Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Statistical Planning and Inference 131 (2005) 175–190

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A comparison of regression calibration approaches for designs with internal validation data

Sally W. Thurston[a,*], Paige L. Williams[b], Russ Hauser[c], Howard Hu[c,f], Mauricio Hernandez-Avila[d], Donna Spiegelman[b,e]

[a] *Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Avenue, P.O. Box 630, Rochester, NY 14642, USA*
[b] *Department of Biostatistics, Harvard School of Public Health, USA*
[c] *Department of Environmental Health, Harvard School of Public Health, USA*
[d] *Centro de Investigaciones en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Morelos, Mexico*
[e] *Department of Epidemiology, Harvard School of Public Health, USA*
[f] *Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, USA*

Received 21 January 2003; accepted 21 December 2003

Abstract

We compare the asymptotic relative efficiency of several regression calibration methods of correcting for measurement error in studies with internal validation data, when a single covariate is measured with error. The estimators we consider are appropriate in main study/hybrid validation study designs, where the latter study includes internal validation and may include external validation data. Although all of the methods we consider produce consistent estimates, the method proposed by Spiegelman et al. (Statistics in Medicine, 20 (2001) 139) has an asymptotically smaller variance than the other methods. The methods for measurement error correction are illustrated using a study of the effect of in utero lead exposure on infant birth weight.

Keywords: Bone lead; Errors-in-variables; Internal validation study; Measurement error

* Corresponding author. Tel.: +1-5852756695; fax: +1-5852731031.
E-mail address: thurston@bst.rochester.edu (S.W. Thurston).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.12.015

1. Introduction

Exposure measurement error is a common problem in many disciplines for which a "gold standard" exposure measure is difficult or expensive to obtain. In occupational epidemiology, for example, quantifying workplace exposure using a "proxy" measure, such as job title or an exposure measurement from an area sampler is more feasible and less expensive to obtain than a personal exposure measurement for each worker. Similarly, in reproductive epidemiology an exposure measure taken from the newborn tissue or blood may be easier to obtain but give a less accurate estimate of in utero exposure than a measurement from the mother which reflects long term exposure. In both cases, using the less accurate measure of exposure in an outcome model introduces bias in the point and interval estimates of the exposure–response relationship.

To adjust for measurement error, information about the measurement error process is needed. Consistent estimates of the exposure effect may be obtained in this setting if validation data is available, in which both the "gold standard" exposure and the "proxy" measure of exposure are measured on the same subjects. Methods considered in this paper require such validation data, as well as data from a main study in which the outcome and the "proxy" exposure measure are obtained. An internal validation study includes measurements of the outcome, as well as both "gold standard" and "proxy" measures of exposure, on a subset of study subjects. Subjects in an external validation study do not have measurements of the outcome, but have both "gold standard" and "proxy" exposure measurements. We use the term "hybrid validation" study to refer to validation data which includes internal validation data and may optionally also include external validation data. This paper addresses study designs which include both a main study and a hybrid validation study, i.e. a main study/hybrid validation study design. Of course, the main study/internal validation study design is a subset of this.

Several methods have been developed to adjust for the bias resulting when one or more regression model covariates is measured with error. Regression calibration is one method which is widely used and easy to implement. It has been discussed and applied in a variety of contexts, including logistic regression (Rosner, Willett and Spiegelman, henceforth RSW, 1989, 1990, 1992; Spiegelman et al., 2001), survival analysis (Prentice 1982; Spiegelman et al., 1997), and others (Armstrong, 1985; Carroll and Stefanski, 1990; Carroll, Ruppert and Stefanski, henceforth CRS, 1995, Chapter 3). When validation data are available, a measurement error model is fit from the regression of the "gold standard" exposure on the "proxy" exposure measure in the validation data. Adjustment for measurement error is achieved either by regressing the outcome on estimated conditional mean values from the measurement error model and then adjusting the variances (CRS, 1995), or by regressing the outcome on the "proxy" measure (RSW, 1989, 1990) and then using information from the measurement error model to correct the coefficients and their variances for bias due to covariate measurement error. For main study/external validation study designs, Thurston et al. (2003) demonstrated that the CRS and RSW methods give identical adjustments for the coefficients and their asymptotic variance under broad conditions. The focus of this paper, in contrast, is on designs which include internal validation data in the form of a hybrid validation study.

We assume linear regression models for relating both the outcome and the "gold standard" to the "proxy" measure. We further assume that the measurement error process is non-differential. We discuss four methods of correcting for measurement error with main study/hybrid validation designs. We compare the asymptotic relative efficiency of the estimators when the validation data is all internal. We motivate and illustrate these comparisons from a study of the effect of in utero lead exposure on infant birth weight. For these data, we give results using both asymptotic methods and the bootstrap (Efron and Tibshirani, 1993), the latter of which may give valid inference in situations when sample sizes are too small to allow valid asymptotic inference.

We start by presenting the notation and models in Section 2. In Section 3, we present the four estimators of the effect of the exposure on the outcome, along with their corresponding asymptotic variances and describe inference for these estimators using bootstrap methods. In Section 4 we compare the asymptotic relative efficiencies of the estimators. The various approaches are illustrated by applying them to the study of the effect of lead exposure on infant birth weight in Section 5. We conclude in Section 6 with discussion of these approaches.

2. Notation and models

For the ith subject, let y_i be the health outcome of interest, x_i be the true exposure of interest, w_i be the "proxy" exposure measured with error, and $Z_i=(z_{i1} \ \cdots \ z_{ip})$ be the p covariates assumed to be measured without error. For the main study/hybrid validation study design the data are of four possible types. The main study, with corresponding indicator $I_{\mathrm{m},i}=1$, consist of (y_i, w_i, Z_i). The internal validation study with indicator $I_{\mathrm{iv},i}=1$ may include observations of two types: (y_i, x_i, Z_i) and (y_i, x_i, w_i, Z_i). The optional external validation study with indicator $I_{\mathrm{ev},i}=1$ consists of (x_i, w_i, Z_i). With n_{m} main study subjects, n_{iv} internal validation subjects, and n_{ev} external validation subjects, the total number of subjects is $n=n_{\mathrm{m}}+n_{\mathrm{iv}}+n_{\mathrm{ev}}$. The methods of this paper assume $n_{\mathrm{m}}>0$ and $n_{\mathrm{iv}}>0$, whereas n_{ev} can equal 0. Methods for external validation study designs assume $n_{\mathrm{ev}}>0$, $n_{\mathrm{m}}>0$ and $n_{\mathrm{iv}}=0$. A study with pure internal validation data assumes $n_{\mathrm{ev}}=0$, which is the case with the lead data discussed in this paper.

We refer to the validation study with indicator $I_{\mathrm{v},i}=1$ as including any observation which has measurements of (x_i, w_i, Z_i). Let $I_{\mathrm{m}}=\mathrm{diag}\{I_{\mathrm{m},i}\}$, $I_{\mathrm{v}}=\mathrm{diag}\{I_{v,i}\}$, and $\mathbf{I}_{\mathrm{iv}}=\mathrm{diag}\{I_{\mathrm{iv},i}\}$ where diag{arg} indicates a diagonal matrix with arg on the diagonals. Let $W_j=(1 \quad w_i \quad Z_i)$ and $X_i=(1 \quad x_i \quad Z_i)$. Let $\mathbf{y}$ be the n-dimensional vector of y_i, $\mathbf{x}$ be the n-dimensional vector of x_i, $\mathbf{Z}$ be the $n\times p$ dimensional matrix of Z_i, $\mathbf{W}$ be the $n\times(p+2)$-dimensional matrix consisting of the n observations of W_i, and $\mathbf{X}$ be the $n\times(p+2)$-dimensional matrix consisting of the n observations of X_i. Note that some entries of $\mathbf{y}$ and $\mathbf{x}$ will be missing, since y_i or x_i are unobserved for some observations.

We assume that the data follow what we call the linear/linear model with one covariate measured with error as defined by

$$y_i=\beta_0+\beta_1 x_i+Z_i\beta_z+v_i=X_i\boldsymbol{\beta}+v_i, \tag{1}$$

$$x_i=\alpha_0+\alpha_1 w_i+Z_i\boldsymbol{\alpha}_z+u_i=W_i\boldsymbol{\alpha}+u_i, \tag{2}$$

where v_i has mean 0 and variance σ_v^2, and u_i has mean 0 and variance σ_u^2. In the models above, we implicitly define $\boldsymbol{\alpha} = (\alpha_0 \ \alpha_1 \ \boldsymbol{\alpha}_z^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\alpha}_z = (\alpha_{z1} \ \alpha_{z2} \ \cdots \ \alpha_{zp})^{\mathrm{T}}$, and $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \boldsymbol{\beta}_z^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_z = (\beta_{z1} \ \beta_{z2} \ \cdots \ \beta_{zp})^{\mathrm{T}}$. We distinguish between different estimators of $\boldsymbol{\beta}$ by a subscript. The primary objective of the analysis is to estimate β_1, the effect of the true exposure x_i on the health outcome y_i, adjusted for covariates Z_i.

In addition to the above notation, let $\hat{X}_i = (1 \ \hat{x}_i \ Z_i)$, where $\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 w_i + Z_i \hat{\boldsymbol{\alpha}}_z = W_i \hat{\boldsymbol{\alpha}}$, and estimation of $\boldsymbol{\alpha}$ is the standard least-squares estimator, as discussed in Appendix A. Let $\hat{\mathbf{X}}$ be the matrix of $\hat{X}_i$ for the n observations. Note that $\hat{X}_i = W_i \hat{\boldsymbol{\alpha}}_{\mathrm{mat}}$, where

$$\boldsymbol{\alpha}_{\mathrm{mat}} = \begin{pmatrix} 1 & \alpha_o & 0_{1\times p} \\ 0 & \alpha_1 & 0_{1\times p} \\ 0_{p\times 1} & \boldsymbol{\alpha}_z & I_{p\times p} \end{pmatrix}.$$

and $\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}$ is obtained by substituting $\hat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$ in $\boldsymbol{\alpha}_{\mathrm{mat}}$. We will also make use of the following relation:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + Z_i \boldsymbol{\beta}_z + \varepsilon_i = \hat{X}_i \boldsymbol{\beta} + \varepsilon_i = W_i \hat{\boldsymbol{\alpha}}_{\mathrm{mat}} \boldsymbol{\beta} + \varepsilon_i, \tag{3}$$

where ε_i has mean 0 and variance σ_ε^2.

3. Estimators of $\boldsymbol{\beta}$ and their asymptotic variances

We consider four methods for estimating $\boldsymbol{\beta}$ in main study/hybrid validation study designs.

1. $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ ("as external"): Ignore the information about the relationship between $\mathbf{y}$ and $\mathbf{X}$ in the internal validation sample, and treat the study as a main study/external validation study design (CRS 1995, RSW 1989, 1990, Thurston et al., 2003).
2. $\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}$ ("SCK"): Use an inverse-variance-weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ estimated directly from the regression of $\mathbf{y}$ on $\mathbf{X}$ from the internal validation study and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ as above (Spiegelman et al., 2001).
3. $\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}$ ("same intercept A"): Use the true measured x_i for those in the internal validation data, and use the predicted $\hat{x}_i = W_i \hat{\boldsymbol{\alpha}}$ otherwise, in a single regression using the main study and internal validation study combined.
4. $\hat{\boldsymbol{\beta}}_{\mathrm{S,B}}$ ("same intercept B"): This is similar to the "same intercept A" method, but uses the estimator derived from a quasi-likelihood score approach in which the contribution to the score function from each observation is weighted by its inverse variance.

We use the notation $\beta_{\mathrm{E},1}, \beta_{\mathrm{SCK},1}, \beta_{\mathrm{S,A},1}$, and $\beta_{\mathrm{S,B},1}$ to denote the regression coefficient corresponding to the bias-corrected exposure effect in each of the four methods, respectively.

3.1. The estimators

In Appendix A, we show that all the estimators can be expressed explicitly as a weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$, the "as external" estimator, and $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$. The estimator of $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ is obtained by fitting model (1) to just the data from the internal validation study, with the following results:

$$\hat{\boldsymbol{\beta}}_{\mathrm{I}} = (\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{y}, \tag{4}$$

$$V_{\mathrm{I}} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \sigma_v^2(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})^{-1} \tag{5}$$

The "as external" estimator $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ and its asymptotic variance are

$$\hat{\boldsymbol{\beta}}_{\mathrm{E}} = \hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{-1}(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{y}, \tag{6}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}}) = \boldsymbol{\alpha}_{\mathrm{mat}}^{-1}\{\beta_{\mathrm{E},1}{}^2\sigma_\mu^2(\mathbf{W}^{\mathrm{T}}\mathbf{I}_v\mathbf{W})^{-1} + \sigma_\varepsilon^2(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\}\boldsymbol{\alpha}_{\mathrm{mat}}^{-\mathrm{T}}, \tag{7}$$

where $\hat{V}_{\mathrm{I}} = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{I}})$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})$ are obtained by substituting estimates for $\boldsymbol{\alpha}_{\mathrm{mat}}$, $\beta_{\mathrm{E},1}$, σ_u^2, σ_v^2 and σ_ε^2 in the above equation. We now briefly discuss each of the four estimators, and refer the reader to Appendix A for further details.

"*As external*": For the estimator of $\boldsymbol{\beta}$ and its asymptotic variance using the "as external" method, the internal validation study is treated as if external. This means that measurements of the outcome $\mathbf{y}$ from the internal validation study are not used so the information about $\boldsymbol{\beta}$ in the internal validation dataset is ignored and $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ is given a weight of zero. $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ and its asymptotic variance can be derived by methods described by CRS (1995) or by RSW (1989, 1990), and some further discussion is given in Appendix A.1.

"**SCK**": The "SCK" method for estimating $\boldsymbol{\beta}$ (Spiegelman et al., 2001) combines the two estimators of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ as given in (4), and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ as given in (6), weighting each estimator by its inverse variance, given in (5) and (7), respectively. $\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}$ and its asymptotic variance are

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{SCK}} &= \{\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{I}})^{-1} + \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})^{-1}\}^{-1}\{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{I}})^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{I}} + \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{E}}\},\\
\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}) &= \{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{I}})^{-1} + \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})^{-1}\}^{-1},\\
&= \left[\frac{1}{\sigma_v^2}(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X}) + \boldsymbol{\alpha}_{\mathrm{mat}}^{\mathrm{T}}\{\beta_{\mathrm{E},1}^2\sigma_u^2(\mathbf{W}^{\mathrm{T}}\mathbf{I}_v\mathbf{W})^{-1} + \sigma_\varepsilon^2(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\}^{-1}\boldsymbol{\alpha}_{\mathrm{mat}}\right]^{-1}.
\end{aligned} \tag{8}$$

All weighted estimators are consistent, as long as the component estimates are consistent. Estimation of the weights, $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_I)^{-1}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_E)^{-1}$, may give an underestimate of the variance in small samples (Carroll and Ruppert, 1988, p. 70), but asymptotically, the variance estimator is consistent.

"*Same intercept A*": For this method $\hat{x}_i$ is substituted for x_i in the regression of y_i on x_i in model (1), when the latter is not available. We fit the model

$$y_i = \beta_0 + \beta_1(I_{\mathrm{iv},i}x_i + I_{\mathrm{m},i}\hat{x}_i) + Z_i\boldsymbol{\beta}_z + \delta_i \tag{9}$$

The estimator and its asymptotic variance are derived in Appendix A.2 and are

$$\hat{\boldsymbol{\beta}}_{\mathrm{S,A}} = (\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}} + \mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})^{-1}(\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}\hat{\boldsymbol{\beta}}_{\mathrm{E}} + \mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{I}})$$

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X} + \boldsymbol{\alpha}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}})^{-1}\{&(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{I}})(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})\\ &+(\boldsymbol{\alpha}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}})\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})(\boldsymbol{\alpha}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}})\}(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X}\\ &+\boldsymbol{\alpha}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}})^{-1}.\end{aligned}$$

Both the "same intercept A" and "same intercept B" estimators and asymptotic variances can be written more compactly in terms of the naive asymptotic variance of $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$, which we call V_{E1}, where

$$V_{\mathrm{E1}} = \sigma_{\varepsilon}^2\boldsymbol{\alpha}_{\mathrm{mat}}^{-1}(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\boldsymbol{\alpha}_{\mathrm{mat}}^{-\mathrm{T}} = \sigma_{\varepsilon}^2(\hat{\mathbf{X}}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\hat{\mathbf{X}})^{-1}. \tag{10}$$

Note that V_{E1} is one of the two terms of $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})$. We can then rewrite $\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{S,A}})$ as:

$$\hat{\boldsymbol{\beta}}_{\mathrm{S,A}} = (\hat{\sigma}_{\varepsilon}^2\hat{V}_{\mathrm{E1}}^{-1} + \hat{\sigma}_{v}^2\hat{V}_{\mathrm{I}}^{-1})^{-1}(\hat{\sigma}_{\varepsilon}^2\hat{V}_{\mathrm{E1}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{E}} + \hat{\sigma}_{v}^2\hat{V}_{\mathrm{I}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{I}}),$$

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}) = (\sigma_{\varepsilon}^2V_{\mathrm{E1}}^{-1} + \sigma_{v}^2V_{\mathrm{I}}^{-1})^{-1}&(\sigma_{\varepsilon}^4V_{\mathrm{E1}}^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})V_{\mathrm{E1}}^{-1} + \sigma_{v}^4V_{\mathrm{I}}^{-1})\\ &(\sigma_{\varepsilon}^2V_{\mathrm{E1}}^{-1} + \sigma_{v}^2V_{\mathrm{I}}^{-1})^{-1}.\end{aligned}$$

"*Same intercept B*": This estimator can be motivated from the quasi-likelihood score function (Wedderburn, 1974, discussed in Godambe, 1991, p. 51), as discussed in Appendix A.3. The estimator and asymptotic variance are

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{S,B}} &= \left(\frac{1}{\hat{\sigma}_{\varepsilon}^2}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}} + \frac{1}{\hat{\sigma}_{v}^2}\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X}\right)^{-1}\left(\frac{1}{\hat{\sigma}_{\varepsilon}^2}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{y} + \frac{1}{\hat{\sigma}_{v}^2}\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{y}\right)\\ &= (V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1})^{-1}(V_{\mathrm{I}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{I}} + V_{\mathrm{E1}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{E}})\end{aligned} \tag{11}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{S,B}}) = (V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1})^{-1}(V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})V_{\mathrm{E1}}^{-1})(V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1})^{-1}$$

Table 1 presents a summary of each estimator and its asymptotic variance. Note that $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ can each be written as weighted averages by simply assigning zero weight to the other estimator. By expressing the estimators of the exposure effect $\hat{\boldsymbol{\beta}}$ in this format, it is clear that the weights corresponding to the SCK method are minimum variance, and thus any other estimator must have greater asymptotic variance.

3.2. Bootstrap variance estimation and confidence interval construction

In addition to the asymptotic variances, we use the "resampling vectors" method for bootstrapping, as described in CRS (1995, p. 277). We sample with replacement n_{m} observations from the main study, n_{ev} observations from the external validation study, and n_{iv} observations from the internal validation study to produce a single bootstrap sample.

For each bootstrap sample, we estimated $\boldsymbol{\beta}$ using four methods ($\hat{\boldsymbol{\beta}}_{\mathrm{E}}, \hat{\boldsymbol{\beta}}_{\mathrm{SCK}}, \hat{\boldsymbol{\beta}}_{\mathrm{S,A}}, \hat{\boldsymbol{\beta}}_{\mathrm{S,B}}$). We calculated 95% empirical bootstrap confidence intervals by sorting the estimates

Table 1
A comparison of $\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ for regression calibration using internal validation data

Estimator	Multiplier for		
	$\hat{\boldsymbol{\beta}}_{\mathrm{E}}$	$\hat{\boldsymbol{\beta}}_{\mathrm{I}}$	$\mathrm{Var}(\hat{\boldsymbol{\beta}})$
$\hat{\boldsymbol{\beta}}_{\mathrm{I}}$	0	1	$V_{\mathrm{I}} = \sigma_v^2(\mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})^{-1}$
$\hat{\boldsymbol{\beta}}_{\mathrm{E}}$	1	0	$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}}) = V_{\mathrm{E1}} + \boldsymbol{\alpha}_{\mathrm{mat}}^{-1}\{\hat{\beta}_{\mathrm{E},1}^2\sigma_u^2(\mathbf{W}^{\mathrm{T}}\mathbf{I}_v\mathbf{W})^{-1}\}\boldsymbol{\alpha}_{\mathrm{mat}}^{-\mathrm{T}}$
$\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}$	$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_E)^{-1}$	$\hat{V}_{\mathrm{I}}^{-1}$	$(V_{\mathrm{I}}^{-1} + \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})^{-1})^{-1}$
$\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}$	$\sigma_\varepsilon^2\hat{V}_{\mathrm{E1}}^{-1}$	$\hat{\sigma}_v^2\hat{V}_{\mathrm{I}}^{-1}$	$(\sigma_\varepsilon^2 V_{\mathrm{E1}}^{-1} + \sigma_v^2 V_{\mathrm{I}}^{-1})^{-1}$ $\times((\sigma_\varepsilon^2)^2 V_{\mathrm{E1}}^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})V_{\mathrm{E1}}^{-1} + (\sigma_v^2)^2 V_{\mathrm{I}}^{-1})(\sigma_\varepsilon^2 V_{\mathrm{E1}}^{-1} + \sigma_v^2 V_{\mathrm{I}}^{-1})^{-1}$
$\hat{\boldsymbol{\beta}}_{\mathrm{S,B}}$	$\hat{V}_{\mathrm{E1}}^{-1}$	$\hat{V}_{\mathrm{I}}^{-1}$	$(V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1})^{-1}(V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{E}})V_{\mathrm{E1}}^{-1})(V_{\mathrm{I}}^{-1} + V_{\mathrm{E1}}^{-1})^{-1}$

Each estimator is a weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$, where each weight is the multiplier as indicated, premultiplied by the inverse of the sum of the two weights. $V_{\mathrm{E1}} = \sigma_\varepsilon^2\boldsymbol{\alpha}_{\mathrm{mat}}^{-1}(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\boldsymbol{\alpha}_{\mathrm{mat}}^{-\mathrm{T}}$ is the naive variance of the "as external" method.

of $\hat{\beta}_1$ and defining the empirical confidence interval as the range of the central 95% of the estimates.

4. Asymptotic relative efficiencies of the methods

Since each estimator is consistent for $\boldsymbol{\beta}$ under the stated assumptions, it remains only to compare their asymptotic relative efficiency.

$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{SCK}})$ is the minimum variance weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$, as shown in (8). $\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}$ is also a weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ (Eq. (10)). Hence, it is asymptotically inefficient relative to $\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}$. Similar arguments apply to $\hat{\boldsymbol{\beta}}_{\mathrm{S,B}}$, whose weights are shown in (11), and to $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$, which gives zero weight to $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$.

To better understand when the three less efficient estimators are relatively efficient or inefficient relative to $\hat{\boldsymbol{\beta}}_{\mathrm{SCK}}$, we first write the asymptotic variances of each estimator in terms of correlations and variances. Although this can be done for any number of covariates, for simplicity we consider the case when there are no additional covariates $\mathbf{Z}$ included in the model. We assume all data are centered within the main and validation studies separately, by subtracting the sample mean from each individual measurement, within each study. We also assume the variance of w_i in the main dataset is the same as the variance of w_i in the validation dataset, and that the validation data are purely internal, i.e. $n_{\mathrm{iv}} = n_v$ so $n_{\mathrm{ev}} = 0$ and n_v/n_{m} is the ratio of the number of validation to main study observations.

We use the notation σ^2 with a subscript to denote the variance of the subscripted variable, and ρ^2 with two subscripts to denote the squared correlation between the two subscripted variables. We rewrite the asymptotic variances of the four estimators as shown below. The derivation of these results can be obtained upon request from the first author.

$$\mathrm{Var}(\hat{\beta}_{\mathrm{E},1}) = \frac{\sigma_y^2}{n_{\mathrm{m}}\sigma_x^2}\left\{\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)(1+\frac{n_{\mathrm{m}}}{n_v}) + (1-\rho_{y,x}^2)}{\rho_{x,w}^2}\right\},$$

$$\mathrm{Var}(\hat{\beta}_{\mathrm{SCK},1}) = \frac{\sigma_y^2}{n_{\mathrm{m}}\sigma_x^2}\left\{\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)(1+\frac{n_{\mathrm{m}}}{n_v})+(1-\rho_{y,x}^2)}{\rho_{x,w}^2+\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)}{(1-\rho_{y,x}^2)}\left(1+\frac{n_v}{n_{\mathrm{m}}}\right)+\frac{n_v}{n_{\mathrm{m}}}}\right\},$$

$$\mathrm{Var}(\hat{\beta}_{\mathrm{S,A},1}) = \frac{\sigma_y^2}{n_{\mathrm{m}}\sigma_x^2}\left\{\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)(1+\frac{n_{\mathrm{m}}}{n_v})\rho_{x,w}^2+(1-\rho_{y,x}^2)(\rho_{x,w}^2+\frac{n_v}{n_{\mathrm{m}}})}{\left(\frac{n_v}{n_{\mathrm{m}}}+\rho_{x,w}^2\right)^2}\right\},$$

$$\mathrm{Var}(\hat{\beta}_{\mathrm{S,B},1}) = \frac{\sigma_y^2}{n_{\mathrm{m}}\sigma_x^2}\left\{\frac{\rho_{x,w}^2[\rho_{y,x}^2(1-\rho_{x,w}^2)(1+\frac{n_{\mathrm{m}}}{n_v})+(1-\rho_{y,x}^2)]+\frac{n_{\mathrm{v}}}{n_{\mathrm{m}}}(1-\rho_{y,x}^2)\left(\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)}{(1-\rho_{y,x}^2)}+1\right)^2}{[\rho_{x,w}^2+\frac{n_{\mathrm{v}}}{n_{\mathrm{m}}}(\frac{\rho_{y,x}^2(1-\rho_{x,w}^2)}{(1-\rho_{y,x}^2)}+1)]^2}\right\}$$

Comparisons of the asymptotic variance of $\hat{\beta}_1$ for the four methods are provided in Figs. 1 and 2 for two values of n_v/n_{m} and varying correlations. As measurement error decreases ($\rho_{x,w} \to 1$), the asymptotic efficiencies of the estimators relative to $\hat{\beta}_{\mathrm{SCK},1}$ approach 1, regardless of other factors such as $\rho_{y,x}$ or n_v/n_{m}. As measurement error

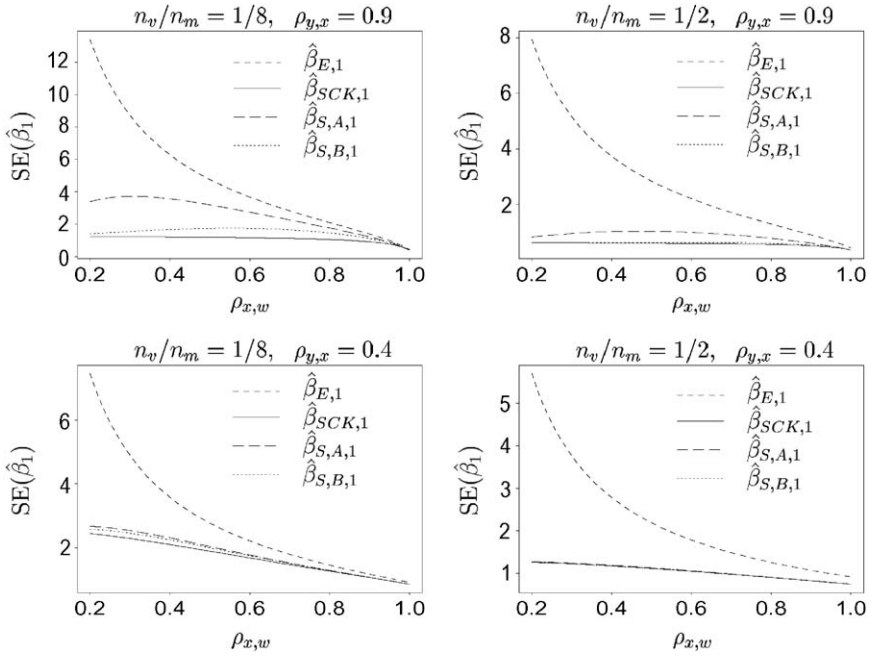

Fig. 1. A comparison of the asymptotic standard error of $\hat{\beta}_1$ as a function of the correlation between the true exposure, x, and the exposure measured with error, w, for two values of the percentage of subjects in the internal validation study, and two values of the correlation between the outcome, y, and the true exposure, x. Plots were constructed assuming no additional covariates and equal variances of the true exposure and the "proxy" exposure.

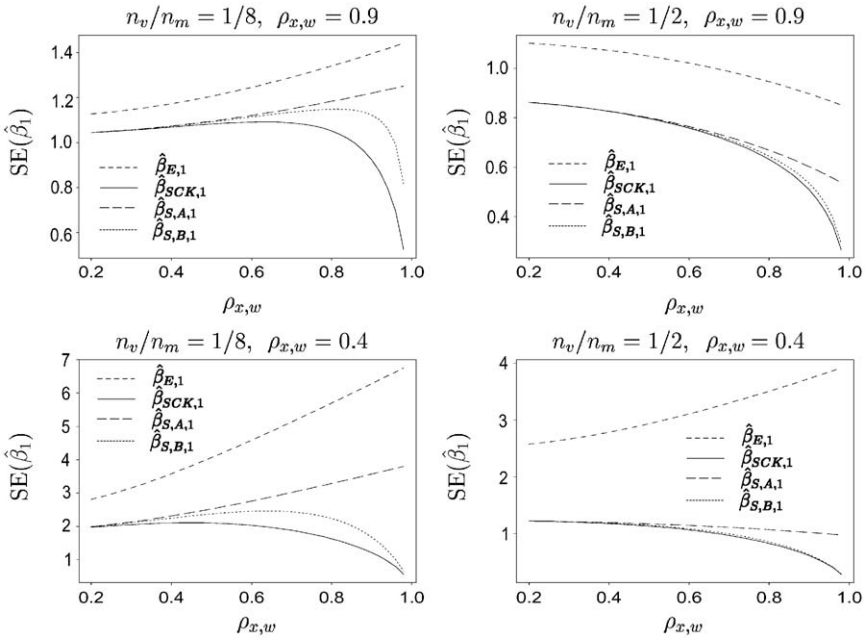

Fig. 2. A comparison of the asymptotic standard error of $\hat{\beta}_1$ as a function of the correlation between the outcome, y, and the true exposure, x, for two values of the percentage of subjects in the internal validation study, and two values of the correlation between the true exposure, x, and the "proxy" exposure, w. Plots were constructed assuming no additional covariates and equal variances of the true exposure and the "proxy" exposure.

increases ($\rho_{x,w} \to 0$), the asymptotic efficiency of $\hat{\beta}_{\mathrm{E},1}$ relative to $\hat{\beta}_{\mathrm{SCK},1}$ approaches 0, while the asymptotic relative efficiencies of $\hat{\beta}_{\mathrm{S,A},1}$ and $\hat{\beta}_{\mathrm{S,B},1}$ do not, and when the proportion of data in the validation study (n_v/n_{m}) is large and the strength of the true regression relationship is weak ($\rho_{y,x}$ is small), $\hat{\beta}_{\mathrm{S,A},1}$, $\hat{\beta}_{\mathrm{S,B},1}$, and $\hat{\beta}_{\mathrm{SCK},1}$ have nearly the same asymptotic variance while the variance of $\hat{\beta}_{\mathrm{E},1}$ is larger. When the strength of the true regression relationship is strong ($\rho_{y,x} \to 1$), the asymptotic efficiency of $\hat{\beta}_{\mathrm{S,B},1}$ relative to $\hat{\beta}_{\mathrm{SCK},1}$ approaches 1, whereas the asymptotic efficiencies of $\hat{\beta}_{\mathrm{E},1}$ and $\hat{\beta}_{\mathrm{S,A},1}$ do not. When $\rho_{y,x} \to 1$ and n_v/n_{m} is small, the asymptotic efficiency of $\hat{\beta}_{\mathrm{E},1}$ relative to $\hat{\beta}_{\mathrm{SCK},1}$ approaches 0.

5. Illustrative example: the relationship between birth weight and lead exposure

5.1. Study description

Several studies have linked lead exposure during the in utero period with a decline in infant birth weight, which in turn is associated with increased infant mortality and

morbidity. Lead accumulated over many years is stored in bone, which comprise 90–95% of adult lead burden (Hu, 1998). Bone lead levels reflect cumulative lead burden, whereas blood lead levels are more indicative of recent lead exposure. One source of fetal lead exposure is the mobilization of maternal bone lead during pregnancy. Measurements of blood lead levels from the newborn umbilical cord are the most commonly used method for estimating in utero lead exposure. However, bone lead can be mobilized into plasma without affecting blood lead levels, and thus maternal bone lead is believed to be a better biomarker of infant lead exposure than umbilical cord blood lead (Gonzalez-Cossio et al., 1997). In contrast to blood lead measurements however, obtaining measurements of bone lead levels requires specialized equipment available in only a handful of research laboratories, and is relatively expensive.

Gonzalez-Cossio et al. (1997) examined the relationship between lead exposure and infant birth weight among a cohort of 272 mother-infant pairs in Mexico City. Measurements of lead exposure were taken on subsets of the cohort from the umbilical cord blood in the newborn, and from two bones in the mother's leg: the tibia and the patella. Bone lead measurements were taken one month postpartum using K-X-ray fluorescence (see Hu, 1998, for a description of this process), and blood lead was analyzed by atomic absorption spectrophotometry. Gonzalez-Cossio et al. (1997) found that maternal tibial bone lead levels were more closely associated with a decreased birth weight than either maternal patella lead or umbilical cord blood lead levels.

5.2. Analysis of the data

We examined the relationship between tibia bone lead and infant birth weight among eligible women who consented to be in the original study, while adjusting for bias due to measurement error when cord blood is used as a surrogate for tibial bone lead levels. These data are a main study/internal validation study design, with 577 observations (mother-infant pairs) in the main study and 585 in the internal validation study. Among the internal validation observations, 100 did not have a measurement of blood lead, leaving 485 observations in the validation study with which to estimate the relationship between bone and blood lead. An advantage of the newer methods is that we do not require observations to have measurements on both lead levels, so our effective sample size was considerably larger than the 272 mother–infant pairs considered in the Gonzalez-Cossio et al. study.

We calculated $\hat{\beta}_1$ and $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ under the four methods described in Section 3. For each method, we calculated asymptotic confidence intervals as well as bootstrap confidence intervals using 5000 bootstrap samples. We estimated β_1 in the model which adjusts for four covariates: maternal nutritional status as measured by mother's calf circumference (cm), parity (1 or 2+), education (years), and gestational age (weeks). These covariates were important predictors of birth weight in both the Gonzalez-Cossio et al. study and in our analysis. Specifically, the models we fit are given in (1)–(3), where Z_i consists of these four covariates. The estimator of $\boldsymbol{\beta}_{\mathrm{I}}$ is obtained directly from fitting model (1) to the internal validation data. The estimator of $\boldsymbol{\beta}_{\mathrm{E}}$ can be obtained by fitting model (3) to the main study data, where estimation of $\hat{X}$ in (3) requires estimation of $\boldsymbol{\alpha}_{\mathrm{mat}}$ from the validation data, as discussed previously. The estimated

exposure effect in the "as external" method, $\hat{\beta}_{E,1}$, is simply the ratio of the "proxy" exposure effect from the uncorrected covariate-adjusted regression of y on w and Z estimated from the main study data (Table 2, line 2), to $\hat{\alpha}_1$ in model (2) estimated from the validation data (Thurston et al., 2003). In the lead data, $\hat{\alpha}_1 = 0.54$, so it is not surprising that $\hat{\beta}_{E,1} = -7.86$ was much farther from zero than $\hat{\beta}_{I,1} = -3.76$.

This study had many observations in the internal validation study, which is often not the case in other situations. With n_v/n_m greater than $\frac{1}{2}$, $\hat{\rho}_{y,x} = -0.057$ and $\hat{\rho}_{x,w} = 0.19$, as expected from Figs. 1 and 2, the "SCK", "same intercept A", and "same intercept B" methods gave very similar estimates of $\hat{\beta}_1$ and $\mathrm{SE}(\hat{\beta}_1)$. In this dataset, $\mathrm{Var}(\hat{\beta}_{E,1})$ was more than 25 times larger than $\mathrm{Var}(\hat{\beta}_{I,1})$ so these three estimators for β_1, all of which are weighted averages of $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_I$, were all much closer to $\hat{\beta}_{I,1}$ than to $\hat{\beta}_{E,1}$.

Gonzalez-Cossio et al. also adjusted for current smoking status. We were unable to adjust for this variable because smoking information was only obtained among participants who also had a measure of tibia lead (i.e. when the mother was seen at the one-month postpartum visit), and thus smoking information was missing for everyone in the main study. The adjusted effect (SE) of tibia lead on birth weight reported in the Gonzalez-Cossio et al. (1997) paper was -7.29 g/μg lead/g bone mineral (2.45 g/μg lead/g bone mineral), while the smoking-unadjusted effect (SE) is -7.45 g/μg lead/g bone mineral (2.47 g/μg lead/g bone mineral) for the same $n = 272$ subjects who had measurements of both tibia and patella lead. Although smoking is an independent determinant of lower birth weight, it does not appear to materially confound the estimated effect of lead. We did not require subjects to have a measurement of patella lead, and our estimate of the adjusted effect (SE) of tibia lead on birth weight was -3.76 g/μg lead/g bone mineral (1.55 g/μg lead/g bone mineral) among the $n = 585$ subjects in the internal validation study, indicating that the inclusion or exclusion of subjects with patella lead has a substantial impact on the estimated exposure effect.

6. Discussion

Exposure measurement error is a common limitation in epidemiologic studies. Data on the most accurate measure of exposure is often only collected on a subset of study participants since it is infeasible or too expensive to measure it on all subjects. Statistical methods for correcting bias due to exposure measurement error have infrequently been used in reproductive or occupational epidemiology, although estimation of the effect of an exposure on a health outcome would be improved by use of measurement error corrections such as those discussed in this paper.

Another estimator for $\boldsymbol{\beta}$ which we did not discuss here, was suggested by CRS (1995, p. 46). This estimator, which one might call $\hat{\boldsymbol{\beta}}_D$ for the "different intercept" method, is the same as $\hat{\boldsymbol{\beta}}_{S,A}$, but allows the observations from the main study corrected for measurement error to have a different intercept from the observations from the internal validation study (CRS, 1995, p. 46). For this method, instead of fitting model (9), one would fit the following model:

$$y_i = \beta_v I_{iv,i} + \beta_0 + \beta_1 (I_{iv,i} x_i + I_{m,i} \hat{x}_i) + Z_i \boldsymbol{\beta}_z + \varepsilon_i.$$

Table 2
Estimates of the effect of tibia bone lead (μg lead/g bone mineral) on birth weight (grams) in the Gonzalez-Cossio et al. (1997) study

	Sample size		Effect of lead		95% confidence interval for lead effect		Effect of other model covariates			
	n_m	n_{iv}	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	Asymptotic	Bootstrap	Calf circumference	Parity	Education	Gestational age
Uncorrected	1062[a]	0	−4.13	3.24			23.27	111.24	16.12	57.56
Uncorrected	577	0	−4.25	4.31			22.35	84.70	10.32	38.07
Interval validation	0	585	−3.76	1.55	(−6.80, −0.73)	(−6.99, −0.57)	22.77	132.60	20.24	76.68
As external ($\hat{\boldsymbol{\beta}}_E$)	577	485[b]	−7.86	8.16	(−23.84, 8.13)	(−28.60, 8.30)	21.76	106.41	12.30	39.10
Weighted (SCK) ($\hat{\boldsymbol{\beta}}_{SCK}$)	577	585	−3.88	1.52	(−6.86, −0.90)	(−7.05, −0.67)	22.29	115.28	16.31	58.96
Same intercept A ($\hat{\boldsymbol{\beta}}_{S,A}$)	577	585	−3.89	1.52	(−6.87, −0.90)	(−7.11, −0.68)	22.27	114.78	16.21	58.40
Same intercept B ($\hat{\boldsymbol{\beta}}_{S,B}$)	577	585	−3.88	1.52	(−6.87, −0.90)	(−7.10, −0.67)	22.27	114.89	16.23	58.51

[a]Includes information relating blood lead and birth weight from observations in internal validation study.
[b]Number of observations used to estimate the relationship between blood lead and bone lead only.

Uncorrected results estimate the effect of blood lead (μg lead/dL) on birth weight (grams). Results are from a multivariate model adjusted for mother's calf circumference (cm), parity (1 or 2+), education (years), and gestational age (weeks). Here n_m and n_{iv} are the number of observations used to estimate the relationship between blood lead and birth weight, and between bone lead and birth weight, respectively. All estimators which correct for measurement error used 485 observations to estimate the relationship between blood lead and bone lead.

The addition of the $\beta_v I_{\mathrm{iv},i}$ term means that the data from the internal validation study and the data from the adjusted main study are allowed to have different intercepts. Since $\mathbf{x}$ is not observed before sampling, the sampling mechanism cannot depend on $\mathbf{x}$. Under the assumption of non-differential measurement error and the assumption that sampling into the internal validation dataset may depend on $\mathbf{w}$ or $\mathbf{Z}$, but does not depend on $\mathbf{y}$, it can be shown that $f(y_i \mid I_{\mathrm{iv},i}, x_i, Z_i) = f(y_i \mid x_i, Z_i)$, and therefore that $\beta_v = 0$. Since correct specification of $E(y_i \mid X_i)$ can be obtained with or without the inclusion of β_v in the model, both $\hat{\boldsymbol{\beta}}_{\mathrm{S,A}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{D}}$ will be consistent for $\boldsymbol{\beta}$. If there was some reason to believe that the models generating the data in the internal validation study and in the main study have different intercepts but the same slope, consideration of the "different intercept" method would be warranted.

In this paper, we compared four regression calibration estimators appropriate for main study/hybrid validation study designs, all of which are consistent for the true parameter value. The "as external" method, which ignores the measured outcome in the internal validation study, has the largest variance. The "same intercept A", "same intercept B" and the "SCK" methods have similar variances under some conditions, such as the lead study discussed in this paper, but the asymptotic variance of the "SCK" method is never larger than the other methods, and in some cases is considerably smaller than the "same intercept A" method. When measurement error is large (i.e. the "gold standard" and "proxy" exposures are not highly correlated), when the outcome is highly correlated with the "gold standard" exposure, and when the fraction of observations in the internal validation study is small, the "SCK" method is asymptotically much more efficient than either the "same intercept A" or the "as external" methods. The improved efficiency of the "SCK" method may be important in studies with considerable measurement error and a small fraction of observations in the internal validation study, such as is typically the case in nutritional studies. For example in the Nurses Health Study there were 173 observations in the internal validation study (Willett et al., 1985), and the main study consisted of over 89,000 participants (Willett et al., 1987). A SAS macro which corrects for measurement error using the "SCK" method is available from Spiegelman (`stdls@channing.harvard.edu`).

Acknowledgements

This publication was made possible, in part, by Grants no. 5 F32 ES05834, ES07981, ES09411, ES05497, ES00002, R01 ES07821 and P42 ES-05947 Project 1 from the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Grant no. CA74112 from the National Cancer Institute (NCI), and Grant no. OH00152 from the US National Institute for Occupational Safety and Health (NIOSH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH, NCI, or NIOSH. This paper reflects work done while the first author was a Research Fellow and Research Associate in the Department of Biostatistics at Harvard School of Public Health. The authors thank Adrienne Ettinger for her help in constructing the dataset discussed in this paper, and an anonymous reviewer for helpful comments.

Appendix A.

In deriving the asymptotic variances of the four estimators, certain conventions were followed. First, in evaluating the relationship between the true exposure and the proxy, all methods used the same estimator for $\boldsymbol{\alpha}$: $\hat{\boldsymbol{\alpha}} = (\mathbf{W}^{\mathrm{T}}\mathbf{I}_v\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{I}_v\mathbf{x}$. In addition, the asymptotic variances of the estimators relied on the variance components, σ_u^2, σ_v^2, and σ_e^2. However, as shown by Thurston et al. (2003) for external validation data, the estimator for the regression coefficients and their asymptotic variance are unchanged whether or not the variance coefficients are estimated jointly. In all methods, we therefore estimated the variance components as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n_{\mathrm{m}}}(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}\hat{\boldsymbol{\beta}}_{\mathrm{E}})^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}\hat{\boldsymbol{\beta}}_{\mathrm{E}}), \hat{\sigma}_v^2 = \frac{1}{n_{\mathrm{iv}}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{I}})^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{I}}),$$

$$\hat{\sigma}_u^2 = \frac{1}{n_v}(\mathbf{x} - \mathbf{W}\hat{\boldsymbol{\alpha}})^{\mathrm{T}}\mathbf{I}_v(\mathbf{x} - \mathbf{W}\hat{\boldsymbol{\alpha}}).$$

A.1. "As external"

In this method, we fit model (3) to the main study only. Because $\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}$ is estimated only from the validation study (Thurston et al., 2003), this is equivalent to regressing y_i on $W_i\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}$, which gives the least squares estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{E}} = \hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{-1}(\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{y}.$$

In the RSW approach (first discussed by Fuller (1987) for linear models), y_i is regressed on W_i in the main study, using the following model:

$$y_i = \gamma_0 + \gamma_1 W_i + \gamma_z Z_i = W_i\boldsymbol{\gamma} + \varepsilon_i,$$

which gives $\hat{\boldsymbol{\gamma}} = (\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}y$. Then $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ is obtained as $\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{-1}\hat{\boldsymbol{\gamma}}$, which is identical to the estimator above. This estimator can be derived from any of the following three score equations: (a) $\psi(\boldsymbol{\theta}) = \mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}(\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}}\boldsymbol{\beta})$; (b), $\psi(\boldsymbol{\theta}) = \boldsymbol{\alpha}_{\mathrm{mat}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}(\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}}\boldsymbol{\beta})$; or (c) $\psi(\boldsymbol{\theta}) = (1/\sigma_\varepsilon^2)\boldsymbol{\alpha}_{\mathrm{mat}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}(\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_{\mathrm{mat}}\boldsymbol{\beta})$, all of which, when set to 0, have the same solution, $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$.

The variance of $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ is obtained by the multivariate delta method for RSW, and can be obtained from stacked estimating functions for CRS. Both methods give the same estimated asymptotic variance (Thurston et al., 2003).

A.2. "Same intercept A"

The estimator for this method can be derived by introducing U, where $(\mathbf{1}\ \tilde{\mathbf{x}}\ \mathbf{Z})$, and $\tilde{\mathbf{x}} = \mathbf{I}_{\mathrm{iv}}\mathbf{x} + \mathbf{I}_{\mathrm{m}}\hat{\mathbf{x}}$. Then $\hat{\boldsymbol{\beta}}_{\mathrm{S,A}} = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}y$. Since we can rewrite U as

$$U = \begin{pmatrix} \mathbf{I}_{\mathrm{iv}}\mathbf{X} \\ \mathbf{I}_{\mathrm{m}}\hat{\mathbf{X}} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{\mathrm{iv}}\mathbf{X} \\ \mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}} \end{pmatrix},$$

it follows that

$$\hat{\boldsymbol{\beta}}_{\mathrm{S,A}} = (\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}} + \mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X})^{-1}(\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{W}\hat{\boldsymbol{\alpha}}_{\mathrm{mat}}\hat{\boldsymbol{\beta}}_{\mathrm{E}} + \mathbf{X}^{\mathrm{T}}\mathbf{I}_{\mathrm{iv}}\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{I}}).$$

This estimator of $\boldsymbol{\beta}$ can also be obtained as the solution to the unbiased estimating function

$$\psi(\boldsymbol{\theta}) = \boldsymbol{\alpha}_{\text{mat}}^{\text{T}} \mathbf{W}^{\text{T}} \mathbf{I}_{\text{m}} (\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_{\text{mat}}\boldsymbol{\beta}) + \mathbf{X}^{\text{T}} \mathbf{I}_{\text{iv}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0. \tag{A.1}$$

A.3. "Same intercept B"

The estimating function used for the "same intercept A" method, given in (A.1), is different from the quasi-likelihood score function (Wedderburn, 1974, discussed in Godambe, 1991, p. 51), which would take the form

$$\psi(\boldsymbol{\theta}) = \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (y_i - \mu_i),$$

where μ_i and V_i are the mean and variance of y_i, respectively. This suggests the following unbiased estimating function:

$$\psi(\boldsymbol{\theta}) = \frac{1}{\sigma_\varepsilon^2} \boldsymbol{\alpha}_{\text{mat}}^{\text{T}} \mathbf{W}^{\text{T}} \mathbf{I}_{\text{m}} (\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_{\text{mat}}\boldsymbol{\beta}) + \frac{1}{\sigma_v^2} \mathbf{X}^{\text{T}} \mathbf{I}_{\text{iv}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The estimator for $\boldsymbol{\beta}$ which solves this function when set to 0, which we call the "same intercept B" method is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{S,B}} &= \left(\frac{1}{\hat{\sigma}_\varepsilon^2} \hat{\boldsymbol{\alpha}}_{\text{mat}}^{\text{T}} \mathbf{W}^{\text{T}} \mathbf{I}_{\text{m}} \mathbf{W} \hat{\boldsymbol{\alpha}}_{\text{mat}} + \frac{1}{\hat{\sigma}_v^2} \mathbf{X}^{\text{T}} \mathbf{I}_{\text{iv}} \mathbf{X} \right)^{-1} \left(\frac{1}{\hat{\sigma}_\varepsilon^2} \hat{\boldsymbol{\alpha}}_{\text{mat}}^{\text{T}} \mathbf{W}^{\text{T}} \mathbf{I}_{\text{m}} \mathbf{y} + \frac{1}{\hat{\sigma}_v^2} \mathbf{X}^{\text{T}} \mathbf{I}_{\text{iv}} \mathbf{y} \right) \\ &= (V_{\text{I}}^{-1} + V_{\text{E1}}^{-1})^{-1} (V_{\text{I}}^{-1} \hat{\boldsymbol{\beta}}_{\text{I}} + V_{\text{E1}}^{-1} \hat{\boldsymbol{\beta}}_{\text{E}}). \end{aligned}$$

References

Armstrong, B., 1985. Measurement error in generalized linear models. Comm. Statist. Ser. B 14, 529–544.

Carroll, R.J., Stefanski, L.A., 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. J. Amer. Statist. Assoc. 85, 652–663.

Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman & Hall, London.

Carroll, R.J., Ruppert, D., Stefanski, L.A., 1995. Measurement Error in Nonlinear Models. Chapman & Hall, London.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Godambe, V.P., 1991. Estimating Functions. Clarendon Press, Oxford.

Gonzalez-Cossio, T., Peterson, K.E., Sanin, L.-H., Fishbein, E., Palazuelos, E., Antonio, A., Hernandez-Avila, M., Hu, H., 1997. Decrease in birth weight in relation to maternal bone-lead burden. Pediatrics 100 (5), 856–862.

Hu, H., 1998. Bone lead as a new biologic marker of lead dose: recent findings and implications for public health. Environ. Health Perspective 106 (Suppl 4), 961–967.

Prentice, R.L., 1982. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 69, 331–342.

Rosner, B., Willett, W.C., Spiegelman, D., 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statist. Med. 8, 1051–1069.

Rosner, B., Spiegelman, D., Willett, W.C., 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Amer. J. Epidemiol. 132, 734–745.

Rosner, B., Spiegelman, D., Willett, W.C., 1992. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Amer. J. Epidemiol. 136, 1400–1413.

Spiegelman, D., McDermott, A., Rosner, B., 1997. The regression calibration method for correcting measurement error bias in nutritional epidemiology. Amer. J. Clin. Nutr. 65 (suppl), 1179S–1186S.

Spiegelman, D., Carroll, R., Kipnis, V., 2001. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Statist. Med. 20, 139–160.

Thurston, S.W., Spiegelman, D., Ruppert, D., 2003. Equivalence of regression calibration methods for external validation data. J. Statist. Planning Inference 113, 527–539.

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61, 439–448.

Willett, W.C., Sampson, L., Stampfer, M.J., Rosner, B., Bain, C., Witschi, J., Hennekens, C.H., Speizer, F.E., 1985. Reproducibility and validity of a semiquantitative food questionnaire. Amer. J. Epidemiol. 122, 51–65.

Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A., Hennekens, C.H., Speizer, F.E., 1987. Dietary fat and the risk of breast cancer. New England J. Med. 316, 22–28.