# Synthesis analysis of regression models with a continuous outcome

**Xiao-Hua Zhou**[1,2,*,†], **Nan Hu**[2], **Guizhou Hu**[3], and **Martin Root**[3]

[1] HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98101, U.S.A

[2] Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A

[3] BioSignia, Inc., 1822 East NC Highway 54, Suite 350, Durham, NC 27713, U.S.A

## SUMMARY

To estimate the multivariate regression model from multiple individual studies, it would be challenging to obtain results if the input from individual studies only provide univariate or incomplete multivariate regression information. Samsa *et al.* (*J. Biomed. Biotechnol.* 2005; **2**:113–123) proposed a simple method to combine coefficients from univariate linear regression models into a multivariate linear regression model, a method known as synthesis analysis. However, the validity of this method relies on the normality assumption of the data, and it does not provide variance estimates. In this paper we propose a new synthesis method that improves on the existing synthesis method by eliminating the normality assumption, reducing bias, and allowing for the variance estimation of the estimated parameters.

### Keywords

synthesis analysis; meta-analysis; linear models

## 1. INTRODUCTION

Meta-analysis is a statistical technique for amalgamating, summarizing, and reviewing previous quantitative research. A typical meta-analysis is to summarize all the research results on one topic and to discuss reliability of this summary. It is based on the condition that each individual study reports the same finding for the same research question. The potential advantage of meta-analysis is the increase in the sample size and the validity of statistical inference. It would be difficult to utilize meta-analysis methodologies if individual studies only provide partial findings.

In a practical example, meta-analysis could be used to build a comprehensive and multivariate prediction model for the risk of chronic diseases such as coronary heart disease (CHD). A wide range of CHD risk factors have been reported in the literature, but a comprehensive multivariate CHD prediction model has yet to be found. The Framingham CHD model is widely considered the most comprehensive model, although many well-known CHD risk factors, such as body mass index (BMI), family history of CHD, and c-reactive protein, are not included in the model [1–3].

---

*Correspondence to: Xiao-Hua Zhou, HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98101, U.S.A.
†azhou@u.washington.edu

We propose a new process to solve several of the problems presented above. This novel multivariate meta-analysis modeling method is called synthesis analysis. Using multiple study results reported in the scientific and medical literature, the objective of our synthesis analysis is to estimate the multivariate relations between multiple predictors ($X$s) and an outcome variable ($Y$) from the univariate relation of each $X$ with $Y$ and the two-way correlations between each pair of $X$s. All the inputs may come from various studies in the literature, while a cross-sectional population survey may provide correlations of all $X$s. We reported the first method of synthesis analysis (the Samsa-Hu-Root or SHR method) in which the partial regression coefficients were calculated using the following matrix equation:

$$B=(R^{-1}(Bu\#S))/S$$

where $B$ is the vector of partial (excluding the intercept, $B_0$) regression coefficients, $Bu$ is the vector of univariate regression coefficients, $R$ is the vector of Pearson correlation coefficients among all independent variables, $S$ is the vector of standard deviations of the independent variables, # stands for element-wise multiplication, and/stands for element-wise division. The intercept, $B_0$, can be calculated using the resulting multivariate formula, the mean of the predictors and outcome, and the newly calculated partial regression coefficient for each predictor.

In the present study, we propose an improvement to the existing synthesis analysis. Compared with the previous method, this method has at least two advantages: (1) it includes a method to compute the variances for predicted outcomes and estimated regression coefficients and (2) the estimates of predicted outcomes and regression coefficients can be more robust when the independent variables are not normally distributed.

Our paper is organized as follows. In Section 2, we describe our new method. In Section 3, we report a simulation study on finite-sample performance of the proposed method in comparison with the existing synthesis method. In Section 4, we illustrate the use of the proposed method in a real-life example from the 1999–2000 National Health and Nutritional Examination Survey. Finally, in Section 5, we conclude our paper with a discussion on some extensions.

## 2. NEW METHOD FOR SYNTHESIS ANALYSIS

### 2.1. Estimation of synthesized parameters

Suppose that we know the individual relationships between an outcome $Y$ and each of $p$ risk factors, $X_1$, $X_2$, …, and $X_p$, which are given as follows:

$$E[Y|X_i]=\gamma_0^i+\gamma_1^i X_i \tag{1}$$

where $i = 1,2, …, p$. In addition, we assume that we know the mean relationships between any two pairs among the $p$ risk factors:

$$E[X_j|X_i]=\alpha_0^{ij}+\alpha_1^{ij}X_i \tag{2}$$

where $i, j = 1,2, …, p$ and $i \neq j$.

The goal of synthesis analysis is to determine the multivariate linear regression model between $Y$ and the $p$ risk factors:

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{3}$$

Note that the linear regression assumption (1) automatically holds under assumptions (2) and (3).

Taking the conditional expectation of the both sides of (3) given $X_i$, we obtain the following equation:

$$E(Y|X_i=x) = \beta_0 + \beta_1 E(X_1|X_i=x) + \cdots + \beta_{i-1} E(X_{i-1}|X_i=x) + \beta_i x + \cdots + \beta_p E(X_p|X_i=x) \tag{4}$$

for $i = 1, \ldots, p$. Combining (1), (2), and (4), we obtain the following result:

$$\gamma_0^i + \gamma_1^i x = \beta_0 + (\beta_1 \alpha_0^{i1} + \cdots + \beta_{i-1}\alpha_0^{i(i-1)} + \beta_{i+1}\alpha_0^{i(i+1)} + \cdots + \beta_p\alpha_0^{ip}) + (\beta_1\alpha_1^{i1} + \cdots + \beta_{i-1}\alpha_1^{i(i+1)} + \beta_i + \beta_{i+1}\alpha_1^{i(i+1)} + \cdots + \beta_p\alpha_1^{ip})x$$

for all $x$, where $i = 1, \ldots, p$. Therefore, we obtain the following two sets of equations:

$$\gamma_0^1 = \beta_0 + (\beta_2\alpha_0^{11} + \cdots + \beta_p\alpha_0^{1p})$$
$$\gamma_0^i = \beta_0 + (\beta_1\alpha_0^{i1} + \cdots + \beta_{i-1}\alpha_0^{i(i-1)} + \beta_{i+1}\alpha_0^{i(i+1)} + \cdots + \beta_p\alpha_0^{ip}) \tag{5}$$

for $i = 2, \ldots, p$; and

$$\gamma_1^1 = \beta_1 + \beta_2\alpha_1^{12} + \cdots + \beta_p\alpha_1^{1p}$$
$$\gamma_1^i = \beta_1\alpha_1^{i1} + \cdots + \beta_{i-1}\alpha_1^{i(i-1)} + \beta_i + \beta_{i+1}\alpha_1^{i(i+1)} + \cdots + \beta_p\alpha_1^{ip} \tag{6}$$

for $i = 2, \ldots, p$.

Let $\mathbf{M}$ be a $p \times p$ matrix with diagonal elements 1, and element $\alpha_1^{ij}$ when $i \neq j$; let $\boldsymbol{\beta} = (\beta_k, k = 1, \ldots, p)$, and $\boldsymbol{\gamma}_1 + (\gamma_1^k, k=1, \ldots, p)$. From (6), we obtain the following $p$ equations for the $p$ unknown slope parameters, $\beta_1, \ldots, \beta_p$:

$$\mathbf{M}\boldsymbol{\beta} = \boldsymbol{\gamma}_1 \tag{7}$$

Using Cramer's rule, we can easily solve the above $p$ simultaneous linear equations. Let us define the following determinants:

$$\mathbf{D}=\begin{vmatrix} 1 & \alpha_1^{12} & \alpha_1^{13} & \cdots & \alpha_1^{1p} \\ \alpha_1^{21} & 1 & \alpha_1^{23} & \cdots & \alpha_1^{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{p1} & \alpha_1^{p2} & \alpha_1^{p3} & \cdots & 1 \end{vmatrix}$$

$$\mathbf{D_1}=\begin{vmatrix} \gamma_1^2 & \alpha_1^{12} & \alpha_1^{13} & \cdots & \alpha_1^{1p} \\ \gamma_1^2 & 1 & \alpha_1^{23} & \cdots & \alpha_1^{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_1^p & \alpha_1^{p2} & \alpha_1^{p3} & \cdots & 1 \end{vmatrix} \quad \text{and}$$

$$\mathbf{D_p}=\begin{vmatrix} 1 & \alpha_1^{12} & \alpha_1^{13} & \cdots & \gamma_1^1 \\ \alpha_1^{21} & 1 & \alpha_1^{23} & \cdots & \gamma_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{p1} & \alpha_1^{p2} & \alpha_1^{p3} & \cdots & \gamma_1^p \end{vmatrix}$$

Cramer's rule gives us the following unique solution to the system of equations (8):

$$\beta_k = \frac{\mathbf{D_k}}{\mathbf{D}} \tag{8}$$

where $k = 1, \ldots, p$.

After obtaining estimates of the vector of slope parameters, $\beta$, we can derive an estimate for the intercept parameter, $\beta_0$, using any one of the $p$ equations given in (6). Hence, we have the following $p$ equations for the unknown intercept parameter $\beta_0$:

$$\beta_0 + 0 + \alpha_0^{12}\beta_2 + \alpha_0^{13}\beta_3 + \cdots + \alpha_0^{1(p-1)}\beta_{p-1} + \alpha_0^{1p}\beta_p = \gamma_0^1$$
$$\beta_0 + \alpha_0^{12}\beta_1 + 0 + \alpha_0^{23}\beta_3 + \cdots + \alpha_0^{2(p-1)}\beta_{p-1} + \alpha_0^{2p}\beta_p = \gamma_0^2$$
$$\vdots$$
$$\beta_0 + \alpha_0^{p1}\beta_1 + \alpha_0^{p2}\beta_2 + \alpha_0^{p3}\beta_3 + \cdots + \alpha_0^{p(p-1)}\beta_{p-1} + 0 = \gamma_0^p$$

Although there are $p$ equations for the parameter $\beta_0$, we show that the solution of $\beta_0$ is unique in Appendix A. We give a detailed description of our solution for the two-covariate case in Appendix B, and in Appendix C, we give an explicit formula for our synthesized parameters in cases with three and four covariates.

## 2.2. Variance estimation

The variance can be estimated using the delta method by assuming that the univariate parameter estimates $\gamma_0^{(i)}$ and $\gamma_1^{(i)}(i=1,\ldots,p)$ from individual univariate linear regression models, given by (1), are independent of each other [4]. Let $\alpha=(\alpha_0^{(ij)}, \alpha_1^{(ij)}, i, j=1,\ldots,p)$ and $\gamma=(\gamma_0^{(k)}, \gamma_1^{(k)}, k=1,\ldots,p)$.

By the well-known result from simple linear regression, we know:

$$n^{1/2}[(\boldsymbol{\alpha},\boldsymbol{\gamma})^{\mathrm{T}} - (\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)^{\mathrm{T}}] \rightarrow_d \mathbf{N}(\mathbf{0}, \sum)$$

where $\alpha_0$ and $\gamma_0$ are the true expected values of $\alpha$ and $\gamma$,

$$\sum = \begin{pmatrix} \Sigma_\alpha & 0 \\ 0 & \Sigma_\gamma \end{pmatrix}$$

Here

$$\sum\nolimits_\alpha = (\sigma_{\alpha_i^{kl}\alpha_j^{k'l'}}, i, j = 0, 1; k, l, k', l' = 1, 2, \ldots, p)$$

where $\sigma_{\alpha_i^{kl}\alpha_j^{k'l'}} (i, j = 0, 1; k, l, k', l' = 1, 2, \ldots, p)$ is the covariance between $\alpha_i^{(kl)}$ and $\alpha_j^{(k'l')}$, and

$$\sum\nolimits_\gamma = \begin{pmatrix} \sigma_{\gamma_0^1\gamma_0^1} & 0 & \cdots & 0 \\ 0 & \sigma_{\gamma_1^1\gamma_1^1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \ddots & \sigma_{\gamma_1^p\gamma_1^p} \end{pmatrix}$$

is the covariance matrix of the estimated parameters $\hat{\gamma}$.

The synthesized parameter estimates $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are functions of $\alpha$'s and $\gamma$'s, which can be expressed mathematically as:

$$\beta = \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$$

If the function $\mathbf{g}$ is differentiable, then the delta method gives the asymptotic variance of $\beta$ as follows:

$$\sum\nolimits_\beta = \nabla\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})^{\mathrm{T}} \sum \nabla\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \tag{9}$$

where $\nabla\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is the vector of derivatives of function $\mathbf{g}$ with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. We give an explicit formula for $\nabla\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ when $p = 2$ in Appendix B. Many programs, such as Mathematica, can perform derivatives symbolically, thereby making the variance calculation much easier, since the derivation of the exact form of the $\nabla\mathbf{g}$ is not required before the calculation.

## 2.3. Variance of predicted value

Once the estimates of parameters and their variances have been derived, we can calculate the covariance matrix of predicted values as follows:

$$\mathrm{Cov}(Y|X) = \mathrm{Cov}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}|\mathbf{X}) = \mathbf{X}^{\mathrm{T}} \sum\nolimits_\beta \mathbf{X}$$

where $\mathbf{X}^T$ is the transpose of the $\mathbf{X}$ matrix, and $\mathbf{\Sigma}_\beta$ is the covariance matrix of $\beta$, given by (9).

## 2.4. Mean-squared error of the predicted value and correlation between predicted and observed values

The mean-squared error (MSE) of the predicted value is given by

$$\mathrm{MSE}_{\widehat{Y}} = \frac{\sum_{i=1}^n (\widehat{Y}_i - Y_i)}{n}$$

where $\hat{Y}_i$ and $Y_i$ are the predicted and observed value of subject $i$, respectively. The correlation coefficient between $\hat{Y}_i$ and $Y_i$, $\rho$, can be calculated by

$$\rho = \frac{\mathrm{Cov}(\widehat{Y}_i, Y_i)}{\sqrt{\mathrm{Var}(\widehat{Y}_i)\mathrm{Var}(Y_i)}}$$

where $\mathrm{Cov}(\hat{Y}_i, Y_i)$ is the covariance between predicted and observed values.

## 3. SIMULATION STUDY

We conducted a simulation study to assess the performance of the proposed method in comparison with our previous method [5], denoted by SHR. We simulated data with two, three, and four predictor variables. For simplicity of presentation, we only reported the results for the two-predictors here, because the results for three-predictor and four-predictor cases are similar to those in the two-predictor case.

In each of these cases, we simulated independent variables from (1) a multivariate normal distribution, (2) a multivariate log-normal distribution, (3) a multivariate exponential distribution, and (4) a multivariate gamma distribution. We chose the variances of all the independent variables to be 1 and correlations for pairs of the independent variables to be 0.5. After simulating the independent variables $X$, we generated the dependent variable $Y$ by adding random normal errors to the mean model:

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i + \varepsilon \quad (p=2,3,4)$$

(10)

where $\varepsilon$ is a random error following the standard normal distribution.

We set the true regression parameters as follows: $(\beta_0, \beta_1, \beta_2) = (-5, 5, 3)$ for the two-variable setting, $(\beta_0, \beta_1, \beta_2, \beta_3) = (-5, 1, 3, 5)$ for the three-variable setting, and $(\beta_0, \beta_1, \beta_2, \beta_3 \beta_4) = (-5, 5, 4, 3, 1)$ for the four-variable setting. We divided each data set into $C_2^{p+1}$ $(p = 2,3,4)$ subsets with equal sample sizes. Here, $C_2^{p+1}$ denoted the total number of combinations of choosing 2 items from $(p + 1)$ items. In simulated data, each subset contained only one pair of variables chosen from $Y, X_1, \ldots, X_p$. The sample size (the total number of observations) used in simulation was 300 and 3000 (with equal size for each subset). For each of the above settings, we simulated a total number of 1000 data sets. As the results for the data from the skewed log-normal distribution were similar to those from the other skewed distributions, we only reported the

results for the normal and log-normal distributions. We reported the mean bias and MSE for estimated parameters in Tables I and II.

In order to evaluate the accuracy of predicted values using the new model, we simulated two data sets with equal sample sizes. One was used as the training set for model derivation, while the other was used as the validation data set. To evaluate prediction performance, we reported mean bias, MSE, and the mean of standard error estimates (SEEs) for predicted values in Tables III and IV. The SEEs were derived using the method developed in Sections 2.2 and 2.3. The correlations between predicted and observed values were also reported in the two tables.

Simulation results for the regression parameters showed that the mean bias and MSE of the estimated regression parameters using our new method were, in general, better than those using the SHR method, across all of the distributions and sample sizes considered here. The results also indicated that when the distributions of independent variables $X$ were heavily skewed (log-normal distribution), the bias and MSE of the estimated regression parameters using both methods were large, especially when sample sizes were small. Nonetheless, the results from our new method were much better than those from the SHR method under this situation.

The results for predicted values indicated that both the new method and the SHR method had similar correlations between observed and predicted values across all sample sizes and distributions. However, mean bias and MSE for predicted values derived from our new method were much smaller than those from the SHR method.

## 4. EXAMPLE

In this section, we analyzed a real-world example and compared the results using our new synthesis method and the SHR method. The data came from the 1999–2000 National Health and Nutritional Examination Survey [6]. There were five variables in this data set, including one outcome $Y$, systolic blood pressure, and four predictors, $X_1$, $X_2$, $X_3$, and $X_4$, which represented age, body mass index (BMI), serum total cholesterol level, and the natural log of serum triglycerides, respectively. First, we fitted a multivariate regression model to this data set, which would serve as the gold standard for this analysis. Next, we randomly divided the data set into the five mutually exclusive subsets with approximately equal sample sizes. The first four subsets included the outcome $Y$ and each of the four covariates, $X_1$, $X_2$, $X_3$, and $X_4$, respectively. The last subset contained all four covariates, which was used to derive pairwise correlations among the covariates. We applied the two synthesis methods to these five subsets to obtain estimated parameters in the multivariate regression model and reported the results in Table V. For comparison purposes, we also included the estimated parameters in the multivariate regression models obtained by the gold standard model in Table V.

The estimated parameters and their standard errors (SEs) from the gold standard and from both our new method and SHR method were listed in Table V (SE was not available by the SHR method). From these results, we observed that the new method produced the coefficient estimates that were comparable to those derived using the gold standard. However, the estimates for Intercept and LOGTRIG from the SHR method were varied somewhat from those derived using the gold standard method. As an illustration, the predicted value for a 65-year-old subject with the BMI of 19, the serum total cholesterol level of 190, and the serum triglycerides of 160 would be 134, 135, and 136, using the gold standard method, the new method, and the SHR method, respectively.

## 5. DISCUSSION

In this paper, we provided several enhancements to the existing SHR synthesis analysis methodology. These improvements allow for more robust estimates of the regression

parameters and predicted values when covariates are not normally distributed. Additionally, the new method allows for estimation of the variance of the resulting parameters and predicted outcomes.

Both the previously reported SHR method and our improved method allow for the building of multivariate regression models using univariate regression coefficients and two-way correlation coefficient data that are derived from different data sources. The underlying assumption is that each individual study is representative of the target population. However, the validity of the previously reported SHR synthesis analysis methodology relies on the normality assumption of the data. Although synthesis analysis is related to both meta-analysis and analysis of missing data, it is also different from these two traditional analyses in two important ways. First, while the goal of traditional meta-analysis is to combine the multivariate regression models with the same covariates from different studies, the goal of synthesis analysis is to create a multivariate linear regression model from univariate linear regression models on different covariates. Although the statistical problem that synthesis analysis address may be considered as one particular type of missing-data problem, unlike a traditional analysis, synthesis analysis does not require individual level data; rather, synthesis analysis only requires coefficient estimates of univariate linear regression models between the outcome and a covariate and between any two covariates.

Although the proposed method was developed to synthesize different univariate linear regression models with different covariates into multivariate linear regression models, it can be easily extended to the setting in which several studies are available for some (or all) of the univariate regression models. In this case, there would be variation among the parameter estimates. For example, if there are five studies available for the linear model, $E(Y \mid X_1)$, and six studies for the linear model, $E(X_1 \mid X_2)$, then we would have the five sets of estimates for the intercept and slope of the linear model of $Y$ on $X$, denoted by $\gamma_0^{j1}$ and $\gamma_1^{j1}$, for $j = 1, \ldots, 5$, and the six sets of estimates for the intercept and slope of the linear model of $X_1$ on $X_2$, denoted by $\alpha_0^{k21}$ and $\alpha_1^{k21}$, for $k = 1, \ldots, 6$.

In this case, we propose to first combine the results on the same univariate regression model from different studies into the one univariate regression model using the weighted mean of $\alpha_i^{jk}$ and $\gamma_i^j$, with the weight being the inverse sample size; that is,

$$\gamma_0^1 = \sum_{j=1}^{5} \frac{N_j}{N} \gamma_0^{j1}, \quad \gamma_1^1 = \sum_{j=1}^{5} \frac{N_j}{N} \gamma_1^{j1}$$

where $N_j$ is the sample size for the $j$th univariate model between $Y$ and $X_1$, and $N = \sum_{j=1}^{5} N_j$. Then, we apply the proposed synthesis method in Section 2 to obtain the multivariate regression model.

We performed a simulation study to assess the performance of the modified method in the two independent variables case, with one independent variables following a normal distribution and another following a log-normal distribution. We also compared this modified method with other combining methods, including mean, median, minimum, and maximum of multiple estimates for a same regression parameter. From these simulation results, we concluded that parameter estimates using the weighted mean had the smallest bias and MSE, and were very close to the bias and MSE using the gold standard. In addition, the predicted value using the weighted mean had the smallest bias, MSE, and SEE. We give a detailed description on our

simulation study and results in Appendix D. The computer software for implementing the proposed method is available at http://faculty.washington.edu/azhou.

## Acknowledgments

## References

1. Hackam DG, Anand SS. Emerging risk factors for atherosclerotic vascular disease. A critical review of the evidence. Journal of the American Medical Association 2003;290:932–940. [PubMed: 12928471]

2. Fruchart-Najib J, Bauge E, Niculescu LS, Pham T, Thomas B, Rommens C, Majd Z, Brewer B, Pennacchio LA, Fruchart JC. Mechanism of triglyceride lowering in mice expressing human apolipoprotein. Biochemical and Biophysical Research Communications 2004;319:397–404. [PubMed: 15178420]

3. Vasan RS. Biomarkers of cardiovascular disease: molecular basis and practical considerations. Circulation 2006;113:2335–2362. [PubMed: 16702488]

4. Casella, G.; Berger, RL. Statistical Inference. 2. Thomson Learning; Pacific Grove, CA: 2002.

5. Samsa G, Hu G, Root M. Combining information from multiple data sources to create multivariable risk models: illustration and preliminary assessment of a new method. Journal of Biomedicine and Biotechnology 2005;2:113–123. [PubMed: 16046816]

6. National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES). 1999–2000. Available from: http://www.cdc.gov/nchs/about/major/nhanes/

## APPENDIX A: SKETCH PROOF FOR UNIQUENESS OF INTERCEPT COEFFICIENT

Here we show that there is a unique solution for the intercept term $\beta_0$ with the $p$ equations (5), meaning that we need to show that the following $p$ solutions are equivalent:

$$\beta_0^{(1)} = \gamma_0^1 - (\alpha_0^{12}\beta_2 + \alpha_0^{13}\beta_3 + \cdots + \alpha_0^{1,p-1}\beta_{p-1} + \alpha_0^{1p}\beta_p)$$
$$\beta_0^{(2)} = \gamma_0^2 - (\alpha_0^{21}\beta_1 + 0 + \alpha_0^{23}\beta_3 + \cdots + \alpha_0^{2,p-1}\beta_{p-1} + \alpha_0^{2p}\beta_p)$$
$$\vdots$$
$$\beta_0^{(p)} = \gamma_0^p - (\alpha_0^{p1}\beta_1 + \alpha_0^{p2}\beta_2 + \alpha_0^{p3}\beta_3 + \cdots + \alpha_0^{p,p-1}\beta_{p-1} + 0)$$

Without losing generality, we only show that the solutions of the first two equations are equal, that is, $\beta_0^{(1)} = \beta_0^{(2)}$. The proof for other solutions is similar.

In order to show

$$\gamma_0^1 - \alpha_0^{12}\beta_2 - \alpha_0^{13}\beta_3 - \cdots - \alpha_0^{1,p-1}\beta_{p-1} - \alpha_0^{1p}\beta_p = \gamma_0^2 - \alpha_0^{21}\beta_1 - \alpha_0^{23}\beta_3 - \cdots - \alpha_0^{2,p-1}\beta_{p-1} - \alpha_0^{2p}\beta_p \tag{A1}$$

we add $E(X_1)\beta_1 + E(X_2)\beta_2 + \cdots + E(X_p)\beta_p$ to both sides of (A1), and then the left side of (A1) becomes

$$\gamma_0^1 + E(X_1)\beta_1 + (E(X_2) - \alpha_0^{12})\beta_2 + \cdots (E(X_{p-1}) - \alpha_0^{1,p-1})\beta_{p-1} + (E(X_p) - \alpha_0^{1p})\beta_p \tag{A2}$$

Because $E(X_j|X_i) = \alpha_0^{ij} + \alpha_1^{ij}X_i$, we can get the following result:

$$E(X_j) = E(E(X_j|X_i)) = \alpha_0^{ij} + \alpha_1^{ij}E(X_i) \tag{A3}$$

Hence, we can replace $(E(X_j) - \alpha_0^{ij})$ with $\alpha_1^{1j}E(X_1)$ in (A2) and obtain the following result:

$$\gamma_0^1 + E(X_1)\beta_1 + \alpha_1^{12}\beta_2 E(X_1) + \alpha_1^{1,p-1}\beta_{p-1}E(X_1) + \alpha_1^{1p}\beta_p E(X_1) = \gamma_0^1 + (\beta_1 + \alpha_1^{12}\beta_2 + \cdots + \alpha_1^{1p}\beta_p)E(X_1) \tag{A4}$$

Because $\beta_1$, …, and $\beta_p$ are the solutions of $M\beta = \gamma_1$, we can obtain the following result:

$$\beta_1 + \alpha_1^{12}\beta_2 + \cdots + \alpha_1^{1p}\beta_p = \gamma_1^1 \tag{A5}$$

Hence, the right side of (A4) becomes $\gamma_0^1 + \gamma_1^1 E(X_1)$, which equals to $E(Y)$ because $E(Y) = E(E(Y|X_1)) = E(\gamma_0^1 + \gamma_1^1 X_1) = \gamma_0^1 + \gamma_1^1 E(X_1)$.

Similarly, we can proof the right side of (A1) plus $E(X_1)\beta_1 + E(X_2)\beta_2 + \cdots + E(X_p)\beta_p$ is also equal to $E(Y)$. This completes the proof.

## APPENDIX B: SOLUTION FOR TWO PREDICTORS CASE

When $p = 2$, we can also have an explicit formula for the derivative of $\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ with respect to agr; and $\gamma$, $\nabla \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, for the two independent variables case. Here, $\nabla \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is used to calculate the variance of $\boldsymbol{\beta}$ and predicted values.

$$\nabla g(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \begin{pmatrix} \frac{\partial \widehat{\beta}_0}{\partial \alpha_0^{12}} & \frac{\partial \widehat{\beta}_1}{\partial \alpha_0^{12}} & \frac{\partial \widehat{\beta}_2}{\partial \alpha_0^{12}} \\ \frac{\partial \widehat{\beta}_0}{\partial \alpha_1^{12}} & \frac{\partial \widehat{\beta}_1}{\partial \alpha_1^{12}} & \frac{\partial \widehat{\beta}_2}{\partial \alpha_1^{12}} \\ \frac{\partial \widehat{\beta}_0}{\partial \alpha_0^{21}} & \frac{\partial \widehat{\beta}_1}{\partial \alpha_0^{21}} & \frac{\partial \widehat{\beta}_2}{\partial \alpha_0^{21}} \\ \frac{\partial \widehat{\beta}_0}{\partial \alpha_1^{21}} & \frac{\partial \widehat{\beta}_1}{\partial \alpha_1^{21}} & \frac{\partial \widehat{\beta}_2}{\partial \alpha_1^{21}} \\ \frac{\partial \widehat{\beta}_0}{\partial \gamma_0^{1}} & \frac{\partial \widehat{\beta}_1}{\partial \gamma_0^{1}} & \frac{\partial \widehat{\beta}_2}{\partial \gamma_0^{1}} \\ \frac{\partial \widehat{\beta}_0}{\partial \gamma_1^{1}} & \frac{\partial \widehat{\beta}_1}{\partial \gamma_1^{1}} & \frac{\partial \widehat{\beta}_2}{\partial \gamma_1^{1}} \\ \frac{\partial \widehat{\beta}_0}{\partial \gamma_0^{2}} & \frac{\partial \widehat{\beta}_1}{\partial \gamma_0^{2}} & \frac{\partial \widehat{\beta}_2}{\partial \gamma_0^{2}} \\ \frac{\partial \widehat{\beta}_0}{\partial \gamma_1^{2}} & \frac{\partial \widehat{\beta}_1}{\partial \gamma_1^{2}} & \frac{\partial \widehat{\beta}_2}{\partial \gamma_1^{2}} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{\gamma_1^2 - \alpha_1^{21}\gamma_1^1}{1 - \alpha_1^{12}\alpha_1^{21}} & 0 & 0 \\ \frac{\alpha_0^{12}\alpha_1^{21}}{(1-\alpha_1^{12}\alpha_1^{21})^2} & -\frac{\gamma_1^2}{1-\alpha_1^{12}\alpha_1^{21}} + \frac{\alpha_1^{21}(\gamma_1^1 - \alpha_1^{21}\gamma_1^2)}{(1-\alpha_1^{12}\alpha_1^{21})^2} & \frac{\alpha_1^{21}(\gamma_1^2 - \alpha_1^{21}\gamma_1^1)}{1-\alpha_1^{12}\alpha_1^{21}} \\ 0 & 0 & 0 \\ -\alpha_0^{12}\left[-\frac{\gamma_1^1}{1-\alpha_1^{12}\alpha_1^{21}} - \frac{\alpha_1^{12}(\gamma_1^2 - \alpha_1^{21}\gamma_1^1)}{(1-\alpha_1^{12}\alpha_1^{21})^2}\right] & \frac{\alpha_1^{12}(\gamma_1^1 - \alpha_1^{12}\gamma_1^2)}{(1-\alpha_1^{12}\alpha_1^{21})^2} & -\frac{\gamma_1^1}{1-\alpha_1^{12}\alpha_1^{21}} - \frac{\alpha_1^{21}(\gamma_1^2 - \alpha_1^{21}\gamma_1^1)}{(1-\alpha_1^{12}\alpha_1^{21})^2} \\ 1 & 0 & 0 \\ \frac{\alpha_0^{12}\alpha_1^{21}}{1-\alpha_1^{12}\alpha_1^{21}} & \frac{1}{1-\alpha_1^{12}\alpha_1^{21}} & -\frac{\alpha_1^{21}}{1-\alpha_1^{12}\alpha_1^{21}} \\ 0 & 0 & 0 \\ -\frac{\alpha_0^{12}}{1-\alpha_1^{12}\alpha_1^{21}} & -\frac{\alpha_1^{12}}{1-\alpha_1^{12}\alpha_1^{21}} & \frac{1}{1-\alpha_1^{12}\alpha_1^{21}} \end{pmatrix}$$

## APPENDIX C: SOLUTION FOR THREE AND FOUR PREDICTORS

When there are three predictors in the model, $\mathbf{D}$ and $\mathbf{D}_i$, $(i = 1, 2, 3)$ are given as follows:

$$\mathbf{D} = \begin{vmatrix} 1 & \alpha_1^{12} & \alpha_1^{13} \\ \alpha_1^{21} & 1 & \alpha_1^{23} \\ \alpha_1^{31} & \alpha_1^{32} & 1 \end{vmatrix} = (1 + \alpha_1^{12}\alpha_1^{23}\alpha_1^{31} + \alpha_1^{13}\alpha_1^{21}\alpha_1^{32}) - (\alpha_1^{12}\alpha_1^{21} + \alpha_1^{13}\alpha_1^{31} + \alpha_1^{23}\alpha_1^{32})$$

$$\mathbf{D}_1 = \begin{vmatrix} \gamma_1^1 & \alpha_1^{12} & \alpha_1^{13} \\ \gamma_1^2 & 1 & \alpha_1^{23} \\ \gamma_1^3 & \alpha_1^{32} & \alpha_1^{33} \end{vmatrix} = (\gamma_1^1\alpha_1^{33} + \alpha_1^{12}\alpha_1^{23}\gamma_1^3 + \alpha_1^{13}\gamma_1^2\alpha_1^{32}) - (\alpha_1^{13}\gamma_1^3 + \alpha_1^{12}\gamma_1^2\alpha_1^{33} + \gamma_1^1\alpha_1^{23}\alpha_1^{32})$$

$$\mathbf{D}_2 = \begin{vmatrix} 1 & \gamma_1^1 & \alpha_1^{13} \\ \alpha_1^{21} & \gamma_1^2 & \alpha_1^{23} \\ \alpha_1^{31} & \gamma_1^3 & \alpha_1^{33} \end{vmatrix} = (\gamma_1^2\alpha_1^{33} + \gamma_1^1\alpha_1^{23}\alpha_1^{31} + \alpha_1^{13}\alpha_1^{21}\gamma_1^3) - (\alpha_1^{13}\gamma_1^2\alpha_1^{31} + \gamma_1^1\alpha_1^{21}\alpha_1^{33} + \alpha_1^{23}\gamma_1^3)$$

and

$$\mathbf{D}_3 = \begin{vmatrix} 1 & \alpha_1^{12} & \gamma_1^1 \\ \alpha_1^{21} & 1 & \gamma_1^2 \\ \alpha_1^{31} & \alpha_1^{32} & \gamma_1^3 \end{vmatrix} = (\gamma_1^3 + \alpha_1^{12}\gamma_1^2\alpha_1^{31} + \gamma_1^1\alpha_1^{21}\alpha_1^{32}) - (\gamma_1^1\alpha_1^{31} + \alpha_1^{12}\alpha_1^{21}\gamma_1^3 + \gamma_1^2\alpha_1^{32})$$

If there are four predictors in the regression model, the $\mathbf{D}$ and $\mathbf{D}_i$, $(i = 1, 2, 3, 4)$ are as follows:

$$\mathbf{D}=\begin{vmatrix} 1 & \alpha_1^{12} & \alpha_1^{13} & \alpha_1^{14} \\ \alpha_1^{21} & 1 & \alpha_1^{23} & \alpha_1^{24} \\ \alpha_1^{31} & \alpha_1^{32} & 1 & \alpha_1^{34} \\ \alpha_1^{41} & \alpha_1^{42} & \alpha_1^{43} & 1 \end{vmatrix}=[(1+\alpha_1^{23}\alpha_1^{34}\alpha_1^{42})+\alpha_1^{24}\alpha_1^{32}\alpha_1^{43})-(\alpha_1^{23}\alpha_1^{32}+\alpha_1^{24}\alpha_1^{42}+\alpha_1^{34}\alpha_1^{43})]$$
$$-\alpha_1^{12}[(\alpha_1^{21}+\alpha_1^{23}\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\alpha_1^{43})-(\alpha_1^{24}\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}+\alpha_1^{21}\alpha_1^{34}\alpha_1^{43})]$$
$$+\alpha_1^{13}[(\alpha_1^{21}\alpha_1^{32}+\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\alpha_1^{42})-(\alpha_1^{24}\alpha_1^{32}\alpha_1^{41}+\alpha_1^{21}\alpha_1^{34}\alpha_1^{42}+\alpha_1^{31})]$$
$$-\alpha_1^{14}[(\alpha_1^{21}\alpha_1^{32}\alpha_1^{43}+\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}\alpha_1^{42})-(\alpha_1^{23}\alpha_1^{32}\alpha_1^{41}+\alpha_1^{31}\alpha_1^{43}+\alpha_1^{21}\alpha_1^{42})]$$

$$\mathbf{D}_1=\begin{vmatrix} \gamma_1^1 & \alpha_1^{12} & \alpha_1^{13} & \alpha_1^{14} \\ \gamma_1^2 & 1 & \alpha_1^{23} & \alpha_1^{24} \\ \gamma_1^3 & \alpha_1^{32} & 1 & \alpha_1^{34} \\ \gamma_1^4 & \alpha_1^{42} & \alpha_1^{43} & 1 \end{vmatrix}=\gamma_1^1[(1+\alpha_1^{23}\alpha_1^{34}\alpha_1^{42})+\alpha_1^{24}\alpha_1^{32}\alpha_1^{43})-(\alpha_1^{23}\alpha_1^{32}+\alpha_1^{24}\alpha_1^{42}+\alpha_1^{34}\alpha_1^{43})]$$
$$-\alpha_1^{12}[(\gamma_1^2+\alpha_1^{23}\alpha_1^{34}\gamma_1^4+\alpha_1^{24}\gamma_1^3\alpha_1^{43})-(\alpha_1^{24}\gamma_1^4+\alpha_1^{23}\gamma_1^3+\alpha_1^{34}\alpha_1^{43}\gamma_1^2)]$$
$$+\alpha_1^{13}[(\gamma_1^2\alpha_1^{32}+\alpha_1^{34}\gamma_1^4+\alpha_1^{24}\gamma_1^3\alpha_1^{42})-(\alpha_1^{24}\alpha_1^{32}\gamma_1^4+\gamma_1^3+\alpha_1^{34}\alpha_1^{42}\gamma_1^2)]$$
$$-\alpha_1^{14}[(\gamma_1^2\alpha_1^{32}\alpha_1^{43}+\gamma_1^4+\alpha_1^{23}\gamma_1^3\alpha_1^{42})-(\alpha_1^{23}\alpha_1^{32}\gamma_1^4+\alpha_1^{43}\gamma_1^3+\alpha_1^{42}\gamma_1^2)]$$

$$\mathbf{D}_2=\begin{vmatrix} 1 & \gamma_1^1 & \alpha_1^{13} & \alpha_1^{14} \\ \alpha_1^{21} & \gamma_1^2 & \alpha_1^{23} & \alpha_1^{24} \\ \alpha_1^{31} & \gamma_1^3 & 1 & \alpha_1^{34} \\ \alpha_1^{41} & \gamma_1^4 & \alpha_1^{43} & 1 \end{vmatrix}=[(\gamma_1^2+\alpha_1^{23}\alpha_1^{34}\gamma_1^4+\alpha_1^{24}\gamma_1^3\alpha_1^{43})-(\alpha_1^{24}\gamma_1^4+\alpha_1^{23}\gamma_1^3+\alpha_1^{34}\alpha_1^{43}\gamma_1^2)]$$
$$-\gamma_1^1[(\alpha_{21}+\alpha_1^{23}\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\alpha_1^{43})-(\alpha_1^{24}\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}+\alpha_1^{21}\alpha_1^{34}\alpha_1^{43})]$$
$$+\alpha_1^{13}[(\alpha_1^{21}\gamma_1^3+\gamma_1^2\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\gamma_1^4)-(\alpha_1^{24}\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}+\alpha_1^{21}\alpha_1^{34}\gamma_1^4)]$$
$$-\alpha_1^{14}[(\alpha_1^{21}\gamma_1^3\alpha_1^{43}+\gamma_1^2\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}\gamma_1^4)-(\alpha_1^{23}\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}\alpha_1^{43}+\alpha_1^{21}\gamma_1^4)]$$

$$\mathbf{D}_3=\begin{vmatrix} 1 & \alpha_1^{12} & \gamma_1^1 & \alpha_1^{14} \\ \alpha_1^{21} & 1 & \gamma_1^2 & \alpha_1^{24} \\ \alpha_1^{31} & \alpha_1^{32} & \gamma_1^3 & \alpha_1^{34} \\ \alpha_1^{41} & \alpha_1^{42} & \gamma_1^4 & 1 \end{vmatrix}=[(\gamma_1^3+\gamma_1^2\alpha_1^{34}\alpha_1^{42}+\alpha_1^{24}\alpha_1^{32}\gamma_1^4)-(\alpha_1^{24}\alpha_1^{42}\gamma_1^3+\gamma_1^2\alpha_1^{32}+\alpha_1^{34}\gamma_1^4)]$$
$$-\alpha_1^{12}[(\alpha_1^{21}\gamma_1^3+\gamma_1^2\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\gamma_1^4)-(\alpha_1^{24}\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}+\alpha_1^{21}\alpha_1^{34}\gamma_1^4)]$$
$$+\gamma_1^1[(\alpha_1^{21}\alpha_1^{32}+\alpha_1^{34}\alpha_1^{41}+\alpha_1^{24}\alpha_1^{31}\alpha_1^{42})-(\alpha_1^{24}\alpha_1^{32}\alpha_1^{41}+\alpha_1^{31}+\alpha_1^{21}\alpha_1^{34}\alpha_1^{42})]$$
$$-\alpha_1^{14}[(\alpha_1^{21}\alpha_1^{32}\gamma_1^4+\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}\alpha_1^{42})-(\gamma_1^2\alpha_1^{32}\alpha_1^{41}+\alpha_1^{31}\gamma_1^4+\alpha_1^{21}\gamma_1^3\alpha_1^{42})]$$

and

$$\mathbf{D}_4=\begin{vmatrix} 1 & \alpha_1^{12} & \alpha_1^{13} & \gamma_1^1 \\ \alpha_1^{21} & 1 & \alpha_1^{23} & \gamma_1^2 \\ \alpha_1^{31} & \alpha_1^{32} & 1 & \gamma_1^3 \\ \alpha_1^{41} & \alpha_1^{42} & \alpha_1^{43} & \gamma_1^4 \end{vmatrix}=[(\gamma_1^4+\alpha_1^{23}\gamma_1^3\alpha_1^{42})+\gamma_1^2\alpha_1^{32}\alpha_1^{43})-(\gamma_1^2\alpha_1^{42}+\alpha_1^{23}\alpha_1^{32}\gamma_1^4+\gamma_1^3\alpha_1^{43})]$$
$$-\alpha_1^{12}[(\alpha_1^{21}\gamma_1^4+\alpha_1^{23}\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}\alpha_1^{43})-(\gamma_1^2\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}\gamma_1^4+\alpha_1^{21}\gamma_1^3\alpha_1^{43})]$$
$$+\alpha_1^{13}[(\alpha_1^{21}\alpha_1^{32}\gamma_1^4+\gamma_1^3\alpha_1^{41}+\gamma_1^2\alpha_1^{31}\alpha_1^{42})-(\gamma_1^2\alpha_1^{32}\alpha_1^{41}+\alpha_1^{31}\gamma_1^4+\alpha_1^{21}\gamma_1^3\alpha_1^{42})]$$
$$-\gamma_1^1[(\alpha_1^{21}\alpha_1^{32}\alpha_1^{43}+\alpha_1^{41}+\alpha_1^{23}\alpha_1^{31}\alpha_1^{42})-(\alpha_1^{23}\alpha_1^{32}\alpha_1^{41}+\alpha_1^{31}\alpha_1^{43}+\alpha_1^{21}\alpha_1^{42})]$$

## APPENDIX D: SIMULATION STUDY ON THE MODIFIED SYNTHESIS

We performed a simulation study to assess the performance of the modified method, as described in the discussion section, for the two independent-variable case when the vector of two covariates follows a bivariate normal distribution or bivariate log-normal distribution. We also compared this modified method with the other combining methods, including mean, median, minimum, and maximum of multiple estimates for a same regression parameter. For each of the three univariate linear models, $E(Y / X_1)$, $E(Y / X_2)$, and $E(X_1 / X_2)$, there were the estimates from five different studies. We selected the sample size for each of the five studies

for each univariate model to be equal (1000 and 100) or unequal (100, 200, 500, 1200, 3000) or (10, 20, 50, 120, 300). We assessed the performance of the modified synthesis method using the weighted mean, mean, median, minimum, and maximum of combing results from the five studies.

Since our results on the simulated data from the bivariate normal distribution are similar to those on the simulated data from the bivariate log-normal distribution, we only report the results on the bivariate normal distribution case. Tables DI–DIV show the bias and MSE for each of the regression parameters $\beta_0, \beta_1, \beta_2$ as well as the mean bias, MSE, correlation, and SEE (mean of SE estimates) for the predicted values.

**Table I**

Mean bias and MSE of estimated regression parameters with two independent variables following a normal distribution.

| Sample size, method | Mean bias | | | MSE | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)$, New | −0.190 | −0.016 | 0.041 | 14.808 | 1.708 | 2.763 |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)$, SHR | 0.486 | −0.033 | −0.090 | 26.897 | 0.939 | 1.527 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)$, New | 0.031 | 0.000 | −0.007 | 1.346 | 0.033 | 0.067 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)$, SHR | 0.050 | −0.004 | −0.009 | 2.628 | 0.079 | 0.139 |

*The sample size for subsets with only outcome $Y$ and predictor $X_1$.

†The sample size for subsets with only outcome $Y$ and predictor $X_2$.

‡The sample size for subsets with only predictors $X_1$ and $X_2$.

**Table II**

Mean bias and MSE of estimated regression parameters with two independent variables following a log-normal distribution.

| Sample size, method | Mean bias | | | MSE | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)$, New | 0.146 | −0.081 | −0.042 | 42.032 | 3.676 | 4.799 |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)$, SHR | 10.377 | −1.104 | −1.412 | 933.764 | 82.249 | 80.029 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)$, New | −0.051 | −0.004 | 0.010 | 1.259 | 0.033 | 0.063 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)$, SHR | −0.015 | −0.013 | 0.006 | 2.349 | 0.080 | 0.126 |

[*] The sample size for subsets with only outcome $Y$ and predictor $X_1$.

[†] The sample size for subsets with only outcome $Y$ and predictor $X_2$.

[‡] The sample size for subsets with only predictors $X_1$ and $X_2$.

**Table III**

Mean bias, MSE, correlation and SE for predicted values with two independent variables following a normal distribution.

| Sample size, method | Mean bias | MSE | Correlation | SEE |
|---|---|---|---|---|
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)_{,\ \text{New}}$ | 0.0108 | 0.8046 | 0.9949 | 6.0496 |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)_{,\ \text{SHR}}$ | 14.1519 | 221.1321 | 0.9900 | — |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)_{,\ \text{New}}$ | −0.0092 | 0.0723 | 0.9996 | 1.8656 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)_{,\ \text{SHR}}$ | 14.0304 | 209.9250 | 0.9954 | — |

*Note:* Correlation is the mean correlation between observed and predicted values across simulations. SEE is the mean of standard error estimates for predicted values.

[*] The sample size for a subset with only outcome $Y$ and predictor $X_1$.

[†] The sample size for a subset with only outcome $Y$ and predictor $X_2$.

[‡] The sample size for a subset with only predictors $X_1$ and $X_2$.

**Table IV**

Mean bias, MSE, correlation and SE for predicted values with two independent variables following a log-normal distribution.

| Sample size, method | Mean bias | MSE | Correlation | SEE |
|---|---|---|---|---|
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)_{,\text{New}}$ | −10.2079 | 199764.1000 | 0.9376 | 254.6255 |
| $n=300(m_1^*=m_2^\dagger=m_3^\ddagger=100)_{,\text{SHR}}$ | 85.9998 | 47835.6600 | 0.9335 | — |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)_{,\text{New}}$ | 1.0546 | 17442.6700 | 0.9918 | 71.3051 |
| $n=3000(m_1^*=m_2^\dagger=m_3^\ddagger=1000)_{,\text{SHR}}$ | 66.5488 | 12226.2700 | 0.9328 | — |

*Note:* Correlation is the mean correlation between observed and predicted values across simulations. SEE is the mean of standard error estimates for predicted values.

[*] The sample size for subset with only outcome $Y$ and predictor $X_1$.

[†] The sample size for subset with only outcome $Y$ and predictor $X_2$.

[‡] The sample size for subset with only predictors $X_1$ and $X_2$.

**Table V**

Parameter estimates (SE) for the NHANES blood pressure example.

| Variables | Gold standard $\tilde{\beta}$ | NEW method $\beta_{NEW}$ | SHR method$^*\beta_{SHR}$ |
|-----------|-------------------------------|--------------------------|---------------------------|
| Intercept | 76.207 (2.556) | 73.482 (4.531) | 83.401 |
| AGE | 0.601 (0.017) | 0.634 (0.050) | 0.681 |
| BID | 0.379 (0.045) | 0.403 (0.128) | 0.337 |
| TCHOL | 0.024 (0.007) | 0.029 (0.018) | 0.006 |
| LOGTRIG | 1.374 (0.529) | 1.506 (0.931) | 0.160 |

$^*$ Cannot calculate SE using this method.

**Table DI**

Bias and MSE for estimated parameters with equal sample sizes.

| Method | Bias | | | MSE | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
| *Total sample size N = 1000 × 3 × 5 (equal sample size) = 15000* | | | | | | |
| Weighted mean (Mean) | 0.0023 | 0.0005 | −0.0005 | 0.2126 | 0.0026 | 0.0068 |
| Median | −0.0055 | −0.0016 | 0.0007 | 0.3792 | 0.0099 | 0.0183 |
| Minimum | 0.0219 | 0.0075 | −0.0036 | 0.5250 | 0.0140 | 0.0266 |
| Maximum | −0.0428 | −0.0084 | 0.0083 | 0.8344 | 0.0214 | 0.0399 |
| *Total sample size N = 100 × 3 × 5 (equal sample size) = 1500* | | | | | | |
| Weighted mean (Mean) | 0.1066 | 0.0107 | −0.0272 | 2.8586 | 0.0708 | 0.1509 |
| Median | 0.1781 | 0.0286 | −0.0433 | 4.2857 | 0.1156 | 0.2228 |
| Minimum | −0.2240 | −0.0181 | 0.0502 | 5.4686 | 0.1158 | 0.2820 |
| Maximum | −0.1285 | −0.0037 | −0.0373 | 11.4781 | 0.3338 | 0.5221 |

**Table DII**

Mean Bias, MSE, Correlation and SEE for predicted values with equal sample sizes.

| Method | Mean bias | MSE | Correlation | SEE |
|---|---|---|---|---|
| *Total sample size N = 1000 × 3 × 5 equal sample size) = 15 000* | | | | |
| Weighted mean (Mean) | 0.0019 | 0.0301 | 0.9998 | 0.9109 |
| *Total sample size N = 100 × 3 × 5 (equal sample size) = 1500* | | | | |
| Weighted mean (Mean) | 0.0126 | 0.3741 | 0.9956 | 3.0272 |

**Table DIII**

Bias and MSE for estimated parameters with unequal sample sizes.

| Method | Bias | | | MSE | | |
| --- | --- | --- | --- | --- | --- | --- |
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_0$ | $\beta_1$ | $\beta_2$ |
| *Total sample size N =(100+200+500+1200+3000) × 3 = 15000* | | | | | | |
| Weighted mean | 0.0196 | 0.0049 | −0.0056 | 0.5540 | 0.0251 | 0.0496 |
| Mean | −0.0231 | 0.0067 | −0.0076 | 0.8445 | 0.0567 | 0.0875 |
| Median | 0.0208 | 0.0073 | −0.0082 | 0.6676 | 0.0680 | 0.0329 |
| Minimum | −0.0538 | 0.0211 | −0.0103 | 3.0387 | 0.0733 | 0.1526 |
| Maximum | −0.0236 | 0.0040 | −0.0123 | 5.8060 | 0.1549 | 0.2748 |
| *Total sample size N = (10+20+50+120+300) × 3 = 1500* | | | | | | |
| Weighted mean | 0.1147 | 0.0268 | −0.0283 | 3.0217 | 0.3488 | 0.3621 |
| Mean | 0.2007 | 0.0234 | 0.0322 | 4.4266 | 0.3396 | 0.4212 |
| Median | 0.1583 | 0.0283 | −0.0379 | 7.2861 | 0.4095 | 0.3714 |
| Minimum | −2.8130 | −0.4905 | 0.6229 | 73.6571 | 2.0423 | 3.8998 |
| Maximum | −0.5346 | 0.1130 | 0.0830 | 529.7432 | 96.6978 | 61.0214 |

### Table DIV

Mean Bias, MSE, Correlation and SEE for predicted values with unequal sample sizes.

| Method | Mean bias | MSE | Correlation | SEE |
|---|---|---|---|---|
| *Total sample size N* = (100+200+500 +1200+3000) × 3 = 15 000 | | | | |
| Weighted mean | 0.0201 | 0.0994 | 0.9886 | 1.1105 |
| Mean | −0.0219 | 0.1134 | 0.9825 | 1.2773 |
| *Total sample size N* = (10+20+50+120+300) × 3 = 1500 | | | | |
| Weighted mean | −0.01580 | 0.3394 | 0.9900 | 4.1135 |
| Mean | 0.1993 | 0.3550. | 0.9789 | 4.3768 |