Published in final edited form as:

Epidemiology. 2010 July ; 21(Suppl 4): S35–S43. doi:10.1097/EDE.0b013e3181ce9eed.

Use of Multiple Assays Subject to Detection Limits With Regression Modeling in Assessing the Relationship Between Exposure and Outcome

Paul S. Albert^a, Ofer Harel^b, Neil Perkins^c, and Richard Browne^d

^aBiostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, MD

^bDepartment of Statistics, University of Connecticut, Storrs, CT

^cEpidemiology Branch, Division of Epidemiology, Statistics, and Prevention, Eunice Kennedy Shriver National Institute of Child Health, Rockville, MD

^dDepartment of Clinical Laboratory Sciences, University at Buffalo, State University of New York, Buffalo, NY.

Abstract

Background—The goal of many studies in environmental epidemiology is to assess the relationship between chemical exposure and disease outcome. Often various assays can be used to measure a particular environmental exposure, with some assays being more invasive or expensive than others.

Methods—We consider the situation in which 2 assays can be used to measure an environmental exposure. The first assay has measurement error and is subject to a lower detection limit (LOD), and the second assay has less measurement error and is not subject to a lower LOD. In this situation, the first assay is less invasive or less expensive and is measured in all study participants, whereas the second assay is more invasive or more expensive and is only measured in a subset of individuals. We develop a flexible class of regression models that incorporates both sets of assay measurements and allows for continuous or binary outcomes. We explore different design strategies for selecting the subset of patients in whom to measure the second assay. One design strategy is to measure the second more invasive or expensive assay only when the first assay is below LOD. We compare these designs with a simple design in which the second assay is measured in a random subset of patients without regard to the results of the first assay.

Results—We develop estimation approaches for these regression models. We demonstrate through simulations that there are efficiency advantages of measuring the second assay in at least a fraction of cases in which the first assay is above LOD. We illustrate the methodology by using data from a study examining the effect of environmental polychlorinated biphenyl exposure on the risk of endometriosis.

Conclusion—The proposed methodology has good statistical properties and will be a useful methodological technique for studying the effect of exposure on outcome when exposure assays are subject to LOD.

Biomarkers are commonly used as outcomes or measures of exposure. These biomarkers are obtained with assays that are known to have measurement error and known limits of detection (LOD). For some biomarkers, there are multiple assays measuring the same exposure, with one assay potentially being more invasive or more expensive than another. For example, in a recent study^{1,2} examining the relationship between polychlorinated biphenyl (PCB) environmental agents and the risk of endometriosis, serum measurements of numerous PCB congeners were obtained in all study participants before laparoscopic surgery to diagnosis endometriosis. In addition, in about 20% of participants, PCB exposure was measured from adipose tissue that was sampled during laparoscopic surgery. Assay measurements of PCB exposure obtained from adipose tissue are considered to have less measurement error than serum-based assays. A complicating factor is that the less invasive serum assays are subject to LODs, whereas assays measured from more invasive adipose tissue samples are not. The focus of this article is to propose new statistical models that incorporate exposure data measured from assays on the 2 different sample sources.

In the PCB exposure/endometriosis study, adipose tissue was sampled in the first 15 patients (of 79 total study participants). Aside from possible accrual effects, this can be viewed as a completely random sample. Alternative designs are possible in which the probability of measuring the tissue sample depends on the value of the serum-based assay. In this paper, we will compare designs in which we observe the tissue sample (1) only when the serum-based assay is outside detectable limits, (2) when the serum-based assay is outside detectable limits and with a fixed probability in other cases, and (3) completely at random. All these designs focus on the situation for which the sampling mechanism of the tissue sample is fixed by design. Namely, we could obtain the tissue sample in all patients, but choose not to, due to the fact that we wish to limit the proportion of patients who will undergo an invasive procedure.

This proposed methodology is closely related to the classic measurement error in covariates problem, where covariates are measured with error and repeated measurements of surrogates are available³; the proposed model is a structural model in which a parametric model is assumed for the true covariate values (compared with a functional model in which the true covariates are assumed fixed or are assumed random with only minimal assumptions made about the distribution of the covariate). The methodology extends this framework by incorporating (1) LODs in the assays and (2) modeling systematic differences between the multiple assays.

The outline for the paper is as follows. In second section entitled “Modeling Framework,” we propose a regression modeling approach that includes multiple linear regression and logistic regression for continuous and binary outcome variables, respectively. We analyze data from the PCB/endometriosis study in “Example.” Simulations evaluating the efficiency for different designs are presented in the fourth section on “Simulations: Evaluation of Designs.” A discussion follows.

MODELING FRAMEWORK

We will assume that the measurements from 2 assays, denoted as X_{1i}^* and X_{2i}^* , have measurement error δ_{1i} and δ_{2i} , where the measurement errors have normal distributions with mean 0 and variance σ_1^2 and σ_2^2 , respectively. The value X_{1i}^* will be from an inexpensive or noninvasive assay (serum-based assay in PCB/endometriosis study), whereas the value X_{2i}^*

will be from an expensive or invasive assay (tissue-based assay). For simplicity, these 2 assays will be described as the inexpensive and expensive assay, respectively. We assume that the means of X_{i1}^* and X_{i2}^* are $\Delta(X_i)$ and X_i , respectively, where X_i is the “true” measurement without measurement error, and $\Delta(X)$ is a function of X that characterizes the calibration between the 2 assay values. We assume that X_i is normally distributed with mean μ_x and variance σ_x^2 . When $\Delta(X) = X + \Delta$, then both assays differ only by a scalar factor; in this case, both assay measurements are parallel to each other. Another function we consider is $\Delta(X) = \Delta_0 + \Delta_1 X$, where the functional relationship between the 2 assays has both a scalar and a multiplicative factor.

In this paper, we focus on the situation in which the initial assay is subject to a lower LOD, and the more refined expensive or invasive assay is not subject to an LOD. We assume that $X_{i1}^* = \Delta(X_i) + \delta_{1i}$ when $\Delta(X_i) + \delta_{1i} > C$ and $X_{i1}^* = C$ when $\Delta(X_i) + \delta_{1i} \leq C$, where C is the lower LOD for the inexpensive assay. Although we refer to these limits as LOD, others have called these limits quantification limits.

Simple Linear Regression

The relationship between the underlying true biologic measurement X_i and a continuous outcome variable Y_i can be examined using the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

where ε_i is a residual error term that is assumed to be normally distributed with mean zero and variance σ_ε^2 . We assume that ε_i , δ_{1i} , δ_{2i} , and X_i are all independent of each other.

When neither of the 2 assays is subject to LOD, the likelihood can be formulated by noting that $(Y_i, X_{i1}^*, X_{i2}^*)$ is multivariate normal with mean $\underline{\mu}$ and variance $\underline{\Sigma}$, where $\underline{\mu} = (\beta_0 + \beta_1 \mu_x, E[\Delta(X)], \mu_x)'$ and where

$$\underline{\Sigma} = \begin{pmatrix} \beta_1^2 \sigma_x^2 + \sigma_\varepsilon^2 & \beta_1 \text{Cov}(X, \Delta(X)) & \beta_1 \sigma_x^2 \\ \beta_1 \text{Cov}(X, \Delta(X)) & \text{Var}(\Delta(X)) + \sigma_{\delta_1}^2 & \text{Cov}(X, \Delta(X)) \\ \beta_1 \sigma_x^2 & \text{Cov}(X, \Delta(X)) & \sigma_x^2 + \sigma_{\delta_2}^2 \end{pmatrix},$$

where $E[\Delta(X)] = \Delta_0 + \Delta_1 \mu_x$, $\text{Cov}(X, \Delta(X)) = \Delta_1 \sigma_x^2$, and $\text{Var}(\Delta(X)) = \Delta_1^2 \sigma_x^2$ when $\Delta(X) = \Delta_0 + \Delta_1 X$.

The likelihood that we need to maximize to obtain maximum likelihood estimates of the

model parameters is $L = \prod_{i=1}^I L_i$, where the individual contribution to the likelihood is $L_i = P(Y_i, X_{i1}^*, X_{i2}^*)$. This likelihood was maximized using the quasi-Newton–Raphson procedure in GAUSS.⁴

When X_{i1}^* is subject to an LOD, the appropriate likelihood to maximize is

$L = \prod_{i=1}^I P(Y_i, X_{i1}^*, X_{i2}^*)^{X_{i1}^* > C} \{P(Y_i, X_{i2}^*) P(X_{i1}^* < C | Y_i, X_{i2}^*)\}^{X_{i1}^* \leq C}$ where $P(Y_i, X_{i2}^*)$ is a bivariate normal distribution with mean given by the first and third element in $\underline{\mu}$ and variance given

by the first and third row and column in $\underline{\Sigma}$, and where $P(X_{i1}^* < C | Y_i, X_{i2}^*)$ is a univariate Gaussian cumulative distribution function formed by noting that the conditional distribution of a jointly normal random variable is a cumulative normal distribution.

This likelihood assumes that the second assay is performed at all measurements. In the PCB/endometriosis example, the second assay is more invasive and is only performed in a subset of patients. We consider the situation in which the initial assay is performed in all participants and the second assay is performed in only a subset of patients. Define Z_i as an indicator of whether the second assay is performed in the i th patient (ie, the second assay is performed when $Z_i = 1$ and is not performed when $Z_i = 0$). This article considers the case in which the second assay could be observed in all patients, but we only perform the second assay in a subset of these patients due to cost or difficulty in performing this assay. For this type of “missing by design,” the missing data mechanism is either missing completely at random (ie, we observe the second assay without regard to the results of the first assay) or missing at random (ie, the probability of measuring the second assay depends on the value of the first assay). Under a missing data mechanism in which the second assay is observed either completely at random or at random, the individual contribution of the likelihood can be written as

$$L_i = P(Y_i, X_{i1}^*, X_{i2}^*, Z_i) = P(Y_i, X_{i1}^*, X_{i2}^*) P(Z_i | Y_i, X_{i1}^*, X_{i2}^*) \propto P(Y_i, X_{i1}^*, X_{i2}^*). \quad (2)$$

The likelihood corresponding to this design can be written as

$$L = \prod_{i=1}^I P(Y_i, X_{i1}^*, X_{i2}^*)^{(X_{i1}^* \geq C)} Z_i P(Y_i, X_{i1}^*)^{(X_{i1}^* \geq C)(1-Z_i)} \times P(Y_i, X_{i1}^* < C, X_{i2}^*)^{(X_{i1}^* < C)} Z_i P(Y_i, X_{i1}^* < C)^{(X_{i1}^* < C)(1-Z_i)}, \quad (3)$$

where $P(Y_i, X_{i1}^* < C, X_{i2}^*) = P(Y_i, X_{i2}^*) P(X_{i1}^* < C | Y_i, X_{i2}^*)$, and where $P(Y_i, X_{i1}^* < C) = P(Y_i) P(X_{i1}^* < C | Y_i)$. The cumulative probability $P(X_{i1}^* < C | Y_i)$ is a cumulative normal because (Y_i, X_{i1}) is bivariate normal, and therefore, $X_{i1}^* | Y_i$ has a normal distribution.

Polynomial Regression

The linear relationship between exposure and outcome may not be appropriate in certain applications. Equation (1) can be generalized to allow for a more flexible relationship between exposure and outcome,

$$Y_i = \beta_0 + g(X_i) + \varepsilon_i, \quad (4)$$

where g may be a linear ($g(x) = \beta_1 x$), quadratic ($g(x) = \beta_1 x + \beta_2 x^2$), cubic ($g(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$), or cubic spline function ($g(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K \gamma_k (x - c_k)^3$, where K is the number of knot points and $c_k, k = 1, 2, \dots, K$ are the positions of the knot points). Unfortunately, unlike for the simple linear regression model presented previously, the joint distribution of $(Y_i, X_{i1}^*, X_{i2}^*)$ does not have a multivariate normal distribution when the 2 assay measurements are not subject to LODs. However, the model parameters can be estimated by maximum likelihood by using (3) where

$$P(Y_i, X_{i1}^*, X_{i2}^*) = \int P(Y_i | X) P(X_{i1}^* | X) P(X_{i2}^* | X) P(X) dX, \quad (5)$$

$$P(Y_i, X_{i1}^*) = \int P(Y_i | X) P(X_{i1}^* | X) P(X) dX, \quad (6)$$

$$P(Y_i, X_{i1}^* < C, X_{i2}^*) = \int P(Y_i | X) F(X_{i1}^* < C | X) P(X_{i2}^* | X) P(X) dX, \quad (7)$$

and

$$P(Y_i, X_{i1}^* < C) = \int P(Y_i | X) F(X_{i1}^* < C | X) P(X) dX, \quad (8)$$

where $P(\dots | X)$ denotes $P(\dots | X_i = X)$, $P(X)$ denoting $P(X_i = X)$ is a normal density with μ_x and variance σ_x^2 , $P(Y_i | X)$ is a normal density with mean $\beta_0 + g(X)$ and variance σ^2 , $P(X_{i2}^* | X)$ is a normal density with mean X and variance $\sigma_{\delta_2}^2$, and $F(X_{i1}^* < C | X)$ is a cumulative normal distribution function with mean $\Delta(X)$ and variance $\sigma_{\delta_1}^2$. We evaluated these integrals numerically by using a simple trapezoidal rule with 400 function values between -5 and 5 . This approach worked well in simulations.

Logistic Regression

The model can be generalized to allow for a nonnormal outcome. We focus our attention on the case, where Y_i is a binary random variable. We consider the logistic regression model,

$$\text{logit } P(Y_i=1) = \beta_0 + \beta_1 X_i. \quad (9)$$

As before, X_i is measured with error using 2 assays (X_{i1}^* and X_{i2}^*), both subject to measurement error and the first subject to a lower LOD. Maximum likelihood estimation follows using (3) and (5)–(8) with $P(Y_i | X)$ being the probability of either a positive or negative outcome with $P(Y_i = 1 | X) = \exp\{\beta_0 + \beta_1 X\} / (1 + \exp\{\beta_0 + \beta_1 X\})$.

Example

Louis et al¹ have recently associated PCB exposure (as measured by 62 individual PCB congeners) with the risk of endometriosis in 79 patients. Specifically, they showed a positive relationship between total PCB exposure measured by serum assays and the risk of endometriosis by using logistic regression. Regardless of whether a serum measurement was above or below the LOD, all measurements were used in the analysis. In addition to serum measurements used in their analysis, assays determined from adipose tissue were also available in a subset of 15 of the 79 patients.² Adipose tissue was collected in the first 15 patients during laparoscopic surgery. They also measured total lipids (using gravitational techniques) that, along with the presence or absence of endometriosis, will be used as outcome variables in our analyses. By using methodology proposed in this paper, we will incorporate assays from both serum and adipose tissue in this analysis.

For illustrative purposes, we focus on the PCB congener 153, which has 44% (35 of 79) of measurements below LODs (lower LOD = 0.142). Figure 1 shows the distribution of all 79 log-transformed serum values (including values that are below the lower LODs, which is graphically demonstrated with a dark bar) for congener 153; the plot suggests that the data are approximately normal on the log-scale. Figure 1 also shows that the distribution of total

lipids is approximately normally distributed on the log-scale. Figure 2 shows the relationship between log-transformed serum and adipose measurements for the 15 patients in whom both assays were measured. The figure demonstrates a strong positive relationship between the serum and adipose tissue assay values for this small group of participants ($r = 0.68$, $P = .005$).

Although the values were available from the serum assays, which were below LODs, we did not incorporate this information into the analysis (ie, the models discussed previously assume that values below LOD are not available for analysis). Table 1 shows the maximum likelihood estimates for model (1) where the dependent variable Y_i is log-transformed total lipids, the independent variable X_i is log-transformed PCB congener 153, and the relationship between the 2 assays is described by $\Delta(X) = \Delta_0 + \Delta_1 X$. Due to the small sample sizes in this study (recall that only 15 adipose tissue assays were performed in 79 study participants), a parametric bootstrap⁵ with 800 bootstrap samples was performed for assessing the variability of the parameter estimates in this study. For many of the model parameters, histograms of the bootstrap estimates were not close to being normally distributed, so we present 95% confidence intervals (CIs) for each of the parameter estimates using the percentile method. The results in Table 1 show that the estimated trend is

positive ($\widehat{\beta}_1 = 0.73$) with the 95% CI being very close to excluding 0. Thus, there appears to be some evidence for a positive increase in log-transformed total lipids with increasing log-transformed PCB congener 153. The β_1 coefficient can be interpreted as the average change in log-transformed total lipids per unit change in log-transformed PCB congener 153. A scale invariant quantity is the change in mean log-transformed total lipids from the first to the third quartile of log-transformed PCB congener 153 true values. This quantity is $2(0.675)\beta_1 \times \sigma_x = 2(0.675)(0.726)(0.449) = 0.44$ with the 95% CI = -0.02 to 0.92). The results in Table 1 also show that the measurement error for the initial serum assay is larger than that for the second assay that was performed with samples from adipose tissue. Further, the estimates of Δ_0 and Δ_1 suggest that the 2 assays can be calibrated by a constant shift (ie, Δ_1 is not significantly different from 1). We also fit a model in which the measurements below the lower LOD were used in the analysis (data not shown). The results showed similar estimates with slightly smaller 95% CIs to those presented in Table 1. For example, $\widehat{\beta}_1 = 0.60$ (95% CI = 0.00 – 1.74) and the estimated change in the mean log-transformed total lipids from the first to the third quartile was 0.46 (0.00 – 0.85). Thus, there appears to be a small increase in precision in incorporating the actual values of the serum assay when they are below LOD. In the next section, we will explore this in more details with simulations.

Table 2 shows the maximum likelihood estimates for a simple logistic regression model (model [3]) with the dependent variable Y_i representing the occurrence of endometriosis during laparoscopic surgery and X_i representing the log-transformed PCB congener 153 value. The results in Table 2 show a positive effect ($\widehat{\beta}_1 = 0.96$) of log-transformed PCB congener 153 on the probability of endometriosis. However, this effect was not statistically different from zero (ie, 95% CI contained zero). The change in the log odds of endometriosis from the first to the third quartile of values of PCB congener 153 is 0.63 (95% CI = -0.26 – 2.08). As with the continuous total lipids outcome (Table 1), the measurement error is substantially smaller for the adipose tissue assay compared with the serum assay. Also, estimates of Δ_0 and Δ_1 suggest that there is a near constant shift between the 2 assays.

In this study, the adipose tissue assays were performed in the first 15 patients in the study. An alternative design would have been to sample the adipose tissue with a probability depending on the value of the serum assay. For example, a design in which the adipose

tissue assay is performed only when the initial serum assay is below LOD could be considered. We explore various designs with simulations in the next section.

SIMULATIONS: EVALUATION OF DESIGNS

We examine designs through simulations. Specifically, we compare designs in which (1) the second more precise assay is measured only when the first assay is below LOD, (2) the second more precise assay is measured when the first assay is below LOD and with probability P when the first assay is above LOD, and (3) when the second assay is measured completely at random (ie, without regard to the value of the initial assay). In the following subsections, we compare these designs for the different models previously discussed. All simulations presented were conducted with $I = 1000$ to mimic asymptotic results for comparing different designs. However, comparisons with smaller sample sizes of $I = 100$ gave similar results for most parameter estimates.

Simple Linear Regression

Table 3 shows simulations under a simple linear model (model given by [1]). The parameters were chosen such that (1) one-half the values of the initial assay were below LOD, (2) the measurement error for the first assay was 4 times that of the second assay, and (3) the calibration between the first assay and the second is specified by the linear function $\Delta_0 + \Delta_1 X$. The columns corresponding to Design A show the mean (standard deviation [SD]) of the parameter estimates for a design for which the second more precise assay is measured when the initial less precise assay is below LOD and with probability P when the initial assay is above LOD. The columns corresponding to Design B show the mean (SD) parameter estimates for a design in which the second assay is measured completely at random (without regard to the value of the initial assay) with probability P . A comparison of these 2 designs shows the efficiency advantage of Design A with $P = 0$ compared with Design B with $P = .05$ (note that both of these designs have the same overall 50% probability of measuring the second assay). For example, the relative efficiency for estimating β_1 for Design B with $P = .05$ compared with Design A with $P = 0$ is 14.27 ($= [0.136/0.036]^2$). Thus, designs in which we only observe the second more expensive assay when the first assay is below LOD are highly inefficient relative to sampling the second assay completely at random in a comparable proportion of participants. Table 3 also shows the large efficiency advantages of taking a small proportion of second-assay measurements when the first assay is above LOD compared with a design that only measures the second assay when the initial assay is below LOD. For example, there is a large efficiency gain in using Design A with $P = .05$ compared with using Design A with $P = 0$ (in this case, the relative efficiency for estimating β_1 is $[0.136/0.052]^2 = 6.84$). Although Design B is more efficient than Design A when the second assay is measured in a relatively small fraction of patients, both designs have similar efficiency when a larger proportion of second assays are sampled. For example, the relative efficiencies when using Design A with $P = .50$ and Design B with $P = .75$ are nearly identical (both these designs have the same overall 75% probability of measuring the second assay).

The assay measurements below the LOD are not incorporated in the approaches taken in this paper. We assume that those values are left censored at the lower LOD (in many cases, the values below LOD are not provided by the laboratory). In a simulation with the same true parameters sample size as those described in Table 3 and with the second assay measured completely at random in 50% of patients, we estimated the mean (SD) estimated parameter value for each of the 9 parameters (data not shown). There was only a moderate efficiency gain in incorporating these values. For example, the efficiency gain for estimating β_1 was 27% ($[0.036/0.032]^2$) when, as in Table 3, 50% of initial values were below LOD. The

efficiency gain will be substantially smaller when there are fewer initial values that are below LOD.

The proposed methodology is completely parametric and inferences may be sensitive to these parametric assumptions. We examined the robustness of the methodology to normal assumption for X_i and ε_i with 2 simulations for Design A with $P = .50$. First, we generated data as in Table 3, with X_i being generated as a t distribution with 3 degrees of freedom instead of the assumed normal distribution. Estimates of β_0 and β_1 were nearly unbiased with this moderate departure from normality (data not shown). Second, we generated data as in Table 3, with ε_i being a t distribution with 3 degrees of freedom. Estimates of β_0 and β_1 were also nearly unbiased (data not shown). These results suggest that the methodology is robust to moderate departure from normality for the variance of the true covariate and for the residual variance in the simple linear regression.

Polynomial Regression

Table 4 shows simulations under a cubic model as described in the second section. As in Table 3, we present simulation results for a design in which the second assay is measured when the first assay is below LOD and with probability P when the first assay is above LOD (Design A) and for a design in which the second assay is measured completely at random with probability P (Design B). In addition to presenting results for each model parameter, the table shows the mean change in Y from the first to the third quartile in X . The results demonstrate nearly unbiased estimation with a sample size of $I = 1000$. Results are similar to those presented for the linear regression model in Table 3. The simulation results demonstrate the advantage of performing at least a small fraction of second assays when the first assay is above LOD. For Design A with $P = .05$ compared with $P = 0$, the relative efficiency for estimating the change in the mean response from the first to the third quartile in X is 1.58 $((0.048/0.037)^2)$. There is a substantially less efficiency gain in performing a large fraction of second assays above LOD. Further, completely random verification is more efficient than only measuring the second assay when the initial assay is below LOD (eg, for estimating the mean response from the first to the third quartile, the efficiency of Design B with $P = .50$ compared with Design A with $P = 0$ is $([0.048/0.028]^2 = 2.94)$. As for simple linear regression model (Table 3), Designs A and B provide more similar efficiency when a large proportion of second assays are sampled (eg, the efficiency of Design B with $P = .75$ compared with Design A with $P = .50$ is $([0.025/0.028]^2 = 0.80)$.

Logistic Regression

This section examines the statistical properties of model (9) under different designs. Table 5 shows simulations for the logistic regression model under both Designs A and B with varying P . With the exception of $\sigma_{\delta 2}$, which is lightly upwardly biased, the estimates are nearly unbiased with a sample size of $I = 1000$. The bias in estimating $\sigma_{\delta 2}$ is reduced with larger sample sizes (data not shown). As in other simulations, the efficiency advantages of measuring the second more expensive assay on at least a subset of measurements when the initial assay is above LOD. Also, completely random verification is more efficient than only sampling when the initial assay is above LOD (compare Design B with $P = .50$ to Design A with $P = 0$).

DISCUSSION

This paper proposed methodology for estimating the regression relationship between exposure and outcome when the exposure variable is assessed by multiple assays that are potentially subject to detection limits. We developed models and evaluated designs for both continuous and binary outcome variables and considered the case in which the relationship

between outcome and exposure followed a polynomial or smoothing spline relationship. The PCB/endometriosis study measured a serum assay in all study participants and the more definitive “gold standard” assay on only a subset of the total patients. In this study, the second gold standard assay was measured from adipose tissue sampled during laparoscopy surgery and can be considered to be sampled completely at random. We explored the alternative design in which the second assay is performed when the first assay is below LOD and with probability P when the initial assay is above LOD. We found that the efficiency of a design in which the second assay is measured only when the initial assay is below LOD is substantially improved when only a small percentage of second assays are measured when the initial assay is above LOD. Additionally, designs in which we only observe the second assay when the first assay is below LOD is highly inefficient relative to sampling the second assay completely at random in a comparable proportion of patients.

In the PCB/endometriosis study, the second assay was an invasive, tissue-based assay, whereas the initial assay was a serum-based assay. In other settings, the second assay may be more expensive than the initial assay, and designs may be compared based on minimizing study costs. In these cases, rather than comparing relative efficiencies as was done in this paper, we can compare designs based on minimizing cost functions. Such an approach would require specifying the relative cost of the 2 assays, and could easily be implemented within the framework discussed in this paper.

For the PCB/endometriosis study, the serum assays were subject to LODs, whereas the gold standard adipose tissue assay was not subject to LOD. Thus, the model development and design considerations were for the situation in which the inexpensive/noninvasive assay measured in all individuals was subject to lower LOD, whereas the gold standard expensive/invasive assay measured on a subset of individuals was not subject to LOD. The model and simulations could be easily altered to allow for LOD on the second assay. Furthermore, the model could be easily altered additionally to allow for an upper LOD for the initial assay. Our investigation is focused on the situation where there are 2 assays. The methodology can easily be extended to incorporate more than 2 assays measuring a single exposure. We expect that the design results presented in this paper will be similar for these alternative formulations.

Most of the models presented in this article assume that values below LOD are not available and treat these values as being left censored. In many instances, values outside LODs are not reported by the laboratory or, for a particular assay, values below a certain limit cannot be quantified. However, as a general principle, these values should be collected when possible, and depending on the type of assay, should be used in the analysis.⁶ For the PCB/endometriosis study, where the value of serum assay measurements below LOD were recorded, we were able to examine the advantages of using the actual measurements below the LOD rather than treating these values as left censored. We found small efficiency gains in incorporating these values in the analysis. Further, simulation studies confirmed these efficiency gains in situations in which the proportion of initial assays below LOD was sizable. In the analyses and simulations, we assumed that measurement error was constant and did not depend on the actual true measurement of the exposure variable. More complex models that allow for a measurement error process that depends on the true value may be more appropriate when we are analyzing actual measurements below LODs.⁶ The development of such models is an area of future research.

We proposed models for cross-sectional data such as the PCB/endometriosis study in which single exposure and outcome measurements were obtained. These models could be extended to the longitudinal setting in which both exposure and outcome are measured longitudinally in time. Albert⁷ proposed such a model when the outcome in a longitudinal study is

measured with multiple assays. A similar model could be developed in which the exposure variable is measured with multiple assays. This is an area of future research.

This article focuses on the situation in which the second assay is missing by design. Namely, an investigator could measure the second assay in all study participants, but because of cost or feasibility issues only performs the second assay in a subset of study participants. In other applications, investigators may attempt to measure the second assay in all participants, but may have missing data. For example, if the second assay requires a tissue specimen, then it may be missing when the tissue sample cannot be obtained due to the participants refusal to undergo the procedure or due to the surgeon failure to obtain the sample. If the reason for missing data mechanism is missing at random or missing completely at random,⁸ then the proposed methodology is appropriate. However, if the probability of missing the second assay depends on the value of the second assay had it been observed (this type of missing data mechanism has been referred to as nonignorable missingness) then the methodology presented in this article may result in biased inference. Extensions of the models for nonignorable missingness is an area of future research.

Acknowledgments

We thank the 2 reviewers for constructive comments that lead to a significant improvement of this article. We thank Enrique Schisterman, Brian Whitticomb, and Germaine Louis for providing the PCB/endometriosis study data and for helpful discussions. The research used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>).

Supported partially by the Long Range Initiative of the American Chemistry Council and funding from the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development; National Institutes of Health.

REFERENCES

1. Louis GM, Weiner JM, Whitcomb BW, et al. Environmental PCB exposure and risk of endometriosis. *Hum Reprod.* 2005; 20:279–285. [PubMed: 15513976]
2. Whitcomb BW, Schisterman EF, Buck G, Weiner JM, Greizerstein H, Kostyniak PJ. Relative concentrations of organochlorines in adipose tissue and serum among reproductive age women. *Environ Toxicol Pharmacol.* 2005; 19:203–213.
3. Carroll, R.J.; Ruppert, D.; Stefanski, L.A. *Measurement Error in Nonlinear Models.* Chapman & Hall; New York: 1995.
4. Aptech Systems. Gauss Systems, Version 3.0. Aptech Systems; Ravensdale, WA: 1992.
5. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap.* Chapman and Hall; New York: 1993.
6. Guo, Y.; Harel, O.; Little, R.J. How Well Quantified is the Limit of Quantification. University of Connecticut Department of Statistics; Storrs, CT: 2008. Technical Report 20
7. Albert PS. Modeling longitudinal biomarker data from multiple assays that have different known detection limits. *Biometrics.* 2008; 64:527–537. [PubMed: 17764483]
8. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data.* John Wiley; New York: 1987.

FIGURE 1.
Histograms of log-transformed serum PCB congener 153 and log-transformed total lipids.
The solid bar denotes the lower LOD for serum PCB congener 153.

FIGURE 2.
Relationship between log-transformed PCB congener 153 measurements from adipose tissue and from serum. The line represents a least-squares fit through the data.

TABLE 1Maximum Likelihood Estimates of the Model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Parameter	Estimate	95% CI
μ_x	4.37	(4.16 to 4.55)
β_0	1.82	(−3.45 to 5.15)
β_1	0.73	(−0.04 to 1.91)
σ_x	0.45	(0.24 to 6.75)
σ_ε	1.24	(1.00 to 1.40)
σ_{δ_1}	0.26	(0.00 to 0.41)
σ_{δ_2}	0.20	(0.00 to 0.35)
Δ_0	−6.54	(−11.46 to −4.69)
Δ_1	1.06	(0.65 to 2.16)
First–third quartile	0.44	(−0.02 to 0.92)

Maximum likelihood estimates of the model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where Y_i is log-transformed total lipids and X_i is the PCB exposure for congener 153. The lower LOD for the serum measurement was −1.95. Bootstrap standard errors (800 bootstrap samples) are presented.

TABLE 2Maximum Likelihood Estimates of the Model, $\text{logit } P(Y_i = 1) = \beta_0 + \beta_1 X_i$

Parameter	Estimate	95% CI
μ_x	4.35	(4.16 to 4.53)
β_0	-4.74	(-17.0 to 1.35)
β_1	0.96	(-0.45 to 3.51)
σ_x	0.50	(0.25 to 0.66)
σ_{δ_1}	0.31	(0.00 to 0.45)
σ_{δ_2}	0.01	(0.00 to 0.36)
Δ_0	-5.73	(-10.67 to 4.45)
Δ_1	0.88	(0.58 to 2.01)
First-third quartile	0.63	(-0.26 to 2.08)

Maximum likelihood estimates of the model, $\text{logit } P(Y_i = 1) = \beta_0 + \beta_1 X_i$, where Y_i is the occurrence of endometriosis during surgery and X_i is the PCB exposure for congener 53. The lower LOD for the serum measurement was -1.95. Bootstrap standard errors (800 bootstrap samples) are presented.

TABLE 3

Simulation Under the Model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Parameter	Truth	Design A				Design B			
		$P = 1$	$P = .50$	$P = .05$	$P = 0$	$P = .75$	$P = .50$	$P = .75$	$P = .50$
μ_x	0	0.00 (0.063)	0.00 (0.066)	0.00 (0.089)	0.00 (0.161)	0.00 (0.066)	0.00 (0.070)	0.00 (0.066)	0.00 (0.070)
β_0	1	1.00 (0.049)	1.00 (0.054)	1.00 (0.084)	1.03 (0.204)	1.00 (0.053)	1.00 (0.058)	1.00 (0.053)	1.00 (0.058)
β_1	1	1.00 (0.030)	1.00 (0.033)	1.00 (0.052)	1.02 (0.136)	1.00 (0.031)	1.00 (0.036)	1.00 (0.031)	1.00 (0.036)
σ_x	2	2.00 (0.052)	2.00 (0.057)	2.00 (0.090)	1.98 (0.166)	2.00 (0.056)	2.00 (0.064)	2.00 (0.056)	2.00 (0.064)
σ_ε	1.5	1.50 (0.039)	1.50 (0.040)	1.50 (0.048)	1.48 (0.112)	1.50 (0.041)	1.50 (0.044)	1.50 (0.041)	1.50 (0.044)
σ_{δ_1}	1	1.00 (0.050)	1.00 (0.058)	1.00 (0.080)	1.01 (0.120)	1.00 (0.054)	1.00 (0.063)	1.00 (0.054)	1.00 (0.063)
σ_{δ_2}	0.5	0.49 (0.068)	0.49 (0.086)	0.48 (0.133)	0.48 (0.200)	0.49 (0.077)	0.49 (0.096)	0.49 (0.077)	0.49 (0.096)
Δ_0	0	0.00 (0.056)	0.00 (0.063)	0.00 (0.093)	0.03 (0.198)	0.00 (0.059)	0.00 (0.068)	0.00 (0.059)	0.00 (0.068)
Δ_1	1.1	1.10 (0.034)	1.10 (0.041)	1.10 (0.058)	1.11 (0.093)	1.10 (0.037)	1.10 (0.043)	1.10 (0.037)	1.10 (0.043)

Simulation under the model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $X_{1i}^* = \Delta_0 + \Delta_1 X_i | \delta_{1i}$ and $X_{2i}^* = X_i + \delta_{2i}$, and where $\varepsilon_i, \delta_{1i}, \delta_{2i}$ are normally distributed with mean 0 and variance $\sigma_\varepsilon^2, \sigma_{\delta_1}^2$, and $\sigma_{\delta_2}^2$. Further, X_i is assumed to be normally distributed with mean μ_x and variance σ_x^2 . We assume that the first assay has a lower LOD of $C = 0$ and the second assay has no LOD. We assume that $I = 1000$ and that there are 1000 simulations. Columns 3–6 evaluate designs (Design A) in which the second assay is measured when the first assay is below LOD and with probability P when the first assay is above LOD. Columns 7–8 evaluate designs (Design B) in which the second assay is measured on a subset of participants completely at random with probability P . Means (SD) of parameter estimates are presented.

TABLE 4

Simulation Under the Model, $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$

Parameter	Truth	Design A					Design B		
		$P = 1$	$P = .50$	$P = .25$	$P = .05$	$P = 0$	$P = .75$	$P = .50$	$P = .50$
μ_x	0	0.00 (0.031)	0.00 (0.033)	0.00 (0.033)	0.00 (0.034)	0.01 (0.038)	0.00 (0.032)	0.00 (0.033)	0.00 (0.033)
β_0	0	0.00 (0.011)	0.00 (0.013)	0.00 (0.014)	0.00 (0.017)	0.00 (0.018)	0.00 (0.013)	0.00 (0.016)	0.00 (0.016)
β_1	1	1.00 (0.019)	1.00 (0.021)	1.00 (0.022)	1.00 (0.029)	1.00 (0.035)	1.00 (0.020)	1.00 (0.024)	1.00 (0.024)
β_2	-0.05	-0.03 (0.0096)	-0.03 (0.011)	-0.03 (0.014)	-0.03 (0.022)	-0.03 (0.033)	-0.03 (0.011)	-0.03 (0.013)	-0.03 (0.013)
β_3	0.20	0.20 (0.0069)	0.20 (0.0076)	0.20 (0.0083)	0.20 (0.011)	0.20 (0.014)	0.20 (0.0072)	0.20 (0.0091)	0.20 (0.0091)
σ_x	1	1.00 (0.022)	1.00 (0.023)	1.00 (0.023)	1.00 (0.025)	1.00 (0.030)	1.00 (0.024)	1.00 (0.024)	1.00 (0.024)
σ_ϵ	0.25	0.25 (0.010)	0.25 (0.012)	0.25 (0.014)	0.25 (0.015)	0.25 (0.016)	0.25 (0.012)	0.25 (0.014)	0.25 (0.014)
σ_{δ_1}	0.5	0.50 (0.016)	0.50 (0.018)	0.50 (0.018)	0.50 (0.018)	0.50 (0.019)	0.50 (0.017)	0.50 (0.017)	0.50 (0.017)
σ_{δ_2}	0.1	0.10 (0.0096)	0.10 (0.011)	0.10 (0.013)	0.10 (0.014)	0.10 (0.015)	0.10 (0.011)	0.10 (0.014)	0.10 (0.014)
Δ_0	0	0.00 (0.026)	0.00 (0.027)	0.00 (0.028)	0.00 (0.030)	0.00 (0.031)	0.00 (0.026)	0.00 (0.026)	0.00 (0.026)
Δ_1	1.0	1.00 (0.028)	1.00 (0.028)	1.00 (0.030)	1.00 (0.036)	1.00 (0.043)	1.00 (0.028)	1.00 (0.028)	1.00 (0.028)
1st-3rd quartile	1.47	1.47 (0.022)	1.47 (0.025)	1.47 (0.027)	1.47 (0.037)	1.47 (0.048)	1.47 (0.023)	1.47 (0.028)	1.47 (0.028)

Simulation under the model, $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$, where $X_{1i}^* = \Delta_0 + \Delta_1 X_i + \delta_{1i}$ and $X_{2i}^* = X_i + \delta_{2i}$, and where ϵ_i , δ_{1i} , δ_{2i} are normally distributed with mean 0 and variance σ_ϵ^2 , $\sigma_{\delta_1}^2$, and $\sigma_{\delta_2}^2$. Further, X_i is assumed to be normally distributed with mean μ_x and variance σ_x^2 . We assume that the first assay has a lower LOD of $C = 0$ and the second assay has no LOD. We assume that $I = 1000$ and that there are 1000 simulations. Design A is a design in which the second assay is measured when the first assay is below detection limits and with probability P otherwise. Design B is a design in which the second assay is measured completely at random with probability P .

TABLE 5

Simulation Under the Logistic Regression Model, $\text{logit } P(Y_i = 1) = \beta_0 + \beta_1 X_i$

Parameter	Truth	Design A					Design B		
		$P = 1$	$P = .50$	$P = .25$	$P = .05$	$P = 0$	$P = .75$	$P = .50$	$P = .50$
μ_x	0	0.00 (0.064)	0.00 (0.065)	0.00 (0.067)	0.00 (0.075)	-0.01 (0.107)	0.00 (0.066)	0.00 (0.070)	0.00 (0.070)
β_0	0	0.00 (0.082)	0.00 (0.085)	0.00 (0.085)	0.00 (0.089)	0.01 (0.116)	0.10 (0.087)	0.00 (0.088)	0.00 (0.088)
β_1	1	1.01 (0.067)	1.01 (0.067)	1.01 (0.068)	1.01 (0.074)	1.02 (0.094)	1.01 (0.070)	1.01 (0.074)	1.01 (0.074)
σ_x	2	2.00 (0.047)	2.00 (0.048)	1.99 (0.051)	1.99 (0.063)	1.98 (0.100)	1.99 (0.049)	1.99 (0.055)	1.99 (0.055)
σ_{δ_1}	0.5	0.48 (0.039)	0.48 (0.048)	0.47 (0.065)	0.46 (0.091)	0.46 (0.115)	0.48 (0.044)	0.47 (0.054)	0.47 (0.054)
σ_{δ_2}	0.05	0.12 (0.104)	0.13 (0.112)	0.14 (0.126)	0.15 (0.138)	0.16 (0.151)	0.13 (0.107)	0.13 (0.120)	0.13 (0.120)
Δ_0	0.25	0.25 (0.027)	0.25 (0.033)	0.25 (0.039)	0.25 (0.051)	0.27 (0.091)	0.25 (0.031)	0.25 (0.037)	0.25 (0.037)
Δ_1	1.0	1.00 (0.017)	1.00 (0.022)	1.01 (0.027)	1.01 (0.041)	1.02 (0.062)	1.01 (0.019)	1.01 (0.023)	1.01 (0.023)

Simulation under the logistic regression model, $\text{logit } P(Y_i = 1) = \beta_0 + \beta_1 X_i$, where $X_{1i}^* = \Delta_0 + \Delta_1 X_i + \delta_{1i}$ and $X_{2i}^* = X_i + \delta_{2i}$, and where δ_{1i} and δ_{2i} are normally distributed with mean 0 and variance $\sigma_{\delta 1}^2$ and $\sigma_{\delta 2}^2$. Further, X_i is assumed to be normally distributed with mean μ_x and variance σ_x^2 . We assume that the first assay has a lower LOD of $C = 0$ and the second assay has no LOD. We assume that $I = 1000$ and that there are 1000 simulations.