Peng Ding* and Luke W. Miratrix

To Adjust or Not to Adjust? Sensitivity Analysis of *M*-Bias and Butterfly-Bias

Abstract: "*M*-Bias," as it is called in the epidemiologic literature, is the bias introduced by conditioning on a pretreatment covariate due to a particular "*M*-Structure" between two latent factors, an observed treatment, an outcome, and a "collider." This potential source of bias, which can occur even when the treatment and the outcome are not confounded, has been a source of considerable controversy. We here present formulae for identifying under which circumstances biases are inflated or reduced. In particular, we show that the magnitude of *M*-Bias in linear structural equation models tends to be relatively small compared to confounding bias, suggesting that it is generally not a serious concern in many applied settings. These theoretical results are consistent with recent empirical findings from simulation studies. We also generalize the *M*-Bias setting (1) to allow for the correlation between the latent factors to be nonzero and (2) to allow for the collider to be a confounder between the treatment and the outcome. These results demonstrate that mild deviations from the *M*-Structure tend to increase confounding bias more rapidly than *M*-Bias, suggesting that choosing *to* condition on any given covariate is generally the superior choice. As an application, we re-examine a controversial example between Professors Donald Rubin and Judea Pearl.

Keywords: causality, collider, confounding, controversy, covariate

DOI 10.1515/jci-2013-0021

1 Introduction

The hallmark of an observational study is selection bias [1–3]. Many statisticians believe that "there is no reason to avoid adjustment for a variable describing subjects before treatment" in observational studies [4, pp. 76], because "typically, the more conditional an assumption, the more generally acceptable it is" [5]. This advice, recently dubbed the "pretreatment criterion" [6], is widely used in empirical studies, as more covariates generally seem to make the ignorability assumption, i.e. the assumption that conditionally on the observed pretreatment covariates, treatment assignment is independent of the potential outcomes [7], more plausible. And, as the validity of causal inference in observational studies relies strongly on this (untestable) assumption [7], it seems reasonable to make all efforts to render it plausible.

However, other researchers [8–12], mainly from the causal diagram community, do not accept this view because of the possibility of a so-called *M*-Structure, illustrated in Figure 1(c). In sharp contrast to Rubin and Rosenbaum's advice, Pearl [9, 13] warns practitioners that spurious bias may arise due to adjusting for a collider *M* in an *M*-Structure, *even if it is a pretreatment covariate*. This form of bias, typically called *M*-Bias, a special version of so-called "collider bias," has since generated considerable controversy and confusion.

We attempt to resolve some of these debates by an analysis of *M*-Bias under the causal diagram or directed acyclic graph (DAG) framework. For readers unfamiliar with the terminologies from the DAG (or Bayesian Network) literature, more details can be found in Pearl [8, 14] . We here use only a small part of this larger framework. Arguably the most important structure in the DAG, and certainly the one at root of

***Corresponding author: Peng Ding,** Department of Statistics, Harvard University, Cambridge, MA 02138, USA, E-mail: pengding@fas.harvard.edu

Luke W. Miratrix, Department of Statistics, Harvard University, Cambridge, MA 02138, USA, E-mail: lmiratrix@stat.harvard.edu

almost all controversy, is the "V-Structure" illustrated in Figure 1(b). Here, U and W are marginally independent of a common outcome M, which shapes a "V" with the vertex M being called a "collider." From a data-generation viewpoint, one might imagine Nature generating data in two steps: she first picks independently two values for U and W from two distributions and then she combines them (possibly along with some additional random variable) to create M. Given this, conditioning on M can cause a spurious correlation between U and W, which is known as the collider bias [15], or, in epidemiology, Berkson's Paradox [16]. Conceptually, this correlation happens because if one cause of an observed outcome is known to have not occurred, the other cause becomes more likely. Consider an automatic-timer sprinkler system where the sprinkler being on is independent of whether it is raining. Here, the weather gives no information on the sprinkler. However, given wet grass, if one observes a sunny day, one will likely conclude that the sprinklers have recently run. Correlation has been induced.

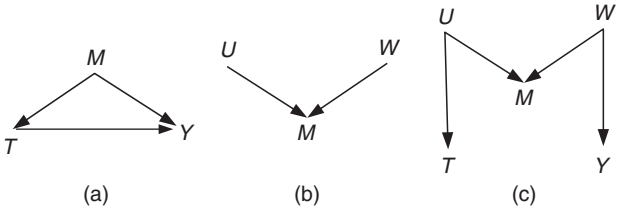


Figure 1: Three DAGs: (a) A simple DAG, (b) V-Structure, and (c) *M*-Structure.

Where things get interesting is when this collider is made into a *pretreatment* variable. Consider Figure 1(c), an extension of Figure 1(b). Here U and W are now also causes of the treatment T and the outcome Y, respectively. Nature, as a last, third step generates T as a function of U and some randomness and Y as a function of W and some randomness. This structure is typically used to represent a circumstance where a researcher observes T, Y, and M in nature and is attempting to derive the causal impact of T on Y. U and W are unobserved, or latent. Clearly, the causal effect of T on Y is zero, which is also equal to the marginal association between T and Y. If a researcher regressed Y onto T, he or she would obtain a zero in expectation, which is correct for estimating the causal effect. But perhaps there is a concern that M, a pretreatment covariate, may be a confounder that is masking a treatment effect. Typically, one would then "adjust" for M to take this possibility into account, e.g. by including M in a regression or by matching units on similar values of M. If we do this in this circumstance, however, then we will not find a zero causal effect, in expectation. This is the so-called "*M*-Bias," and this special structure is called the "*M*-Structure" in the DAG literature.

Previous qualitative analysis for binary variables shows that collider bias generally tends to be small [15], and simulation studies [17] again demonstrate that *M*-Bias is small in many realistic settings. While mathematically describing the magnitudes of *M*-Bias in general models is intractable, it is possible to derive exact formulae of the biases as functions of the correlation coefficients in linear structural equation models (LSEMs). The LSEM has a long history in statistics [18, 19] to describe dependence among multiple random variables. Sprites [20] uses linear models to illustrate *M*-Bias in observational studies, and Pearl [21] also utilizes the transparency of such linear models to examine various types of causal phenomena, biases, and paradoxes. We here extend these works and provide exact formulae for biases, allowing for a more detailed quantitative analysis of *M*-Bias.

While *M*-Bias does exist when the true underlying data generating process (DGP) follows the exact *M*-Structure, it might be rather sensitive to various deviations from the exact *M*-Structure. Furthermore, some might argue that an exact *M*-Structure is unlikely to hold in practice. Gelman [22], for example, doubts the exact independence assumption required for the *M*-Structure in the social sciences by arguing that there are "(almost) no true zeros" in this discipline. Indeed, since U and W are often latent characteristics of the same individual, the independence assumption $U \perp\!\!\!\perp W$ is a rather strong structural assumption. Furthermore, it might be plausible that the pretreatment covariate M is also a confounder between, i.e.

has some causal impact on both, the treatment and the outcome. We extend our work by accounting for these departures from a pure *M*-Structure and find that even slight departures from the *M*-Structure can dramatically change the forms of the biases.

This paper theoretically compares the bias from conditioning on an M to not under several scenarios and finds that *M*-Bias is indeed small relative to other concerns unless there is a strong correlation structure for the variables. We further show that these findings extend to a binary treatment regime as well. This argument proceeds in several stages. First, in Section 2, we examine a pure *M*-Structure and introduce our LSEM framework. We then discuss the cases when the latent variables U and W may be correlated and M may also be a confounder between the treatment T and the outcome Y. In Section 3, we generalize the results in Section 2 to a binary treatment. In Section 4, we illustrate the theoretical findings using a controversial example between Professors Donald Rubin and Judea Pearl [13, 23]. Section 5 discusses the relevance of our findings by examining *M*-Bias in actual practice and by comparing asymptotic to finite sample properties. We conclude with a brief discussion and present all technical details in the Appendix.

2 *M*-Bias and Butterfly-Bias in LSEMs

We begin by examining pure *M*-Bias in a LSEM. As our primary focus is bias, we assume data are ample and that anything estimable is estimated with nearly perfect precision. In particular, when we say we obtain a result from a regression, we implicitly mean we obtain that result in expectation; in practice an estimator will be near the given quantities. We do not compare relative uncertainties of different estimators given the need to estimate more or fewer parameters. There are likely degrees-of-freedom issues that would implicitly advocate using estimators with fewer parameters, but in the circumstances considered here these concerns are likely to be minor as all the models have few parameters.

A causal DAG can be viewed as a hierarchical DGP. In particular, any variable on the graph can be viewed as a function of its parents and some additional noise, e.g. if R had parents $A, B,$ and C, we would have

$$R = f(A, B, C, \varepsilon_R) \text{ with } \varepsilon_R \perp\!\!\!\perp (A, B, C).$$

Generally noise terms such as ε_R are considered to be independent from each other, but they can also be given an unknown correlation structure corresponding to earlier variables not explicitly included in the diagram. This is typically represented by drawing the dependent noise terms jointly from some multivariate distribution. This framework is quite general; we can represent any distribution that can be factored as a product of conditional distributions corresponding to a DAG (which is one representation of the Markov Condition, a fundamental assumption for DAGs).

LSEMs are special cases of the above with additional linearity and additivity constraints. For simplicity, and without loss of generality, we also rescale all primary variables (U, W, M, T, Y) to have zero mean and unit variance. For example, consider this DGP corresponding to Figure 1(a):

$$\begin{cases} M, \varepsilon_T, \varepsilon_Y & \overset{\text{ind}}{\sim} & [0,1], \\ T & = & aM + \sqrt{1-a^2}\varepsilon_T, \\ Y & = & bT + cM + \sqrt{1-b^2-c^2}\varepsilon_Y, \end{cases}$$

where we use $A \sim [0,1]$ to denote a random variable with mean zero and variance one.

In the causal DAG literature, we think about causality as reaching in and fixing a given node to a set value, but letting Nature take her course otherwise. For example, if we were able to set T at t, the above data-generation process would be transformed to:

$$\begin{cases} M, \varepsilon_T, \varepsilon_Y & \overset{\text{ind}}{\sim} & [0,1], \\ T & = & t, \\ Y & = & bt + cM + \sqrt{1-b^2-c^2}\varepsilon_Y. \end{cases}$$

The previous cause, M, of T has been broken, but the impact of T on Y remains intact. This changes the distribution of Y but not M. More importantly, this results in a distribution distinct from that of *conditioning* on $T = t$. Consider the case of positive $a, b,$ and c. If we *observe* a high T, we can infer a high M (as T and M are correlated) and a high Y due to both the bT and the cM terms in Y's equation. However, if we *set* T to a high value, M is unchanged. Thus, while we will still have the large bT term for Y, the cM term will be 0 in expectation. Thus, the expected value for Y will be less.

This setting as compared to conditioning is represented with the "do" operator. Given the "do" operator, we define a local causal effect of T on Y at $T = t$ as:

$$\tau_t = \frac{\partial \mathbb{E}\{Y|\text{do}(T = t)\}}{\partial t}.$$

For linear models, the local causal effect is a constant, and thus we do not need to specify t. We use "do" here purely to indicate the different distributions. For a more technical overview, see Pearl [8, 14]. Our results, with more formality, can easily be expressed in this more technical notation.

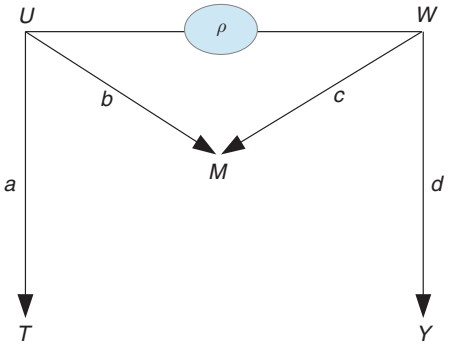


Figure 2: *M*-Structure with possibly correlated hidden causes.

If we extend the *M*-Structure in Figure 1(c) by allowing possible correlation between the two hidden causes U and W, we obtain the DAG in Figure 2. This in turn gives the following DGP:

$$\begin{cases} \varepsilon_M, \varepsilon_T, \varepsilon_Y & \overset{\text{ind}}{\sim} & [0,1], \\ (U, W) & \sim & [0,0;1,1,\rho], \\ M & = & bU + cW + \sqrt{1-b^2-c^2-2bc\rho}\,\varepsilon_M, \\ T & = & aU + \sqrt{1-a^2}\varepsilon_T, \\ Y & = & dW + \sqrt{1-d^2}\varepsilon_Y, \end{cases}$$

where we use $(A, B) \sim [0,0;1,1,\rho]$ to denote a bivariate random vector with means zero, variances one, and correlation coefficient ρ.

Here, the true causal effect of T on Y is zero, namely, $\tau_t = 0$ for all t. The unadjusted estimator for the causal effect obtained by regressing Y onto T is the same as the covariance between T and Y:

$$\text{Bias}_{\text{unadj}} = \text{Cov}(T, Y) = \text{Cov}(aU, dW) = ad\text{Cov}(U, W) = ad\rho.$$

The adjusted estimator (see Lemma 2 in Appendix for a proof) obtained by regressing Y onto (T, M) is

$$\text{Bias}_{\text{adj}} = \frac{ad\rho(1-b^2-c^2-bc\rho) - abcd}{1-(ab+ac\rho)^2}.$$

The results above and some of the results discussed later in this paper can be obtained directly from traditional path analysis [18, 19]. However, we provide elementary proofs, which can easily be extended to binary treatment, in the Appendix. If we allowed for a treatment effect, our results would remain essentially

unchanged; the only difference would be due to restrictions on the correlation terms needed to maintain unit variance for all variables.

The above can also be expressed in the potential outcomes framework [24, 25]. In particular, for a given unit let Nature draw $\varepsilon_M, \varepsilon_T, \varepsilon_Y, U,$ and V as before. Let T be the "natural treatment" for that unit, i.e. what treatment it would receive without intervention. Then calculate $Y(t)$ for any t of interest using the "do" operator. These are what we would see if we set $T = t$. How $Y(t)$ changes for a particular unit defines that unit's collection of potential outcomes. Then $\mathbb{E}\{Y(t)\}$ for some t is the expected potential outcome over the population for a particular t. We can examine the derivative of this function as above to get a local treatment effect. This connection is exact: the findings in this paper are the same as what one would find using this DGP and the potential outcomes framework. We here examine regression as the estimator. Note that matching would produce identical results as the amount of data grew (assuming the DGP ensures common support, etc.).

2.1 Exact *M*-Bias

The *M*-Bias originally considered in the literature is the special case where the correlation coefficient between U and W is $\rho = 0$. In this case, the unadjusted estimator is unbiased, and the absolute bias of the adjusted estimator is $|abcd|/\{1-(ab)^2\}$. With moderate correlation coefficients a, b, c, d the denominator $1-(ab)^2$ is close to one, and the bias is close to $-abcd$. Since $abcd$ is a product of four correlation coefficients, it can be viewed as a "higher order bias." For example, if $a = b = c = d = 0.2$, then $1-(ab)^2 = 0.9984 \approx 1$, and the bias of the adjusted estimator is $-abcd/\{1-(ab)^2\} = -0.0016 \approx 0$; if $a = b = c = d = 0.3$, then $1-(ab)^2 = 0.9919 \approx 1$, and the bias of the adjusted estimator is $-abcd/\{1-(ab)^2\} = -0.0082 \approx 0$. Even moderate correlation results in little bias.

In Figure 3, we plot the bias of the adjusted estimator as a function of the correlation coefficients, and let these coefficients change to see how the bias changes. In the first subfigure, we assume all the correlation coefficients have the same magnitude ($a = b = c = d$), and we plot the absolute bias of the adjusted estimator versus a. The constraints on variance and correlation only allow for some combinations of values for $a, b, c,$ and d which limits the domain of the figures. In this case, for example, $|a| \leq \sqrt{2}/2$ due to the requirement that $b^2 + c^2 = 2a^2 \leq 1$. Other figures have limited domains due to similar constraints. In the second subfigure of Figure 3, we assume that M is more predictive to the treatment T than to the outcome Y, with $a = b = 2c = 2d$. In the third subfigure of Figure 3, we assume that M is more predictive to the outcome Y, with $2a = 2b = c = d$. The biases are generally very small within wide ranges of the feasible regions of the correlation coefficients. However, the biases do blow up when the correlation coefficients are extremely large. Near the boundary of the feasible regions in Figure 3, the *M*-Structure is approximately deterministic, which is rare in social sciences. Pearl [9] does not exclude the worst cases, and thus he considers *M*-Bias as a severe problem.

In Figure 4(a), we assume $a = b$ and $c = d$ and examine a broader range of relationships. Here, the gray area satisfies $|\text{Bias}_{\text{adj}}| < \min(|a|, |c|)/20$. For example, when the absolute values of the correlation coefficients are smaller than 0.5 (the square with dashed boundary in Figure 4(a)), the corresponding area is almost gray, implying small bias.

Due to the four-dimensional sensitivity parameters (a, b, c, d), a full exploration and graphical illustration over all possible values of the sensitivity parameters is formidably hard. In the absence of prior knowledge about the DAG, our sensitivity analysis here is based on some simplifications (e.g. $a = b$ and $c = d$), which may reflect some real situations. Using the bias formulae in this paper, we can easily conduct sensitivity analysis for other parameter combinations, depending on our practical problem and background knowledge about the DAG.

As a side note, Pearl [21] notices a surprising fact: the stronger the correlation between T and M, the larger the absolute bias of the adjusted estimator, since the absolute bias is monotone increasing in $|ab|$. From the second and the third subfigures of Figure 3, we see that when M is more predictive of the treatment, the biases of the adjusted estimator indeed tend to be larger.

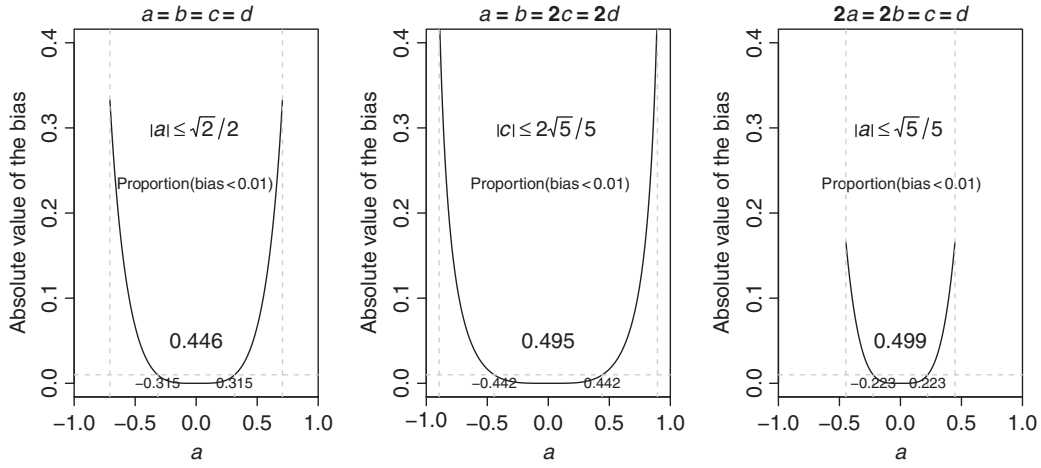


Figure 3: *M*-Bias with independent *U* and *W*. The three subfigures correspond to the cases when (U, W) are equally/more/less predictive to the treatment than to the outcome. In each subfigure, we show the proportions of the areas where the adjusted estimator has a bias smaller than 0.01.

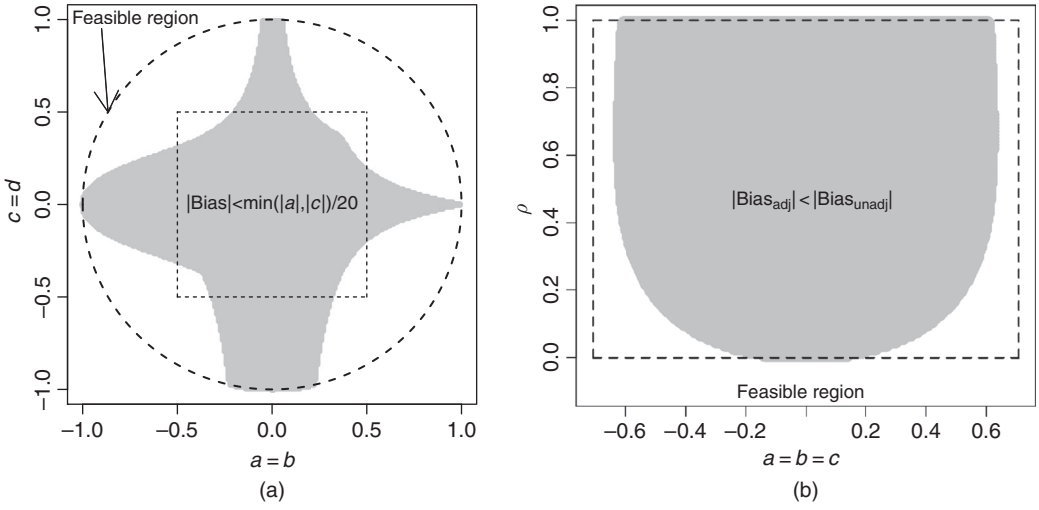


Figure 4: *M*-Bias under different scenarios. (a) Pure *M*-Bias. Within the gray region, the absolute bias of the adjusted estimator is less than 1/20 of the minimum of $|a|(=|b|)$ and $|c|(=|d|)$. (b) *M*-Bias with correlated *U* and *W*. Within the gray region, the adjusted estimator is superior.

2.2 Correlated latent variables

When the latent variables *U* and *W* are correlated with $\rho > 0$, *both* the unadjusted and the adjusted estimators may be biased. The question then becomes: which is worse? The ratio of the absolute biases is

$$\left|\frac{\text{Bias}_{\text{adj}}}{\text{Bias}_{\text{unadj}}}\right| = \left|\frac{\rho(1-b^2-c^2-bc\rho)-bc}{\rho\left\{1-(ab+ac\rho)^2\right\}}\right|,$$

which does not depend on *d* (the relationship between *W* and *Y*). For example, if the correlation coefficients a, b, c, ρ all equal 0.2, the ratio above is 0.714; in this case, the adjusted estimator is superior to the unadjusted one by a factor of 1.4. Figure 4(b) compares this ratio to 1 for all combinations of ρ and $a(=b=c)$. Generally, the adjusted estimator has smaller bias except when $a, b,$ and c are quite large. One of the editors pointed out the interesting fact that, if $\rho < 0$ and $a = b = c$, the unadjusted estimator is

always less biased than the adjusted one, something not shown in Figure 4(b). We can see this in the bias ratio formula; in particular it is due to the sign of the $bc\rho$ term. In fact, it is not just the sign of ρ itself, but the relationship of the signs of ρ, b, and c that is key. One can always replace the latent U by $-U$, and/or W with $-W$ to make ρ is positive, but not change the structure of the problem or relative biases of the estimators. Such replacements will simply change the signs of a, b, etc. In general, when $bc\rho > 0$, we have Figure 4(b); when $bc\rho < 0$, adjusting is more biased than not adjusting. The question is then whether this latter relationship is likely in practice. A negative $bc\rho$ is due to a pair of positively associated U and W that have opposite impacts on M. We did not find such relationships in any of the cases discussed later in Sections 4 and 5, but this is admittedly an important area of further investigation.

Manipulating the signs of coefficients with Butterfly-Bias, discussed next, also allows for increasing and decreasing levels of bias. Here, however, it is even less of a concern as the signs result only in manipulating "higher order" terms that are usually smaller in magnitude than other, dominating terms. We feel, therefore, that the associated discussion is broadly applicable. Overall, we advocate examining the bias ratio formulae and assessing the potential impact of changing the signs of the coefficients.

In Figure 5, we again assume $a = b = c = d$ and investigate the absolute biases as functions of a for fixed ρ at $0.1, 0.2,$ and 0.4. When the correlation coefficients $a(= b = c = d)$ are not dramatically larger than ρ, the adjusted estimator has smaller bias than the unadjusted one.

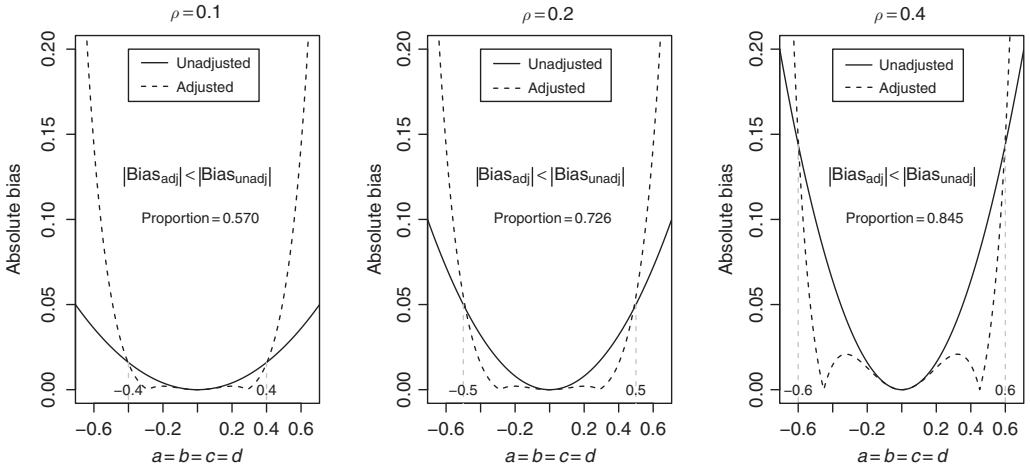


Figure 5: *M*-Bias with correlated U and W when $\rho \neq 0$ with $a = b = c = d$. In each subfigure, we show the proportion of the areas where the adjusted estimator has a smaller bias than the unadjusted estimator.

2.3 The disjunctive cause criterion

In order to remove biases in observational studies, VanderWeele and Shpitser [6] propose a new "disjunctive cause criterion" for selecting confounders, which requires controlling for all the covariates that are either causes of the treatment, causes of the outcome, or causes of both. According to the "disjunctive cause criterion," when $\rho \neq 0$, we should control for (U, W) if possible. Unfortunately, neither of (U, W) is observable. However, controlling the "proxy variable" M for (U, W) may reduce bias when ρ is relatively large. In the special case with $b = 0$, the ratio of the absolute biases is

$$\left|\frac{\text{Bias}_{\text{adj}}}{\text{Bias}_{\text{unadj}}}\right| = \frac{1 - c^2}{1 - (ac\rho)^2} \leq 1;$$

in another special case with $c = 0$, the ratio of the absolute biases is

$$\left|\frac{\text{Bias}_{\text{adj}}}{\text{Bias}_{\text{unadj}}}\right| = \frac{1 - b^2}{1 - (ab)^2} \leq 1. \tag{1}$$

Therefore, if either U or W is not causative to M, the adjusted estimator is always better than the unadjusted one.

2.4 Butterfly-Bias: M-Bias with confounding bias

Models, especially in the social sciences, are approximations. They rarely hold exactly. In particular, for any covariate M of interest, there is likely to be some concerns that M is indeed a confounder, even if it is also a possible source of M-Bias. If we let M both be a confounder and the middle of an M-Structure, we obtain a "Butterfly-Structure" [21] as shown in Figure 6. In this circumstance, conditioning will help with confounding bias, but hurt with M-Bias. Ignoring M will not resolve any confounding, but will avoid M-Bias. The question then becomes that of determining which is the lesser of the two evils.

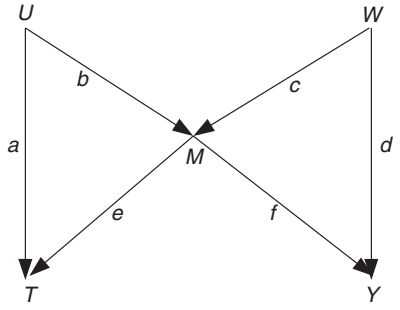


Figure 6: Butterfly-Structure.

We can examine this trade-off for a LSEM corresponding to Figure 6. The DGP is given by the following equations:

$$\begin{cases} U, W, \varepsilon_M, \varepsilon_T, \varepsilon_Y & \overset{\text{ind}}{\sim} & [0,1], \\ M & = & bU + cW + \sqrt{1-b^2-c^2}\varepsilon_M, \\ T & = & aU + eM + \sqrt{1-a^2-e^2}\varepsilon_T, \\ Y & = & dW + fM + \sqrt{1-d^2-f^2}\varepsilon_Y. \end{cases}$$

Again, the true causal effect of T on Y is zero. The unadjusted estimator obtained by regressing Y onto T is the covariance between T and Y:

$$\text{Bias}_{\text{unadj}} = \text{Cov}(T, Y) = abf + cde + ef.$$

It is not, in general, zero, implying bias. The adjusted estimator (see Lemma 3 in Appendix for a proof) obtained by regressing Y onto (T, M) has bias

$$\text{Bias}_{\text{adj}} = -\frac{abcd}{1-(ab+e)^2}.$$

If the values of e and f are relatively high (i.e. M has a strong effect on both T and Y), the confounding bias is large and the unadjusted estimator will be severely biased. For example, if $a, b, c, d, e,$ and f all equal 0.2, the bias of the unadjusted estimator is 0.056, but the bias of the adjusted estimator is only −0.0017, an order of magnitude smaller. Generally, the largest term for the unadjusted bias is the second-order term of ef, while the adjusted bias only has, ignoring the denominator, a fourth-order term of $abcd$. This suggests adjustment is generally preferable and that M-Bias is in some respect a "higher order bias."

Detailed comparison of the ratio of the biases is difficult, since we can vary six parameters (a, b, c, d, e, f). In Figure 7(a), we assume all the correlation coefficients have the same magnitude and plot bias for both estimators as a function of the correlation coefficient within the feasible region, defined by the restrictions $-\sqrt{2}/2 < a < (-1+\sqrt{5})/2$, due to the restrictions

$$b^2 + c^2 < 1, a^2 + e^2 < 1, d^2 + f^2 < 1, \text{ and } |a^2 + e| < 1. \tag{2}$$

Within 74.9% of the feasible region, the adjusted estimator has smaller bias than the unadjusted one. The unadjusted estimator only has smaller bias than the adjusted estimator when the correlation coefficients are extremely large. In Figure 7(b), we assume $a = b = c = d$ and $e = f$ and compare $|\text{Bias}_{\text{adj}}|$ and $|\text{Bias}_{\text{unadj}}|$ within the feasible region of (a, e) defined by eq. (2). We can see that the adjusted estimator is superior to the unadjusted one for 71% (colored in gray in Figure 7(b)) of the feasible region. In the area satisfying $|e| > |a|$ in Figure 7(b), where the connection between M to T and Y is stronger than the other connections, the area is almost entirely gray suggesting that the adjusted estimator is preferable. This is sensible because here the confounding bias has larger magnitude than the M-Bias. In the area satisfying $|a| < |e|$, where M-Bias is stronger than confounding bias, the unadjusted estimator is superior for some values, but still tends to be inferior when the correlations are roughly the same size.

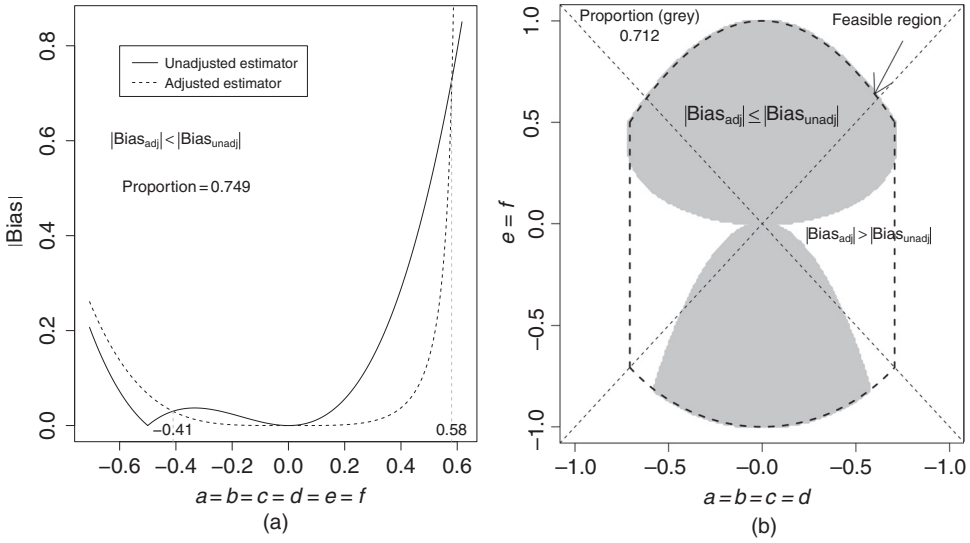


Figure 7: Butterfly-Bias. (a) Absolute biases of both estimators with $a = b = c = d = e = f$. (b) Comparison of the absolute biases with $a = b = c = d$ and $e = f$. Within 71.2% (in gray) of the feasible region, the adjusted estimator has smaller bias than the unadjusted one.

3 Extensions to a binary treatment

One might worry that the conclusions in the previous section are not applicable for a binary treatment. It turns out, however, that they are. In this section, we extend the results in Section 2 to binary treatments by representing the treatment through a latent Gaussian variable as shown in Figure 8.

3.1 Correlated latent variables

We extend Figure 2 to Figure 8(a). Here, T^* is the old T. The generating equations for T and T^* become

$$T = I(T^* \geq \alpha), \text{ and } T^* = aU + \sqrt{1 - a^2}\varepsilon_T.$$

Other variables and noise terms remain the same. Although it might be relaxed, we make reference to the normal assumption of the error terms for mathematical simplicity. The intercept α determines the proportion of the individuals receiving the treatment: $\Phi(-\alpha) = P(T = 1)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. When $\alpha = 0$, the number of individuals exposed to the treatment and control are balanced; when $\alpha < 0$, more individuals are exposed to the treatment; when $\alpha > 0$, the reverse.

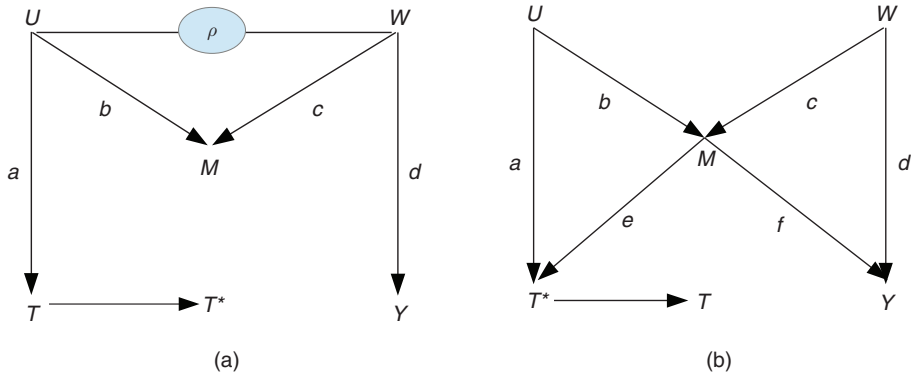


Figure 8: DAGs for binary treatment. (a) Correlated hidden causes and (b) Butterfly-Structure.

The true causal effect of T on Y is again zero. Let $\phi(\cdot) = \Phi'(\cdot)$ and $\eta(\alpha) \equiv \phi(\alpha)/\{\Phi(\alpha)\Phi(-\alpha)\}$. Then Lemma 6 in Appendix shows that the unadjusted estimator has bias

$$\text{Bias}_{\text{unadj}} = ad\rho\eta(\alpha),$$

and the adjusted estimator has bias

$$\text{Bias}_{\text{adj}} = \frac{ad\eta(\alpha)\{\rho(1 - b^2 - c^2 - bc\rho) - bc\}}{\rho\left\{1 - (ab + ac\rho)^2\phi(\alpha)\eta(\alpha)\right\}}.$$

When $\rho = 0$, the unadjusted estimator is unbiased, but the adjusted estimator has bias

$$-\frac{abcd\eta(\alpha)}{1 - (ab)^2\phi(\alpha)\eta(\alpha)}.$$

When $\rho \neq 0$, the ratio of the absolute biases is

$$\left|\frac{\text{Bias}_{\text{adj}}}{\text{Bias}_{\text{unadj}}}\right| = \left|\frac{\rho(1 - b^2 - c^2 - bc\rho) - bc}{\rho\{1 - (ab + ac\rho)^2\phi(\alpha)\eta(\alpha)\}}\right|,$$

which again does not depend on d.

The patterns for a binary treatment do not differ much from a continuous treatment. As before, if the correlation coefficient is moderately small, the *M*-Bias also tends to be small. As shown in Figure 9 (analogous to Figure 5), when $|\rho|$ is comparable to $|a|(= |b| = |c| = |d|)$, the adjusted estimator is less biased than the unadjusted estimator. Only when $|a|$ is much larger than $|\rho|$ is the unadjusted estimator superior.

3.2 Butterfly-Bias with a binary treatment

We can extend the LSEM Butterfly-Bias setup to binary treatment just as we extended the *M*-Bias setup. Compare Figure 8(b) to Figure 6. T becomes T^* and T is built from T^* as above. The structural equations for T and T^* for Butterfly-Bias in the binary case are then

$$T = I(T^* \geq \alpha), \text{ and } T^* = aU + eM + \sqrt{1 - a^2 - e^2}\varepsilon_T.$$

The other equations and variables are the same as before.

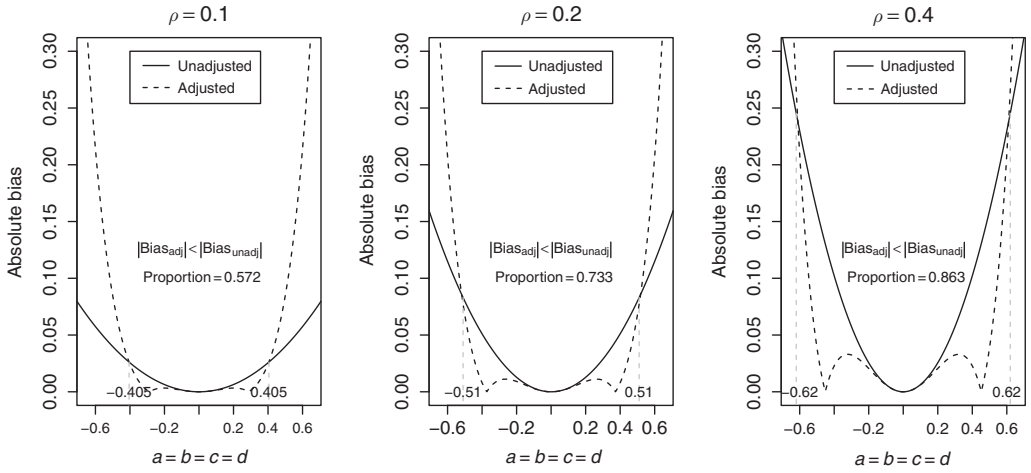


Figure 9: *M*-Bias with correlated (U, W) and binary treatment. Compare to Figure 5.

Although the true causal effect of T on Y is zero, Lemma 7 in Appendix shows that the unadjusted estimator has bias

$$\text{Bias}_{\text{unadj}} = (cde + abf + ef)\eta(\alpha), \tag{3}$$

and the adjusted estimator has bias

$$\text{Bias}_{\text{adj}} = -\frac{abcd\eta(\alpha)}{1-(ab+e)^2\phi(\alpha)\eta(\alpha)}. \tag{4}$$

Complete investigation of the ratio of the biases is intractable with seven varying parameters $(a, b, c, d, e, f, \alpha)$. However, in the very common case with $\alpha = 0$, which gives equal-sized treatment and control groups, we again find trends similar to the continuous treatment case. See Figure 10. As before, only in the cases with very small $e(= f)$ but large $a(= b = c = d)$, does the unadjusted estimator tend to be superior. Within a reasonable region of α, these patterns are quite similar.

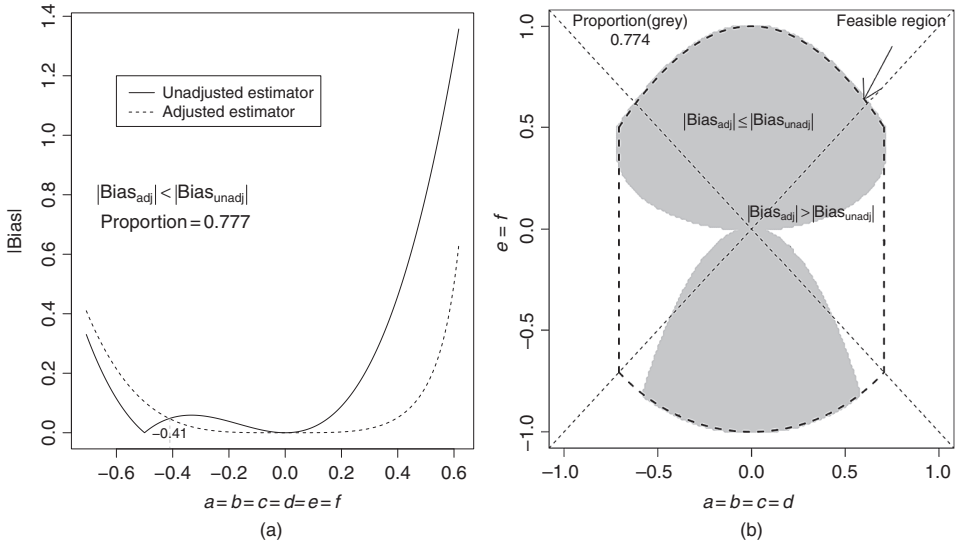


Figure 10: Butterfly-Bias with a binary treatment. (a) Absolute biases with $a = b = c = d = e = f$. (b) Comparison of the absolute bias with $a = b = c = d$ and $e = f$. Within 74.4% (in gray) of the feasible region, the adjusted estimator is better than the unadjusted estimator.

4 Illustration: the Rubin–Pearl controversy

Pearl [13] cites Rubin [23]'s example about the causal effect of smoking habits (T) on lung cancer (Y) and argues that conditioning on the pretreatment covariate "seat-belt usage" (M) would introduce spurious associations, since M could be reasonably thought of as an indicator of a person's attitudes toward societal norms (U) as well as safety and health related measures (W). Assuming all the analysis is already conditioned on other observed covariates, we focus our discussion on the five variables (U, W, M, T, Y), of which the dependence structure is illustrated in Figure 11. Since the patterns with a continuous treatment and a binary treatment are similar, we focus our discussion on LSEMs.

As Pearl [13] points out,

> If we have good reasons to believe that these two types of attitudes are marginally independent, we have a pure M-structure on our hand.

In the case with $\rho = 0$, conditioning on M will lead to spurious correlation between T and Y under the null and will bias the estimation of the causal effect of T on Y. However, Pearl [13] also recognizes that the independence assumption seems very strong in this example, since U and W are both background variables about the habit and personality of a person. Pearl [13] further argues:

> But even if marginal independence does not hold precisely, conditioning on "seat-belt usage" is likely to introduce spurious associations, hence bias, and should be approached with caution.

Although we believe most things should be approached with caution, our work, above, suggests that even mild perturbations of an M-Structure can switch which of the two approaches, conditioning or not conditioning, is likely to remove more bias. In particular, Pearl [13] is correct in that the adjusted estimator indeed tends to introduce more bias than the unadjusted one when an exact M-Structure holds and thus the general advice "to condition on all observed covariates" may not always be sensible. However, in the example of Rubin [23], the exact independence between a person's attitude toward societal norms U and safety and health related measures W is questionable, since we have good reasons to believe that other hidden variables such as income and family background will affect both U and W simultaneously, and thus Pearl's fears may be unfounded.

To examine this further, we consider two possible deviations from the exact M-Structure and investigate the biases of the unadjusted and adjusted estimators for each.

(a) (Correlated U and W) Assume the DGP follows the DAG in Figure 11, with an additional correlation between the attitudes U and W as shown in Figure 2. If we then assume that all the correlation coefficients have the same positive magnitude, earlier results demonstrate that the adjusted estimator is preferable as it strictly dominates the unadjusted estimator except for extremely large values of the correlation coefficients.

    Furthermore, in Rubin [23]'s example, attitudes toward societal norms U are more likely to affect the "seat-belt usage" variable M than safety and health related measures W, which further strengthens the case for adjustment. If we were willing to assume that c is zero but ρ is not, eq. (1) again shows that the adjusted estimator is superior.

(b) (An arrow from W to T) Pearl's example seems a bit confusing on further inspection, even if we accept his independence assumption $U \perp\!\!\!\perp W$. In particular, one's "attitudes towards safety and health related measures" likely impact one's decisions about smoking. Therefore, we might reasonably expect an arrow from W to T. In Figure 11(a), we remove the correlation between U and W, but we allow an arrow from W to T, i.e. the generating equation for T becomes $T = aU + gW + \sqrt{1 - a^2 - g^2}\varepsilon_T$. Lemma 8 in Appendix gives the associated formulae for biases of the adjusted and unadjusted estimators. Figure 11 (b) shows that, assuming $a = b = c = d = g$ (i.e. equal correlations), the adjusted estimator is uniformly better.

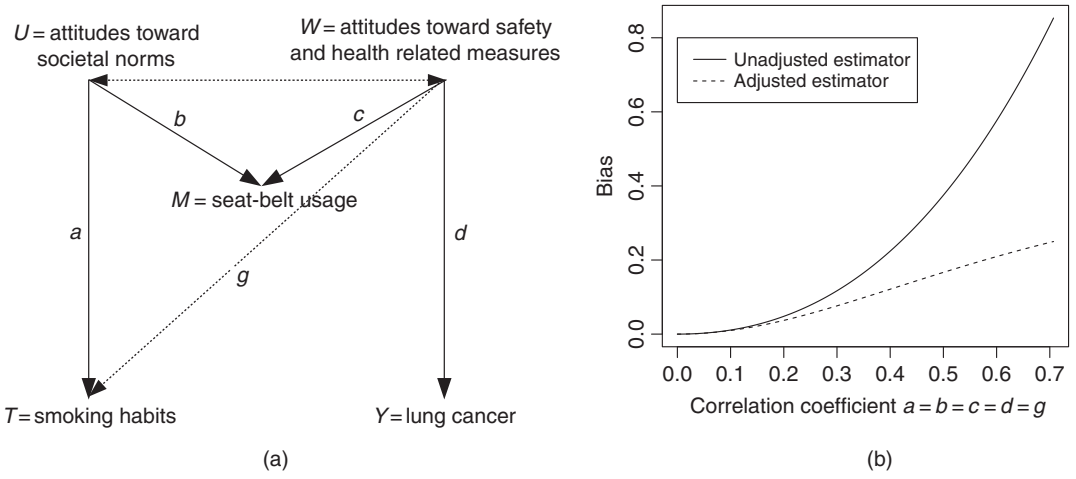


Figure 11: Sensitivity analysis of Pearl [13]'s critique on Rubin [23]. (a) *M*-Bias, with possible deviations: correlated *U* and *V*, and an additional arrow from *W* to *T*. (b) Biases of resulting unadjusted and adjusted estimators with $a = b = c = d = g$.

5 Two further issues

One controversy about *M*-Bias is whether *M*-Structure is rare or not in practice, and we go through several examples to discuss this issue. In the second part of this section, we make a distinction between asymptotic and finite sample properties of *M*-Bias.

5.1 Is *M*-Structure rare?

Although Pearl [13] argues that *M*-Bias is a structural property, Rubin [5] claims that *M*-Bias is a rare phenomenon such as "trying to balance a multidimensional cone on its point with no external supports in some visible directions." As mentioned in Section 1, Gelman [22] argues that, in social sciences, "true zeros" are rare and consequently the independence structure in the exact *M*-Structure is also rare. Section 4 revisits the controversial example between Professors Pearl and Rubin, and Figure 11(a) illustrates two possible deviations from the exact *M*-Structure. Both deviations suggest conditioning is a superior choice. In the following, we review three other examples of *M*-Bias in the current literature and investigate the plausibility of the exact *M*-Structure.

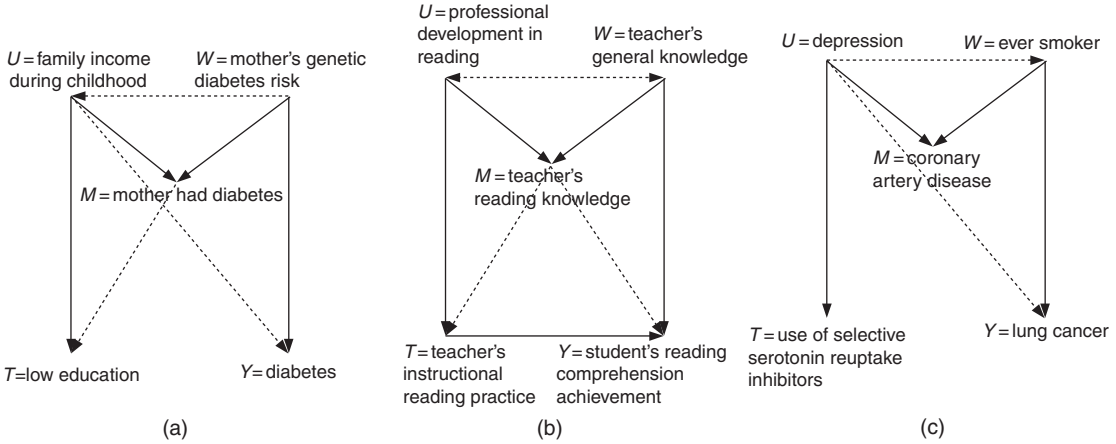


Figure 12: Three *M*-Structures with possible deviations in dashed arrows. (a) Glymour [4], (b) Kelcey and Carlisle [8], and (c) Liu et al. [9].

As shown in Figure 12(a), Glymour [26] postulates a possible *M*-Structure with exposure "low income," outcome "diabetes," and *M* variable "mother had diabetes," where "family income during childhood" affects both exposure and *M*, and "mother's genetic diabetes risk" affect both outcome and *M*. However, this *M*-Structure is subject to several plausible deviations: "mother's genetic diabetes risk" may affect "family income during childhood; "mother had diabetes" may affect "low education;" and "family income during childhood" may affect "diabetes."

Kelcey and Carlisle [27] have (T, Y, M, U, W) as "teacher's instructional reading practice," "student's reading comprehension achievement," "teacher's reading knowledge," "professional development in reading," and "teacher's general knowledge." See Figure 12(b). However, this *M*-Structure is dubious because of the possible correlation between the latent U and W, and the confounding effect of M on the relationship between T and Y.

Figure 12(c) is a possible *M*-Structure investigated by Liu et al. [17], where (T, Y, M, U, W) are "use of selective serotonin reuptake inhibitors (SSRIs)," "lung cancer," "coronary artery disease," "depression," and "ever smoker." Although it is plausible that "coronary artery disease" is not a confounder between "use of SSRIs" and "lung cancer," it is very likely that "depression" affects both "ever smoker" and "lung cancer."

In summary, while all the examples above were quite useful to illustrate *M*-Bias in theoretical research, it is unwise to believe that these *M*-Structures are exact based on our background knowledge. Therefore, we suggest researchers conduct sensitivity analyses, such as illustrated earlier, according to their scientific knowledge about the structure of the DAG and the associated parameters.

5.2 Asymptotic versus finite sample properties

The discussion in the previous sections is mainly based on asymptotic theory assuming large samples. As argued by Pearl [9], this approach allows for investigating the existence of bias in a certain DAG, and asymptotic analysis helps reveal the structural property of a DAG. A referee pointed out that asymptotic theory is quite different from more practical finite sample theory. In finite sample data analysis, practitioners, often interested in interval estimation and hypothesis testing, are typically more interested in whether associated confidence intervals cover the true causal parameters at nominal rates and whether tests for null hypotheses about the causal effect have valid size. These questions are related to the asymptotic property of the DAGs, but also depend on the sample size, the procedure used for constructing ones confidence interval, and choice of test statistic. Theoretical discussion of the finite sample theory is unfortunately more difficult. Simulation study, however, is an alternative tool for these questions. Some studies exist. In particular, Liu et al. [17] simulate large cohort studies under an *M*-Structure corresponding to their science question of interest and find that the impact of *M*-Bias is small for most of their 178 scenarios unless the association between *M* and the unmeasured confounders is very large.

6 Discussion

For objective causal inference, Rubin and Rosenbaum suggest balancing all the pretreatment covariate in observational studies to parallel with the design of randomized experiments [4, 5, 23, 28], which is called the "pretreatment criterion" [6]. However, Pearl and other researchers [9, 10–13] criticize the "pretreatment criterion" by pointing out that this criterion may lead to biased inference in presence of a possible *M*-Structure even if the treatment assignment is unconfounded. We investigate this controversy in detail for LSEMs, ideally providing a template for future research about more general DAGs (e.g. nonparametric and nonlinear models). While we agree that Pearl's warning is very insightful, our asymptotic theory shows that, at least for LSEMs, this conclusion is quite sensitive to various deviations from the exact *M*-Structure,

e.g. to circumstances where latent causes may be correlated or the M variable may also be a confounder between the treatment and the outcome. We also go through several candidate M-Structures in the existing literature and find that exact M-Structure is likely to be rare with various deviations typically being more plausible. Overall, this suggests that for linear systems, except in some extreme cases, adjusting for all the pretreatment covariates is in fact a reasonable choice.

Acknowledgments: The authors thank all the participants in the "Causal Graphs in Low and High Dimensions" seminar at Harvard Statistics Department in Fall, 2012, and thank Professor Peter Sprites for sending us his slides [20]. Comments from the associate editor and two reviewers greatly improved the quality of our paper. During the proof stage, Dr. Elias Bareinboim and one of the editors of *Journal of Causal Inference* pointed out the problem with negative association between the latent U and W in Section 2.2, which lead to our further explanation on this issue.

Appendix: Lemmas and Proofs

Lemma 1 *In the linear regression model $Y = \beta_0 + \beta_T T + \beta_M M + \varepsilon$ with $\varepsilon \perp\!\!\!\perp$ (T,M) and $E(\varepsilon) = 0$, we have*

$$\beta_T = \frac{\mathrm{Cov}(Y,T)\mathrm{Var}(M) - \mathrm{Cov}(Y,M)\mathrm{Cov}(M,T)}{\mathrm{Var}(T)\mathrm{Var}(M) - \mathrm{Cov}^2(M,T)}.$$

Proof. Solve for (β_T, β_M) using the following moment conditions:

$$\begin{cases} \mathrm{Cov}(Y,T) &= \beta_T \mathrm{Var}(T) + \beta_M \mathrm{Cov}(M,T), \\ \mathrm{Cov}(Y,M) &= \beta_T \mathrm{Cov}(M,T) + \beta_M \mathrm{Var}(M). \end{cases}$$

Lemma 2 *Under the model generated by Figure 2, the regression coefficient of T by regressing Y onto (T, M) is*

$$\beta_T = \frac{ad\rho(1 - b^2 - c^2 - bc\rho) - abcd}{1 - (ab + ac\rho)^2}.$$

Proof. We apply Lemma 1, where all variance terms such as $\mathrm{Var}(M)$ are 1, and the covariance terms are easily calculated. For example, we have $\mathrm{Cov}(Y,T) = \mathrm{Cov}(dW, aU) = ad\rho$ and $\mathrm{Cov}(Y,M) = \mathrm{Cov}(dW, bU + cW) = bd\rho + cd$.

Lemma 3 *Under the model generated by Figure 6, the regression coefficient of T by regressing Y onto (T, M) is*

$$\beta_T = -\frac{abcd}{1 - (ab + e)^2}.$$

Proof. Similar to the proof of Lemma 2, we apply Lemma 1.

Lemma 4 *Assume that (X_1, X_2) follows a bivariate normal distribution with means zero, variances one, and correlation coefficient r. Then $\mathbb{E}(X_1|X_2 \geq z) - \mathbb{E}(X_1|X_2 < z) = r\eta(z)$, where $\eta(z) = \phi(z)/\{\Phi(z)\Phi(-z)\}$.*

Proof. Since $X_1 = rX_2 + \sqrt{1 - r^2}Z$ with $Z \sim N(0,1)$ and $Z \perp\!\!\!\perp X_2$, we have
$\mathbb{E}(X_1|X_2 \geq z) = r\mathbb{E}(X_2|X_2 \geq z) = \frac{r}{\Phi(-z)} \int_z^\infty x\phi(x)dx = -\frac{r}{\Phi(-z)} \int_z^\infty d\phi(x) = r\frac{\phi(z)}{\Phi(-z)}$.

Similarly, we have $\mathbb{E}(X_1|X_2 < z) = \mathbb{E}(X_1| - X_2 > -z) = -r\phi(-z)/\Phi(z) = -r\phi(z)/\Phi(z)$.
Therefore, $\mathbb{E}(X_1|X_2 \geq z) - \mathbb{E}(X_1|X_2 < z) = r\phi(z)\{1/\Phi(-z) + 1/\Phi(z)\} = r\eta(z)$.

Lemma 5 *The covariance between X and B∼ Bernoulli*(p) *is*

$$\mathrm{Cov}(X,B) = p(1-p)\{\mathbb{E}(X|B=1) - \mathbb{E}(X|B=0)\}.$$

Proof. It follows from the definition of the covariance.

Lemma 6 *Under the model generated by Figure 8(a), the regression coefficient of T by regressing Y onto* (T,M) *is*

$$\beta_T = \frac{ad\eta(\alpha)\{\rho(1-b^2-c^2-bc\rho)-bc\}}{\rho\{1-(ab+ac\rho)^2\phi(\alpha)\eta(\alpha)\}}.$$

Proof. We have the following joint normality of (Y,M,T^*):

$$\begin{pmatrix} Y \\ M \\ T^* \end{pmatrix} \sim N_3\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & bd\rho+cd & ad\rho \\ bd\rho+cd & 1 & ab+ac\rho \\ ad\rho & ab+ac\rho & 1 \end{pmatrix} \right\}.$$

From Lemma 4, we have

$$\mathbb{E}(M|T=1) - \mathbb{E}(M|T=0) = \mathbb{E}(M|T^* \geq \alpha) - \mathbb{E}(M|T^* < \alpha) = (ab+ac\rho)\eta(\alpha),$$

$$\mathbb{E}(Y|T=1) - \mathbb{E}(Y|T=0) = \mathbb{E}(Y|T^* \geq \alpha) - \mathbb{E}(Y|T^* < \alpha) = ad\rho\eta(\alpha).$$

Therefore, from Lemma 5, the covariances are $\mathrm{Cov}(M,T) = \Phi(\alpha)\Phi(-\alpha)(ab+ac\rho)\eta(\alpha)$ and $\mathrm{Cov}(Y,T) = \Phi(\alpha)\Phi(-\alpha)ad\rho\eta(\alpha)$. According to Lemma 1, the regression coefficient β_T is

$$\beta_T = \frac{\Phi(\alpha)\Phi(-\alpha)ad\rho\eta(\alpha) - (bd\rho+cd)\Phi(\alpha)\Phi(-\alpha)(ab+ac\rho)\eta(\alpha)}{\Phi(\alpha)\Phi(-\alpha) - \Phi^2(\alpha)\Phi^2(-\alpha)(ab+ac\rho)^2\eta^2(\alpha)} = \frac{ad\eta(\alpha)\{\rho(1-b^2-c^2-bc\rho)-bc\}}{\rho\{1-(ab+ac\rho)^2\phi(\alpha)\eta(\alpha)\}}.$$

Lemma 7 *Under the model generated by Figure 8(b), the regression coefficient of T by regressing Y onto* (T,M) *is*

$$\beta_T = -\frac{abcd\eta(\alpha)}{1-(ab+e)\phi(\alpha)}.$$

Proof. We have the following joint normality of (Y,M,T^*):

$$\begin{pmatrix} Y \\ M \\ T^* \end{pmatrix} \sim N_3\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & cd+f & cde+abf+ef \\ cd+f & 1 & ab+e \\ cde+abf+ef & ab+e & 1 \end{pmatrix} \right\}.$$

From Lemma 4, we have

$$\mathbb{E}(M|T=1) - \mathbb{E}(M|T=0) = \mathbb{E}(M|T^* \geq \alpha) - \mathbb{E}(M|T^* < \alpha) = (ab+e)\eta(\alpha),$$

$$\mathbb{E}(Y|T=1) - \mathbb{E}(Y|T=0) = \mathbb{E}(Y|T^* \geq \alpha) - \mathbb{E}(Y|T^* < \alpha) = (cde+abf+ef)\eta(\alpha).$$

From Lemma 5, we obtain their covariances $\mathrm{Cov}(M,T) = \Phi(\alpha)\Phi(-\alpha)(ab+e)\eta(\alpha)$ and $\mathrm{Cov}(Y,T) = \Phi(\alpha)\Phi(-\alpha)(cde+abf+ef)\eta(\alpha)$. According to Lemma 1, the regression coefficient β_T is

$$\beta_T = \frac{\Phi(\alpha)\Phi(-\alpha)(cde + abf + ef)\eta(\alpha) - (cd + f)\Phi(\alpha)\Phi(-\alpha)(ab + e)\eta(\alpha)}{\Phi(\alpha)\Phi(-\alpha) - \Phi^2(\alpha)\Phi^2(-\alpha)(ab + e)^2\eta^2(\alpha)} = -\frac{abcd\eta(\alpha)}{1 - (ab + e)^2\phi(\alpha)\eta(\alpha)}.$$

Lemma 8 *Under the model generated by* Figure 11(a) *with an arrow from W to T, the unadjusted estimator has bias* $ad\rho + dg$*, and the adjusted estimator has bias*

$$\frac{dg - (cd)(ab + cg)}{1 - (ab + cg)^2}.$$

Proof. The unadjusted estimator is $\mathrm{Cov}(T, Y) = ad\rho + dg$. Expanding Lemma 1 gives the above as the regression coefficient of T by regressing Y onto (T, M).

References

1. Copas JB, Li HG. Inference for non-random samples (with discussion). J R Stat Soc Ser B 1997;59:55–95.
2. Heckman JJ. Sample selection bias as a specification error. Econometrica 1979;47:153–61.
3. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615–25.
4. Rosenbaum PR. Observational studies. New York: Springer, 2002.
5. Rubin DB. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat Med 2009;28:1420–3.
6. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics 2011;67:1406–13.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
8. Pearl J. Causality: models, reasoning and inference, 2nd ed. Cambridge: Cambridge University Press, 2009.
9. Pearl J. Letter to the editor. Stat Med 2009;28:1415–16.
10. Shrier I. Letter to the editor. Stat Med 2008;27:2740–1.
11. Shrier I. Propensity scores. Stat Med 2009;28:1315–18.
12. Sjölander A. Propensity scores and M-structures. Stat Med 2009;28:1416–20.
13. Pearl J. Myth, confusion, and science in causal analysis. Technical Report, 2009. Available at: http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf
14. Pearl J. Causal diagrams for empirical research. Biometrika 1995;82:669–88.
15. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2002;14:300–6.
16. Berkson J. Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin 1946;2:47–53.
17. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of m bias in epidemiologic studies: a simulation study. Am J Epidemiol 2012;176:938–48.
18. Wright S. Correlation and causation. J Agric Res 1921;20:557–85.
19. Wright S. The method of path coefficients. Ann Math Stat 1934;5:161–215.
20. Sprites P. Presented at WNAR/IMS meeting, Los Angeles, CA, June2002.
21. Pearl J. Linear models: a useful microscope for causal analysis. J Causal Inference 2013;1:155–70.
22. Gelman A. Causality and statistical learning. Am J Sociol 2011;117:955–66.
23. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26:20–36.
24. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci 1923/1990;5:465–72.
25. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688–701.
26. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes M, Kaufman J, editors. Methods in social epidemiology. San Francisco, CA: Jossey-Bass, 2006:393–428.
27. Kelcey B, Carlisle J. The threshold of embedded m collider bias and confounding bias. Society for research on educational effectiveness conference, 2011. Available at: http://files.eric.ed.gov/fulltext/ED519118.pdf
28. Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat 2008;2:808–40.