Published in final edited form as:

Epidemiology. 2017 January ; 28(1): 47–53. doi:10.1097/EDE.0000000000000554.

Collinearity and causal diagrams – a lesson on the importance of model specification

Enrique F. Schisterman¹, Neil J. Perkins¹, Sunni L. Mumford¹, Katherine A. Ahrens¹, and Emily M. Mitchell¹

¹Epidemiology Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD

Abstract

Background—Correlated data are ubiquitous in epidemiologic research, particularly in nutritional and environmental epidemiology where mixtures of factors are studied. Our objective is to demonstrate how highly correlated data arise in epidemiologic research and provide guidance on how to proceed analytically when faced with highly correlated data utilizing a directed acyclic graph approach.

Methods—We identified three fundamental structural scenarios in which high correlation between a given variable and the exposure can arise: intermediates, confounders, and colliders. For each of these scenarios we evaluated the consequences of increasing correlation between the given variable and the exposure on the bias and variance for the total effect of the exposure on the outcome using unadjusted and adjusted models. We derived closed form solutions for continuous outcomes using linear regression and empirically present our findings for binary outcomes using logistic regression.

Results—For models properly specified, total effect estimates remained unbiased even when there was almost perfect correlation between the exposure and a given intermediate, confounder, or collider. In general, as the correlation increased the variance of the parameter estimate for the exposure in the adjusted models increased, while in the unadjusted models it increased to a lesser extent or decreased.

Conclusion—Our findings highlight the importance of considering the causal framework under study when specifying regression models. Strategies that do not take into consideration the causal structure may lead to biased effect estimation for the original question of interest, even under high correlation.

Keywords

collider bias; collinearity; confounding; directed acyclic graphs; overadjustment bias

Corresponding Author and individual responsible for reprint requests: Dr. Enrique F. Schisterman, Ph.D., Epidemiology Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development 6100 Executive Blvd., 7B03, Rockville, MD 20852, Tel: (301) 435-6893, Fax: (301) 402-2084, schistee@mail.nih.gov.

Conflicts of Interest: The authors have no conflicts of interest to disclose.

INTRODUCTION

Exposure to environmental factors, such as chemicals or diet, often consists of exposure not only to a single factor but rather to an array of related individual factors. These individual factors may be highly correlated with each other, causing researchers to be concerned about the impact of collinearity when they attempt to identify individual effects. The biostatistical literature provides several methods for handling collinear variables such as scaling, ridge regression, principal components analysis, and the “leave one variable out” approach.¹ These methods are effective for providing unbiased and efficient estimators, and they operate under the assumption that the model is correctly specified. These approaches work especially well when comparing estimators from different correctly specified models to the true parameter. However, under highly correlated data, correct specification of the model may be in direct contrast to achieving stable and efficient estimators, highlighting the dilemma faced by epidemiologists using regression models to answer research questions about specific exposure-outcome relationships.

Whereas in biostatistics the emphasis tends to be placed on discovering new methods for estimating consistent and more efficient estimators, in epidemiology the consistency of the estimator is assumed. Thus, any observed bias is thought to arise from model misspecification due to improper omission or inclusion of covariates, or its functional form. Moreover, the correct specification of the model is often based on the hypothesized causal framework, and directed acyclic graphs (DAGs) are recommended as a tool for documenting this framework.² These two distinct philosophical approaches can lead to different modeling strategies for handling collinear variables depending on whether the motivation is primarily biostatistical or epidemiological.

The objective of our paper is to evaluate the effects of highly correlated data in epidemiologic research when the focus is on correct specification of the model, while assuming statistical consistency. Specifically, we use DAGs to identify three fundamental structural scenarios in which high correlation between a given variable and the exposure can arise. We then evaluate the consequences of increasing correlation between a given variable and the exposure using both linear and non-linear models. We derive closed-form solutions for the bias and variance for the total effect of the exposure on the outcome using linear regression, and we use simulations to empirically investigate the bias and variance of the exposure effect estimate in non-linear models. Lastly, we discuss a real data example of highly correlated variables in the context of reproductive epidemiology.

METHODS

Definition of Collinearity

Two variables are defined as collinear if one can be expressed as an exact or near linear combination of the other.³ For example, X_1 and X_2 are collinear if $X_1 = a + \lambda * X_2$, where a and λ are constants. In general, collinearity between variables is described by the magnitude of their correlation or the degree to which they approach being collinear.

Causal Scenarios Where Collinearity Can Arise

We identified three fundamental causal scenarios in which high correlation between a given variable and the exposure can arise: 1) intermediates, 2) confounders, and 3) colliders (Figure 1). Given each of these scenarios, we modeled the effect of an exposure (E) on the outcome (D), first not adjusting and then adjusting for the given third variable, identified in Figure 1 by an outlined box and denoted as (V) in the equations below. For all scenarios, the total effect of E on D was of primary interest. Equations (1) and (2) represent these unadjusted and adjusted models, respectively.

$$D = \beta_1 E + \varepsilon \quad (1)$$

$$D = \beta_2 E + \beta_3 V + \varepsilon \quad (2)$$

Depending on the direction of the association, V may refer to an intermediate (I) or confounding (C) variable. Conditional on the presence of confounding between I and D, I might also be a collider. The coefficient β_1 represents the marginal linear effect of the exposure (E) on the outcome (D). The interpretation of β_2 and β_3 depend on the causal scenario, which will be discussed below. All regression coefficients correspond to centered and standardized variables.

Intermediate variable—In the first causal scenario (Structure 1), we are interested in estimating the total effect of E on D in the presence of increasing correlation between E and I. The magnitude of the effect of E on I is represented by ρ , I on D by a , and E on D by b . In this case, the unadjusted model (equation 1) is correctly specified and, according to DAG theory, should yield an unbiased estimate of the total effect of E on D. The adjusted model (equation 2) is misspecified and therefore may lead to overadjustment bias.³ The magnitude of overadjustment bias under increasing correlation between I and E (and the effects on precision) will be the focus of the analysis for this causal scenario.

Confounder variable—The second causal scenario (Structure 2) evaluates the effect of increasing correlation between a confounder C and the exposure on the relationship between E and D. The magnitude of the effect of C on E is represented by ρ , E on D by a , and C on D by b . We would expect the correctly specified model for the total effect of E on D would adjust for C (equation 2); therefore, the unadjusted model, which is misspecified (equation 1), would have the potential to be biased. The bias under increasing correlation between E and C will be the focus of the analysis for this causal scenario.

Collider variable—The third causal scenario (Structure 3) depicts the effect of adjusting for an intermediate collider variable which is correlated with the exposure.^{4,5} In this scenario an unmeasured variable, U, is a common cause of both D and I (which is a descendent of E), with a representing the direct effect of I on D. Adjusting for I creates a spurious association between E and U, potentially biasing the total effect of E on D. The correctly specified model for the total effect of E on D would thus be the unadjusted model (equation 1). The

adjusted model (equation 2) is misspecified and may lead to collider stratification bias.⁶ The bias under increasing correlation between E and I will be the focus of the analysis for this causal scenario.

Effects of Collinearity on Bias and Precision

Given these three causal scenarios, we investigated the effect of increasing correlation (ρ) between the exposure (E) and the confounding or intermediate variable (V) on parameter estimation using both correctly specified and misspecified models. We examined these scenarios with a continuous outcome measure using linear regression and with a binary outcome using logistic regression. Although many techniques are available to perform non-linear regression, here we illustrate logistic regression due to its common use for binary outcomes in epidemiologic analyses. Again, the goal of this analysis is the estimation of the total effect of E on D.

Continuous Outcome—Assuming a continuous outcome, we derived the closed form of the bias, variance, and covariance for each causal scenario under the normality assumption (for analytical details see Appendix 1). Closed forms were verified empirically via simulation (data available upon request). Relative bias was calculated for $\hat{\beta}_1$ and $\hat{\beta}_2$ in relation to the total effect of E on D in the correctly specified model. For example, in

Structure 1 modeled under equation 1, the relative bias for $\hat{\beta}_1$ is $\frac{\hat{\beta}_1 - \beta_1}{\beta_1}$ since the correctly

specified model is equation 1. Whereas, in Structure 2 the relative bias for $\hat{\beta}_1$ is $\frac{\hat{\beta}_1 - \beta_2}{\beta_2}$ since the correctly specified model is equation 2. Relative changes in standard error (SE) for $\hat{\beta}_1$ and $\hat{\beta}_2$ were calculated for each respective parameter estimate across the range of possible correlations ($\rho=0.1$ to 0.99) in comparison to the case where the correlation was set to be low ($\rho=0.1$) in order to evaluate the impact of collinearity relative to an ideal scenario when essentially no collinearity is present. For example, in Structure 1, the relative SE for $\hat{\beta}_1$ is

$$\frac{SE(\hat{\beta}_1|\rho=0.1, \dots, 0.99)}{SE(\hat{\beta}_1|\rho=0.1)}, \text{ and for } \hat{\beta}_2 \text{ is } \frac{SE(\hat{\beta}_2|\rho=0.1, \dots, 0.99)}{SE(\hat{\beta}_2|\rho=0.1)}.$$

As shown in Figure 1, the bias due to model misspecification observed for each causal scenario is a function of the degree of correlation present, represented by ρ . The closed form solutions from Figure 1 are shown graphically in Figure 2 as relative changes in bias and SE as a function of increasing ρ , with a held fixed at 0.5. Specifically, for the intermediate variable scenario (Structure 1), β_1 is equivalent to $b + a\rho$, which is the total effect (indirect plus direct) of E on D, and when estimating the total effect of E on D, β_2 will be biased by an amount equivalent to $-a\rho$. Similarly, for the confounding scenario (Structure 2), the unadjusted model is biased by $b\rho$, and for the collider scenario (Structure 3) the adjusted

model is biased by $-\frac{b\gamma\rho}{1-\rho^2} - a\rho$. In these scenarios, ρ can be positive or negative, which can in turn affect the degree of bias of the effect of interest. Consequently, collinearity can exacerbate any model misspecification bias. On the other hand, collinearity does not induce any measurable bias when the model is correctly specified and correlations are less than

0.99. This is especially important under the confounding variable scenario, where correctly specifying the model calls for adjusting for a potentially collinear variable. Under the unlikely scenario of extreme collinearity ($\rho > 0.999$), the numerical matrix inversion required for estimation fails, resulting in unstable and biased parameters.

Efficiency for each causal scenario depends on both the degree of correlation present and whether the model is correctly specified. In the overadjustment and collider structures (Structures 1 and 3, respectively) where $\hat{\beta}_1$ is the correctly specified parameter, with the number of observations held fixed, we observe that efficiency increases (as noted by the decrease in relative SE) as the level of correlation increases. This trend is due to the improved predictive capacity of the exposure on the outcome, which thereby reduces the model's random error. In other words, the amount of variation in D attributable to I is now better captured in the relationship between E and D, due to the increasing similarity between E and I. When the model is misspecified, by adjusting for I, the precision drastically decreases as correlation increases, especially for high levels of correlation, resulting in both biased and imprecise estimates.

In the confounding variable scenario (Structure 2), although the misspecified model (equation 1) produces apparent gains in precision as the correlation increases, the corresponding increase in bias can result in invalid inference, by implicating a highly precise but inaccurate estimate of the regression parameter of interest. Figure 3 illustrates this scenario under a null effect of E on D, at different levels of confounding. Although the estimate of β_2 under the correctly specified model (adjusted for C) is less precise than the unadjusted model, the probability of a false positive is nominal at 5%. The unadjusted model, on the other hand, gives precise but biased estimates of β_2 , resulting in an inflated Type I Error and potentially invalid inference.

Thus, while the correctly specified model results in a decrease in precision as correlation increases, especially at high levels of correlation, the ability of the correctly specified model to maintain an unbiased estimate at all levels of collinearity validates the preference of this accurate, although imprecise, model over the misspecified version. In addition, only for extreme values of ρ close to 1 is $\hat{\beta}_2$ non-identifiable due to numerical instability. Under all other values of ρ , the correctly specified model produces the most desirable (e.g. unbiased) estimates of the total effect of the exposure on the outcome.

Binary Outcome—For binary outcomes, equations 1 and 2 were modified to represent logistic regressions:

$$\text{logit}(\pi) = \beta_1 * E + \varepsilon \quad (3)$$

$$\text{logit}(\pi) = \beta_2 * E + \beta_3 * V + \varepsilon \quad (4)$$

Equations 3 and 4 represent unadjusted and adjusted models, respectively, for Structures 1–3 similarly to the continuous outcome scenarios. The probability of outcome D is represented by π . The coefficient β_1^* represents the marginal log odds effect of an exposure on the outcome, and the interpretation of β_2^* and β_3^* depend on the causal scenario.

The bias and precision under logistic regression models were estimated using simulated data with the same node to node relations as in the linear regression models, with the exception of D being a binary variable with a marginal outcome probability of p (see Appendix).

Coefficients in equations 3 and 4 were estimated from the average values from 10,000 simulations of 10,000 observations for each set of parameters in the three causal scenarios. Parameter estimates were evaluated among combinations of the following direct effects: $\rho = (0.1 \text{ to } 0.9, 0.99)$, $a = (0.5)$, and $b = (0.5, 1.0)$. Relative bias and SE were calculated in the same manner as for the continuous outcome models.

The simulated expected values of $\hat{\beta}_1^*$, $\hat{\beta}_2^*$ and $\hat{\beta}_3^*$ for Structures 1, 2, and 3 (Figure 4) were remarkably close to the corresponding closed form solutions from linear regression shown in Figure 1. Subsequently, the bias from model misspecification was similar to that of the linear models in direction and magnitude. For Structure 1, the unadjusted model remained unbiased, while the bias in $\hat{\beta}_2^*$ increased slightly as ρ increased, with little deviation due to changes in the strength of E-D (b) relationship. In Structure 2, $\hat{\beta}_2^*$ is unbiased, while $\hat{\beta}_1^*$ bias increased with ρ and also with the C-D (b) relation. For Structure 3, $\hat{\beta}_2^*$ has a relative bias that was driven by stronger U-D (b) relations evident when $\rho < 0.5$. The bias of $\hat{\beta}_2^*$ in Model 3 increased gradually as ρ increased up to $\rho = 0.8$, where the bias then began to increase dramatically. For Model 3, $\hat{\beta}_1^*$ remained unbiased.

Variance estimates for binary outcome models did not exhibit trends identical to those found in simulations performed under the linear regression scenario. For Structures 1–3, in most cases $SE(\hat{\beta}_1^*)$ and $SE(\hat{\beta}_2^*)$ increased as ρ increased, with $SE(\hat{\beta}_1^*)$ increasing to a lesser extent than $SE(\hat{\beta}_2^*)$. Indeed, once ρ exceeded 0.6 the $SE(\hat{\beta}_2^*)$ increased markedly, more than doubling when $\rho = 0.9$ compared to $\rho = 0.1$.

Real Data Example: Leptin and Estrogen

As is often the case in practice, investigators are faced with the decision of how to handle two highly correlated variables. Studies examining the effects of leptin on estrogen levels provide a natural setting for a discussion of collinearity because of the high correlation found between leptin and measures of body fat. In fact, correlations of 0.76 were observed in the BioCycle Study.^{7,8} Under such strong correlations the traditional approach would be to only include body fat or leptin in the model, and not both, or to include less correlated proxies of adiposity or composite variables,^{9–12} to avoid potential consequences of collinearity without regard to the underlying causal structure. For our study, we were interested in the effect of leptin on estrogen concentrations on the expected day of ovulation holding adiposity fixed. Based on *a priori* knowledge of the associations between leptin, estrogen, and body fat, we hypothesized that body fat was a confounder of the leptin-

estrogen relationship because body fat is a source of both leptin and estrogen production (common cause).¹³ This relationship is represented in Structure 2. Under this causal structure, body fat should be included in the adjusted model to obtain unbiased estimates and we can estimate the expected absolute bias due to model misspecification if we did not adjust for body fat in our analyses using the equations in Figure 1. Under the unadjusted model, $\hat{\beta}_1=0.154$ and $SE(\hat{\beta}_1)=0.062$. In the adjusted model, the effect of leptin on estrogen was $\hat{\beta}_2=0.256$ (equivalent to *a*) with $SE(\hat{\beta}_2)=0.095$, and the effect of body fat on estrogen was $\hat{\beta}_3=-0.135$ (equivalent to *b*) with $SE(\hat{\beta}_3)=0.095$. According to the formulas from Figure 1 (and derived in the appendix), the expected absolute bias in the unadjusted model would be $b\rho$, which was $-0.135*0.76=-0.102$. Algebraically, this was equivalent to $\hat{\beta}_1 - \hat{\beta}_2$ and represents a relative bias of -40% . In addition, we further observed that the variance of the adjusted model was greater than the unadjusted model ($SE(\hat{\beta}_1)<SE(\hat{\beta}_2)$). Although adjustment for body fat increased the variance estimate due to correlation between the predictors, failure to account for the confounding effect of body fat on this relationship resulted in a biased estimate of this association ($\hat{\beta}_1=0.154$) which was more precisely estimate than the true estimate ($\hat{\beta}_2=0.254$). Thus, we can conclude from this analysis that, under the assumed causal structure, leptin is positively and significantly associated with estrogen after conditioning on body fat (p-value = 0.007). This example highlights the need to clearly specify the causal structure to determine the appropriate modeling strategy.

DISCUSSION

Our findings highlight the importance of considering the causal framework under study when specifying regression models in the presence of highly correlated data. Here we evaluated the effects of collinearity in epidemiologic research under three fundamental causal scenarios and found that when models were correctly specified based on the underlying causal structure that parameter estimates of the effect of interest remained unbiased. This was the case even when there was almost perfect collinearity of the exposure with another variable under intermediate, confounder, or collider scenarios. Incorrectly specifying the models, such as adjusting for an intermediate or collider variable, resulted in biased effect estimates, with the bias being magnified as correlation increased.

There is often the concern that including two highly correlated variables in a model will result in numerical instability. Indeed, as discussed in Kleinbaum,¹ the potential danger of collinearity stems from the potential singularity of the Hessian matrix, which is obtained by taking the second derivative of the log-likelihood. When the correlation between predictors is perfect ($=1$), inverting the Hessian to obtain standard errors is equivalent to dividing by 0, which produces infinite or undefined standard errors. This issue is similar to the problem of complete separation in logistic regression, which violates the consistency assumption. This real concern about numerical instability has somewhat been alleviated by the advancement of computing technology as the ability of computer programs to accurately invert nearly singular matrices has significantly improved. In addition, the probability of selecting distinct predictors that are perfectly or nearly perfectly correlated in practice is highly unlikely. As described in Kleinbaum,¹ we observed that numerical instability occurs when correlation

exceeds 0.99, equivalent to an R^2 value of greater than 0.95, or a Variance Inflation Factor > 10 . This can arise naturally in situations where there are different powers of the same variable, such as age + age² or when multiple interaction terms are included in the model. In epidemiological studies, however, this degree of correlation is unlikely to occur between distinct variables even when variables are known to be highly associated such as spousal characteristics like age, body mass index, or education.^{14–15} If such a degree of correlation does occur, it is likely that two variables are essentially measuring the same thing in two different ways. Thus, misspecifying the causal structure of the model in order to allay collinearity concerns is much more likely to create bias rather than to be of practical use in avoiding numerical instability.

In the statistical literature, concerns about collinearity stem from the desire to improve parameter estimates' precision, which can be inflated when highly correlated covariates are included in the model. Epidemiologists, on the other hand, tend to be interested in a specific effect from a correctly specified causal model, even in the presence of highly correlated data, despite a potential loss of statistical efficiency.² Under such a viewpoint, bias due to misspecification of the causal associations is a more devastating result than reduced precision of parameter estimates correctly estimated. Further, a precisely but incorrectly estimated parameter may lead to invalid inferences by increasing the risk of false positive findings.

In this study, we have analytically derived the bias and relative efficiency due to model misspecification in the linear regression setting, i.e., the bias induced to the total effect by overadjusting or adjusting for a variable that induces collider stratification bias. To our knowledge, this is the first such derivation existing under the causal structures of intermediate and collider variables. In addition, we used simulation studies to empirically assess the potential magnitude of the bias and loss of precision under various levels of collinearity in the nonlinear regression setting. While we have demonstrated that high correlation between variables can inflate standard errors, the causal relationship between the variables plays a more important role in delivering valid inference and should be the primary consideration for choosing which covariates to include in the regression formulation. To illustrate these findings applied to actual data, as well as to promote the application of these methods, we conducted a step-by-step demonstration using an example from the BioCycle study regarding the association between leptin and estrogen levels.

While previous research has explored the potential for bias when inappropriately adjusting for a collider or mediator,^{1,16,17} our study generalizes this idea to incorporate continuous variables, as well as to explicitly quantify the change in precision due to collinearity in these circumstances. As with any etiological study, however, the validity of the inferences is a reflection of the accuracy of the assumptions; a faulty DAG could produce flawed inference, regardless of the impact of collinearity. Since only a limited number of scenarios were considered in this study, alternative associations among the variables are likely to produce varied results, although the main conclusions are expected to remain the same. In addition, while we focused on total effects of an exposure on an outcome of interest, the same principles may apply if the estimation of the direct or indirect effects were of interest. In conclusion, highly correlated or collinear data is present under a variety of causal structures.

Our paper highlights the danger of adjusting or not adjusting for variables without considering their causal framework, and selecting models based solely on how precisely parameters can be estimated may lead to biased effect estimation for the original question of interest.

Acknowledgments

Source of Funding: This research was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health

APPENDIX

Variable Construction

Nodes in Models 1–3 were assumed to be or generated in following manner:

Continuous Outcome

Model 1: Intermediate— $E \sim \text{Normal}(\mu = 0, \sigma = 1)$ because it had no parent node.

$I = \rho E + \sqrt{1 - \rho^2} \varepsilon$ and $D = aI + bE + \sqrt{1 - a^2 - b^2 - 2ab\rho} \varepsilon$, where $\varepsilon \sim \text{Normal}(\mu = 0, \sigma = 1)$ and ρ and a were between 0 and 1 and $b < -a\rho + \sqrt{a^2\rho^2 + 1 - a^2}$ to maintain a constant variance of 1 at all nodes.

Model 2: Confounder— $C \sim \text{Normal}(\mu = 0, \sigma = 1)$ because it had no parent node.

$E = \rho C + \sqrt{1 - \rho^2} \varepsilon$ and $D = aE + bC + \sqrt{1 - a^2 - b^2 - 2ab\rho} \varepsilon$, where $\varepsilon \sim \text{Normal}(\mu = 0, \sigma = 1)$ and ρ and a were between 0 and 1 and $b < -a\rho + \sqrt{a^2\rho^2 + 1 - a^2}$ to maintain a constant variance of 1 at all nodes.

Model 3: Collider— $E \sim \text{Normal}(\mu = 0, \sigma = 1)$ and $U \sim \text{Normal}(\mu = 0, \sigma = 1)$ because they had no parent nodes.

$I = \rho E + \gamma U + \sqrt{1 - \rho^2 - \gamma^2} \varepsilon$ and $D = aI + bU + \sqrt{1 - a^2 - b^2 - 2ab\gamma} \varepsilon$, where $\varepsilon \sim \text{Normal}(\mu = 0, \sigma = 1)$ and ρ and a were between 0 and 1 and $\gamma = 0.9 \sqrt{1 - \rho^2}$ and $b < -a\gamma + \sqrt{a^2\gamma^2 + 1 - a^2}$ to maintain a constant variance of 1 at all nodes.

Binary Outcome

For Models 1–3, nodes E, I, C and U were generated similar to the continuous outcome scenarios. The outcome was $D \sim \text{Binominal}(p)$, where p changed between models.

Model 1: Intermediate

$$p = 1 / (1 + \exp(-aI - bE))$$

Model 2: Confounder

$$p=1/(1+\exp(-aE - bC))$$

Model 3: Collider

$$p=1/(1+\exp(-aI - bU))$$

Expected Values and Variances

Based on the assumed scenarios above, the following closed form solutions were derived for linear regression coefficients as well as accompanying variance and covariance terms. Given

the linear regression model $D = \vec{\beta}^T \vec{X}$ the effect estimators $\widehat{\vec{\beta}} = (X^T X)^{-1} (X^T D)$ had the expected value of $E[\widehat{\vec{\beta}}] = Cov(\vec{X}, \vec{X})^{-1} Cov(\vec{X}, D)$ and covariance

$Cov[\widehat{\vec{\beta}}] = (W^2/(n-p)) Cov(\vec{X}, \vec{X})^{-1}$ where n is the sample size, p is the number of covariates and $W^2 = 1 - Cov(\vec{X}, D)^T E[\widehat{\vec{\beta}}]$.

Model 1: Intermediate

$$E[\widehat{\beta}_1] = Cov(\vec{X}, \vec{X})^{-1} Cov(\vec{X}, D) = [1]^{-1} (b+a\rho) = b+a\rho$$

$$\begin{aligned} E[\widehat{\beta}_{23}] &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{pmatrix} b+a\rho \\ a+b\rho \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{pmatrix} b+a\rho \\ a+b\rho \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{pmatrix} b(1-\rho^2) \\ a(1-\rho^2) \end{pmatrix} \\ &= \begin{pmatrix} b \\ a \end{pmatrix} \end{aligned}$$

$$\begin{aligned} Cov[\widehat{\beta}_1] &= \left(\frac{1}{(n-p)} W^2 \right) Cov(\vec{X}, \vec{X})^{-1} = \frac{1}{(n-2)} (1 - (b+a\rho)^2) [1]^{-1} \\ &= \frac{1}{(n-2)} (1 - (b+a\rho)^2) \end{aligned}$$

$$Cov[\widehat{\beta}_{23}] = \frac{1}{(n-3)} (1 - (a^2 + b^2 + 2ab\rho)) \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

Model 2: Confounder

$$E[\widehat{\beta}_1] = Cov(\vec{X}, \vec{X})^{-1} Cov(\vec{X}, D) = [1]^{-1}(a+b\rho) = a+b\rho$$

$$E[\widehat{\beta}_{23}] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{pmatrix} a+b\rho \\ b+a\rho \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{pmatrix} a+b\rho \\ b+a\rho \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{pmatrix} a(1-\rho^2) \\ b(1-\rho^2) \end{pmatrix} \\ = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$Cov[\widehat{\beta}_1] = \left(\frac{1}{(n-p)} W^2 \right) Cov(\vec{X}, \vec{X})^{-1} = \frac{1}{(n-2)} (1 - (a+b\rho)^2) [1]^{-1} \\ = \frac{1}{(n-2)} (1 - (a+b\rho)^2)$$

$$Cov[\widehat{\beta}_{23}] = \frac{1}{(n-3)} (1 - (a^2 + b^2 + 2ab\rho)) \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}^{-1}$$

Model 3: Collider

$$E[\widehat{\beta}_1] = Cov(\vec{X}, \vec{X})^{-1} Cov(\vec{X}, D) = [1]^{-1}(a\rho) = a\rho$$

$$E[\widehat{\beta}_{23}] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{pmatrix} a\rho \\ a+b\gamma \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{pmatrix} a\rho \\ a+b\gamma \end{pmatrix} \\ = \frac{1}{(1-\rho^2)} \begin{pmatrix} -b\gamma\rho \\ a(1-\rho^2) + b\gamma \end{pmatrix}$$

$$Cov[\widehat{\beta}_1] = \left(\frac{1}{(n-p)} W^2 \right) Cov(\vec{X}, \vec{X})^{-1} = \frac{1}{(n-2)} (1 - (a\rho)^2) [1]^{-1} = \frac{1-a^2\rho^2}{(n-2)}$$

$$Cov[\widehat{\beta}_{23}] = \frac{1}{(n-3)} \left((1 - a^2 - 2ab\gamma) - \frac{b^2\gamma^2}{(1-\rho^2)} \right) \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}^{-1}, \text{ where}$$

$$W^2 = 1 - Cov(\vec{X}, D)^T E[\widehat{\beta}] = 1 - [a\rho \ a+b\gamma] \frac{1}{(1-\rho^2)} \begin{bmatrix} -b\gamma\rho \\ a(1-\rho^2) + b\gamma \end{bmatrix} \\ = \frac{1}{(1-\rho^2)} ((1-\rho^2) - 2ab\gamma(1-\rho^2) - a^2(1-\rho^2) - b^2\gamma^2) \\ = 1 - a^2 - 2ab\gamma - \frac{b^2\gamma^2}{(1-\rho^2)}$$

References

1. Kleinbaum, DGKLL., Muller, KE., Nizam, A. Applied regression analysis and other multivariable methods. 3rd. Pacific Grove: Duxbury Press; 1998. p. 237-248.
2. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol. 2002; 155:176-84. [PubMed: 11790682]

3. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009; 20:488–95. [PubMed: 19525685]
4. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004; 15:615–25. [PubMed: 15308962]
5. Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of colliderstratification bias and the birthweight paradox. *Paediatr Perinat Epidemiol*. 2009; 23:394–402. [PubMed: 19689488]
6. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010; 39:417–20. [PubMed: 19926667]
7. Ahrens K, Mumford SL, Schliep KC, et al. Serum leptin levels and reproductive function during the menstrual cycle. *Am J Obstet Gynecol*. 2014; 210:248 e1–9. [PubMed: 24215851]
8. Wactawski-Wende J, Schisterman EF, Hovey KM, et al. BioCycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatr Perinat Epidemiol*. 2009; 23:171–84. [PubMed: 19159403]
9. Ajala OM, Ogunro PS, Elusanmi GF, Ogunyemi OE, Bolarinde AA. Changes in serum leptin during phases of menstrual cycle of fertile women: relationship to age groups and fertility. *Int J Endocrinol Metabol*. 2013; 11:27–33.
10. Asimakopoulos B, Milousis A, Gioka T, et al. Serum pattern of circulating adipokines throughout the physiological menstrual cycle. *Endocrine Journal*. 2009; 56:425–33. [PubMed: 19225215]
11. Cella F, Giordano G, Cordera R. Serum leptin concentrations during the menstrual cycle in normal-weight women: effects of an oral triphasic estrogen-progestin medication. *European J Endocrinol*. 2000; 142:174–8. [PubMed: 10664527]
12. Hardie L, Trayhurn P, Abramovich D, Fowler P. Circulating leptin in women: a longitudinal study in the menstrual cycle and during pregnancy. *Clin Endocrinol*. 1997; 47:101–6.
13. Ahima RS, Flier JS. Adipose tissue as an endocrine organ. *Trends Endocrinol Metab*. 2000; 11:327–32. [PubMed: 10996528]
14. Price RA, Vandenberg SG. Spouse similarity in American and Swedish couples. *Behav Genet*. 1980; 10:59–71. [PubMed: 7191703]
15. Knuiman MW, Divitini ML, Bartholomew HC, Welborn TA. Spouse correlations in cardiovascular risk factors and the effect of marriage duration. *Am J Epidemiol*. 1996; 143:48–53. [PubMed: 8533746]
16. Howards PP, Schisterman EF, Heagerty PJ. Potential confounding by exposure history and prior outcomes: an example from perinatal epidemiology. *Epidemiology*. 2007; 18:544–51. [PubMed: 17879426]
17. Greenland S. Quantifying biases in causal models: classical confounding vs colliderstratification bias. *Epidemiology*. 2003; 14:300–6. [PubMed: 12859030]

Causal structure	Parameter	Expected Value	Expected Absolute Bias for Total Effect of E on D	Expected Variance
Structure 1: Intermediate (Overadjustment)	β_1	$b + a\rho$	0	$\frac{1 - (b + a\rho)^2}{n - 2}$
	β_2	b	$-a\rho$	$\frac{1 - (a^2 + b^2 + 2ab\rho)}{(n - 3)(1 - \rho^2)}$
Structure 2: Confounding	β_1	$a + b\rho$	$b\rho$	$\frac{1 - (a + b\rho)^2}{n - 2}$
	β_2	a	0	$\frac{1 - (a^2 + b^2 + 2ab\rho)}{(n - 3)(1 - \rho^2)}$
Structure 3: Collider	β_1	$a\rho$	0	$\frac{1 - a^2\rho^2}{n - 2}$
	β_2	$-\frac{b\gamma\rho}{1 - \rho^2}$	$-\frac{b\gamma\rho}{1 - \rho^2} - a\rho$	$\frac{1}{(n - 3)(1 - \rho^2)} \left(1 - a^2 - 2ab\gamma - \frac{b^2\gamma^2}{(1 - \rho^2)} \right)$

Figure 1.

Expected value, bias, and variance of parameter estimates from linear regression models with β_1 and β_2 representing unadjusted and adjusted effect estimates of exposure on outcome, respectively.

a , b , and γ represent direct effects of standardized factors, ρ represents both correlation and direct effect between E and I/C

n = number of observations

Figure 2.

Relative bias and standard error for parameter estimates β_1 (unadjusted model) and β_2 (adjusted model) with increased correlation ρ between the given variable and exposure for continuous outcomes.

SE, standard error

See Figure 1 for diagrams of causal structures. All other associations between variables were held fixed (see appendix for details). Relative bias is in reference to 0 (no bias). Relative SE is calculated in reference to SE for each parameter estimate when the correlation between the given variable and exposure is 0.1.

Figure 3.

Point estimates for b (effect of C on D) and β_2 (effect of E on D) for a null effect under Structure 2 (Confounding) under various levels of confounding ($b = 0.01, 0.1, 0.5$) and correlation (top row: 0.990, bottom row: 0.999); a significant estimate for β_2 represents a Type I Error.

SE, standard error

See Figure 1 for diagrams of causal structures. The relationship between the exposure, confounder, and confounder of the collider-outcome relationship is represented by b for Structures 1–3, respectively, and was held at two values: 0.5 (weaker) and 1.0 (stronger). All other associations between variables were held fixed (see appendix for details). Relative bias is in reference to 0 (no bias). Relative SE is calculated in reference to SE for each parameter estimate when the correlation between the given variable and exposure is 0.1.

Figure 4. Relative bias and standard error for parameter estimates β_1 (unadjusted model) and β_2 (adjusted model) with increased correlation ρ between the given variable and exposure for binary outcomes.