and $\hat{\beta}_X^{(2)}$ will have increased uncertainty from the added variability in U, as well as from the uncertainty in the estimated parameters $\hat{\alpha}_0$ and $\hat{\alpha}_Y$. Alternatively, bootstrapping can be used. An efficient estimator of β_X is then given by the 'best weighted combination'

$$\hat{\beta}_X^C = (J^T\Sigma^{-1}J)^{-1}J^T\Sigma^{-1}\left(\hat{\beta}_X^{(1)T}, \hat{\beta}_X^{(2)T}\right)^T$$

where J=(I,I) and I is the identity matrix with the same number of rows as there are elements of β_X. The above result is general and extends to regressions with multiple covariates having a vector of regression coefficients β_X. In the simple setting of a single covariate X, the efficient estimator is

$$\hat{\beta}_X^C = \frac{\hat{\beta}_X^{(1)}\sigma^2_{(2)} + \hat{\beta}_X^{(2)}\sigma^2_{(1)} - \left(\hat{\beta}_X^{(1)} + \hat{\beta}_X^{(2)}\right)\sigma_{(12)}}{\sigma^2_{(1)} + \sigma^2_{(2)} - 2\sigma_{(12)}},$$

where $\Sigma = \begin{pmatrix} \sigma^2_{(1)} & \sigma_{(12)} \\ \sigma_{(12)} & \sigma^2_{(2)} \end{pmatrix}$ is the variance-covariance matrix. This approach, based on a weighted combination of estimates, extends to the setting in which a replicates study is available instead of a validation study (see Part 1, Section 4.2). Suppose that two repeats of an unbiased measurement (e.g. biomarkers) $Y_1^{**} = Y + U_1$ and $Y_2^{**} = Y + U_2$, with U_1 and U_2 independent, are available for a subset of individuals. A consistent estimate of β_X, again denoted $\hat{\beta}_X^{(1)}$, can be estimated in the replicates sub-study from a regression of $(Y_1^{**} + Y_2^{**})/2$ on X. The parameters α_0 and α_Y can be estimated in the replicates sub-study using a method of moments approach (Carroll et al[1], Section 15).

If the values of parameters α_0 and α_Y are not known from a previous study, the need to estimate them should be accommodated at the design stage by incorporating plans and resources to conduct a validation or replicates sub-study. Further research is needed to establish methods for optimal design of such studies, including the incorporation of information on the relative cost of the systematic-error-prone and biomarker measures.

5.4.3 Differential error in a continuous outcome—In some studies, the outcome measure is prone to differential error. This can arise in intervention studies with a self-reported outcome when participants are aware of their intervention group. We consider a measurement error model of the form $Y^* = \alpha_{0X} + \alpha_{YX}Y + U$ for two groups X = 0 and 1. This is a generalization of the linear measurement error model considered above. Differential error gives rise to biased estimates of the intervention effect; additional information is needed to estimate the form of the differential error so as to obtain consistent estimates of the intervention effect. Keogh et al[115] described methods for analysis in the

setting of dietary intervention trials in which the main differential-error-prone outcome measure is from a self-report and unbiased biomarkers are available in a replicates sub-study, based on the Buonaccorsi[112] approach outlined above. The Buonaccorsi method extends directly to the differential error setting, with $\hat{\beta}_X^{(2)}$ being based on a regression of $\frac{Y^* - \hat{\alpha}_{0X}}{\hat{\alpha}_{YX}}$ on X. In particular, Keogh et al[115] investigated the contribution of $\hat{\beta}_X^{(2)}$ to the estimator $\hat{\beta}_X^C$ (the weighted combination of $\hat{\beta}_X^{(1)}$ and $\hat{\beta}_X^{(2)}$) under different assumptions. It was shown theoretically that in the case of non-differential error the combined estimator will be more precise than $\hat{\beta}_X^{(1)}$, while in the case of differential error nearly all the information about the intervention effect comes from the validation or replicates study and that $\hat{\beta}_X^{(2)}$ adds little in large samples. However, via simulation studies Keogh et al[115] found that in finite samples, it is advantageous to use the self-report data in addition to the replicate biomarkers to estimate the intervention effect when the reliability of self-report measurements is comparable to that of the biomarker.

5.4.4 Misclassification of a binary outcome—Section 2.5 noted that "matrix methods" can be used for handling a misclassified binary exposure when the outcome is also binary. Matrix methods can also be applied directly when instead it is the outcome that is misclassified, but they work only in very simple settings. Binary outcomes are more typically analyzed using logistic regression, and methods for correcting the impact of outcome misclassification in logistic regression analysis have also been devised.

For a study of *n* individuals, the full likelihood can be written:

$$L = \prod_{i=1}^{n} \Pr(Y^* = y_i^* | X = x_i) = \prod_{i=1}^{n} \sum_{y=0}^{1} \Pr(Y = y | X = x_i) \Pr(Y^* = y_i^* | Y = y, X = x_i), \quad (11)$$

where the logistic model of interest is $\text{logit}(\Pr(Y = 1|X = x_i)) = \beta_0 + \beta_X x_i$. The misclassification probabilities in the second term of the likelihood, $\Pr(Y^* = y_i^* | Y = y, X = x_i)$, can be expressed in terms of $\Pr(Y = y|X = x_i)$ and the sensitivity (*Sn*) and specificity (*Sp*) of Y*. Magder and Hughes[116] described estimation of β_0 and β_X using an EM algorithm. They described approaches for when the sensitivity and specificity are known, when they can be estimated from a previous validation study, and when they need to be estimated from the data. In the last situation, there are issues of identifiability if the model is saturated, and smoothing assumptions concerning the relation between covariates and outcome are needed to proceed. The authors caution against using this approach unless the smoothing assumptions are strongly believed. The probabilities of Y=y given Y* and X are estimated in the E-step, and the logistic regression parameters β_0 and β_X are estimated in the M-step. Neuhaus,[117] Lyles and Lin [2010],[118] and Lyles et al[119] instead used direct maximum likelihood estimation based on (11). They used the result that for the case of non-differential misclassification, the likelihood can be expressed in terms of *Sn* and *Sp*, as follows:

$$\mathrm{L} = \prod_{i=1}^{n} \begin{array}{l} \{(1-Sp)\Pr(\mathrm{Y}=0|\mathrm{X}=x_i) + Sn\,\Pr(\mathrm{Y}=1|\mathrm{X}=x_i)\}^{y_i^*} \times \\ \{Sp\Pr(\mathrm{Y}=0|\mathrm{X}=x_i) + (1-Sn)\Pr(\mathrm{Y}=1|\mathrm{X}=x_i)\}^{1-y_i^*} \end{array} \quad (12)$$

They considered sensitivity analyses assuming particular values for *Sn* and *Sp*, and making use of internal or external validation data. *SAS* code was provided.

A number of other authors have also considered sensitivity analyses for investigating the impact of outcome misclassification, incorporating uncertainty in the specified values for sensitivities and specificities. Fox et al[120] described probabilistic sensitivity analyses that involve simulating the data that would have been observed if there were no misclassification, given sensitivities and specificities. They focused on misclassified exposures but noted that the methods could also be applied for a misclassified outcome. Lyles and Lin[118] described a 'predictive value weighting' for handling misclassified exposures, which can also be applied for misclassified outcomes, and in the more complex scenario of misclassification in both outcome and exposures. This has been implemented in the *pvw* module in *Stata*.[121]

Edwards et al[122] applied multiple imputation to handle misclassified outcome data when there is an internal validation study. A Bayesian approach can also be taken by assigning priors to the sensitivity and specificity. Some special considerations are needed when Y represents case or control status in a case-control study.[116,119,123]

5.5 Misclassification due to categorizing continuous exposures measured with error

We have presented in Sections 2.2 and 3.2 of Part 1, and in Section 2.5 of this second part, problems arising from and methods for dealing with misclassified categorical variables. In this section, we discuss the special case where the categorical variable has been formed by categorizing an observed continuous variable. Despite the resulting loss of information, in epidemiologic analyses, continuous exposure variables are often categorized using either pre-specified cut-points or estimated quantiles of the variable's distribution. Flegal et al[124] published a key result showing that dichotomization of a continuous exposure that is subject to non-differential measurement error leads to a binary exposure that has differential misclassification. Later work of Brenner and Blettner[125] and Delpizzo and Borghes[126] also stress this point. Although differential measurement error and misclassification may, in general, lead to bias in the estimated regression coefficient in any direction (see Part 1, Section 3), the simulations of Flegal et al[124] demonstrated relative risk estimates that were attenuated. These simulations were based on a univariate linear logistic regression with a continuous exposure X prone to classical measurement error and dichotomization using a pre-specified cut-point.

Considering the same assumptions regarding the exposure X, Gustafson and Le[127] extended the results of Flegal et al. First, they provided analytic expressions for linear regression in addition to numerical results for linear logistic regression. Second, they considered the effect of changing the pre-specified cut-point c. Third, they considered the inclusion of a second precisely measured continuous covariate Z in the regression and the effect of correlation ρ between X and Z. Finally, they considered situations where the true regression of the

outcome Y was linear in Z and a weighted average of the continuous and dichotomized X, i.e.,

$$E(Y|X,Z) = \beta_0 + \beta_1\{(1-\omega)X + \omega I(X > c)\} + \beta_2 Z. \quad (13)$$

This form of regression allowed a more general investigation of the effects of covariate dichotomization, i.e., it considers the situation where the truth lies somewhere `ìn between" the extremes of dichotomization leading to a completely right model specification versus a completely wrong model specification. Their results demonstrated that, when the true regression contains a linear exposure on the continuous scale ($\omega<1$), it can be beneficial to dichotomize imprecise continuous exposures, as this can reduce bias from the analysis on the continuous scale. For example, in the case of the linear regression (with $\omega=0$),

$$E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z, \quad (14)$$

where X~N(0,1), X*~N(0,1 + var(U)), Z~N(0,1), and ρ = cor(X,Z), the multiplicative biases in the estimated regression coefficients arising in the analysis with continuous X* and categorized W* = I(X* > c) are given by, respectively,

$$\frac{\beta_{1X^*}}{\beta_1} = \frac{1}{1+\text{var}(U)/(1-\rho^2)} \text{and } \frac{\beta_{1W^*}}{\beta_{1W}} = \theta\left\{\frac{R(c)-\rho^2\phi(c)}{R(\theta c)-\rho^2\theta\phi(\theta c)}\right\}, \quad (15)$$

where $\theta = \frac{1}{\sqrt{1+\text{var}(U)}}$, $R(c) = \frac{\Phi(c)\{1-\Phi(c)\}}{\phi(c)}$ and $\phi(.),\Phi(.)$ are the probability density and cumulative probability functions of the standard normal distribution. Some values of var(U), ρ and c lead to $\frac{\beta_{1W^*}}{\beta_{1W}} > \frac{\beta_{1X^*}}{\beta_1}$. Gustafson and Le[127] point out that no general statements can be made about how the bias due to dichotomization depends on the choice of threshold or the strength of correlation between predictors. The authors produce several graphs of attenuations comparing continuous and dichotomized observed main exposure for $0 < \text{var}(U)^{1/2} < 1.2$, c = 1,2,3, and ρ = 0,0.3,0.6,0.9. In all those cases, attenuation was greater in the continuous case, i.e. $\frac{\beta_{1W^*}}{\beta_{1W}} > \frac{\beta_{1X^*}}{\beta_1}$. However, this inequality also depends on the nature of the underlying relationship between the outcome variable and the exposure. In situations where the dichotomized true exposure ($\omega=1$) produces a better fitting model, the bias in its regression coefficient tends to be larger than that in the coefficient of the continuous exposure, unless var(U) gets close to or exceeds the var(X). In general, it is important to remember that dichotomizing a continuous exposure loses information when the relationship between the outcome variable and the exposure is truly continuous, and changes the interpretation of the corresponding regression coefficient. Thus, even in situations where dichotomization does reduce bias in the estimated regression coefficient, this advantage could be outweighed by loss of information and/or degradation of model fit.

Recently Keogh et al[128] investigated several methods of adjusting for misclassification due to dichotomizing a continuous error-prone predictor. In contrast to previous work, a measured continuous exposure was specified to follow a linear measurement error model,

thereby allowing for systematic error. In addition to methods using estimated misclassification probabilities, the authors considered applying two regression calibration (RC) based methods, multiple imputation (MI) and moment reconstruction (MR) to the continuous exposure followed by dichotomization, and also a SIMEX method. Simulation studies were used to compare the methods when either the true exposure or reference measurements with classical error were available in a validation subsample. The underlying relationship between the continuous exposure and the outcome in the simulations was a univariate linear logistic regression, and dichotomization was based on a pre-defined cutpoint. In that study, regression calibration and SIMEX methods failed to correct adequately for bias because both methods assume non-differential error (the failure of regression calibration was also confirmed by Dalen et al[129]). However, MI and MR performed well. Methods using estimated misclassification probabilities also performed well, provided differential misclassification was assumed (see also Dalen et al[130]). It is important to note that the latter methods are restricted to estimating odds ratios, while MI and MR could, in principle, be used with different regression models, with quantile-based categorization, and could also accommodate covariate adjustment. Extending MI and MR to those cases as well as to the case of regression of the outcome on a non-linear function of the exposure remains an important area for further research.

6. External, imperfect or missing reference instruments

While Part 1 and earlier sections of this paper have made it clear that both theory and software are available for handling errors in measurement or classification of variables, ultimately their use hinges on the knowledge and data available regarding the measurement error or misclassification model that relates the imperfect observed exposure X* to the true value X. Unfortunately, all too often, information about the error model is incomplete or missing. In this section, we describe how one might approach the problem of measurement error or misclassification when there is lack of knowledge about such a model.

The ideal setting for applying methods to address measurement error or misclassification of variables is one in which an internal validation study has been done, which would directly relate the imperfect observed exposure or outcome to its true value in the population of interest. We consider analysis options when such a validation study is not available. In this case, there may be data from another cohort, namely an external validation study, or there may be an imperfect reference instrument that can provide partial information about the nature of the measurement error. We also discuss approaches for settings where there is little to no information available regarding the error-prone data.

6.1 Using an external validation study

Internal validation studies are the ideal because their data can be used directly with all the methods of measurement error analysis described previously. In their absence, external studies can be used to specify the measurement error model (or its associated regression calibration model) in the external data and to estimate its parameters. However, the use of this external study data in the analysis of the primary study is then reliant on the assumption of transportability for the model of X*|X,Z (see also Part 1, Section 4.2). Transportability

means that the specified model relating the error prone X* with the true data (X,Z) holds with the same parameter values in both studies, and that the relevant parameter estimates and their standard errors obtained in the external study can be used without bias in analysis of the primary study. See, for example, Guo et al[131] who provide a multiple imputation method that addresses covariate measurement error in a regression model using information from an external validation study that provided information regarding only the covariate error, without measurement of the outcome or other study variables. Further discussion of this article appears in Liao et al.[132] Buonaccorsi[2] also provides some discussion of error correction methods that rely on external data. In this case, also assuming non-differential error, one can use the external data to inform the regression calibration approach, as described in Part 1, Section 6.1; however, the calculation of the standard errors would need to be different, particularly if the original external data were not available.

The assumption of transportability may be reasonable if external independent data come from a similar population with measurements obtained with the same or a very similar instrument. Consider, for example, a nutritional study with dietary intake measured by a food frequency questionnaire (FFQ). If the external validation study provides independent data from the same population with the same FFQ plus a reference instrument, the error model that is estimated in the validation study could be assumed transportable to the primary study and used for adjusting its results for measurement error. If, however, the same FFQ is used in a somewhat different population, or a different version of FFQ is used in that population, the transportability assumption may not be fully justified.

A different example of possible problems with transportability includes the situation when the distribution of X* given (X, Z), the measurement error model, is the same in both primary and external studies, but the distributions of true exposure X given Z are different. Since by Bayes' theorem the regression of X given (X*, Z) depends on both of these distributions, the regression calibration model is not transportable. Carroll et al[1] (Section 2.2.5) give an example of this phenomenon related to blood pressure measurement.

Since adopting the correct error model is critical for an appropriate adjustment for measurement error, whenever there is doubt about the transportability of an external validation study, it is advisable to conduct a sensitivity analysis by considering some possible variation in the relevant error parameters and their effect on the results of the primary study. Thus, while external validation studies can undoubtedly provide worthwhile information about the measurement error, they often do not entirely exempt the investigator from conducting a sensitivity analysis. However, in comparison to the situations in Sections 6.2 and 6.3 that follow, where there is less information regarding the measurement error, in the case of external validation studies, the sensitivity analysis could involve a more restricted range of parameter values.

6.2 Methods that use an imperfect reference instrument

Often, exposures in epidemiological studies are known to be measured with substantial error, but the corresponding measurement error model is not known due to absence of appropriate reference instruments. Typical examples include most dietary exposures (see Part 1, Section 4.3.1) and characteristics of physical activity such as measures of moderate to vigorous

activity (see Part 1, Section 4.3.2). In some such cases, an instrument less biased than the main study instrument, but nevertheless biased, is used as the reference instrument (calibration study with imperfect reference instrument). Examples in nutrition and physical activity studies are given in Part 1, Sections 4.3.1 and 4.3.2.

As mentioned later in Section 6.3, in the absence of knowledge of the measurement error model, a bias or sensitivity analysis is recommended using a plausible set of parameters (or their distribution) for the model. When the measurement error model is estimated using an imperfect reference, sensitivity analysis is also recommended. Although the error model parameters are imperfectly estimated, they may nevertheless be used together with supplementary information to choose the range of parameters for the sensitivity analysis. Thiébaut et al[133] provided a good example of this approach. They reported the estimated relative risk (RR) for breast cancer associated with a two-fold increase in fat density (percent of total energy provided by fat). In the Results section, they first report the unadjusted RR estimate of 1.15 (95% CI 1.05–1.26) based on food frequency questionnaire data. They then report the RR adjusted for measurement error based on a 24-hour recall (imperfect) reference validation study: 1.32 (95% CI 1.11–1.58). Finally, in the Discussion section, they use data from the OPEN validation study[7] to adjust for the bias that may have occurred due to use of an imperfect reference instrument for fat density. The adjustment for bias is calculated by comparing attenuation factors based on the imperfect reference (24-hour recall) with that based on the perfect reference (recovery biomarkers) for protein density that has a recovery biomarker, and transporting the ratios of these estimates to the case of fat density, which does not have a recovery biomarker. This procedure, which was justified by the substantial correlation between protein and fat intake, gave an estimated RR of 1.46 (no confidence limits for this estimate were provided). Although they did not perform a formal sensitivity analysis in the last analysis, it is clear that their approach was moving in that direction.

This example includes the elements of how a sensitivity analysis may be constructed from knowledge of measurement error in exposures similar to that being considered. For exposure X (e.g. fat intake), the measured exposure X* (using instrument $I_{Q,}$ e.g. a food frequency questionnaire) is compared to an imperfect reference instrument X^{**}_{Imp} (using instrument I_R, e.g. a 24-hour recall) to obtain an estimated measurement error model M*. For some similar exposure X_1 (e.g., protein intake), information is available on the measured exposure X_1* (using instrument I_Q, the food frequency questionnaire), its imperfect reference instrument $X^{**}_{1,Imp}$ (using instrument I_R, the 24-hour recall), and also an unbiased reference measurement X^{**}_1 (e.g., 24-hour urinary nitrogen excretion). The availability of both the imperfect reference measurement and an unbiased reference measurement for X_1 enables one to learn about the relationship between the true measurement error model M_1 estimated using measurement X^{**}_1 and the model M_1* estimated using the imperfect reference measure $X^{**}_{1,Imp}$. This information about M_1 versus M_1* is then applied to the estimated measurement error model M* for the exposure of real interest X, to yield the desired range of parameters for the true measurement error model M for measured exposure X*.

In this approach, the choice of the "similar" exposure X_1 will, of course, depend on the context. For dietary intakes, it will be the intake of another dietary component, one that has an unbiased reference measurement; for physical activity measures such as moderate or vigorous activity measured by a physical activity diary and compared to an accelerometer reference, it could be total energy expenditure that can be measured unbiasedly by doubly labeled water.

A different approach to dealing with studies using imperfect reference instruments starts with the question of whether using an imperfect reference instrument to adjust for measurement error is preferable to making no adjustment whatsoever. In other words, if one uses an imperfect reference to estimate the measurement error model and then uses this model to adjust risk parameter estimates in the health outcome model, would these adjusted estimates, even if biased, still have less bias than unadjusted ones? If that could be demonstrated, it could motivate the use of these imperfectly adjusted estimates in preference to the unadjusted ones. This approach is less demanding than conducting a sensitivity analysis, since it involves applying the measurement error adjustment for just one measurement error model, but it is also less complete.

The issue has been studied in nutritional epidemiology. Freedman et al[134] published the results of analyzing such a question using data from the OPEN study, and more recently updated their results using data from the five validation studies included in the Validation Studies Pooling Project.[135] They concluded that, on average, 24-hour recall-based calibration of a food frequency questionnaire reduced, but did not eliminate, the bias in the risk estimates in multivariate risk models that included energy, and protein, potassium and sodium intake densities, in comparison with unadjusted estimates. Although those results, as well as similar results using linear measurement error models in the sensitivity analysis conducted by Buonaccorsi et al,[136] indicate that using a 24-hour recall as a reference instrument to adjust for measurement error would improve the analysis of studies in nutritional epidemiology, there remain some doubts. The improvement has been demonstrated in only a handful of nutrients (those which have unbiased biomarkers), and may not transfer to all other dietary components, especially episodically consumed dietary components, for which the measurement error model is highly non-linear.[137]

A general limitation of this approach is that even if the resulting estimates are less biased than unadjusted estimates, they are nevertheless biased. Therefore, presenting them as the best estimates available does not reveal the full extent of the underlying uncertainty, and is a less complete approach than conducting a sensitivity analysis.

6.3 Approaches when there is no reference instrument

If we have no knowledge about the measurement error model, then we have to make assumptions about it. Note that ignoring measurement error is one (incorrect) assumption, which is akin to assuming there was no error in measurement. In this section, we outline alternatives to this naïve approach. Investigations can still be undertaken to understand the potential impact of measurement error or misclassification on study results. This approach involves three steps. First, a measurement error model is posited. Second, study results are produced that are corrected for measurement error, under the assumed measurement error

model. This may involve direct reanalysis of the data, or post-hoc adjustment of estimated outcome model parameters. Finally, assumptions about the parameters in the assumed error model are typically varied in a sensitivity analysis to examine the robustness of study results to a range of assumptions about the measurement error.

Bias analysis, sometimes also referred to as uncertainty analysis or probabilistic sensitivity analysis, follows the above general approach for quantifying the potential effects of measurement error. This method focuses on sources of systematic and random error. Bias analyses have several goals: i) to estimate the direction and magnitude of the bias in study results induced by the errors in the data, ii) to make explicit the sources of suspected errors and the degree of uncertainty that they introduce into study results, and iii) to efficiently guide future research by elucidating what associations are sensitive to the underlying amount of measurement error or misclassification and would best benefit from future replicates or validation studies to estimate that error. [138]

Methods for bias analyses are well established in the statistical and epidemiologic literature and apply to estimating the impact of sources of bias that go beyond just measurement error, such as unmeasured confounding or non-ignorable dropout [Lash et al[138], Greenland et al, [139] Greenland,[140] Fox and Lash,[141] Fox 2009,[142] MacLehose et al,[143] Lash and Ahern[144]]. A fundamental step of bias analysis is to thoroughly review the study's subject selection and retention, methods of data collection, and other opportunities for confounding, selection bias and measurement error.[138] Once those potential sources of bias have been identified, mathematical models are developed for the relationship between the underlying true data with biases removed and the study data. For this endeavor, distributions rather than a single set of parameter values are used to generate a sensitivity analysis for the results of the bias analysis. In the absence of any validation data or other studies to inform the selection of parameters, educated guesses can be used to posit such relationships.[142] In this last step, one option is to assign a prior distribution from which to draw the necessary error parameters, which allows for a Bayesian analysis that naturally integrates the uncertainties coming from the sub-models for exposure, outcome, and measurement. Choice of this prior in the absence of validation studies could similarly be informed by expert option, as discussed in Section 2.2. Lash et al[138] provide a review of best practices for bias analysis. One challenge to this approach is its reliance on proper specification of the mathematical form of the measurement error, such as additive or multiplicative. This choice is likely best informed by validation data but could also be made part of the sensitivity analysis in the absence of such data.

A practical example of bias analysis can be seen in the study by Jurek et al,[145] who sought to quantify the impact that exposure misclassification may have had on a study by Ross et al[146] reporting on the effect of maternal supplement use on the risk of leukemia in children with Down syndrome. Because of a lack of an internal validation study, Jurek et al[145] developed their misclassification models and parameter distributions from a mixture of expert opinion, a literature review of validation studies of similar exposure instruments, and limitations set by the data themselves. Using several error model scenarios, including both differential and non-differential misclassification and a formula to adjust the estimated odds ratio for the underlying exposure misclassification, they conducted a sensitivity analysis for the induced bias. Their bias analyses revealed that data that were corrected for the reporting

bias in supplement use generally yielded a stronger protective effect than the naïve analysis that ignored the misclassification. The uncertainty was increased in all scenarios.

He et al[147] provided an alternative approach for examining the potential effects of measurement error. These authors considered an accelerated failure time model for mortality in the Bussleton Health Study cohort that included two error-prone covariates, serum cholesterol and systolic blood pressure (SBP), as well as other assumed precise covariates. The errors in cholesterol and blood pressure were assumed to be independent and to follow the classical measurement error. Lacking a replicates sub-study, the authors considered several possibilities for the size of the underlying measurement error variance and applied a SIMEX approach to re-estimate the regression parameters for each value of the assumed measurement error variance. With this exercise, the authors were able to conclude that even under small to moderate classical measurement error, the factors determining mortality remained the same, with the most uncertainty about the magnitude of the effect of SBP. Such analyses motivate future replicates studies to gather multiple measures of SBP in similar settings to better understand the magnitude of the measurement error variance and the relationship of SBP to mortality. The decreased sensitivity of results to the measurement error in cholesterol suggest that for this exposure such studies may be of secondary importance to that for SBP.

7. Conclusion

In this two-part tutorial, we have presented basic information needed to understand the impact of measurement error and misclassification on results of epidemiological research studies, and methods available to adjust for such error. In Part 2, we have also presented some more advanced methods to address covariate and outcome measurement error, but our review is not exhaustive. Some notable methods not considered here include conditional score and corrected score methods to address covariate error. For density estimation, there is also only minimal detail regarding deconvolution kernel estimators and other nonparametric density estimation methods. The reader is referred to some recent textbooks (Carroll et al[1], Buonaccorsi[2], Yi[3], Gustafson[13]) for introductions to these and other methods not considered.

Our impression is that the problem of measurement error and misclassification is being seriously neglected in the design of many epidemiologic studies and in the presentation of their results.[148,149] Barriers to satisfactory handling of such problems include lack of validation studies required to quantify the amount and type of error, lack of appreciation and understanding of the effects of such error, and lack of knowledge of the methods and software required to adjust for these effects. Publication of this paper is part of a wider effort by our STRATOS Topic Group to bring these problems to the attention of the biostatistical and epidemiologic communities; and on a broader perspective, our work on publishing this guidance paper is part of the general aim of STRATOS to strengthen the analytic thinking underlying observational studies.[150]

Acknowledgements

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

References

1. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: A Modern Perspective. Second Edition. Chapman and Hall Boca Raton FL 2006.
2. Buonaccorsi JP. Measurement error: models, methods, and applications. CRC Press, Boca Raton, FL 2010.
3. Yi GY. Statistical Analysis with Measurement Error or Misclassification. Springer New York, NY 2017.
4. Little RJA, Rubin DB. Statistical analysis with missing data. Second edition. John Wiley & Sons, Inc Hoboken NJ, 2014.
5. Tanner MA. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Second edition. Springer–Verlag, New York, NY 1993.
6. Geyer CJ, Thompson EA. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). J Roy Stat Soc, Series B 1992; 54: 657–700.
7. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, Sunshine J, Schatzkin A. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. Am J Epidemiol 2003; 158:1–13. [PubMed: 12835280]
8. Selected OPEN Data. https://epi.grants.cancer.gov/past-initiatives/open/ (last accessed September 5, 2019).
9. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990; 85:398–409.
10. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. Stat Med 1993; 12:1703–1722. [PubMed: 8248663]
11. Schmid CH, Rosner B. A Bayesian approach to logistic regression models having measurement error following a mixture distribution. Stat Med 1993; 12:1141–1153. [PubMed: 8210818]
12. Dellaportas P, Stephens DA. Bayesian analysis of errors-in-variables regression models. Biometrics 1995; 51:1085–1095.
13. Gustafson P Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. CRC Press, Boca Raton FL 2004.
14. Bartlett JW, Keogh RH. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Stat Methods Med Res 2016; ISSN 0962–2802 doi: 10.1177/0962280216667764.
15. Sarkar A, Mallick BK, Staudenmayer J, Pati D, Carroll RJ. Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. J Comput Graph Stat 2014a; 23:1101–1125. [PubMed: 25378893]
16. Sarkar A, Mallick BK, Carroll RJ. Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors. Biometrics 2014b; 70:823–834. [PubMed: 24965117]
17. Sinha S, Wang S. Semiparametric Bayesian analysis of censored linear regression with errors-in-covariates. Stat Methods Med Res 2017; 26:1389–1415. [PubMed: 25882297]
18. Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. Biometrics 2004; 60:172–181. [PubMed: 15032787]

19. Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. Stat Med 2008; 27:5195–5216. [PubMed: 18680172]
20. Thomas L, Stefanski L, Davidian M. A moment-adjusted imputation method for measurement error models. Biometrics 2011; 67:1461–1470. [PubMed: 21385161]
21. Thomas L, Stefanski LA, Davidian M. Moment adjusted imputation for multivariate measurement error data with applications to logistic regression. Comput Stat Data An 2013; 67:15–24.
22. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol 2006; 35:1074–1081. [PubMed: 16709616]
23. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. Stat Med 2008; 27:6332–6350. [PubMed: 18937275]
24. Keogh R and White I. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Statistics in Medicine 2014; 33:2137–2155. [PubMed: 24497385]
25. Gray C Use of the Bayesian family of methods to correct for effects of exposure measurement error in polynomial regression models. PhD thesis, London School of Hygiene & Tropical Medicine, 2018 DOI: 10.17037/PUBS.04649757.
26. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model, Stat Methods Med Res 2015; 24: 462–487. [PubMed: 24525487]
27. Bartlett JW and Keogh RH smcfcs: Multiple imputation of covariates by substantive model compatible fully conditional specification. https://CRAN.R-project.org/package=smcfcs. Last accessed August 25, 2019.
28. Shepherd BE, Shaw PA, Dodd LE. Using audit information to adjust parameter estimates for data errors in clinical trials. Clin Trials 2012; 9:721–729. [PubMed: 22848072]
29. Bang H, Chiu YL, Kaufman JS, et al. Bias Correction Methods for Misclassified Covariates in the Cox Model: comparison of five correction methods by simulation and data analysis. J Stat Theory Prac, 2013; 7: 381–400.
30. Shaw PA, He J, and Shepherd BE. Regression calibration to correct correlated errors in outcome and exposure. ArXiv:1811.10147, November 2018.
31. Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics 1977; 33:414–418. [PubMed: 884199]
32. Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. J Clin Epidemiol 1990; 43:941–947. [PubMed: 2213082]
33. Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics 1999; 55:338–344. [PubMed: 11318185]
34. Lyles RH. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. Biometrics 2002; 58:1034–1036. [PubMed: 12495160]
35. Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. J Stat Plan Infer 2008; 138: 528–538.
36. Kosinski AS, Flanders WD. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach. Stat Med 1999; 18: 2795–2808. [PubMed: 10521867]
37. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 1995; 141:263–272. [PubMed: 7840100]
38. Gustafson P, Le ND, Saskin R. Case–control analysis with partial knowledge of exposure misclassification probabilities. Biometrics 2001; 57:598–609. [PubMed: 11414590]
39. Johnson WO, Gastwirth JL, Pearson LM. Screening without a “gold standard”: the Hui-Walter paradigm revisited. Am J Epidemiol 2001; 153:921–924. [PubMed: 11323324]
40. Prescott GJ, Garthwaite PH. A simple Bayesian analysis of misclassified binary data with a validation substudy. Biometrics 2002; 58:454–458. [PubMed: 12071421]
41. Küchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: The misclassification SIMEX. Biometrics 2006 3;62(1):85–96. [PubMed: 16542233]

42. Küchenhoff H, Lederer W, Lesaffre E. Asymptotic variance estimation for the misclassification SIMEX. Comput Stat Data An 2007; 51:6197–6211.
43. Kraus JF, Greenland S, Bulterys M. Risk factors for sudden infant death syndrome in the US Collaborative Perinatal Project. Int J Epidemiol 1989;18:113–120. [PubMed: 2722353]
44. Greenland S Variance estimation for epidemiologic effect estimates under misclassification. Statistics in Medicine 1988; 7:745–57. [PubMed: 3043623]
45. Chu R, Gustafson P, Le N. Bayesian adjustment for exposure misclassification in case–control studies. Stat Med 2010; 29:994–1003. [PubMed: 20087839]
46. Beaton GH, Milner J, Corey P, McGuire V, Cousins M, Stewart E, de Ramos M, Hewitt D, Grambsch PV, Kassim N, Little JA. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Am J Clin Nutr 1979; 32:2546–2559. [PubMed: 506977]
47. Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, Tooze JA, Krebs-Smith SM. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. J Am Diet Assoc 2006; 106:1640–1650. [PubMed: 17000197]
48. Yanetz R, Kipnis V, Carroll RJ, Dodd KW, Subar AF, Schatzkin A, Freedman LS. Using biomarker data to adjust estimates of the distribution of usual intakes for misreporting: application to energy intake in the US population. J Am Diet Assoc 2008; 108:455–464; erratum in: J Am Diet Assoc 2008; 108:890.
49. National Research Council. Nutrient Adequacy: Assessment using food consumption surveys. National Academy Press, Washington, D.C 1986.
50. Nusser SM, Carriquiry AL, Dodd KW and Fuller WA. A semiparametric transformation approach to estimating usual daily intake distributions. J Am Stat Assoc 1996a; 91:1440–1449.
51. Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF, Guenther PM, Carroll RJ, Kipnis V. A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. J Am Diet Assoc 2006; 106:1575–1587. [PubMed: 17000190]
52. Tooze JA, Kipnis V, Buckman DW, Carroll RJ, Freedman LS, Guenther PM, Krebs-Smith SM, Subar AF, Dodd KW. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. Stat Med 2010; 29:2857–2868. [PubMed: 20862656]
53. Dekkers Al, Verkaik-Kloosterman J, van Rossum CT, Ocke MC. SPADE, a new statistical program to estimate habitual dietary intake from multiple food sources and dietary supplements. J Nutr 2014; 144:2083–2091. [PubMed: 25320187]
54. Haubrock J, Nöthlings U, Volatier JL, Dekkers A, Ocké M, Harttig U, Illner AK, Knuppel S, Andersen LF, Boeing H, European Food Consumption Validation Consortium. Estimating usual food intake distributions by using the Multiple Source Method in the EPIC-Potsdam Calibration Study. J Nutr 2011; 141:914–920. [PubMed: 21430241]
55. Dodd KW. Estimating usual intake distributions for dietary components consumed daily by nearly all persons. Measurement Error Webinar Series, Webinar 2 https://epi.grants.cancer.gov/events/measurement-error/mews_webinar2_6_slides.pdf; 2011.
56. Nusser SM, Fuller WA, Guenther PM. Estimation of usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data In: Survey Measurement and Process Quality. Trewin D (ed). Wiley, New York NY; 1996b: 689–709.
57. Gibney MJ, van der Voet H. Introduction to the Monte Carlo project and the approach to the validation of probabilistic models of dietary exposure to selected food chemicals. Food Addit Contam 2003; 20(Suppl 1):S1–S7. [PubMed: 14555353]
58. Freedman LS, Guenther PM, Dodd KW, Krebs-Smith SM, Midthune D. A population's distribution of Healthy Eating Index-2005 component scores can be estimated when more than one 24-hour recall is available. J Nutr 2010; 140:1529–1534. [PubMed: 20573940]
59. Zhang S, Midthune D, Guenther PM, Krebs-Smith SM, Kipnis V, Dodd KW, Buckman DW, Tooze JA, Freedman L, Carroll RJ. A new multivariate measurement error model with zero-inflated dietary data and its application to dietary assessment. Ann Appl Stat 2011; 5:1456–1487. [PubMed: 21804910]

60. Carroll RJ and Hall P Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Association 1988; 83(404): 1184–1186.
61. Stefanski LA and Carroll RJ Deconvoluting kernel density estimators. Statistics 1990; 21: 169–184.
62. Fan J On the optimal rates of convergence for nonparametric deconvolution problems. Annals of Statistics 1991;19(3):1257–1272.
63. Delaigle A and Meister A Density estimation with heteroscedastic error. Bernoulli 2008; 14(2):562–79.
64. Staudenmayer J, Ruppert D and Buonaccorsi JP Density estimation in the presence of heteroscedastic measurement error. Journal of the American Statistical Association 2008; 103 726–736.
65. Masry E Multivariate probability density deconvolution for stationary random processes, IEEE Trans. Inform. Theory IT-37 1991; 37(4): 1105–1115.
66. Sarkar A, Pati D, Chakraborty A, Mallick BK and Carroll RJ Bayesian semiparametric multivariate density deconvolution. Journal of the American Statistical Association 2018; 113(521): 401–416. [PubMed: 30078920]
67. Skrondal A and Rabe-Hesketh S Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models, Section 6.4. Chapman & Hall, 2004.
68. Stefanski LA and Carroll RJ. Covariate measurement error in logistic regression. Annals of Statistics 1985;13: 1335–1351.
69. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000; 10:325–337.
70. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Stat Med 2009; 28:3049–3067. [PubMed: 19630097]
71. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book: A practical introduction to Bayesian analysis. CRC Press, Boca Raton FL 2012.
72. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software 2005; 12(3):1–16.
73. Plummer M rjags: Bayesian Graphical Models using MCMC. R package version 4–6. https://CRAN.R-project.org/package=rjags. 2016.
74. Bartlett JW and Keogh RH Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Statistical Methods in Medical Research 2018; 27(6):1695–1708. [PubMed: 27647812]
75. Stan Development Team. (2016a). Stan modeling language users guide and reference manual, version 2.14.0 [Computer software manual]. Retrieved from http://mc-stan.org/
76. Stan Development Team (2016b). RStan: the R interface to Stan. R package version 2.14.1. http://mc-stan.org/
77. Van Buuren S, Groothuis-Oudshoom K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011; 45(3).
78. Bartlett JW, Morris TP. Multiple imputation of covariates by substantive model compatible fully conditional specification, Stata J 2015; 15: 437–456.
79. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalized linear models with covariate measurement error. Stata J 2003; 3:386–411.
80. Lederer W, Küchenhoff H. A short introduction to the SIMEX and MCSIMEX. R News 2006; 6(4):26–31.
81. Lederer W, Küchenhoff H. Simex: SIMEX- and MCSIMEX-Algorithm for Measurement Error Models. 2013: http://CRAN.R-project.org/package=simex.
82. Reeves GK, Cox DR, Darby SC, Whitley E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. Stat Med 1998; 17:2157–2177. [PubMed: 9802176]
83. Pierce DA, Stram DO, Vaeth M, Schafer D. Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. J Am Stat Assoc 1992: 87:351–359.

84. Mallick B, Hoffman FO, Carroll RJ. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. Biometrics 2002; 58:13–20. [PubMed: 11890308]
85. Li Y, Guolo A, Hoffman FO, Carroll RJ. Shared uncertainty in measurement error problems, with application to Nevada Test Site fallout data. Biometrics 2007; 63:1226–1236. [PubMed: 18078484]
86. Stram DO, Kopecky KJ. Power and uncertainty analysis of epidemiological studies of radiation related disease risk in which dose estimates are based on a complex dosimetry system: some observations. Radiation Research 2003: 160:408–417. [PubMed: 12968933]
87. Hoffman FO, Ruttenber AJ, Greenland S, Carroll RJ. Radiation exposure and thyroid cancer: Letter to the editor. J Am Med Assoc 2006; 296:513.
88. Hoffman FO, Ruttenber AJ, Apostoaei AI, Carroll RJ, Greenland S. The Hanford Thyroid Disease Study: an alternative view of the findings. Health Phys 2007; 92:99–111. [PubMed: 17220711]
89. Kukush A, Shklyar S, Masiuk S, Likhtarov I, Kovgan L, Carroll RJ, Bouville A. Methods for estimation of radiation risk in epidemiological studies accounting for classical and Berkson errors in doses. Int J Biostat 2011; 7(1):15. [PubMed: 21423564]
90. Little MP, Kukush AG, Masiuk SV, Shklyar SV, Carroll RJ, Lubin JH, Kwon D, Brenner AV, Tronko MD, Mabuchi K, Bogdanova TI, Hatch M, Zablotska LB, Tereschenko VP, Ostroumova E, Bouville AC, Drozdovitch V, Chepurny MI, Kovgan LN, Simon SL, Shpak VM, Likhtarev IA. Impact of uncertainties in exposure assessment on thyroid cancer risk among Ukrainian children and adolescents exposed from the Chornobyl accident. PLoS One 2014; 9: e85723. [PubMed: 24489667]
91. Masiuk S, Shklyar S, Kukush A, Carroll RJ, Kovgan L, Likhtarov IA. Estimation of radiation risk in presence of classical additive and Berkson multiplicative errors in exposure doses. Biostatistics 2016; 17:422–436. [PubMed: 26795191]
92. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. J Am Stat Assoc 2000; 95:51–61.
93. White I, Frost C, Tokunaga S. Correcting for measurement error in binary and continuous variables using replicates. Stat Med 2001; 20(22):3441–57. [PubMed: 11746328]
94. Yi GY, Ma Y, Spiegelman D, Carroll RJ. Functional and structural methods with mixed measurement error and misclassification in covariates. J Am Stat Assoc 2015; 109, 681–696.
95. Tsiatis AA, Ma Y. Locally efficient semiparametric estimators for functional measurement error models. Biometrika 2004; 91:835–848.
96. Ma Y, Tsiatis AA. Closed form semiparametric estimators for measurement error models. Stat Sinica 2006; 16:183–193.
97. Zhang X, Wang H, Ma Y, Carroll RJ. Linear Model Selection When Covariates Contain Errors. Journal of the American Statistical Association. 2017;112(520):1553–61. [PubMed: 29416191]
98. Breiman L Heuristics of instability and stabilization in model selection. Ann Stat 1996; 24:2350–2383.
99. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Second Edition. Springer-Verlag, New York NY, 2009.
100. Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. Biometrical J 2018;60(3):431–49.
101. Harrell F Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Springer International Publishing, Switzerland 2015.
102. Kyung M, Gill J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal 2010; 5:369–411.
103. Liang H, Li R. Variable selection for partially linear models with measurement errors. J Am Stat Assoc 2009; 104: 234–248. [PubMed: 20046976]
104. He X, Liang H. Quantile regression estimates for a class of linear and partially linear errors-in-variables models. Stat Sinica 2000; 10:129–140.

105. Ma Y, Li R. Variable selection in measurement error models. Bernoulli 2010; 16: 274–300. [PubMed: 20209020]
106. Datta A, Zou H. Cocolasso for high-dimensional error-in-variables regression. The Annals of Statistics. 2017;45(6):2400–26.
107. Loh PL, Wainwright MJ. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. The Annals of Statistics. 2012; 40(3):1637–64.
108. Sørensen Ø, Frigessi A, Thoresen M. Measurement error in LASSO: Impact and likelihood bias correction. Statistica sinica. 2015 4 1:809–29.
109. Yang H, Xia X. Variable selection for semiparametric varying coefficient partially linear errors-in-Variables (EV) model with missing response. Communications in Statistics-Theory and Methods. 2015 11 2;44(21):4521–39.
110. Tian R, Xue L. Variable selection for semiparametric errors-in-variables regression model with longitudinal data. Journal of Statistical Computation and Simulation 2014;84(8):1654–69.
111. Wang H, Zou G, Wan AT. Adaptive LASSO for varying-coefficient partially linear measurement error models. Journal of Statistical Planning and Inference 2013;143(1):40–54.
112. Buonaccorsi JP. Measurement error, linear calibration and inferences for means. Comput Stat Data An 1991; 11:239–257.
113. Buonaccorsi JP. Measurement error in the outcome in the general linear model. J Am Stat Assoc 1996; 91:633–642.
114. Buonaccorsi JP, Tosteson T. Correcting for nonlinear measurement error in the dependent variable in the general linear model. Commun Stat-Theor M 1993; 22:2687–2702.
115. Keogh RH, Carroll RJ, Tooze JA, Kirkpatrick SI, Freedman LS. Statistical issues related to dietary intake as the outcome variable in intervention trials. Stat Med 2016; 35:4493–4508. [PubMed: 27324170]
116. Magder L, Hughes J. Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol 1997; 146:195–203. [PubMed: 9230782]
117. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. Biometrika 1999; 86:843–855.
118. Lyles R, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. Statistics in Medicine 2010; 29:2297–2309. [PubMed: 20552681]
119. Lyles RH, Tang L, Superak HM, King CC, Celentano DD, Lo Y, Sobel JD. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology 2011; 22:589–597. [PubMed: 21487295]
120. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 2005; 34:1370–1376. [PubMed: 16172102]
121. Bartlett J PVW: Stata module to perform predictive value weighting for covariate misclassification in logistic regression. EconPapers 2014: https://EconPapers.repec.org/RePEc:boc:bocode:s457825.
122. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. Am J Epidemiol 2013; 177:904–912. [PubMed: 24627573]
123. Gilbert R, Martin RM, Donovan J, Lane JA, Hamdy F, Neal DE, Metcalfe C. Misclassification of outcome in case–control studies: methods for sensitivity analysis. Stat Meth Med Res 2016; 25:2377–2393.
124. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991; 134:1233–1244. [PubMed: 1746532]
125. Brenner H, Blettner M. Misclassification bias arising from random error in exposure measurement: implications for dual measurement strategies. Am J Epidemiol 1993; 138:453–461. [PubMed: 8213750]
126. Delpizzo V, Borghesi JL. Exposure measurement errors, risk estimate and statistical power in case-control studies using dichotomous analysis of a continuous exposure variable. International Journal of Epidemiology 1995;24(4):851–62. [PubMed: 8550285]

127. Gustafson P, Le ND. Comparing the effects of continuous and discrete covariate measurement, with emphasis on the dichotomization of mismeasured predictors. Biometrics 2002; 58:878–887. [PubMed: 12495142]
128. Keogh RH, Strawbridge AD, White I. Correcting for bias due to misclassification when error-prone continuous exposures are misclassified. Epidemiologic Methods 2012; 1:9.
129. Dalen I, Buonaccorsi JP, Laake P, Hjartåker A, Thoresen M. Regression analysis with categorized regression calibrated exposure: some interesting findings. Emerging Themes in Epidemiology 2006; 3:6. [PubMed: 16820052]
130. Dalen I, Buonaccorsi JP, Sexton JA, Laake P, Thoresen M. Correction for misclassification of a categorized exposure in binary regression using replication data. Stat Med 2009; 28:3386–3410. [PubMed: 19757445]
131. Guo Y, Little RJ, McConnell DS. On using summary statistics from an external calibration sample to correct for covariate measurement error. Epidemiology 2012; 23(1):165–74. [PubMed: 22157312]
132. Liao X, Spiegelman D, Carroll RJ, Guo Y, Little RJ. Regression calibration is valid when properly applied. Epidemiology 2013; 24(3):466–8. [PubMed: 23549186]
133. Thiébaut AC, Freedman LS, Carroll RJ, Kipnis V. Is it necessary to correct for measurement error in nutritional epidemiology? Ann Intern Med 2007; 146:65–67. [PubMed: 17200225]
134. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 2011; 103:1086–1092. [PubMed: 21653922]
135. Freedman LS, Commins JM, Willett W, Tinker LF, Spiegelman D, Rhodes D, Potischman N, Neuhouser ML, Moshfegh AJ, Kipnis V, Baer DJ, Arab L, Prentice RL, Subar AF. 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; 186:73–82. [PubMed: 28402488]
136. Buonaccorsi JP, Dalen I, Laake P, Hjartaker A, Engeset D, Thoresen M. Sensitivity of regression calibration to non-perfect validation data with application to the Norwegian Women and Cancer Study. Stat Med 2015; 34:1389–1403. [PubMed: 25627982]
137. Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics 2009; 65:1003–1010. [PubMed: 19302405]
138. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014; 43:1969–1985. [PubMed: 25080530]
139. Greenland S, Lash TL. Bias analysis In: Modern Epidemiology. Third edition. Rothman KJ, Greenland S, Lash TL (eds). Lippincott Williams and Wilkins, Philadelphia PA 2008.
140. Greenland S Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996; 25:1107–1116. [PubMed: 9027513]
141. Fox MP, Lash TL. On the need for quantitative bias analysis in the peer-review process. Am J Epidemiol 2017; 185:865–868. [PubMed: 28430833]
142. Fox MP. Creating a demand for bias analysis in epidemiological research. J Epidemiol Commun Health 2009; 63:91.
143. MacLehose RF, Olshan AF, Herring AH, Honein MA, Shaw GM, Romitti PA. Bayesian methods for correcting misclassification: an example from birth defects epidemiology. Epidemiology 2009; 20:27–35. [PubMed: 19234399]
144. Lash TL, Ahern TP. Bias analysis to guide new data collection. Int J Biostat 2012; 8(2):1–23.
145. Jurek A, Maldonado G, Spector J, Ross JA. Periconceptional maternal vitamin supplementation and childhood leukaemia: an uncertainty analysis. J Epidemiol Commun Health 2009; 63:168–171.
146. Ross JA, Blair CK, Olshan AF, Robison LL, Smith FO, Heerema NA, Roesler M. Periconceptional vitamin use and leukemia risk in children with Down syndrome. Cancer 2005; 104:405–410. [PubMed: 15952191]
147. He W, Yi GY, Xiong J. Accelerated failure time models with covariates subject to measurement error. Stat Med 2007; 26:4817–4832. [PubMed: 17436310]

148. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, Van Smeden Maarten. Measurement error is often neglected in medical literature: a systematic review. J Clin Epi 2018; 98: 89–97.

149. Shaw PA, Deffner V, Keogh RH, Tooze JA, Dodd KW, Kuechenhoff H, Kipnis V, Freedman LS. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. Ann Epidemiol 2018; 28(11):821–828. [PubMed: 30316629]

150. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, STRATOS initiative. Strengthening analytical thinking for observational studies: the STRATOS initiative. Statistics in Medicine. 2014 12 30;33(30):5413–32. [PubMed: 25074480]

151. van der Voet H, van Klaveren JD, Boon PE (2003). Validation of Monte Carlo models for estimating pesticide intake of Dutch infants. Report 2003.002. RIKILT, Wageningen. http://edepot.wur.nl/39363

152. Wang XF, Wang B. Deconvolution estimation in measurement error models: The R package decon. J Stat Softw 2011; 39:1–24.

153. Delaigle A Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. Aust NZ J Stat 2014; 56:105–124.

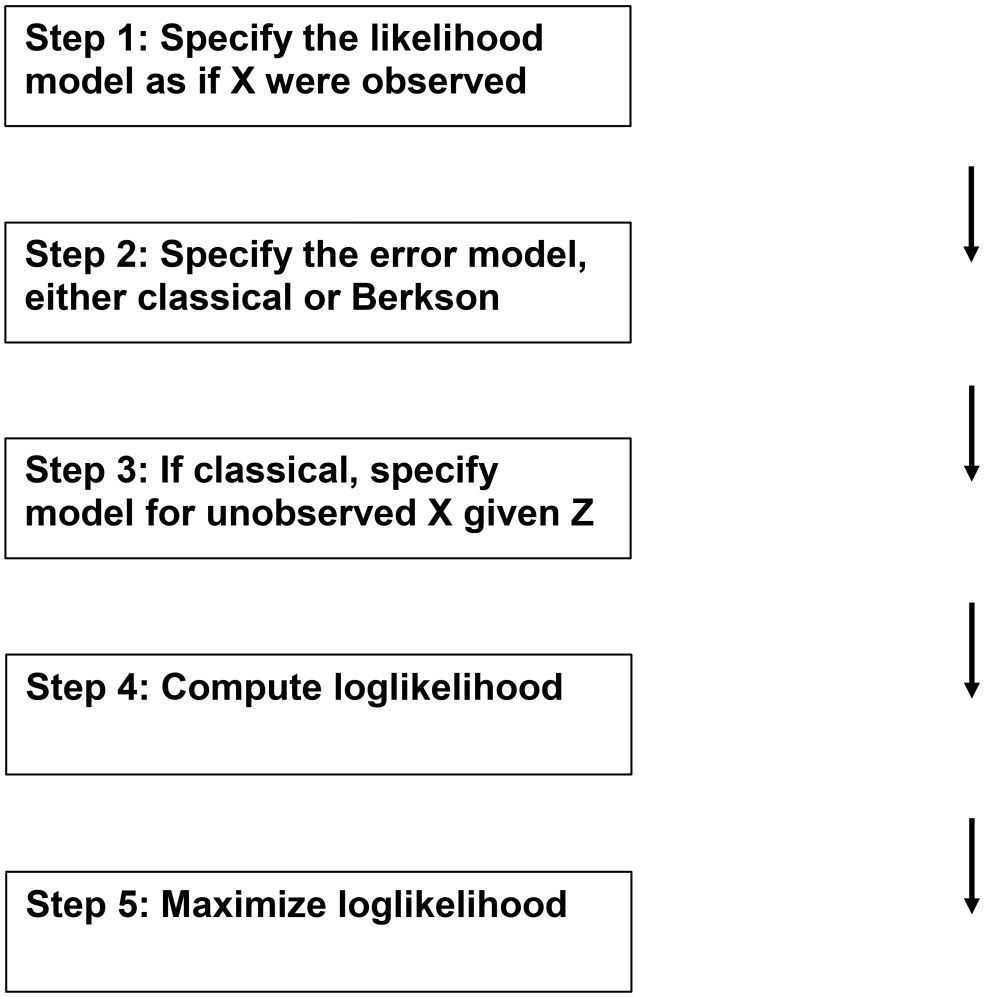

Figure 1:
Flowchart for the steps in a likelihood analysis.

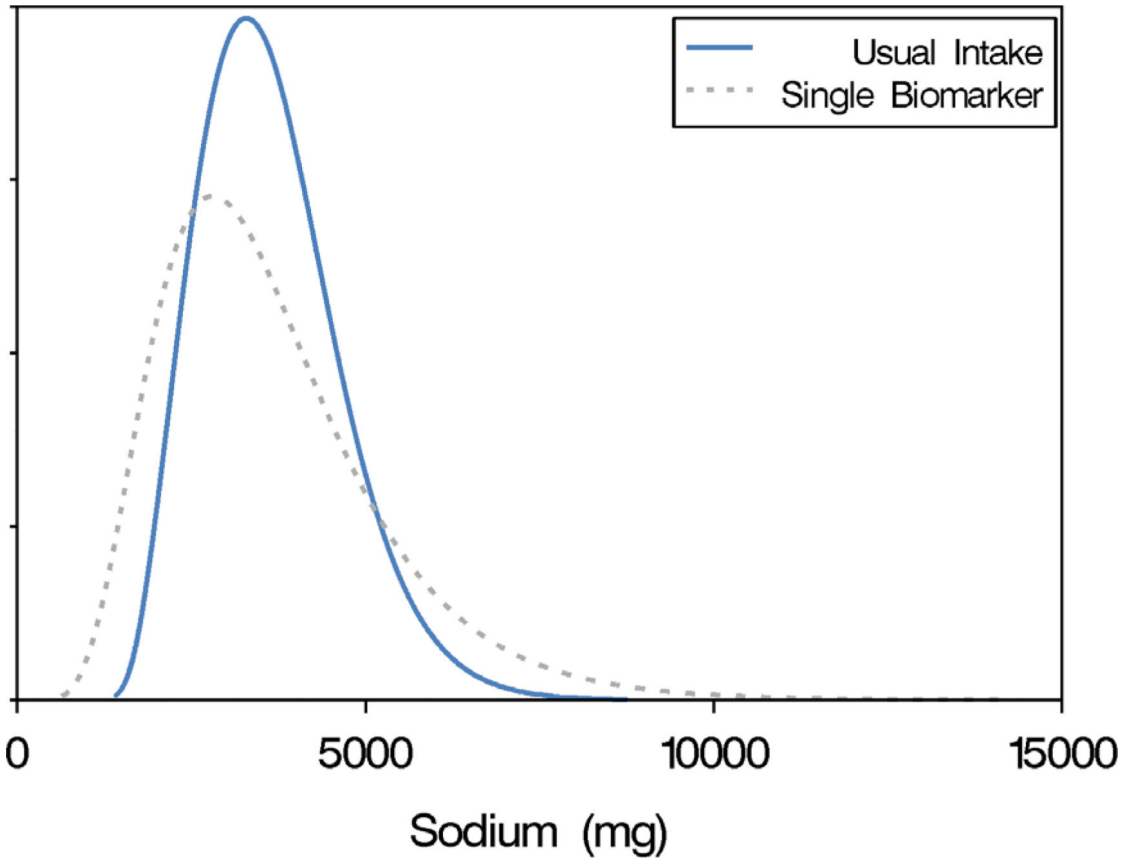

Figure 2:
Smoothed estimated densities of usual sodium intake per day among the OPEN study participants based on (i) a single 24-hour urinary sodium determination (single biomarker) and (ii) the NRC method of adjusting for the measurement error in a single determination (usual intake).

Table 1:

Analyses of the association of log potassium density intake with body mass index, using maximum likelihood estimation and Bayesian methods: data from the 484 participants in the OPEN study

Notation

X = true log potassium density; Z_1 = Sex; Z_2 = Age

X* = FFQ potassium density; X** = biomarker log potassium density

Y = BMI

Models

Outcome Model	$Y\|X,Z_1,Z_2 \sim N(\beta_0 + \beta_X X + \beta_1 Z_1 + \beta_2 Z_2, \sigma_e^2)$
Measurement Error Model 1	$X^*\|X,Z_1,Z_2 \sim N(\alpha_0 + \alpha_X X + \alpha_1 Z_1 + \alpha_2 Z_2, \sigma_U^2)$
Measurement Error Model 2	$X^{**}\|X \sim N(X, \sigma^2_{X^{**}})$
Exposure Model	$X\|Z_1,Z_2 \sim N(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2, \sigma_X^2)$

Maximum Likelihood Results for Outcome Model

Coefficient	Estimated coefficient	Standard Error	95% CI	P-value
log potassium density, β_X	−7.19	1.25	−9.64, −4.74	<0.001
Sex (F v M), β_1	−0.03	0.51	−1.03, 0.97	0.95
Age (years), β_2	0.09	0.03	0.03, 0.15	0.004

Bayesian Results for Outcome Model

Coefficient	Estimated coefficient	Posterior Standard Deviation	95% credible limits	Posterior Probability of being >0
log potassium density, β_X	−6.08	1.43	−9.38, −3.78	<0.005
Sex (F v M), β_1	−0.30	0.51	−1.30, 0.71	0.27
Age (years), β_2	0.08	0.03	0.02, 0.15	>0.995

Table 2:

Analyses of the association of log sodium intake with body mass index, using analyses unadjusted for covariate measurement error, moment reconstruction, and multiple imputation. Data from the 484 participants in the OPEN study.

Part A: Outcome regression parameter estimates			
Variable	Unadjusted Analysis β (SE)	Moment Reconstruction β (SE)	Multiple Imputation β (SE)
log sodium intake	1.13 (0.57)	12.21 (2.70[a])	11.02 (2.65[b])
Sex (F v M)	−0.23 (0.51)	3.16 (1.00[a])	2.75 (1.45[b])
Age (years)	0.037 (0.029)	0.052 (0.035[a])	0.051 (0.070[b])
Part B: Regression models needed for Moment Reconstruction			
B1: Model of biomarker log sodium intake on BMI, sex and age [c]	Estimated coefficient	Standard error	z-value
Intercept	8.03	0.19	43.4
BMI (kg/m^2)	0.031	0.004	7.81
Sex (F v M)	−0.29	0.04	−7.00
Age (years)	−0.0027	0.0025	−1.11
Residual variance	0.0965		
B2: Model of FFQ log sodium intake on BMI, sex and age	Estimated coefficient	Standard error	z-value
Intercept	8.43	0.170	49.53
BMI (kg/m^2)	0.0071	0.0036	1.97
Sex (F v M)	−0.28	0.04	−7.27
Age (years)	−0.0062	0.0023	−2.68
Residual variance	0.1736		
Part C: Regression models needed for Multiple Imputation			
Model of biomarker log sodium intake on FFQ log sodium, BMI, sex and age [c,d]	Estimated coefficient	Standard error	z-value
Intercept	7.54	0.45	16.93
FFQ log sodium intake	0.059	0.049	1.22
BMI (kg/m^2)	0.030	0.004	7.47
Sex (F v M)	−0.27	0.04	−6.22
Age (years)	−0.0022	0.0025	−0.91
Residual variance	0.0963		

[a] From a bootstrap sample of 5000

[b] From 500 multiple imputations

[c] Based on a random subsample of 250 participants

[d] Used to impute biomarker log sodium intake

Table 3:

Analyses of the association of log potassium density intake with body mass index, using analyses unadjusted for measurement error, moment reconstruction and multiple imputation. Data from the 484 participants in the OPEN study.

Variable	Unadjusted Analysis β (SE)	Moment Reconstruction β (SE)	Multiple Imputation β (SE)
log potassium density	−1.69 (0.93)	−8.13 (1.77[a])	−7.28 (2.03[b])
Sex (F v M)	−0.38 (0.49)	0.04 (0.53[a])	0.05 (0.73[b])
Age (years)	0.039 (0.29)	0.101 (0.035[a])	0.094 (0.048[b])

[a]From a bootstrap sample of 5000

[b]Using 500 multiple imputations, where the imputation model for the biomarker log-potassium intake was based on a random subsample of 250 participants and included the self-reported FFQ log-potassium density, sex, age (years), and BMI

Table 4:

Data from a study of risk factors for sudden infant death syndrome [Kraus et al, (1989)[43]]: Y = sudden infant death syndrome (case=1, control=0); X = antibiotic use during pregnancy according to medical record (yes=1, no=0); X* = antibiotic use during pregnancy according to mother's report (yes=1, no=0)

	X=0	X=1	X unobserved	Total
Y=0, X*=0	168	16	479	663
Y=0, X*=1	12	21	101	134
Y=1, X*=0	143	17	442	602
Y=1, X*=1	22	29	122	173

Table 5:

Estimated percentiles of the distribution of usual sodium intake (mg/day) among a population typical of the participants in the OPEN study, using a single measurement of urinary sodium (unadjusted method) versus the NRC method applied to the log value of this measurement

Percentile	Unadjusted method (mg/day)	NRC method (mg/day)
5	1,810	2,233
10	2,150	2,530
25	2,879	3,126
50	3,948	3,928
75	5,322	4,876
90	6,649	5,729
95	7,686	6,363

Table 6:

Software for estimating distributions (see notes at in Section 4.3)

Package/ procedure	Website location or information	Notes	References
SAS code for National Research Council method	https://www.nal.usda.gov/sites/default/files/fnic_uploads//dietary_planning_full_report.pdf Pages 207–208	The National Research Council method was devised for estimating distributions of usual dietary intake from reports of a single day's intake. *SAS* code for implementing the National Research Council method can be found in the source given in the References column.	National Research Council (1986).[49]
PC-SIDE (Iowa State University)	http://www.side.stat.iastate.edu/pc-side.php	Created with the same intention as the National Research Council method. Implements the method of Nusser et al which employs a nonparametric transformation of the data to normality, deconvolutes, and back-transforms to the original scale	Nusser et al (1996).[50]
MCRA 8.2	https://mcra.rivm.nl/	Implements the Monte-Carlo Risk Assessment (MCRA) method devised by researchers at Wageningen University and Research, Netherlands, This method was created for assessing risk from chemicals in the diet and includes the facility of creating a distribution of X from X*.	van der Voet et al (2003)[151]
NCI Method *SAS* macros	https://epi.grants.cancer.gov/diet/usualintakes/macros.html	Created for analyzing dietary data. 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*. X* may be univariate, bivariate or multivariate, and may have excess zeros. Allows inclusion of covariates in the modeling.	Tooze et al (2006);[51] Freedman et al (2010);[58] Zhang et al (2011)[59]
SPADE	http://rivm.nl/en/Topics/S/SPADE/Access_to_SPADE	The Statistical Program to Assess Dietary Exposure (SPADE) is written in R and made available by the Netherlands National Institute for Public Health and Environment. X* is univariate and may have excess zeros.	Dekkers et al (2014)[53]
MSM	https://msm.dife.de/	The Multiple Source Method (MSM) was developed by the German Institute of Human Nutrition for estimating distributions of usual dietary intake. Currently, there is a warning on the website regarding use of the program with covariates.	Haubrock et al (2011)[54]
Matlab and R code	https://researchers.ms.unimelb.edu.au/~aurored/links.html#Code	Nonparametric kernel density methods; including those in Delaigle and Meister 2008; Stefanski and Carroll 1990 and several other methods	Delaigle and Meister 2008;[63] Stefanski and Carroll 1990[61]
Rcode	https://abhrastat.github.io/software.html	Bayesian semiparametric density deconvolution methods, including those in Sarkar et al 2014; Sarkar et al 2018, and other methods	Sakar et al 2014;[15] Sakar et al 2018[66]
Package *decon* in R language	https://cran.r-project.org/web/packages/decon/decon.pdf	Written by Wang XF and Wang B, this package computes the deconvolution kernel density estimator and its bandwidths from X* that has normal or Laplace homoscedastic errors or normal heteroscedastic errors with known variances. There the quality of its "tuning parameter" for (bandwidth) estimation has potential problems (Delaigle, 2014).	Wang and Wang (2011)[152]; Delaigle 2014[153]