Published in final edited form as:

Biometrics. 2011 June ; 67(2): 636–645. doi:10.1111/j.1541-0420.2010.01464.x.

Binary Regression Analysis With Pooled Exposure Measurements: A Regression Calibration Approach

Zhiwei Zhang* and Paul S. Albert**

Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, U.S.A

Summary

It has become increasingly common in epidemiologic studies to pool specimens across subjects in order to achieve accurate quantitation of biomarkers and certain environmental chemicals. In this paper, we consider the problem of fitting a binary regression model when an important exposure is subject to pooling. We take a regression calibration approach and derive several methods, including plug-in methods that use a pooled measurement and other covariate information to predict the exposure level of an individual subject, and normality-based methods that make further adjustments by assuming normality of calibration errors. Within each class we propose two ways to perform the calibration (covariate augmentation and imputation). These methods are shown in simulation experiments to effectively reduce the bias associated with the naive method that simply substitutes a pooled measurement for all individual measurements in the pool. In particular, the normality-based imputation method performs reasonably well in a variety of settings, even under skewed distributions of calibration errors. The methods are illustrated using data from the Collaborative Perinatal Project.

Keywords

Bias correction; Logistic regression; Measurement error; Missing covariates; Pooling; Probit regression

1. Introduction

In this paper we consider the problem of estimating the relationship of a binary outcome with a collection of covariates, including an important exposure that is measured on pooled specimens from different subjects and not for the individual subjects. Our interest in this problem is driven by the following epidemiologic study. The Upstate New York Infant Development Screening Program (Upstate KIDS) is an ongoing study that enrolls infants at three to five months of age and follows them for three years. The study cohort consists of approximately 6000 children in Upstate New York, of which 1500 will be conceived by infertility treatment (CIT). Each CIT child will be matched with three non-CIT children for residence area and plurality of birth. Of particular interest to us is a new component of the study that examines the roles of inflammation, infection and exposure to environmental chemicals, such as polychlorinated biphenyls (PCBs), in infant and child growth and neurodevelopment. Exposure information for this new component will be obtained from neonatal blood spots on the Guthrie cards to be collected at birth as mandated by state law.

* zhiwei.zhang@nih.gov

** albertp@mail.nih.gov

Each card contains five blood spots and each spot can be punched (sampled) for about 10–12 times. Given the limited amounts of specimens available, assay sensitivity is a serious issue with many of the biomarkers and chemicals to be measured from the Guthrie cards. For many of them, it is anticipated that more than 12 punches may be required to obtain reliable assay results. Thus, pooling samples across individuals may be the only reasonable option for conducting this component of the study. Such pooled exposure measurements present a statistical methodological challenge which is the subject of this paper.

The statistical implications of pooling measurements across subjects have been considered for biomarker studies (Vexler et al., 2006; Schisterman and Vexler, 2008; Schisterman et al., 2010), diagnostic testing (Faraggi et al., 2003; Liu and Schisterman, 2003; Mumford et al., 2006; Vexler et al., 2008), and longitudinal outcomes (Albert and Shih, 2009). In the Upstate KIDS study, the variables subject to pooling (e.g., PCB) will likely be treated as covariates in regression models for describing and explaining developmental problems. Weinberg and Umbach (1999) propose a case-control design where exposure measurements are pooled for cases and controls separately, and point out that a logistic regression model for the individual subjects implies a logistic regression model of the same form for pools of subjects with all covariates summed within pools. It is not immediately clear how to extend their approach to the present cohort study. In the present setting, it is straightforward to deal with continuous outcomes for which a linear regression model can be assumed. For then we can simply average all variables (both the outcome and the covariates) in the linear regression model across subjects in the same pool, in the same way that a pooled PCB measurement is obtained, and the pool averages will follow the same linear regression model as for the original measurements, with the same regression coefficients and a smaller error variance (depending on the size of the pool). It seems more challenging to deal with other types of outcomes. In this paper, we focus on binary regression models even though much of the discussion extends easily to other models (e.g., regression models for count data).

The statistical problem here could be cast in terms of errors in variables (Carroll et al., 2006) by treating a pooled PCB measurement as an imprecise measurement of the PCB level of each individual subject in the pool. On the other hand, the pooling problem has a special structure that merits separate consideration. In this paper, we derive and compare several methods for the pooling problem that can be regarded as variants of regression calibration, a general approach to measurement error models (Carroll et al., 2006, Chapter 4). We set up the notation and state basic assumptions in the next section. Then we adapt the regression calibration approach to the pooling problem and derive some specific methods in Sections 3 and 4. The methods are compared through simulations in Section 5 and illustrated with a real example in Section 6. We then consider some more flexible pooling schemes in Section 7 and conclude the paper with a discussion in Section 8.

2. Notation and assumptions

Let us write X for a continuous exposure variable subject to pooling (e.g., PCB), Y for a binary outcome of interest (e.g., abnormal psychomotor development), and $\mathbf{Z}$ for a vector of other covariates to adjust for. Suppose that Y depends on X and $\mathbf{Z}$ through the following regression model:

$$P(Y=1|\mathbf{Z}, X; \beta) = G(\beta_1 + \beta_z' \mathbf{Z} + \beta_x X), \quad (1)$$

where G is a known function such as the probit function, and $\beta = (\beta_1, \beta_z', \beta_x)'$ is a vector of unknown regression parameters we would like to estimate. The problem is that X is not

measured for each individual subject, but rather for pooled specimens only. The pooling mechanism will be discussed shortly.

Following the regression calibration approach (Carroll et al., 2006), we would like to “calibrate” model (1) with the available information that is predictive of X and then recover β from the calibrated model. This can be implemented with the following calibration model:

$$E(X|Z, V; \alpha) = \alpha_1 + \alpha'_w W, \quad (2)$$

where V is an optional collection of variables that may help to predict X and W is an arbitrary (vector-valued) function of Z and V . This model allows us to predict X using all relevant information: both the observed value of Z and possibly some extra information in V . The actual usage of this information in predicting X is quite flexible: W may depend on all or part of Z , in the original form or after a suitable transformation, and it may involve arbitrary interactions between Z and V . What is essential to our approach is the linearity in model (2), which is not really restrictive given the flexibility with W . A particular form of model (2) is given by

$$X = \alpha_1 + \alpha'_w W + \varepsilon, \quad (3)$$

where we assume that ε is independent of W with $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2$.

It is important to note that the information in V , if any, is strictly for predicting X and should play no role in explaining Y given the covariates in model (1). Formally, we assume that

$$P(Y=1|Z, X, V) = P(Y=1|Z, X), \quad (4)$$

that is, V is conditionally independent of Y given (Z, X) . One could think of V as a collection of instrumental variables (Carroll et al., 2006, Chapter 6), but we do not rely on that interpretation in our methodology. Our use of V , if at all, is mainly to enhance the precision in predicting X .

Our consideration of the pooling mechanism will proceed in steps, starting with simplistic assumptions and then moving to more realistic and flexible pooling schemes. To begin, assume that the subjects are pooled into groups of identical size m , that subjects in the same group contribute equally to the pooled measurement of X , and that the pooling is completely random with respect to all variables mentioned above. The last assumption (random pooling) will be relaxed at the end of this section, and the others in Section 7 (uneven pooling). Let us

use the subscripts ij to denote the j th subject in the i th pool, and write $\bar{X}_i = m^{-1} \sum_{j=1}^m X_{ij}$ for the measured value of X in the i th pool. The observed data can then be written as $\{\bar{X}_i, (Y_{ij}, Z_{ij}, W_{ij}), j = 1, \dots, m\}, i = 1, \dots, n$. Random pooling implies that the random vectors $(X_{ij}, Y_{ij}, Z_{ij}, W_{ij})$ are independent across subjects (whether they belong to the same pool or not), and identically distributed as the unsubscripted version (X, Y, Z, W) (for a generic subject chosen randomly from the population). Together with the calibration model (3), this allows us to relate the unobserved X -values to the available data as follows:

$$E(X_i | \bar{X}_i, W_i) = \bar{X}_i + (W_i - \bar{W}_i) \alpha_w, \quad (5)$$

$$\text{var}(X_i|\bar{X}_i, \mathbf{W}_i) = \sigma^2(\mathbf{I}_m - m^{-1}\mathbf{J}_m), \quad (6)$$

where $X_i = (X_{i1}, \dots, X_{im})'$, $\mathbf{W}_i = (\mathbf{W}_{i1}, \dots, \mathbf{W}_{im})'$, $\bar{X}_i = X_i \mathbf{1}_m$, $\bar{\mathbf{W}}_i = m^{-1} \sum_{j=1}^m \mathbf{W}_{ij}$, $\bar{\mathbf{W}}_i = \mathbf{1}_m \bar{\mathbf{W}}_i'$, $\mathbf{1}_m$ is the m -vector of ones, $\mathbf{I}_m$ is the $m \times m$ identity matrix, and $\mathbf{J}_m$ is the $m \times m$ matrix of ones. Similar notation will be used throughout without further comment. Note that the X_{ij} ($j = 1, \dots, m$) are negatively correlated after conditioning on $\bar{X}_i$.

Although we previously made a random pooling assumption, this assumption can be relaxed through covariate-dependent pooling. Specifically, since equations (5) and (6) are both conditional on $\mathbf{W}_i$, we can, and will, allow covariate-dependent pooling where the grouping of subjects may depend on their $\mathbf{W}$ -values but not on other variables given the $\mathbf{W}$'s. A special case of this is stratified pooling where random pooling is performed within each of a few strata defined by $\mathbf{W}$. The main advantage of covariate-dependent pooling is a potential gain in efficiency. Intuitively, it should be easier to predict X_{ij} using $\bar{X}_i$ if the subjects in the same pool tend to be alike.

3. Plug-in methods

Without observing the true values of the X_{ij} , one might be tempted to substitute $\bar{X}_i$ for X_{ij} in fitting model (1). This is unjustified theoretically and can be shown numerically to result in biased estimates. This bias can be reduced and in some cases eliminated using the regression calibration approach (Carroll et al., 2006, Chapter 4), which can be adapted to the pooling problem via the following identity:

$$P(Y_{ij}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = \int G(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x x_{ij}) f(x_{ij}|\bar{X}_i, \mathbf{W}_i) dx_{ij}, \quad (7)$$

where $f(x_{ij}|\bar{X}_i, \mathbf{W}_i)$ denotes the conditional density of X_{ij} given $(\bar{X}_i, \mathbf{W}_i)$. The identity follows from assumptions (1), (2) and (4), and the details are given in Appendix A.

Without additional assumptions we cannot evaluate the integral in (7) exactly. We could, however, approximate the integrand (as a function of x_{ij}) through a first-order Taylor series expansion around the conditional mean $E(X_{ij}|\bar{X}_i, \mathbf{W}_i)$:

$$G(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x x_{ij}) \approx G(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x E(X_{ij}|\bar{X}_i, \mathbf{W}_i)) + \dot{G}(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x E(X_{ij}|\bar{X}_i, \mathbf{W}_i)) \beta_x \{x_{ij} - E(X_{ij}|\bar{X}_i, \mathbf{W}_i)\},$$

where $\dot{G}$ is the first derivative of G . Integrating this with respect to the conditional density $f(x_{ij}|\bar{X}_i, \mathbf{W}_i)$ as in (7), we obtain the approximate model:

$$\begin{aligned} P(Y_{ij}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) &\approx G(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x E(X_{ij}|\bar{X}_i, \mathbf{W}_i)) \\ &= G(\beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x \bar{X}_i + \beta_x \alpha_w' (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)), \end{aligned} \quad (8)$$

where the second step follows by substituting (5). This is actually the exact model for the identity link, i.e., when G is the identity function. For other link functions, the quality of the approximation depends on the magnitude of the calibration error, which is determined by σ^2 : the smaller σ^2 is, the better the approximation tends to work. Methods based on model (8) will be referred to as plug-in methods. It is shown in Appendix B that the Y_{ij} in the same pool are generally correlated with each other given $(\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ because of the correlation

among the X_{ij} ($j = 1, \dots, m$) characterized by (6). The generalized estimating equations (GEE) approach (Liang and Zeger, 1986) may be used to account for this correlation. The negative correlation in (6) suggests, heuristically, that the Y_{ij} ($j = 1, \dots, m$) might be negatively correlated as well, given $(\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$. In view of the simple exchangeable structure in (6), one might be tempted to specify the same correlation structure for the Y_{ij} ($j = 1, \dots, m$) in a GEE analysis. It is not clear, however, that the correlation among the Y_{ij} ($j = 1, \dots, m$), given $(\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$, is exchangeable or follows any simple structure. Our numerical experience suggests that the working correlation structure has little impact on the performance of the procedure.

We propose two ways to configure the covariates for GEE analysis of model (8):

Augmentation

Fit the model with the augmented covariate vector $(1, \mathbf{Z}'_{ij}, \bar{X}_i, (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)')$ and then focus on the first three covariate components for inference about β , essentially treating $(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)$ as a nuisance component. This method is easy to implement but not necessarily efficient because it ignores the information about β_x in the last covariate component and also the available information about α_w .

Imputation

Combine the last two terms in model (8) and fit the model with the covariate vector $(1, \mathbf{Z}'_{ij}, \hat{X}_{ij})'$, where $\hat{X}_{ij} = \bar{X}_i + \hat{\alpha}'_w (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)$ and $\hat{\alpha}_w$ is an estimate of α_w obtained from a least squares regression of $\bar{X}_i$ on $(1, \bar{\mathbf{W}}_i')$. The estimate $\hat{\alpha}_w$ is justified by the relationship

$$\bar{X}_i = \alpha_1 + \alpha'_w \bar{\mathbf{W}}_i + \bar{\epsilon}_i, \quad (9)$$

which is immediate from (3). Note that $\hat{X}_{ij}$ is the predicted value of X_{ij} based on $(\bar{X}_i, \mathbf{W}_i)$ and the estimate $\hat{\alpha}_w$, hence the term imputation. This method does incorporate more information into estimation of β than the augmentation method.

For the augmentation method, inference about β can be based directly on standard errors from the GEE analysis, bearing in mind the approximate nature of model (8). For the imputation method, standard asymptotic theory can be used to account for the data-dependent nature of the $\hat{X}_{ij}$. The least squares estimator $\hat{\alpha}$ based on (9) solves the equation $\sum_{i=1}^n U_\alpha(\bar{\mathbf{W}}_i, \bar{X}_i; \alpha) = \mathbf{0}$ with $U_\alpha(\bar{\mathbf{W}}_i, \bar{X}_i; \alpha) = (1, \bar{\mathbf{W}}_i')' (\bar{X}_i - \alpha_1 - \alpha'_w \bar{\mathbf{W}}_i)$. The GEE estimator $\hat{\beta}$ solves the equation $\sum_{i=1}^n U_\beta(\mathbf{W}_i, \bar{X}_i, \mathbf{Z}_i, \mathbf{Y}_i; \hat{\alpha}, \beta) = \mathbf{0}$ with

$$U_\beta(\mathbf{W}_i, \bar{X}_i, \mathbf{Z}_i, \mathbf{Y}_i; \hat{\alpha}, \beta) = \left(\frac{\partial G_i(\alpha, \beta)}{\partial \beta} \right)' \mathbf{C}_i^{-1} \{ \mathbf{Y}_i - G_i(\alpha, \beta) \}, \quad (10)$$

where $G_i(\alpha, \beta) = G[\beta_1 + \mathbf{Z}_i \beta_z + \beta_x \{ \bar{X}_i + (\mathbf{W}_i - \bar{\mathbf{W}}_i) \alpha_w \}]$ and $\mathbf{C}_i$ is the working covariance matrix. The joint estimating function for $(\hat{\alpha}, \hat{\beta})'$ is therefore $\mathbf{U} = (\mathbf{U}'_\alpha, \mathbf{U}'_\beta)'$. Denote by β^* the value of β that minimizes the Kullback-Leibler divergence of model (8) from the true distribution of the data. If model (8) is correct (e.g., for the identity link), then β^* is just the

true value of β . In general, β^* is the best approximation under model (8) to the true value of β . As $n \rightarrow \infty$ with m fixed, $\sqrt{n}\{(\hat{\alpha}', \hat{\beta}') - (\alpha', \beta^*)'\}$ converges to a normal distribution with mean zero and variance given by $\Sigma = \mathbf{D}\text{var}(\mathbf{U})\mathbf{D}'$, where $\mathbf{D} = [\mathbf{E}\{\partial \mathbf{U} / \partial (\alpha, \beta)\}]^{-1}$, under standard regularity conditions (e.g., van der Vaart, 1998; Stefanski and Boos, 2002). A consistent estimate of Σ could be obtained by deriving explicit expressions for the above display and then substituting parameter estimates and sample means and variances. Alternatively, one could always resort to the bootstrap methodology for variance estimation, treating each pool as a unit for resampling.

4. Normality-based methods

In this section, we consider how to improve the approximate model (8) with a normality assumption. In addition to the previous assumptions concerning model (3), we assume that the error term ε is normally distributed. The normality assumption can be checked through model (9), and it implies that, given $(\bar{X}_i, \mathbf{W}_i)$, $\mathbf{X}_i$ follows a multivariate normal distribution with mean and variance given by (5) and (6), respectively. This makes it possible to carry out a maximum likelihood analysis based on the following decomposition:

$$\begin{aligned} f(\bar{X}_i, \mathbf{Y}_i | \mathbf{Z}_i, \mathbf{W}_i) &= f(\bar{X}_i | \mathbf{Z}_i, \mathbf{W}_i) f(\mathbf{Y}_i | \bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) \\ &= f(\bar{X}_i | \mathbf{W}_i) \mathbf{E}\{f(\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) | \bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i\} \\ &= f(\bar{X}_i | \bar{\mathbf{W}}_i; \alpha, \sigma^2) \int \left\{ \prod_{j=1}^m f(Y_{ij} | X_{ij}, Z_{ij}; \beta) \right\} f(\mathbf{X}_i | \bar{X}_i, \mathbf{W}_i; \alpha, \sigma^2) d\mathbf{X}_i, \end{aligned}$$

where $f(\cdot)$ denotes a generic conditional density or mass function. The full likelihood for $(\alpha, \beta, \sigma^2)$ can be maximized using a numerical integration method (of dimension $m - 1$) or a Monte Carlo EM algorithm. Either way, the computation can be challenging when m is large. Our regression calibration approach based on (7) has a pseudo-likelihood flavor in that we estimate (α, σ^2) from the first part of the likelihood (for $\bar{X}_i$) and then substitute the estimates into the second part of the likelihood (for $\mathbf{Y}_i$). Further, we take a marginal approach to the likelihood for $\mathbf{Y}_i$, working with each individual Y_{ij} and accounting for the correlation using GEEs. The regression calibration approach is computationally simple and, according to Thoresen and Laake (2000) and Fearn et al. (2008), may provide a good alternative to the maximum likelihood approach. For our purpose, the normality in model (3), and hence in equations (5) and (6), is mainly to help characterize the calibrated model (7). This will be done for specific link functions using existing techniques (e.g., Carroll et al., 1984; Burr, 1988; Liang and Liu, 1991; Follman et al., 1999).

The idea may be easier to communicate for the log link. In this case, G is the exponential function, and we can evaluate the integral in (7) explicitly as

$$\begin{aligned} P(Y_{ij}=1 | \bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) &= \exp(\beta_1 + \beta'_z \mathbf{Z}_{ij}) \mathbf{E}\{\exp(\beta'_x \mathbf{X}_{ij}) | \bar{X}_i, \mathbf{W}_i\} \\ &= \exp\{\beta_{\log,1} + \beta'_{\log,z} \mathbf{Z}_{ij} + \beta_{\log,x} \bar{X}_i + \beta_{\log,w} (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)\}, \end{aligned} \quad (11)$$

where $\beta_{\log} = (\beta_{\log,1}, \beta'_{\log,z}, \beta_{\log,x})'$ is identical to β except for the intercept:

$\beta_{\log,1} = \beta_1 + (m - 1)\beta_x^2 \sigma^2 / (2m)$. This follows from the fact that $\mathbf{E}\{\exp(S)\} = \exp(\mu + \tau^2/2)$ if S is normally distributed with mean μ and variance τ^2 . The above expression does resemble model (8) (with $G = \exp$), the only difference being the value of the intercept. For the probit link, it can be shown as in Carroll et al. (1984) that

$$P(Y_{ij}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = \Phi\{\beta_{\text{probit},1} + \beta'_{\text{probit},z} \mathbf{Z}_{ij} + \beta_{\text{probit},x} \bar{X}_i + \beta_{\text{probit},x} \alpha'_w (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)\}, \quad (12)$$

where $\beta_{\text{probit}} = (\beta_{\text{probit},1}, \beta'_{\text{probit},z}, \beta_{\text{probit},x})' = \{1 + (m-1)\beta_x^2 \sigma^2 / m\}^{-1/2} \beta$ and Φ is the standard normal distribution function. Thus the calibrated model is also a probit regression, and β_{probit} is a rescaling of β with all components attenuated by the same factor. As a result of the attenuation, $|\beta_{\text{probit},x}|$ is bounded from above by $\{m/(m-1)\}^{1/2}/\sigma$, and the upper bound decreases to 0 with increasing σ . The logit link does not admit a simple exact expression for (7), unfortunately. In this case, the one-dimensional integral in (7) can be approximated using numerical integration or Monte Carlo techniques, as noted by the Associate Editor. This approach could be applied to any link function and any distribution of calibration errors. An alternative approach, which is specific to the logit link but readily available in standard packages, is based on an approximation of logit^{-1} , the standard logistic distribution function, by $\Phi(\cdot/c)$ with $c = 15\pi/(16\sqrt{3}) \approx 1.70$ (Johnson and Kotz, 1970; Zeger et al., 1988; Liang and Liu, 1991). Under the latter approach, we have

$$\begin{aligned} P(Y_{ij}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) &= \int \text{logit}^{-1}(\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x x_{ij}) f(x_{ij}|\bar{X}_i, \mathbf{W}_i) dx_{ij} \\ &\approx \int \Phi\{(\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x x_{ij})/c\} f(x_{ij}|\bar{X}_i, \mathbf{W}_i) dx_{ij} \\ &= \Phi\left(\frac{\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x \bar{X}_i + \beta_x \alpha'_w (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)}{c\{1 + (m-1)\beta_x^2 \sigma^2 / (mc^2)\}^{1/2}}\right) \\ &\approx \text{logit}^{-1}\{\beta_{\text{logit},1} + \beta'_{\text{logit},z} \mathbf{Z}_{ij} + \beta_{\text{logit},x} \bar{X}_i + \beta_{\text{logit},x} \alpha'_w (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)\}, \end{aligned} \quad (13)$$

where $\beta_{\text{logit}} = (\beta_{\text{logit},1}, \beta'_{\text{logit},z}, \beta_{\text{logit},x})' = \{1 + m^{-1}(m-1)\beta_x^2 \sigma^2 c^{-2}\}^{-1/2} \beta$. As in the probit case, β_{logit} is an attenuation of β by a factor that further depends on c . The approximation here is good for small or moderate σ (Zeger et al., 1988). For large σ , it may be preferable to use a numerical iteration or Monte Carlo method.

Each of the three models (11)–(13) can be fit with GEEs, using either the augmentation method or the imputation method described earlier. In each case, the “working” regression parameter vector β_{link} , where the subscript could be any of the three link functions considered above, is zero if and only if $\beta = \mathbf{0}$, and the same equivalence holds between the corresponding sub-vectors of β_{link} and β , except for the intercept in the log link case. Therefore, in most cases, the null hypothesis that some components of β are zero can be tested (approximately in the logit case) by applying a standard procedure (e.g., Wald test) to the corresponding components of β_{link} . Estimation of β may involve some conversion from a GEE estimate of β_{link} . For the log link, this is needed only for $\beta_1 = \beta_{\text{log},1} - (m-1)\beta_x^2 \sigma^2 / (2m)$. In the probit case, the conversion consists of $\beta_x = \text{sign}(\beta_{\text{probit},x}) \{\beta_{\text{probit},x}^{-2} - (m-1)\sigma^2 / m\}^{-1/2}$ followed by $\beta = \{1 + (m-1)\beta_x^2 \sigma^2 / m\}^{1/2} \beta_{\text{probit}}$. Numerically, this requires that

$$\hat{\beta}_{\text{probit},x}^{-2} > (m-1)\hat{\sigma}^2 / m, \quad (14)$$

where $\hat{\beta}_{\text{probit},x}$ is the GEE estimate and $\hat{\sigma}^2$ is m times the estimated error variance for model (9). When (14) fails, intuition suggests that one might simply set $\hat{\beta}_x = \hat{\beta}_{\text{probit},x}$, although there are other possibilities (Burr, 1988; Tosteson et al., 1989). This would not alter the consistency or asymptotic normality of $\hat{\beta}$. Indeed, under the stated assumptions the

probability of (14) failing tends to 0, and any modification for that case would have no effect asymptotically. The logit case could be handled in a similar way.

Under the assumed models, the estimators derived from models (11)–(13) are consistent (approximately in the logit case) and asymptotically normal. For the augmentation method, this follows directly from the delta method. For the imputation method, we could use a joint estimating function argument as in Section 3, followed by the delta method. Robust variance estimates from an augmented GEE analysis are directly applicable to β_z and β_x under the log link, and require some delta-method-type manipulation under the probit or logit link. In general, a bootstrap procedure with poolwise resampling can be used to obtain standard errors and confidence intervals.

5. Simulation experiments

The proposed methods are evaluated in finite samples using simulated data. In the following experiments, Z and V (both scalars) are generated independently as standard normal variables, and X follows the calibration model (2) with $\mathbf{W} = (Z, V)'$ and $\alpha = (0, 1, -1)'$. The calibration error ε in (3) is independent of $\mathbf{W}$ with mean 0 and standard deviation $\sigma = 0.5$ or 1, and its distribution can be normal or skewed (according to a shifted exponential model to be specified later). The binary outcome Y depends on Z and X but not directly on V , as we assume in (4). The distribution of Y follows model (1) with a logit link, and the value of β will be specified in each experiment. In each sample, a total of 400 subjects are simulated independently according to the above mechanism, and are grouped into pools of size four, first randomly (Tables 1, 2) and then in a stratified fashion (Table 3). The X -values are averaged within pools. In each scenario (combination of parameter values), 1000 samples are generated.

Each simulated sample is analyzed using the following six methods. FD is a logistic regression analysis based on the full data including the value of X for each individual subject. This is not a competing method for pooled data, but it serves as an indicator for the amount of information in the data generated. NV is the naive approach that simply substitutes $\bar{X}_i$ for the unobserved values X_{ij} . The regression calibration methods will be denoted by a letter indicating the main approach (P for plug-in; N for normality-based) followed by another letter indicating the covariate configuration (A for augmentation; I for imputation). Thus, for example, NI denotes the normality-based imputation method, and there are three other combinations. The normality-based methods have been implemented using a numerical integration method as well as the approximation given by (13). The former implementation appears to work better but the difference tends to be small in the scenarios we consider (see the next paragraph). The reported results are based on model (13), which requires less computation. Implementation of the regression calibration methods in the GEE framework also requires specification of a working correlation structure. We have compared several correlation structures (exchangeable, independence and AR1) in preliminary simulations and have found no real impact on the performance of the methods. All numerical results in this paper are obtained under working independence.

The methods are compared in terms of empirical bias and standard deviation for estimating each component of $\beta = (\beta_1, \beta_z', \beta_x)'$. Table 1 presents the results for normally distributed errors in the calibration model (2). The first two scenarios, with $\sigma = 0.5$ and $\beta_x = 1$, clearly demonstrate the relevance of the proposed methods, which effectively remove most of the bias associated with the naive approach. (Given the observed variability, an empirical bias of 0.03 would be significantly different from 0, and so would a bias difference of 0.06 between any two methods.) As expected, the regression calibration estimators are more variable than the FD and NV estimators. In the latter case, the increase in variability seems a reasonable

price to pay for bias reduction. The imputation estimators, designed to improve upon their augmentation counterparts for better efficiency, appear to have achieved that objective. Another general observation is that the normality-based methods tend to be more variable than the plug-in methods, which is not surprising given the additional correction involved in the former approach. It may also be of interest to note that the intercept β_1 appears less susceptible to bias, even under the naive approach, when its value is 0 as opposed to 1. This is not the case for β_z . In the next two scenarios with $\beta_x = 0$, the bias problem is non-existent, not only for β_x but for all of β . This makes intuitive sense because, if Y does not depend on X , then it should not matter how the missing value of X is handled. For this reason, we will set $\beta_x = 1$ in most of the subsequent experiments. With larger errors ($\sigma = 1$), the plug-in methods become visibly biased. This should be expected because the plug-in methods ignore the inevitable errors around the conditional mean when predicting the X_{ij} with the available covariate information. The normality-based methods are still quite effective in these scenarios, which is no surprise either because they are based on correct distributional assumptions about the errors. In fact, using a numerical integration method (mentioned in the last paragraph), a further reduction of 0.01–0.03 in both bias and standard deviation (relative to NI) could be obtained in the last two scenarios ($\sigma = 1$) (results not shown).

The real question concerning the normality-based methods is how they perform when the calibration error ε is not normally distributed. Our numerical experience with logistic and Laplace distributions (results not shown) suggests that bias levels similar to those seen in Table 1 should be expected under symmetric, unimodal distributions. Therefore, a skewed distribution for ε is chosen for the next set of simulations. Specifically, we adopt a shifted exponential model where $\sigma^{-1}\varepsilon + 1$ is exponentially distributed with mean 1. The results are presented in Table 2. Since the distribution of X is no longer symmetric, here we also consider a negative intercept $\beta_1 = -1$ in addition to the values (0, 1) in Table 1. With $\sigma = 0.5$, Table 2 shows that all four regression calibration methods perform well with negligible bias relative to the naive approach. These methods become more biased when $\sigma = 1$, but they clearly outperform the naive approach. Despite serious misspecification of the error distribution, the normality-based methods still perform relatively well, with bias usually less than 0.1.

Table 3 illustrates the effects of stratified pooling in the case of normal calibration errors. Here the pooling is stratified by the signs of Z and V and is implemented as follows. Once a list of 400 subjects is generated randomly, it is ordered by the signs of Z and V into four blocks, each of which is homogeneous with respect to the signs of these variables. We then go through the list and group every four subjects into a pool. Because the sizes of the strata are random, there will be at most three heterogeneous pools which consist of subjects from different strata. In Table 3, the naive method is clearly less biased than in Table 1, which is reasonable to expect because the X_{ij} in a homogeneous pool will tend to be similar to each other and therefore to $\bar{X}_j$. The other methods have not changed much in terms of bias. It is worth noting that the regression calibration methods have generally become more efficient for estimating β_x and β_z . The efficiency gain is quite dramatic for the augmentation methods, which are still less efficient than the imputation methods though the difference is smaller. Simulation results for stratified pooling with skewed calibration errors follow the same pattern observed in Table 3 and are omitted.

We have also experimented with different sample sizes and pool sizes, and the results are qualitatively similar to those reported here.

6. A real example

The Upstate KIDS study motivating this research is still ongoing, and the component concerning PCB exposure is in the design stage, so the data are not yet available to illustrate and compare the various methods. Here we apply the methods to a similar dataset analyzed and reported by Daniels et al. (2003), who evaluated the association between prenatal PCB exposure and children's neurodevelopment at eight months. The data were collected as part of the Collaborative Perinatal Project (CPP), a large cohort study that followed the growth and neurologic development of approximately 55000 children born in 1959 through 1966 in the United States. The PCB measurements in the CPP study were based on third-trimester maternal serum samples, and not on neonatal blood spots as in the Upstate KIDS study. However, Rogan et al. (1986) have shown that PCB levels tend to be strongly correlated between maternal and cord serum at birth ($r > 0.7$). The Bayley Scales of Infant Development was used to measure children's mental and psychomotor development, and the scores were standardized to have population mean 100 and standard deviation 16.

From Daniels et al. (2003), a simple random sample of 1065 children is available for our analysis. Daniels et al. (2003) have found that the association of PCB exposure with psychomotor scores varies by study center, which implies an interaction between Z and X if we adjust for center as a covariate in Z . This makes it difficult to work with model (1), which does not allow interactions between Z and X . One way to address this issue is to fit model (1) to each center separately and interpret the results as a stratified analysis. In what follows we focus on the 222 children at the largest center (Boston, MA). The binary outcome of interest to us is abnormality in psychomotor development, which may be defined as the Bayley Psychomotor Development Index (PDI) being lower than the population mean by more than two standard deviations. Given the standardized nature of the index, this means that a child will be considered abnormal for our purpose if his or her PDI is below 68. This will be regressed through a logit link on PCB level while adjusting for the mother's education and triglyceride level as well as the child's birth order as in Daniels et al. (2003). Without pooling, the log-odds ratio for abnormal psychomotor development due to an increase of 1 $\mu\text{g/liter}$ in PCB level is estimated to be -0.010 with standard error 0.092. This is consistent with the linear regression analysis of Daniels et al. (2003) which found no significant association between PCB level and PDI as a continuous variable.

It is easy to pool the subjects artificially, average the PCB measurements in each pool, and then apply the various methods for pooled data. However, the results based on a single pooled dataset could be arbitrary and difficult to interpret. Thus it is more informative to repeat this pooling process many times in evaluating the different approaches. Table 4 gives a summary based on 1000 repetitions using the same labels for the different methods as in the previous section. Here the regression calibration methods are implemented with the mother's race, education and cholesterol level as predictors for her PCB level. Each method produces, over repeated pooling, 1000 estimates of the log-odds ratio (β_x) relating low PDI to PCB, which are then averaged and compared with the full-data estimate. We also calculate standard deviations (SDs) in an analogous fashion. The pool size is either 3 or 6, and the pooling process is either completely random or stratified by the discrete components of W (i.e., race and education). Specifically, we consider three strata: African American (AA, 24 subjects), non-AA with at most high school education (158), non-AA with education beyond high school (40). The first stratum (AA) appears too small to be further stratified by education. In the case of random pooling, the regression calibration methods lead to estimates that are closer to the full-data estimate on average. The augmentation methods appear more variable than the naive method, but the imputation methods have smaller SDs. There is little difference between the plug-in methods and the normality-based methods, because the data suggest that β_x is very small (which implies that $\beta_{\text{logit}} \approx \beta$).

Surprisingly, the results for stratified pooling are generally inferior to those for random pooling (with some exceptions, of course). We should be careful, though, not to over-interpret the results based a single dataset (even though the pooling can be repeated). We have also performed some simulations for this particular application, using the original covariate values ($\mathbf{Z}$, $\mathbf{W}$) and generating (X, Y) under values of $(\alpha, \beta, \sigma^2)$ that are estimated from the full data. The results are similar to those in the previous section for the case $\beta_x = 0$, and are therefore omitted.

7. Uneven pooling

We have assumed so far that the pool size is constant and that the weighting within each pool is uniform. Real life is more complicated than this. In the Upstate KIDS study, for example, the number of punches available from each subject is variable and difficult to predict, and consequently, the number of subjects to pool together cannot be fixed, and the subjects in the same pool need not contribute equally. We now consider more flexible pooling schemes that allow uneven pooling. We write m_i for the size of the i th pool, and

consider a pooled measurement a weighted average of the form $\tilde{X}_i = \sum_{j=1}^{m_i} p_{ij} X_{ij} = \mathbf{p}_i' \mathbf{X}_i$. The weight vector $\mathbf{p}_i$ can be arbitrary as long as it has positive elements adding to one. We allow $\mathbf{p}_i$ to depend on $\mathbf{W}_i$ stochastically, but will treat the $\mathbf{p}_i$ as fixed in the following discussion, which is entirely conditional on the $\mathbf{W}_i$.

Except for a more flexible pooling scheme, we retain the previous assumptions including (1)–(4) as well as the normality assumption in Section 4. Some implications of the new pooling scheme should be noted before we consider inference for specific link functions. First, estimation of the calibration model is now based on the relationship

$$\tilde{X}_i = \alpha_1 + \alpha_w' \tilde{\mathbf{W}}_i + \tilde{\varepsilon}_i, \quad (15)$$

where $\tilde{\mathbf{W}}_i = \mathbf{W}_i' \mathbf{p}_i$ and $\tilde{\varepsilon}_i = \mathbf{p}_i' \boldsymbol{\varepsilon}_i = \sum_{j=1}^{m_i} p_{ij} \varepsilon_{ij}$. Since $\tilde{\varepsilon}_i$ has variance $\sigma^2(\mathbf{p}_i' \mathbf{p}_i)$, which varies across pools, a weighted least squares procedure that assigns weight $(\mathbf{p}_i' \mathbf{p}_i)^{-1}$ to the i th pool should be used to fit model (15) for efficient estimation of α and consistent estimation of σ^2 . Second, the conditional distribution of $\mathbf{X}_i$ given $(\tilde{X}_i, \mathbf{W}_i)$ is multivariate normal with mean and variance given by

$$\begin{aligned} E(\mathbf{X}_i | \tilde{X}_i, \mathbf{W}_i) &= \alpha_1 \mathbf{1}_{m_i} + \mathbf{W}_i \alpha_w + (\mathbf{p}_i' \mathbf{p}_i)^{-1} (\tilde{X}_i - \alpha_1 - \alpha_w' \tilde{\mathbf{W}}_i) \mathbf{p}_i, \\ \text{var}(\mathbf{X}_i | \tilde{X}_i, \mathbf{W}_i) &= \sigma^2 \{ \mathbf{I}_{m_i} - (\mathbf{p}_i' \mathbf{p}_i)^{-1} \mathbf{p}_i \mathbf{p}_i' \}. \end{aligned}$$

This follows from the joint normal distribution of $(\mathbf{X}_i', \tilde{X}_i)'$ given $\mathbf{W}_i$.

For the identity link, it is easy to see that

$$P(Y_{ij} = 1 | \tilde{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = \beta_1 + \beta_z' \mathbf{Z}_{ij} + \beta_x \tilde{X}_{ij}(\alpha),$$

where $\widehat{X}_{ij}(\alpha) = \alpha_1 + \alpha'_w \mathbf{W}_{ij} + p_{ij}(\mathbf{p}'_i \mathbf{p}_i)^{-1} (\tilde{X}_i - \alpha_1 - \alpha'_w \tilde{\mathbf{W}}_i)$. This suggests an imputation method that replaces each X_{ij} with $\widehat{X}_{ij}(\hat{\alpha})$, where $\hat{\alpha}$ denotes the weighted least squares estimate for model (15). A GEE analysis of model (1) with identity link can then be performed using the imputed values $\widehat{X}_{ij}(\hat{\alpha})$.

For the log link, it can be shown as before that

$$P(Y_{ij}=1|\tilde{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = \exp[\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x \widehat{X}_{ij}(\alpha) + \beta_x^2 \sigma^2 \{1 - p_{ij}^2(\mathbf{p}'_i \mathbf{p}_i)^{-1}\}/2]. \quad (16)$$

In this case, β could be estimated from a standard GEE analysis regressing Y_{ij} on the covariate vector $(1, Z_{ij}, \widehat{X}_{ij}(\hat{\alpha}), \widehat{\sigma}^2 \{1 - p_{ij}^2(\mathbf{p}'_i \mathbf{p}_i)^{-1}\}/2)$, where $\widehat{\sigma}^2$ is also obtained from a weighted least squares analysis of model (15). Note that the last element of the covariate vector depends on (i, j) and cannot be absorbed into the intercept, unless $m_i \equiv m$ and $p_{ij} \equiv 1/m$ as in Section 4. This extra covariate, with regression coefficient β_x^2 , contains relevant information about β_x , which ideally should be incorporated into estimation of β . One way to do that is to modify the GEE code specially for model (16) by computing its derivatives with respect to β , etc.

For the probit link, we have

$$P(Y_{ij}=1|\tilde{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = \Phi \left(\frac{\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x \widehat{X}_{ij}(\alpha)}{[1 + \beta_x^2 \sigma^2 \{1 - p_{ij}^2(\mathbf{p}'_i \mathbf{p}_i)^{-1}\}]^{1/2}} \right).$$

Because the denominator in $\Phi(\cdot)$ also depends on (i, j) , it is no longer feasible to simply “absorb” it into the regression coefficients and apply the conversion formulae of Section 4. A custom GEE program could be developed to fully utilize the special structure of the problem. Once again, the logit case can be handled by numerical integration or by the aforementioned approximation by the probit link. In the latter case, the approximate model is given by

$$P(Y_{ij}=1|\tilde{X}_i, \mathbf{Z}_i, \mathbf{W}_i) \approx \text{logit}^{-1} \left(\frac{\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x \widehat{X}_{ij}(\alpha)}{[1 + \beta_x^2 \sigma^2 \{1 - p_{ij}^2(\mathbf{p}'_i \mathbf{p}_i)^{-1}\}/c^2]^{1/2}} \right).$$

8. Discussion

It has become increasingly common in epidemiologic studies to pool specimens for quantitative measurement of biomarkers and environmental chemicals, often due to limits on assay sensitivity or sample availability. These pooled measurements often provide valuable information about an important exposure, which should be used effectively and efficiently to recover the individual-level regression relationship. Here we take a regression calibration approach to this problem and derive several methods for binary outcome data. Theoretical considerations and numerical results suggest that the regression calibration methods generally perform better than the naive method that simply substitutes a pooled measurement for all individual measurements in the pool. In particular, the normality-based

imputation method appears quite effective in a variety of situations, even with non-normal calibration errors. The methods extend easily to variable pool sizes and unequal weights within pools.

Some issues remain to be addressed. It has been assumed here that pooling is uninformative (possibly dependent on W but independent of everything else given W). While this may be appropriate for the Upstate KIDS study, it is conceivable that pooling can be informative and may even depend on the outcome in some studies. It will be helpful to develop methods that deal with informative pooling and maybe even efficient designs that take advantage of the informativeness. Another issue is that the pooling weights, denoted by p_i in Section 7, may be unobservable in some biomarker studies. It is not yet clear how to deal with such “blinded” pooling.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The authors would like to thank Drs. Enrique Schisterman and Mary Hediger for helpful discussions, Drs. Matthew Longnecker, Zhen Chen and Mark Klebanoff for providing the CPP data, and the Associate Editor and an anonymous referee for insightful comments that improved the manuscript.

References

- Albert, PS.; Shih, JH. Modeling batched Gaussian longitudinal data subject to informative dropout. 2009. Unpublished manuscript
- Burr D. On errors-in-variables in binary regression-Berkson case. *Journal of the American Statistical Association*. 1988; 83:739–743.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2. London: Chapman & Hall; 2006.
- Carroll RJ, Spiegelman CH, Lan KKG, Kent KT, Abbott RD. On error-in-variables for binary regression models. *Biometrika*. 1984; 71:19–25.
- Daniels JL, Longnecker MP, Klebanoff MA, Gray KA, Brock JW, Zhou H, Chen Z, Needham LL. Prenatal exposure to low-level polychlorinated biphenyls in relation to mental and motor development at 8 months. *American Journal of Epidemiology*. 2003; 157:485–492. [PubMed: 12631537]
- Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine*. 2003; 22:2515–2527. [PubMed: 12872306]
- Fearn T, Hill DC, Darby SC. Measurement error in the explanatory variable of a binary regression: Regression calibration and integrated conditional likelihood in studies of residential radon and lung cancer. *Statistics in Medicine*. 2008; 27:2159–2167. [PubMed: 18081195]
- Follman DA, Hunsberger SA, Albert PS. Repeated probit regression when covariates are measured with error. *Biometrics*. 1999; 55:403–409. [PubMed: 11318193]
- Johnson, NL.; Kotz, S. *Distributions in Statistics*. Vol. 2. Boston: Houghton-Mifflin; 1970.
- Liang, K-Y.; Liu, X-H. Estimating equations in generalized linear models with measurement error. In: Godambe, AP., editor. *Estimating Functions*. Oxford: Clarendon Press; 1991.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal*. 2003; 45:631–644.
- Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics*. 2006; 7:585–598. [PubMed: 16531470]
- Rogan WJ, Gladen BC, McKinney JD, Carreras N, Hardy P, Thullen J, Tingelstad J, Tully M. Polychlorinated biphenyls (PCBs) and dichlorodiphenyl dichloroethene (DDE) in human milk:

- effects of maternal factors and previous lactation. *American Journal of Public Health*. 1986; 76:172–177. [PubMed: 3080910]
- Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Pediatric and Perinatal Epidemiology*. 2008; 22:486–496. [PubMed: 18782255]
- Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 29:597–613. [PubMed: 20049693]
- Stefanski LA, Boos DD. The calculus of M-estimation. *The American Statistician*. 2002; 56:29–38.
- Thoresen M, Laake P. A simulation study of measurement error correction methods in logistic regression. *Biometrics*. 2000; 56:868–872. [PubMed: 10985228]
- Tosteson TD, Stefanski LA, Schafer DW. A measurement-error model for binary and ordinal regression. *Statistics in Medicine*. 1989; 8:1139–1147. [PubMed: 2799134]
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 1998.
- Vexler A, Liu A, Schisterman EF. Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biometrical Journal*. 2006; 48:780–791. [PubMed: 17094343]
- Vexler A, Schisterman EF, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine*. 2008; 27:280–296. [PubMed: 17721905]
- Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
- Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988; 44:1049–1060. [PubMed: 3233245]

Appendix A. Derivation for equation (7)

A conditioning argument yields

$$P(Y_{ij}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = E\{P(Y_{ij}=1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i\}. \quad (\text{A.1})$$

Since the pooling depends only on the $\mathbf{W}$'s, we have that, given $\mathbf{W}_i = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{im})'$, the random vector $(X_{ij}, Y_{ij}, \mathbf{Z}_{ij})$ is distributed as $(X, Y, \mathbf{Z}|\mathbf{W} = \mathbf{w}_{ij})$, independently across subjects ($j = 1, \dots, m$). This allows us to write

$$P(Y_{ij}=1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = P(Y_{ij}=1|X_{ij}, \mathbf{Z}_{ij}, \mathbf{W}_{ij}) = G(\beta_1 + \beta'_z \mathbf{Z}_{ij} + \beta_x X_{ij}), \quad (\text{A.2})$$

where the second step follows from assumptions (4) and (1). To evaluate (A.1), we can integrate (A.2) with respect to the conditional distribution of $(X_{ij}|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$. Assumption (2) implies that X_i and $\mathbf{Z}_i$ are conditionally independent given $\mathbf{W}_i$. Thus $(\mathbf{Z}_i|\mathbf{W}_i)$ is distributed as $(\mathbf{Z}_i|\bar{X}_i, \mathbf{W}_i) = (\mathbf{Z}_i|\bar{X}_i, X_i, \mathbf{W}_i)$. This shows that $\mathbf{Z}_i$ and X_i are conditionally independent given $(\bar{X}_i, \mathbf{W}_i)$. Therefore $(X_i|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ is distributed as $(X_i|\bar{X}_i, \mathbf{W}_i)$, and the same relationship holds with X_i reduced to X_{ij} . Using this fact and (A.2), we can now rewrite (A.1) as (7).

Appendix B. Correlation among the Y_{ij} ($j = 1, \dots, m$) given $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$

For distinct subjects j_1, j_2 in the i -th pool, it can be argued as in Appendix A that

$$\begin{aligned}
P(Y_{ij_1}=1, Y_{ij_2}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) &= E\{P(Y_{ij_1}=1, Y_{ij_2}=1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i\} \\
&= E\{P(Y_{ij_1}=1|\mathbf{X}_{ij_1}, \mathbf{Z}_{ij_1}, \mathbf{W}_{ij_1}) \times P(Y_{ij_2}=1|\mathbf{X}_{ij_2}, \mathbf{Z}_{ij_2}, \mathbf{W}_{ij_2})|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i\} \\
&= \int G(\beta_1 + \beta'_z \mathbf{Z}_{ij_1} + \beta_x x_{ij_1}) G(\beta_1 + \beta'_z \mathbf{Z}_{ij_2} + \beta_x x_{ij_2}) \times f(x_{ij_1}, x_{ij_2}|\bar{X}_i, \mathbf{W}_i) dx_{ij_1} dx_{ij_2}.
\end{aligned}$$

which follows from assumptions (1)–(4). Because X_{ij_1} and X_{ij_2} are stochastically dependent given $(\bar{X}_i, \mathbf{W}_i)$, the above integral is generally different from

$$\int G(\beta_1 + \beta'_z \mathbf{Z}_{ij_1} + \beta_x x_{ij_1}) f(x_{ij_1}|\bar{X}_i, \mathbf{W}_i) dx_{ij_1} \int G(\beta_1 + \beta'_z \mathbf{Z}_{ij_2} + \beta_x x_{ij_2}) f(x_{ij_2}|\bar{X}_i, \mathbf{W}_i) dx_{ij_2} = P(Y_{ij_1}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i) P(Y_{ij_2}=1|\bar{X}_i, \mathbf{Z}_i, \mathbf{W}_i),$$

hence the correlation.

Table 1

Empirical bias and standard deviation (SD) for estimating $\beta=(\beta_1, \beta_z, \beta_x)'$ under random pooling with normal calibration errors (see Section 5 for details)

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$					
			β_1	β_z	β_x	β_1	β_z	β_x	β_1	β_z	β_x
0.5	(0, 0, 1)	FD	0	1	1	12	16	14			
		NV	0	72	-47	10	13	15			
		PA	0	0	-1	12	34	22			
		PI	0	1	-2	12	17	14			
		NA	0	-1	3	12	35	25			
		NI	0	0	2	12	18	16			
	(1, 1, 1)	FD	1	2	2	16	21	16			
		NV	-15	59	-47	14	19	19			
		PA	-2	1	-1	15	38	26			
		PI	-2	-2	-1	15	22	16			
		NA	3	4	3	17	38	30			
		NI	2	2	3	16	22	19			
	(0, 1, 0)	FD	0	1	0	11	17	10			
		NV	0	2	0	11	14	16			
		PA	0	2	0	11	31	20			
		PI	0	2	0	11	17	11			
		NA	0	2	0	11	31	20			
		NI	0	2	0	11	17	11			
	(1, 0, 0)	FD	0	1	0	11	15	11			
		NV	0	0	1	11	12	17			
		PA	0	2	1	12	31	21			
		PI	0	1	0	11	16	11			
		NA	1	2	1	12	32	21			
		NI	0	1	0	11	16	11			
1	(0, 1, 1)	FD	1	1	0	14	19	13			
		NV	0	45	-43	12	17	15			
		PA	0	-12	-10	13	33	20			

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$					
			β_1	β_z	β_x	β_1	β_z	β_x	β_1	β_z	β_x
		PI	0	-11	-11	13	21	15			
		NA	1	2	8	15	36	35			
		NI	1	2	5	15	23	26			
	(1, 0, 1)	FD	1	0	1	14	16	13			
		NV	-22	65	-44	11	13	13			
		PA	-11	1	-11	12	30	18			
		PI	-11	0	-11	12	18	14			
		NA	4	-1	5	19	36	30			
		NI	3	-1	5	18	22	25			

Table 2

Empirical bias and standard deviation (SD) for estimating $\beta=(\beta_1, \beta_z, \beta_x)'$ under random pooling with skewed calibration errors (see Section 5 for details)

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$					
			β_1	β_z	β_x	β_1	β_z	β_x	β_1	β_z	β_x
0.5	(0, 0, 1)	FD	0	0	1	12	17	14			
		NV	0	71	-48	10	13	16			
		PA	0	0	-2	11	32	22			
		PI	0	0	-3	11	18	14			
		NA	0	0	2	12	33	26			
		NI	0	0	1	12	19	16			
	(1, 1, 1)	FD	1	2	1	16	21	16			
		NV	-15	59	-47	14	19	19			
		PA	-2	2	-2	16	38	27			
		PI	-3	-1	-2	16	22	17			
		NA	2	6	2	17	39	30			
		NI	1	3	2	17	22	19			
1	(-1, 1, 1)	FD	-2	2	3	16	21	16			
		NV	14	59	-46	14	18	18			
		PA	0	0	-1	16	38	26			
		PI	1	-2	-1	15	22	17			
		NA	-4	4	3	17	39	30			
		NI	-3	2	4	17	22	19			
	(0, 1, 1)	FD	1	1	3	14	20	13			
		NV	-2	47	-43	12	17	15			
		PA	-2	-5	-11	13	32	21			
		PI	-2	-8	-9	13	21	15			
		NA	-2	10	7	15	36	37			
		NI	-2	7	7	15	24	26			
(1, 0, 1)		FD	1	-1	2	15	17	13			
		NV	-24	67	-47	11	13	14			
		PA	-13	9	-16	13	30	19			

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$		
			β_1	β_z	β_x	β_1	β_z	β_x
		PI	-13	3	-11	13	19	14
		NA	1	9	-2	19	35	31
		NI	1	2	5	18	23	24
	(-1, 0, 1)	FD	-1	-1	2	14	16	12
		NV	18	64	-42	11	13	13
		PA	6	-1	-10	12	30	18
		PI	7	-1	-10	12	18	14
		NA	-10	-2	6	19	36	29
		NI	-8	-2	5	17	21	22

Table 3

Empirical bias and standard deviation (SD) for estimating $\beta=(\beta_1, \beta_z, \beta_x)'$ under stratified pooling with normal calibration errors (see Section 5 for details)

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$					
			β_1	β_z	β_x	β_1	β_z	β_x	β_1	β_z	β_x
0.5	(0,0,1)	FD	0	1	1	1	12	17	14		
		NV	0	36	-28	11	14	13			
		PA	0	1	-2	12	19	15			
		PI	0	1	-3	12	17	14			
	(1,1,1)	NA	0	1	2	12	19	17			
		NI	0	1	1	12	18	16			
		FD	1	1	2	16	21	15			
		NV	-8	21	-27	14	18	15			
		PA	-2	-2	-1	15	23	17			
		PI	-2	-2	-1	15	21	16			
		NA	3	2	4	16	23	20			
		NI	2	1	3	16	21	18			
	(0,1,0)	FD	-1	1	0	11	17	10			
		NV	-1	2	-1	11	16	11			
		PA	-1	2	-1	11	20	13			
		PI	-1	1	0	11	18	11			
	(1,0,0)	NA	-1	2	-1	11	20	13			
		NI	-1	1	0	11	18	11			
		FD	1	0	0	11	15	10			
		NV	1	-1	1	11	14	11			
	(0,1,1)	PA	1	0	1	12	18	12			
		PI	1	0	0	11	15	11			
		NA	1	0	1	12	18	12			
		NI	1	0	0	12	15	11			
1	(0,1,1)	FD	0	0	2	14	19	14			
		NV	0	7	-31	12	16	12			
		PA	0	-11	-10	13	19	14			

σ	β'	Method	Bias $\times 10^2$			SD $\times 10^2$		
			β_1	β_z	β_x	β_1	β_z	β_x
(1,0,1)		PI	0	-12	-10	13	19	14
		NA	0	3	6	15	21	23
		NI	0	2	5	15	20	22
		FD	1	0	2	14	16	12
		NV	-17	32	-32	11	13	12
		PA	-11	-1	-10	12	17	14
		PI	-12	0	-11	12	16	13
		NA	4	-1	5	18	20	22
		NI	3	-1	4	17	19	21

Table 4Summary of estimates of β_x over repeated pooling in the example of Section 6

Pool Size	Method	Average	SD
	FD	-0.010	
<i>random pooling</i>			
3	NV	-0.023	0.13
	PA	-0.017	0.14
	PI	-0.013	0.10
	NA	-0.017	0.15
	NI	-0.013	0.10
6	NV	-0.021	0.23
	PA	-0.011	0.24
	PI	-0.003	0.13
	NA	-0.017	0.29
	NI	-0.004	0.14
<i>stratified pooling</i>			
3	NV	-0.036	0.14
	PA	-0.034	0.15
	PI	-0.009	0.13
	NA	-0.036	0.16
	NI	-0.009	0.14
6	NV	-0.041	0.20
	PA	-0.030	0.23
	PI	0.022	0.20
	NA	-0.035	0.28
	NI	0.024	0.22