Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Causal simulation experiments: Lessons from bias amplification

Statistical Methods in Medical Research
2022, Vol. 31(1) 3–46


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280221995963
journals.sagepub.com/home/smm
SAGE

Tyrel Stokes[1], Russell Steele[1] and Ian Shrier[2,3]

Abstract

Recent theoretical work in causal inference has explored an important class of variables which, when conditioned on, may further amplify existing unmeasured confounding bias (bias amplification). Despite this theoretical work, existing simulations of bias amplification in clinical settings have suggested bias amplification may not be as important in many practical cases as suggested in the theoretical literature. We resolve this tension by using tools from the semi-parametric regression literature leading to a general characterization in terms of the geometry of OLS estimators which allows us to extend current results to a larger class of DAGs, functional forms, and distributional assumptions. We further use these results to understand the limitations of current simulation approaches and to propose a new framework for performing causal simulation experiments to compare estimators. We then evaluate the challenges and benefits of extending this simulation approach to the context of a real clinical data set with a binary treatment, laying the groundwork for a principled approach to sensitivity analysis for bias amplification in the presence of unmeasured confounding.

Keywords

Causal simulation, bias amplification, sensitivity analysis, causal inference, simulation experiments

1 Introduction

Causal identification strategies aim to condition on a sufficient set of observables such that the potential outcomes are conditionally independent of the treatment of interest.[1–3] Causal variable selection procedures often assume that at least one subset of the observed variables forms such a sufficient set.[4] The object in causal variable selection then becomes how to separate variables which are necessary for identification of the causal effect from those variables which are extraneous[4,5] in the interest of reducing estimator variance or covariate dimensionality.[4,6]

In non-experimental observational studies, we do not have full access to a sufficient set in many realistic settings, and important confounding pathways remain unblocked.[7,8] This is referred to as unmeasured confounding or endogeneity in the statistics and econometrics literatures, respectively. However, applied researchers currently rely on variable selection techniques such as lasso, step-wise, change-in-estimator selection, and outcome and/or treatment oriented approaches[9] despite violating their underlying assumptions.

Often, variable selection techniques are used to avoid conditioning on negligible confounding pathways without introducing meaningful bias to the estimator. In this paper, we explore how this intuition can break down under even mild violations of the underlying assumptions. In particular, we build on the work of Pearl[10] and explore how treatment prediction-oriented approaches may inadvertently amplify bias if unmeasured

[1]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada
[2]Department of Family Medicine, McGill University, Montreal, QC, Canada
[3]Centre for Clinical Epidemiology, Lady Davis Institute, Montreal, QC, Canada

Corresponding author:
Russell Steele, Department of Mathematics and Statistics, McGill University, Burnside Hall, Room 1005 805, Sherbrooke Street, West Montreal, Quebec H3A 0B9, Canada.
Email: russell.steele@mcgill.ca

confounding pathways remain. We use the lessons from this failure of intuition to better understand the consequences of variable selection in both linear models and partially linear orthogonalized models in the presence of important unmeasured confounding. We then build a principled approach to simulate from such systems of equations which may be used to compare the theoretical properties of different estimators, such as accurately quantifying the change in bias when we modify the causal relationship between a confounder and the treatment, or for the purpose of sensitivity analysis.

First consider data generated from the following directed acyclic graph (DAG) (Figure 1) and set of structural equations

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \epsilon_1 \tag{1}$$

$$\boldsymbol{A} = \alpha_a + \boldsymbol{U}\gamma_u + \sum_{i=1}^{10} \boldsymbol{BAV_i}\gamma_{bav_i} + \epsilon_2 \tag{2}$$

$$\boldsymbol{U} = \alpha_u + \sum_{i=1}^{10} \boldsymbol{BAV_i}\psi_{bav_i} + \epsilon_3 \tag{3}$$

where $\boldsymbol{Y}$ is the outcome, $\boldsymbol{A}$ is the treatment of interest, $\boldsymbol{U}$ is an unmeasured variable, $\boldsymbol{BAV}$ refers to 10 different potential bias amplifying variables that are measured and affect both $\boldsymbol{A}$ and $\boldsymbol{U}$ but have no direct effect on $\boldsymbol{Y}$, γ_x is the coefficient for the effect of the variable on $\boldsymbol{A}$, β_x is the coefficient for the effect of the variable on $\boldsymbol{Y}$, and ϵ_{number} is an error term for the corresponding equation. This model contains one confounding path that cannot be blocked ($\boldsymbol{A} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$) and 10 confounding paths ($\boldsymbol{A} \leftarrow \boldsymbol{BAV_1} \rightarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$; ...; $\boldsymbol{A} \leftarrow \boldsymbol{BAV_{10}} \rightarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$) that can be blocked by including the measured $\boldsymbol{BAV_i}$ variables in the model. However, including any of these $\boldsymbol{BAV_i}$ might also increase bias (potential bias amplifying variables). Our goal is to find the least biased estimator of the average causal effect (ACE) of treatment (β_a).

By including more of the $\boldsymbol{BAV}$ variables in the model, intuition suggests the remaining unmeasured confounding bias should decrease because more potential confounders have been included in the conditioning set. However, as demonstrated in the bias amplification literature,[10–12] conditioning on confounders may still increase bias. For example, suppose further the 10 observable variables account for 90% of the variance in the variable $\boldsymbol{U}$ responsible for unmeasured confounding. The blue violin plot in Figure 2 represents the sampling distribution of the estimator from the true outcome model that includes the treatment and both measured/unmeasured confounding variables included as regressors. As expected, the estimates are approximately normally distributed around the true value $\beta_a = 0.7$. The green violin plot represents the biased estimates from the naive model, the simple regression of the outcome $\boldsymbol{Y}$ on the treatment $\boldsymbol{A}$, which does not include any of the confounders (measured or unmeasured). The red violin plot represents the linear model adjusted for all 10 measured confounders which account for 90% of the unmeasured confounding. The adjusted model performs much worse than the naive model both in terms of bias (0.73 compared to 0.43, interpretable as standard deviations) and variance (standard deviation of 0.1 compared to 0.02). In fact, in 4990 of 5000 simulations the adjusted estimate was farther from the truth than the naive estimate and nearly 65% of the adjusted estimates had the incorrect effect sign.

The purpose of this paper is to explain why model selection intuition fails us in this case and how we can use a combination of data and simulation approaches to improve model selection. We build upon an emerging theoretical literature exploring a class of variables which can amplify existing unmeasured confounding bias,[10,13,14,30,31,32,35] called bias amplifiers. Throughout this text, we will refer to (potential) bias amplifiers as those variables which, upon their inclusion in a model, (may) increase the absolute bias in the estimation of particular target parameters relative to a smaller, nested model. In the above example, all 10 $\boldsymbol{BAV}$s increase the bias due to the path $\boldsymbol{A} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$ and in fact could also increase the bias of any other variable $\boldsymbol{BAV}$ if it were not included in the model, since they are all confounders.

The class of bias-amplifying variables is potentially very large and common in practical applications. Bias amplification occurs when the absolute bias of an estimator for a target parameter is larger than the absolute bias of an estimator for a competing, nested model. In this text, we define the natural model comparison to be the naive regression of the outcome ($\boldsymbol{Y}$) on the treatment ($\boldsymbol{A}$) and bias amplification will be the additional relative bias that results from adding additional predictors. Throughout, the goal of model selection in this context will be to

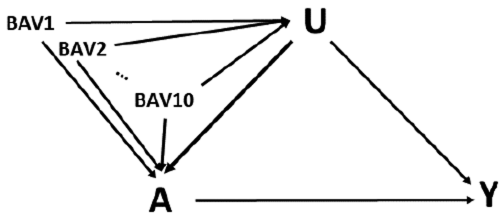

Figure 1. Directed acyclic graph (DAG): Meyers (2011) extended to 10 possible bias amplifying variables BAVs.

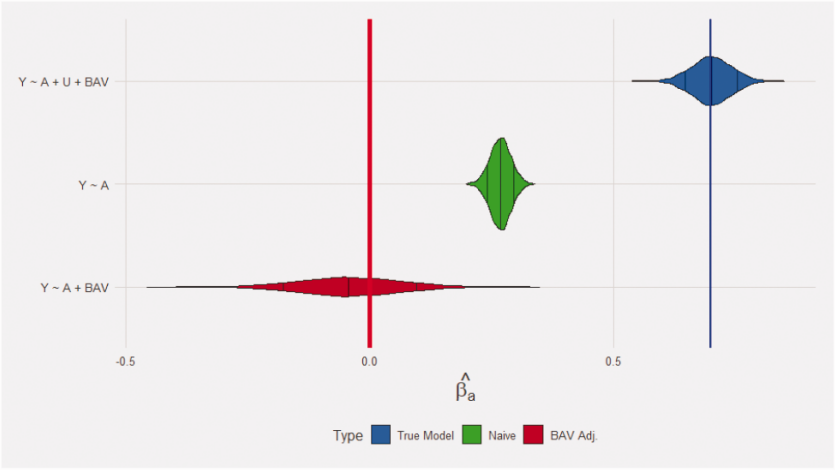

Figure 2. Violin plot from simulations from equations (1) to (3). There were 5000 replication with $n = 5000$. The true effect of interest was $\beta_a = 0.7$, represented by the blue line. The confounding effects were $\beta_u = -0.5$ and $\gamma_u = 0.59$. The vector of coefficients for ***BAV*** on ***U*** was $\psi_{bav} = \{-0.55, -0.45, -0.3, 0.30, .25, 0.20, -0.20, 0.20, -0.15, 0.10\}$, which in general were larger than the impact of ***BAV*** on ***A***, $\gamma_{bav} = \{-.1, -.15, -.1, .21, -.2, .3, -.2, -.15, -.2, .075\}$. $\boldsymbol{U}, \boldsymbol{BAV}, \boldsymbol{A}$, and $\boldsymbol{Y}$ are all standard normal variables. The error terms were normally distributed with standard deviations: $\sigma_1 = 0.93, \sigma_2 = 0.07$, and $\sigma_3 = 0.32$. All intercepts were set to 0. The code to reproduce the plot can be found in the supplementary materials. The red vertical line represents $\widehat{\beta}_a = 0$.

choose the variables and thus the model which minimizes bias in estimating the true average treatment of effect in the presence of unmeasured confounding pathways.

We adopt a matrix notation framework to characterize this problem because (1) we can more easily generalize to a much larger class of directed acyclic graphs and structural equations than previously studied; (2) it offers a unifying geometric explanation for the amplified bias in the context of least squares estimation; and (3) it offers a solid foundation for how to build data-informed model selection procedures. Finally, we develop a procedure for simulating from a more complete parameter space in a way that respects the underlying amplification process. In addition to lending itself better to articulating and answering causal simulation questions, this procedure helps explain why some previous studies have incorrectly concluded that applied investigators need not worry about amplification in practice.[15] We evaluate the challenges of implementing this approach with a real clinical example with binary treatment.

2 Problem formulation

Figure 3 shows a basic directed acyclic graph (DAG) which has both measured and unmeasured confounding. Let $\boldsymbol{Y}$ represent the outcome and let $\boldsymbol{A}$ be the treatment or variable of interest. Let $\boldsymbol{U}$ be an unmeasured confounding variable that we cannot include in a regression model, but which has a functional relationship with both $\boldsymbol{Y}$ and $\boldsymbol{A}$. The bias amplifying variable ($\boldsymbol{BAV}$) in this DAG is analogous to $\boldsymbol{U}$ in that it is a cause of $\boldsymbol{Y}$ and a cause of $\boldsymbol{A}$; however, we are able to measure $\boldsymbol{BAV}$ and not $\boldsymbol{U}$. Intuition from currently recommended causal variable selection techniques would tell us to include $\boldsymbol{BAV}$ in the regression to reduce bias because it is the root of a confounding

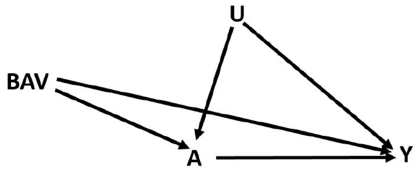

Figure 3. DAG: Two confounding paths, where *A* is the treatment of interest, *Y* is the outcome, *U* is an unmeasured variable and *BAV* is a measured variable.

path ($\boldsymbol{A} \leftarrow \boldsymbol{BAV} \rightarrow \boldsymbol{Y}$). However, as has been demonstrated,[10–12,14] blocking this confounding path can actually increase or amplify the bias relative to the naive estimator that depends only on $\boldsymbol{A}$.

Assume that we now restrict the possible models for the DAG in Figure 3 to only linear associations amongst variables. The data generating model (or causal structural equations) representing Figure 3 under strictly linear association can be written as

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\boldsymbol{\beta_{bav}} + \boldsymbol{\epsilon_1} \tag{4}$$

$$\boldsymbol{A} = \alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\boldsymbol{\gamma_{bav}} + \boldsymbol{\epsilon_2} \tag{5}$$

where α_y and α_a are the intercept terms for $\boldsymbol{Y}$ and $\boldsymbol{A}$, respectively. As noted above, we use the form β_x throughout this paper to denote true structural coefficients for some variable $\boldsymbol{X}$ on the outcome $\boldsymbol{Y}$. For example, the true structural coefficient for $\boldsymbol{U}$ on $\boldsymbol{Y}$ is β_u. Analogously, the true regression parameter for some variable $\boldsymbol{X}$ on the treatment $\boldsymbol{A}$ is represented by γ_x. The estimates of these parameters by OLS are denoted by $\widehat{\beta_x}$, $\widehat{\gamma_x}$ with additional superscripts to clarify which set of estimating equations the estimator is derived from. By assumption $\boldsymbol{\epsilon_1}$ and $\boldsymbol{\epsilon_2}$ are error terms independent of each other we further assume that $E[\boldsymbol{\epsilon_1}|\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{BAV}] = 0$ and $E[\boldsymbol{\epsilon_2}|\boldsymbol{A}, \boldsymbol{BAV}] = 0$, and that the error terms have some finite variance $\sigma^2_{\epsilon_{1,2}}$. In simulation experiments, we typically simulate from normal distributions which are independent from all other variables, but only the above assumptions are necessary for the theoretical results to hold.

We also assume that the target estimand is the ACE. In the linear model case, this is simply the β_a above in equation (4) (See Appendix A.3.I for derivation of the ACE). $\boldsymbol{U}$ is unmeasured and thus we cannot identify the ACE from the observed data if $\beta_u \neq 0$, but we are interested in estimating the quantity with as little bias as possible.

2.1 Matrix notation and probability limits

To tackle the question of model selection, we must derive properties of different $\widehat{\beta_a}$ estimators, with the restriction that the estimators be functions of only observed variables, under different assumed conditional regression models. To this aim, we propose expressing OLS estimates using matrix notation and ideas borrowed from the partial regression literature. Further, we propose considering also the probability limits of the estimators to extend our results to more general and realistic cases of bias amplification (see Appendix A.2). For the naive estimator, we use a simple linear regression to estimate a conditional expectation of the form

$$E[\boldsymbol{Y}|\boldsymbol{A}] = \alpha_y^{naive} + \boldsymbol{A}\beta_a^{naive} + \boldsymbol{\upsilon_1} \tag{6}$$

where $\boldsymbol{\upsilon_1}$ is the error of the regression term. Unbiased estimation of the naive model by OLS, and thus of the true conditional expectation $E[\boldsymbol{Y}|\boldsymbol{A}]$, requires the assumption that $E[\boldsymbol{\upsilon_1}|\boldsymbol{A}] = 0$. However, this of course is not true since according to the data generating equation (4), the error will be a function of the confounding terms, resulting in non-zero bias. The naive estimator bias is a special case of the classic omitted variables problem, where we have two omitted variables which are related to both the treatment and the exposure, $\boldsymbol{U}$ and $\boldsymbol{BAV}$.

Let $\widehat{\beta}_a^{naive}$ be the estimate of β_a from the naive model (6). Throughout this paper, we will consider the matrix $\boldsymbol{Z}$ to be a matrix of all the variables that we include in a regression that are not the variable of interest $\boldsymbol{A}$, in other words control variables in a selection on observables approach. In the naive model, $\boldsymbol{Z} = \boldsymbol{1}$ where throughout $\mathbf{1}$ will

denote an $n \times 1$ vector of 1 s. In matrix notation, applying the Frisch-Waugh-Lovell (FWL) theorem (see Appendix A.1), we can write $\widehat{\beta}_a^{naive}$ as

$$\widehat{\beta}_a^{naive} = \frac{\boldsymbol{A^T M_1 Y}}{\boldsymbol{A^T M_1 A}} \tag{7}$$

where $\boldsymbol{M_1}$ is a centering projection matrix, defined and described in detail in Appendix section A.1. In the case of linear relationships between all the variables, following Pearl,[10] this estimator has the following expectation

$$E[\widehat{\beta}_a^{naive}] = \beta_a + (\beta_u \gamma_u \sigma_u^2 + \beta_{bav} \gamma_{bav} \sigma_{bav}^2) \frac{1}{\sigma_a^2} \tag{8}$$

The absolute bias for the ACE then clearly is $|(\beta_u \gamma_u \sigma_u^2 + \beta_{bav} \gamma_{bav} \sigma_{bav}^2) \frac{1}{\sigma_a^2}|$. Now consider the estimates resulting from further conditioning on the observable $\boldsymbol{BAV}$ variable, i.e. estimating a conditional expectation of the form

$$E[\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{BAV}] = \alpha_y^{naive} + \boldsymbol{A}\beta_a^{bav} + \boldsymbol{BAV}\beta_{bav}^{bav} + \boldsymbol{v_2} \tag{9}$$

We will denote the resulting estimator $\widehat{\beta}_a^{bav}$ which can be written as follows by again applying the FWL theorem

$$\widehat{\beta}_a^{bav} = \frac{\boldsymbol{A^T M_z Y}}{\boldsymbol{A^T M_z A}} \tag{10}$$

where $\boldsymbol{Z} = [\boldsymbol{1}, \boldsymbol{BAV}]$ and $\boldsymbol{M_z}$ is the annihilator projection matrix of the matrix Z (see Appendix A.1 for details and properties). Again following Pearl,[10] the expectation of $\widehat{\beta}_a^{bav}$ is

$$E[\widehat{\beta_a^{bav}}] = \beta_a + (\beta_u \gamma_u \sigma_u^2) \frac{1}{\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2} \tag{11}$$

In Appendix A.4, we explicitly show Pearl's derivation and how it relies on the conditional expectation $E[\boldsymbol{U}|\boldsymbol{A}, \boldsymbol{BAV}]$ being linear in both $\boldsymbol{A}$ and $\boldsymbol{BAV}$. Pearl's derivation is limited in that it is cumbersome and does not generalize well to a broad class of DAGs and functional forms. A simple example where we are unable to use Pearl's method is the case of an interaction term in the exposure structural equation between $\boldsymbol{U}$ and $\boldsymbol{BAV}$. This implies that $E[\boldsymbol{U}|\boldsymbol{A}, \boldsymbol{BAV}]$ is nonlinear in $\boldsymbol{A}$ and $\boldsymbol{BAV}$. This cannot be represented by an unbiased least squares projection of the form $\boldsymbol{U} = \alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav} + \epsilon_{\boldsymbol{3}}$ as required by Pearl's derivation method (see Appendix A.4 for details), where ζ_i represents the true regression coefficient for variable i. If we impose further strict distributional assumptions over all the variables, we may still be able to directly solve the conditional expectation and find an expression for bias in terms of the underlying parameters. However, in many applied cases, these distributional assumptions will not be justified, particularly assuming a distribution for the unmeasured confounding which will always be untestable.

In contrast, if we consider the probability limits, we do not need to assume that $E[\boldsymbol{U}|\boldsymbol{A}, \boldsymbol{BAV}]$ is linear, nor do we have to make any additional distributional assumptions to find meaningful limiting expressions for our estimators in a broad class of clinically relevant circumstances. In addition to giving rise to a meaningful interpretation, the closed form asymptotics we derive allow us to more easily harness domain knowledge about the underlying causal process for the purpose of model selection.

Since we are still interested in the finite sample expectation of the estimators and the bias directly, we report the expectations when appropriate and feasible. The probability limit facilitates insight under weaker assumptions than those necessary to derive exact forms of the expectations. Additionally, in some cases, like the linear model of Pearl,[10] the probability limits for $\widehat{\beta}_a^{naive}$ and $\widehat{\beta}_a^{bav}$ are precisely equal to their expectations (see Appendix A.7).

3 Treatment variance explained and amplifying effects for strictly linear models

In this section we examine the probability limits of the estimators to better understand the mechanics and root causes of bias amplification. First, notice the denominator of equation (11) $(\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2)$ gets smaller as the

causal edge $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$ increases in strength. This is because when we specify the functional form of a system of random variables and conditional independence assumptions, we are also determining a formula for its variance. Under the structural equation we specified for the exposure (equation (5), with the corresponding DAG shown later in Figure 7(a)), $\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2$ is precisely equal to the remaining residual treatment variance in A after having adjusted for $\boldsymbol{BAV}$ in a regression model. When we assume that all the variables are standardized, this term becomes $1 - \gamma_{bav}^2$ as presented in Pearl[10] (under the assumption of standard normal, here we drop the distributional requirement), because the variance of standard normal variables is equal to 1 ($\sigma_a^2, \sigma_{bav}^2 = 1$).

In order to visualize this phenomenon, we use ideas from partial regression plots.[16] By the FWL theorem, we can always pre-multiply an estimating equation by the residual-making variables of a set of regressors and get the same estimates (see Appendix A.1 for further details). For example, the following two regression equations produce the same numerical ordinary least squares estimates of $\widehat{\beta}_a^{naive}$

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{\upsilon_1} \tag{12}$$

$$\boldsymbol{M_1 Y} = \boldsymbol{M_1 A}\beta_a + \boldsymbol{\upsilon_1} \tag{13}$$

Equation (13) is the model for a simple linear regression of a modified outcome, $\boldsymbol{M_1 Y}$, on a modified treatment, $\boldsymbol{M_1 A}$ (see Appendix A.1). There is no intercept term as the mean of the modified treatment must be equal to zero.

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{BAV}\beta_{bav} + \boldsymbol{\upsilon_2} \tag{14}$$

$$\boldsymbol{M_z Y} = \boldsymbol{M_z A}\beta_a + \boldsymbol{\upsilon_2} \tag{15}$$

Similarly equations (14) and (15) above produce equivalent ordinary least squares estimates of $\widehat{\beta}_a$, where $\boldsymbol{Z} = [\boldsymbol{1} \quad \boldsymbol{BAV}]$ is a column of 1 s and the $\boldsymbol{BAV}$ variable. Equation (15) is a single variable regression on a transformed set of variables. The modified $\boldsymbol{Y}$ is produced by taking the residuals from regressing $\boldsymbol{Y}$ on a column of 1 s and $\boldsymbol{BAV}$. In other words, the dependent variable is the residual vector resulting from regressing $\boldsymbol{Y}$ on an intercept column and $\boldsymbol{BAV}$. The independent variable is the residual vector resulting from regressing $\boldsymbol{A}$ on a column of 1 s and $\boldsymbol{BAV}$. Another way to think of the residuals is in the context of orthogonalization techniques, where $\boldsymbol{A^T M_z}$ for example is the part of $\boldsymbol{A}$ orthogonal to linear combinations of the control variables $\boldsymbol{A}$. We then regress the orthogonalized outcome on the orthogonalized treatment, which is implicitly the mechanics of what happens whenever we use least squares estimation. See section 4.2 for more general orthogonalization techniques.

Since we have reduced the multi-variable regression problem to a simple linear regression for two linearly transformed variables, we can easily visualize the amplification process via a partial regression plot. In Figure 4 the left subplot visualizes the naive regression equation (13), whereas the blue subplot on the right visualizes the regression equation (15) that includes $\boldsymbol{BAV}$. The data were simulated from a special case of equations (4) and (5), with $n = 1000$. Details can be found in the appendix (Appendix section A.13).

The unbiased ACE is the slope of the black line ($\beta_a = 0.2$) in these plots. The slope of the blue line (equal to the OLS estimator from the amplifying model) is clearly farther away from the true slope (in black) compared to the slope of the red line from the naive model, and thus the conditional estimator is more biased.

Note first that including $\boldsymbol{BAV}$ in the model reduces the variance in the adjusted treatment, which can be seen by comparing the relative sparsity of points along the x-axis in red compared to the relative density of points along the x-axis in blue. However, if we inspect the spread of points vertically along the y-axis, we can see that the red and blue samples are similarly dispersed in this dimension because conditional on the treatment, linear combinations of $\boldsymbol{BAV}$ explain very little of the variance in the outcome. Most importantly, including only $\boldsymbol{BAV}$ does not change the variance in $\boldsymbol{Y}$ due to $\boldsymbol{U}$, the unmeasured confounder.

When we add the $\boldsymbol{BAV}$ to the regression model, the bias is 0.14 larger in absolute terms (or approximately 65% greater in relative terms) than the naive estimate, even though it blocks a confounding path between the treatment $\boldsymbol{A}$ and the outcome $\boldsymbol{Y}$. More simply, trying to block a confounding path with weak response association can amplify bias in causal effect estimation because it increases the proportion of treatment association due to unmeasured confounding on unblocked paths.

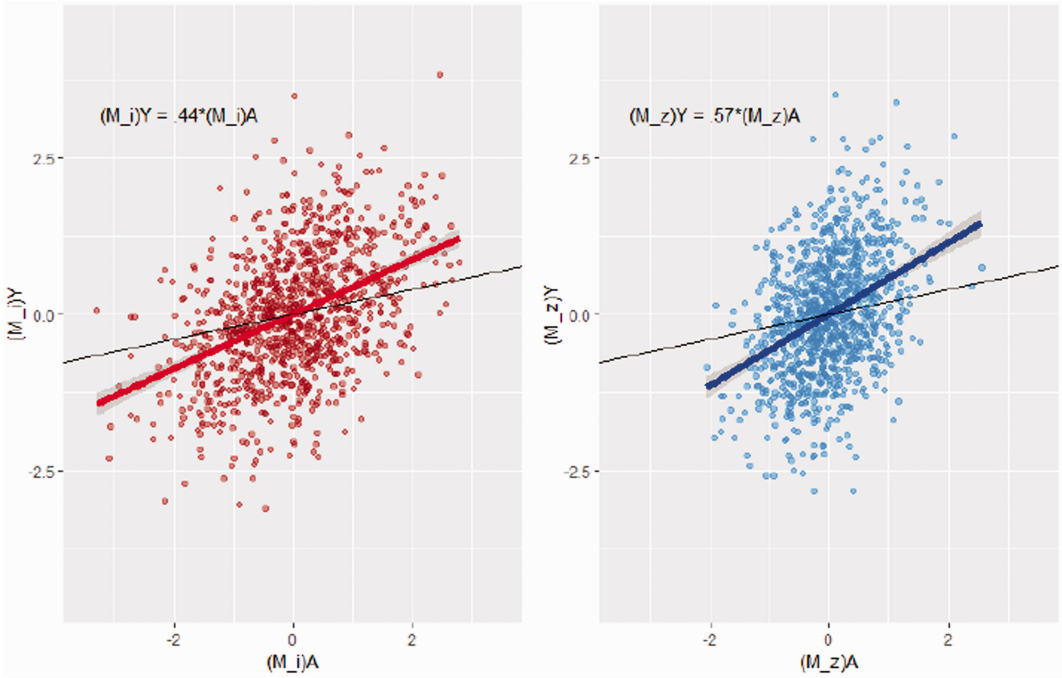

Figure 4. In both panels, the unbiased ACE ($\beta_a = 0.2$) is shown by the dotted black line. In the left panel, the red dots represent the centered treatment ($\boldsymbol{M_tA}$) plotted against the centered outcome ($\boldsymbol{M_tY}$) and the estimated slope $\widehat{\beta}_a^{naive}$ is shown with the bolded red line. This represents the equivalent regressions in equations (12) and (13). In the right panel, the blue dots represent the modified treatment ($\boldsymbol{M_zA}$) plotted against the modified outcome ($\boldsymbol{M_zY}$). The solid blue line represents the treatment estimate from the equivalent regressions (14) and (15).

The magnitude of bias amplification can be potentially very large. From equations (8) and (9), the absolute bias of the ***BAV*** estimator will be larger than the absolute bias of the naive estimator whenever

$$\frac{|\text{Bias}(\widehat{\beta}_a^{bav\prime})|}{|\text{Bias}(\widehat{\beta}_a^{naive})|} = \left(\frac{\sigma_a^2}{\sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2}\right)\left(\frac{|\beta_u\gamma_u\sigma_u^2|}{|\beta_u\gamma_u\sigma_u^2 + \beta_{bav}\gamma_{bav}\sigma_{bav}^2|}\right) > 1 \tag{16}$$

We can rewrite the first term on the right side of the equal sign as $(1 - \mathcal{R}^2_{A|BAV})^{-1}$, which is the inverse of 1 minus percentage variance explained in $\boldsymbol{A}$ adjusting for ***BAV***

$$\frac{|\text{Bias}(\widehat{\beta}_a^{bav\prime})|}{|\text{Bias}(\widehat{\beta}_a^{naive})|} = \left(\frac{1}{1 - \mathcal{R}^2_{A|BAV}}\right)\left(\frac{|\beta_u\gamma_u\sigma_u^2|}{|\beta_u\gamma_u\sigma_u^2 + \beta_{bav}\gamma_{bav}\sigma_{bav}^2|}\right) > 1 \tag{17}$$

Notice that the first term must always be greater than or equal to 1. The more strongly the control variables predict the treatment, the larger the magnitude of the term, increasing monotonically[11] and hyperbolically in the treatment variance explained by ***BAV***.

We now examine the second term after the equal sign. If $\beta_u = 0$ or $\gamma_u = 0$, then there is no unmeasured confounding and $\widehat{\beta}_{bav}$ will be unbiased. If $\beta_u \neq 0$ and $\gamma_u \neq 0$ then there is unmeasured confounding, so by manipulating equation (17), we can write the second term as

$$\frac{|\text{Bias}(\widehat{\beta}_a^{bav\prime})|}{|\text{Bias}(\widehat{\beta}_a^{naive})|} = \left(\frac{1}{1 - \mathcal{R}^2_{A|BAV}}\right)\left(\frac{1}{|1 + \eta\text{sgn}(\beta_u)\text{sgn}(\gamma_u)\text{sgn}(\beta_{bav})\text{sgn}(\gamma_{bav})|}\right) > 1 \tag{18}$$

where $\text{sgn}(\cdot)$ is equal to $+1$, -1, and 0 if the argument is positive, negative, and equal to 0, respectively, and η is defined as the ratio of the absolute strength of the confounding path through $\boldsymbol{BAV}$ ($|\beta_{bav}\gamma_{bav}\sigma^2_{bav}|$) to the absolute strength of the confounding path through $\boldsymbol{U}$ ($|\beta_u\gamma_u\sigma^2_u|$).

$$\eta = \frac{|\beta_{bay}\gamma_{bay}\sigma^2_{bay}|}{|\beta_u\gamma_u\sigma^2_u|}$$

The second term depends on the signs of the structural coefficients for the two confounding paths. This is a dimensionless quantity that does not depend on the scale of the data, which can make it useful for sensitivity analysis. If $\gamma_{bav} = 0$, then $\boldsymbol{BAV}$ does not affect $\boldsymbol{A}$, so it has no effect on the bias because $\eta = 0$ and $\mathcal{R}^2_{A|BAV} = 0$. In the special case that $\beta_{bav} = 0$, $\boldsymbol{BAV}$ is a true instrumental variable under the DAG in Figure 3. If there are no interactions as in the structural equations (4) and (5), then including an instrumental variable will always increase bias in the presence of unmeasured confounding.[3,10,14] We can see this clearly in equation (16) since by definition an instrumental variable does not explain variance in the outcome except through the treatment and must correlate with the treatment. A strong instrument in this case must be a strong amplifier since it will strongly predict variance in the treatment making the first term in equation (16) large and the second term exactly 1.

In the case that none of the structural coefficients are equal to 0, note that $\text{sgn}(\beta_u)\text{sgn}(\gamma_u)\text{sgn}(\beta_{BAV})\text{sgn}(\gamma_{BAV})$ will be equal to $+1$ if there is an even number of positive signs for the structural coefficients and it will be equal to -1 if there is an odd number of signs. We have defined η in terms of absolute values, so it must be non-negative. If $0 \leq \eta < 1$, then the confounding path through $\boldsymbol{U}$ is stronger than the path through $\boldsymbol{BAV}$. If $\eta > 1$, then the confounding path through $\boldsymbol{BAV}$ is stronger than the path through $\boldsymbol{U}$. When $\eta = 1$, the strengths of the two confounding paths are equal.

We can use equation (18) to characterize all possible confounding structures that lead to bias amplification. First assume that there are an even number of positively signed structural coefficients. This implies that the second term of equation (18) is equal to $(1+\eta)^{-1}$ and bias amplification will occur when $(1-\mathcal{R}^2_{A|BAV})^{-1} > (1+\eta)$ or equivalently when $R^2_{A|BAV}(1-R^2_{A|BAV})^{-1} > \eta$. So if there is an even number of positively signed structural coefficients, the larger amount of treatment variance explained by $\boldsymbol{BAV}$, the greater the range of possible η values that will lead to bias amplification.

Next assume that there is an odd number of positively signed structural coefficients, so that the second term will be equal to $|1-\eta|^{-1}$. In this case, if $0 < \eta \leq 1$ (i.e. if the confounding path through $\boldsymbol{U}$ is stronger than the confounding path through $\boldsymbol{BAV}$), then there will always be bias amplification. If $\eta > 1$, then there will be bias amplification if $(1-R^2_{A|BAV})^{-1} > (\eta - 1)$, similar to the case when $\eta > 1$.

In Appendix A.5 we further re-express η in terms of only correlations (or partial correlations if desired) and the coefficients of determination and free of model parameters. This is a dimensionless quantity that does not depend on the scale of the data, which can make it useful for sensitivity analysis. Additionally, the different ways of expressing the quantity can reveal portions of the unknown quantity which are estimable, i.e. functions of only observed data. Depending on the application parameters, it may be easier to express domain knowledge through correlations or percentage variance explained. This allows one to move more easily from one to the other, which should be very useful to applied researchers. We expand on these ideas in sections 5 and develop principled ways to simulate from such systems of equation. In Appendix A.5, we discuss exploiting the fact that correlations and variances are sufficient for determining the whole system of equations.

In summary, variable selection approaches which aggressively target confounding paths with strong associations with treatment and weak associations with outcome are at greater risk of bias amplification as they are much more sensitive to the assumption that a full sufficient set is measurable. Adding controlling variables in proportion to their ability to predict the treatment in linear models is only guaranteed to be bias reducing if the resulting selected variables satisfy ignorability assumptions. When this is not the case, we have shown that unmeasured confounding bias may be severely amplified since the treatment variance term increases both monotonically and hyperbolically. It is often not possible or extremely unlikely to select a sufficient set in many non-experimental settings, i.e. most investigators are not willing to assume their observational study is equivalent to a randomized trial. Thus the potential for bias amplification must be considered carefully.

4 Generalizing to a larger class of causal models

In sections (1) to (3), we showed examples of bias amplification under two different DAGs, each with linear structural equations in both the coefficients and variables and probability limit results under the DAG in Figure 3 and linear structural equations. In section 4.1 we extend bias amplification to the class of structural functions that are additive in the outcome and whose target causal effect is represented by β_a. In other words, we consider structural equations of the form

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + f_1(\boldsymbol{U}) + f_2(\boldsymbol{BAV}) + \epsilon_{\boldsymbol{1}} \tag{19}$$

$$\boldsymbol{A} = \alpha_a + g(\boldsymbol{U}, \boldsymbol{BAV}) + \epsilon_{\boldsymbol{2}} \tag{20}$$

We require minimal assumptions of treatment structural equation (20), in fact it is completely general except that it must be decomposable into a function of $\boldsymbol{U}$ and $\boldsymbol{BAV}$ (i.e. the $E[\boldsymbol{A}|\boldsymbol{U}, \boldsymbol{BAV}] = \alpha_a + g(\boldsymbol{U}, \boldsymbol{BAV})$) and an error term orthogonal to this function ($\epsilon_{\boldsymbol{2}} : E[\epsilon_{\boldsymbol{1}}|\boldsymbol{\imath}, \boldsymbol{U}, \boldsymbol{BAV}] = 0$). For the outcome equation, we require more structure and assumptions on top of an orthogonal decomposition into the conditional expectation and error term ($\epsilon_{\boldsymbol{1}} : E[\epsilon_{\boldsymbol{1}}|\boldsymbol{\imath}, \boldsymbol{A}, \boldsymbol{U}, \boldsymbol{BAV}] = 0$). Specifically we require that the outcome equation (19) is additive in functions of $\boldsymbol{A}$, $\boldsymbol{U}$, and $\boldsymbol{BAV}$ (ruling out interaction effects for example) and that the causal effect of interest is a single parameter β_a ($\frac{\partial E[\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{BAV}]}{\partial \boldsymbol{A}} = \beta_a$).

In section 4.2, we again consider the same structural setting, but extend the results and intuitions developed in sections 3 and 4.1 to more general orthogonalization techniques beyond least squares, namely a specific case of Neyman-Orthogonalization in partially linear models as used in double debiased approaches like in Chernozhukov et al.[17] We show that in both settings, we can decompose bias amplification into a component due to covariance between the treatment and the outcome through the confounding paths and a second term which is a ratio of marginal treatment variance to residual treatment variance after adjusting for other covariates. We can always estimate the residual treatment variance and should report these estimates in observational studies, particularly if variable selection was in part determined by strength of treatment prediction. As the remaining treatment variance decreases, the potential for bias amplification increases and any inferences about the causal effect depend more strongly on the assumption that unmeasured confounding is negligible. Further, similar to the earlier setting, we can directly estimate correlations, coefficients of determination, and conditional expectations composed only of observable variables. While the entire confounding pathways cannot be identified, some of the components can be. These estimable quantities, in turn, place restrictions on what the total confounding pathways can be. In section 5 we show how to use these principles and perform simulations to test the performance of competing estimators. In section 6 we apply these ideas to a real clinical data set.

4.1 Generalized bias amplification in least squares

If $f_2(\boldsymbol{BAV})$ is known or can be well approximated (say by an appropriate basis expansion that grows in dimension as $n \to \infty$) up to a constant of proportionality and intercept, then in the limit there will not be any bias due to misspecification of $f_2(\boldsymbol{BAV})$ (for a look at the case when $f_2(\boldsymbol{BAV})$ is misspecified see Appendix section A.9.1). In such a case, we again can compare two feasible OLS estimators, the first being the Naive estimator ($\widehat{\beta}_a^{naive}$) as before and the second the OLS estimator resulting from including $f_2(\boldsymbol{BAV})$ or its approximation ($\widehat{\beta}_a^{|f_2(bav)}$). Under any DAG, it can be shown that under structural equations (19) and (20), there will be bias amplification when

$$\begin{aligned} \frac{|\text{Bias}(\widehat{\beta}_a^{bav\prime})|}{|\text{Bias}(\widehat{\beta}_a^{naive})|} &= \left(\frac{1}{1 - \mathcal{R}^2_{\boldsymbol{A}|f_2(\boldsymbol{BAV})}}\right) \\ &\times \left(\frac{|COR(\boldsymbol{A}, f_1(\boldsymbol{U})) - COR(\boldsymbol{A}, f_2(\boldsymbol{BAV}))COR(f_1(\boldsymbol{U}), f_2(\boldsymbol{BAV}))|}{|COR(\boldsymbol{A}, f_1(\boldsymbol{U})) + COR(\boldsymbol{A}, f_2(\boldsymbol{BAV}))\kappa|}\right) > 1 \end{aligned} \tag{21}$$

where $\kappa = \sigma_{f_2(\boldsymbol{BAV})}\sigma^{-1}_{f_1(\boldsymbol{U})}$. In the above expression, we assume the case that there is in fact bias in the naive model, i.e. $Bias(\widehat{\beta}_a^{naive}) \neq 0$.

Bias amplification is decomposed into two terms, the ratio of remaining variance in the treatment once the control variables are projected out, and the remaining covariance between the outcome and treatment through the

confounding pathways. Just like in the linear case, the first term $(1-\mathcal{R}^2_{A|f_2(BAV)})^{-1}$ is the ratio of remaining treatment variance once we have projected out linear combinations of the control variables, which is just a constant in the naive case and $f_2(\boldsymbol{BAV})$ in the adjusted case. This first term must be greater than 1 and is a hyperbolic function increasing as $f_2(\boldsymbol{BAV})$ linearly explains more variance in the treatment.

Now consider the second bias term $\left(\frac{|COR(\boldsymbol{A},f_1(\boldsymbol{U}))-COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))COR(f_1(\boldsymbol{U}),f_2(\boldsymbol{BAV}))|}{|COR(\boldsymbol{A},f_1(\boldsymbol{U}))+COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))\kappa|}\right)$ which is the ratio of the covariance between the outcome and treatment remaining through the uncontrolled confounding pathways. The numerator of this expression $(|COR(\boldsymbol{A},f_1(\boldsymbol{U}))-COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))COR(f_1(\boldsymbol{U}),f_2(\boldsymbol{BAV}))|)$ is the magnitude of the correlation between the treatment and $f_1(\boldsymbol{U})$ once we have projected out linear combinations of $\boldsymbol{BAV}$. When $f_1(\boldsymbol{U})$ and $f_2(\boldsymbol{BAV})$ are uncorrelated, or independent as they would be under the DAG in Figure 3 and equation (17) in section 3, then the numerator reduces to simply $|COR(\boldsymbol{A},f_1(\boldsymbol{U}))|$ since $\boldsymbol{BAV}$ does not linearly explain any of the shared variance in the treatment $\boldsymbol{A}$ and $f_1(\boldsymbol{U})$.

In general, when there is correlation between $f_1(\boldsymbol{U})$ and $f_2(\boldsymbol{BAV})$, adjusting for $f_2(\boldsymbol{BAV})$ reduces the confounding bias of $f_2(\boldsymbol{U})$ when the signs of the correlations $COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))$ and $COR(f_1(\boldsymbol{U}),f_2(\boldsymbol{BAV}))$ are the same. Otherwise, this adjustment may further increase the confounding bias. As a simple example, if all of the correlations are positive (i.e. $COR(\boldsymbol{A},f_1(\boldsymbol{U})),COR(\boldsymbol{A},f_2(\boldsymbol{BAV})),COR(f_1(\boldsymbol{U}),f_2(\boldsymbol{BAV}))>0$ then the second term in the amplification expression will be less than or equal to 1 ($\left(\frac{|COR(\boldsymbol{A},f_1(\boldsymbol{U}))-COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))COR(f_1(\boldsymbol{U}),f_2(\boldsymbol{BAV}))|}{|COR(\boldsymbol{A},f_1(\boldsymbol{U}))+COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))\kappa|}\right)\leq 1$). In this case, we reduce bias in two ways. First we remove the bias due to $COR(\boldsymbol{A},\boldsymbol{BAV})$ directly and then we reduce the remaining covariance between $\boldsymbol{A}$ and $f_1(\boldsymbol{U})$ through the part of this correlation explained by $\boldsymbol{BAV}$.

There will be amplification, on the other hand, if the first term (which is always greater than one) is bigger than the reciprocal of the second term (which in this case is less than one). If $COR(\boldsymbol{A},f_1(\boldsymbol{U}))$ and $COR(\boldsymbol{A},\boldsymbol{f_2}(\boldsymbol{BAV}))$ are opposite signed, then when we adjust for $\boldsymbol{BAV}$ and remove bias due to the confounding path $\boldsymbol{A}\leftarrow\boldsymbol{BAV}\rightarrow\boldsymbol{Y}$ we may increase bias since $COR(\boldsymbol{A},f_1(\boldsymbol{U}))$ and $COR(\boldsymbol{A},f_2(\boldsymbol{BAV}))$ would no longer be partially cancelling each other out as they do in the naive model. Similarly, projecting out linear combinations of $\boldsymbol{BAV}$ may increase or decrease the remaining covariance between the treatment and $f_1(\boldsymbol{U})$ as seen in the numerator of the second term. Overall, this second term may be greater than or less than 1, meaning it may contribute to increasing or decreasing the bias relative to the naive estimator. By estimating the correlations and variances which are functions of observables and eliciting domain knowledge about the remaining unobserved quantities in the relative bias expression, we can make informed predictions about the plausibility of bias amplification.

Even in the most general case, where there are interactions and non-separability of $\boldsymbol{U}$ and $\boldsymbol{BAV}$ in the data generating model equation, the second term of the bias term will always be proportional to $1-\mathcal{R}^2_{A|Z}$, where $\boldsymbol{Z}=[\boldsymbol{\iota},h(\boldsymbol{BAV})]$ and $h(\boldsymbol{BAV})$ are whatever function of $\boldsymbol{BAV}$ that we include in a regression. As can be seen in the appendix (section A.9.1), this is even true when we have misspecified $f_2(\boldsymbol{BAV})$. However, there will be an additional bias term due to misspecification, but this will also be amplified with respect to the denominator term. The overall lesson is that when using OLS for causal effect estimation, unmeasured confounding bias (and remaining misspecification bias) will be amplified hyperbolically with respect to how well the controlling variables linearly explain variance in the treatment, i.e. the magnitude of $\mathcal{R}_{A|Z}$, where $\boldsymbol{Z}$ is all of the variables that are not $\boldsymbol{A}$ that we include in our OLS regression. Further, it is important to note that the amplifying factor $((1-\mathcal{R}^2_{A|f_2(BAV)})^{-1})$ is always a function of only observables and can always be estimated by running the regression $\boldsymbol{A}$ on $f_2(\boldsymbol{BAV})$. We can use these estimates for sensitivity analysis or in a model selection procedure. In a case like the simulation in Figure 2, our potential controls explain the large majority of the treatment variance. When this occurs, we should require more confidence that there is no unmeasured confounding after adjusting for our controls before we trust these estimates.

When there are causal pathways between $\boldsymbol{U}$ and $\boldsymbol{BAV}$, we showed that this may help make the first term smaller. However, this will open up an additional pathway for $f_2(\boldsymbol{BAV})$ to explain variance in the treatment (for example $\boldsymbol{BAV}\rightarrow\boldsymbol{U}\rightarrow\boldsymbol{A}$ as in the DAG in Figure 1, but in general the path could be indirect or the direction of causation may be reversed) and thus make the second hyperbolic term larger. This was the source of the dramatic bias amplification in Figure 2 simulated under linear structural equations from the DAG in Figure 1. In that specific case, $\boldsymbol{BAV}$ was a very good proxy for the unmeasured confounding $\boldsymbol{U}$ and this has two effects. First, it means that including $\boldsymbol{BAV}$ eliminates the large majority of the unmeasured confounding through its path to the outcome. Second, this means that $\boldsymbol{BAV}$ linearly explains nearly as much variation in the treatment as including both $\boldsymbol{BAV}$ and $\boldsymbol{U}$, which is nearly all of the variation in the treatment in that simulation case. So the little unmeasured confounding remaining was amplified significantly, more than 25 times $\left(((1-\mathcal{R}^2_{A|BAV})\sigma^2_a)^{-1}=(\sigma^2_a-\sum_{i=1}^{10}(\gamma_{bav\,i}+\gamma_{u\,i}\psi_{bav\,i})^2\sigma^2_{bav\,i})^{-1}=(1-0.96015)^{-1}=25.09725\right)$ under the DAG in Figure 1 and linear

structural equations (1) to (3)). This is because the potential amplifier linearly explained more than 96% of the variance in the treatment.

4.2 Bias amplification in more general semi-parametric models

In the beginning of this paper, we considered bias amplification in the context of linear expectations and linear models. In section 4.1 we extended this to when the structural models are non-linear when the estimator is ordinary least squares. Here we show proof of concept for bias amplification in more general semi-parametric orthogonalization methods. We plan to further develop this work in the future. We again consider the underlying structural equations to be represented by equations (19) and (20).

In the case that $\boldsymbol{f_2(BAV)}$ is unknown, one might turn to more general semi-parametric estimators for the causal effect estimation. A reasonable semi-parametric estimator for these structural equations is Robinson's Double Residual Regression (DRR hereforth)[18] when the outcome is hypothesized to be linear in the treatment (or well approximated by a partial linear function). In section 3 we showed via the FWL theorem that we can think of multiple regression as simple regression of the outcome and treatment once linear combinations of the control variables (which may be non-linear functions themselves as shown in section 4.1) have been projected out. In other words, we orthogonalized the outcome and treatment with respect to the subspace spanned by the linear combinations of all the control variables we included in our regression model ($\boldsymbol{Z}$) and then perform simple linear regression on the orthogonalized variables. The coefficient of interest in DRR, just like in multiple linear regression, is the result of a least squares estimate of an orthogonalized outcome and treatment, but we orthogonalize with respect to a more general subspace using the expectation operator.

Considering some arbitrary random variable $\boldsymbol{Y}$ and collection of controls $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ (each with finite variance), the conditional expectation $E[\boldsymbol{Y}|\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}]$ is the projection of $\boldsymbol{Y}$ onto a closed subspace of $\mathcal{L}^2$ consisting of all Borel functions $\phi(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}) : \mathcal{R}^n \rightarrow \mathcal{R}$, see for example Brockwell and Davis.[19] This is the heart of the familiar result that the conditional expectation is the function of $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ which minimizes the mean-squared error ($\text{argmin}_{f:\mathcal{R}^n \rightarrow \mathcal{R},\ \text{Borel}}\ E[(\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}))^2] = E[\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]$). The set of linear functions is a strict subset of all Borel functions, making conditional expectation a more general projection than the projections implicit in least squares estimation. In other words, if it turns out that the conditional expectations for both $\boldsymbol{Y}$ and $\boldsymbol{A}$ are in fact linear in the same variables, the DRR estimator will be equivalent to OLS in the probability limit.

The first step of DRR is to estimate the conditional expectations, i.e. the projections, $E[\boldsymbol{Y}|\boldsymbol{BAV}]$ and $E[\boldsymbol{A}|\boldsymbol{BAV}]$. We can use our favorite non-parametric estimator(s) as long as it is consistent as $n \rightarrow \infty$. It is common to use a conditional density estimator, but one may choose to use modern regression tree models or machine learning techniques like neural networks as in the context of Chernozhukov et al..[17] Now we construct our orthogonalized outcome and treatment, $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \widehat{E}[\boldsymbol{Y}|\boldsymbol{BAV}]$ and $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}]$. Finally, we perform least squares of the modified outcome ($\tilde{\boldsymbol{Y}}$) on the modified treatment ($\tilde{\boldsymbol{A}}$) to get our semi-parametric estimator denoted $\widehat{\beta}^{semi}$.

Bias amplification will occur in the probability limit when

$$\frac{|Bias(\widehat{\beta}^{semi})|}{|Bias(\widehat{\beta}_a^{naive})|} = \left(\frac{1}{1 - \frac{VAR(E[\boldsymbol{A}|\boldsymbol{BAV}])}{\sigma_a^2}}\right) \times \left(\frac{|E[COV(\boldsymbol{A}, f_1(\boldsymbol{U})|\boldsymbol{BAV})]|}{|COV(\boldsymbol{A}, f_1(\boldsymbol{U})) + COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))|}\right) > 1$$

Once again, we can decompose this relative bias expression into two components. The first term is the ratio of remaining treatment variance once the controlling variables have been projected out of the treatment and the second term $\left(\frac{|E[COV(\boldsymbol{A}, f_1(\boldsymbol{U})|\boldsymbol{BAV})]|}{|COV(\boldsymbol{A}, f_1(\boldsymbol{U})) + COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))|}\right)$ is the ratio of confounding covariances remaining through the outcome paths $\boldsymbol{U} \rightarrow \boldsymbol{A} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{BAV} \rightarrow \boldsymbol{A} \rightarrow \boldsymbol{Y}$. The first term again always contributes to amplification, since $(1 - VAR(E[\boldsymbol{A}|\boldsymbol{BAV}])/\sigma_a^2)^{-1} \geq 1$. This is the direct analogue of the results developed in the previous sections. In this case, $\mathcal{R}^2_{A|BAV}$ is the variation of the treatment explained by linear combinations of $\boldsymbol{BAV}$ (or $h(\boldsymbol{BAV})$ if we have included a different function of $\boldsymbol{BAV}$ in our regression), and $VAR(E[\boldsymbol{A}|\boldsymbol{BAV}])\sigma_a^{-2}$ is the total variation explained by fluctuations in $\boldsymbol{BAV}$.[20] As before, the first term increases hyperbolically as $\boldsymbol{BAV}$ explains more treatment variance.

When we project out $\boldsymbol{BAV}$, the term $COV(\boldsymbol{A}, f_2(\boldsymbol{BAV})$ drops from the second component and $COV(\boldsymbol{A}, f_1(\boldsymbol{U}))$ becomes $E[COV(\boldsymbol{A}, f_1(\boldsymbol{U})|\boldsymbol{BAV})]$. Depending on the signs of the bias and the direction of the correlation $COV(E[\boldsymbol{A}|\boldsymbol{BAV}], E[f_1(\boldsymbol{U})|\boldsymbol{BAV}])$, the second term may be greater or less than 1. This is similar to the first term

in the expression in equation (21) since the biases from the two confounding paths may partially cancel each other out and thus by projecting them both out, one may increase bias.

In other words, the underlying mechanics of bias amplification extends to more general orthogonalization methods, both explicit as in the case of DRR and implicit like that in OLS. In particular, the partially linear regression set-up is the basis for more complicated causal machine learning techniques such as Double Machine Learning residuals on residuals regression,[17] which has become a popular approach for dealing with regularization bias in high dimensional settings. Even in these more complicated settings where we orthogonalize in more sophisticated ways, variables which predict large amounts of variance in the treatment may significantly increase bias if confounding pathways remain. Once again, the hyperbolic remaining treatment variance term is entirely a function of observed variables and thus can be estimated (non-parametrically if desired) to see if one is at risk of bias amplification with respect to priors or sensitivity analysis over expected levels of unmeasured confounding. Similarly, we can use the mean-squared error of our resulting model to evaluate the extent to which unmeasured confounding is possible through the outcome path since no interaction implies that remaining variation must either be independent error or unmeasured confounding. A full treatment of this framework is beyond the scope of this article and will be the subject of future work.

5 Causal simulation experiments: the case of bias amplification

In our experience, simulating bias amplification is challenging in a number of subtle, but important ways. Our context of interest is assessing the potential for bias amplification in an analysis of an observational study in which we have measured several independent variables and the outcome but there might be an unmeasured confounder. We are interested in evaluating the feasible estimators we have developed in the previous sections, $\widehat{\beta}_a^{naive}$ and $\widehat{\beta}_a^{bavi}$ for example, with respect to possible data sets generated by a class of DAGs and structural equations. In this section, we show that if we constrain certain aspects of the simulated data (in particular, the marginal variances of observed quantities), we are better able to articulate and answer causal questions about the effect of bias amplification on proposed estimators. While we discuss the example of bias amplification simulations specifically, this section has implications for simulating data to test causal estimators more broadly. Now, consider the challenge of determining the effect of increasing unmeasured confounding on bias amplification in Figure 5(a). We might, for example, be interested in how large an unmeasured confounder must be, with fixed amplifying variables, to cross some threshold of bias in the adjusted model as part of a sensitivity analysis. To answer such a question, we must define clearly what is meant by the strength of an unmeasured confounder. In Figure 5(a), there are two edges which determine the overall bias due to the unmeasured confounding path through U: the edge from $\boldsymbol{U}$ to $\boldsymbol{A}$ and the edge from $\boldsymbol{U}$ to $\boldsymbol{Y}$. The bias due to the unmeasured confounding path through $\boldsymbol{U}$ in the naive model is simply the product of the weight of these two edges, scaled by the variance of the treatment as shown in equation (7). The extent to which bias can become amplified, however, is not symmetric with respect to the weight of the edges $\boldsymbol{U} \rightarrow \boldsymbol{A}$ and $\boldsymbol{U} \rightarrow \boldsymbol{Y}$, since amplification is the result of variance explained in the treatment as discussed in section 3. There is more potential for amplification of a strong unmeasured confounder (in the sense the product of the confounding edges is large) when the strength is due to $\boldsymbol{U}$ being a strong cause of $\boldsymbol{Y}$ compared to a strong cause of $\boldsymbol{A}$. This is because when $\boldsymbol{U}$ is a strong cause of $\boldsymbol{A}$, the $\boldsymbol{BAV}$ can only explain a small amount of the variance of $\boldsymbol{A}$, limiting the possible amount of bias amplification. Thus, to answer a causal question about the effect of increased unmeasured confounding on bias amplification, we should only vary one of the confounding edges, holding all other edges fixed.

As an example, suppose we are interested in the change in bias amplification when we increase the strength of the edge from $\boldsymbol{U}$ to $\boldsymbol{A}$, holding all else constant. This notion of intervening on a single edge of our DAG while holding the others fixed should be familiar to causal inference practitioners since it is the principle behind counterfactual analysis more broadly. Here we want to ensure that our results from varying a single edge are not confounded by variations in other edges as the result of unintended consequences or induced associations.

Because the goal is to increase the strength of a single edge, holding all else constant, we must specify a metric by which we measure the strength of the edge. In a fully linear system, we might consider the strength of the edge as the regression coefficient itself, γ_u, or the proportion of variance explained by $\boldsymbol{U}$, $\frac{\gamma_u^2 \sigma_u^2}{\sigma_a^2}$ and the sign of γ_u. It is tempting to see the two measures as equivalent with different scalings, but this is only true in the context of simulating a single equation. In the context of a system of linear equations, especially with the potential for bias amplification, we argue the relevant quantity is the proportion of variance explained by each child node of the parent variable. This can be seen most easily by examining the bias formula in equation (11), where the denominator is the remaining variation in $\boldsymbol{A}$ unexplained by the potential bias amplifying variables.

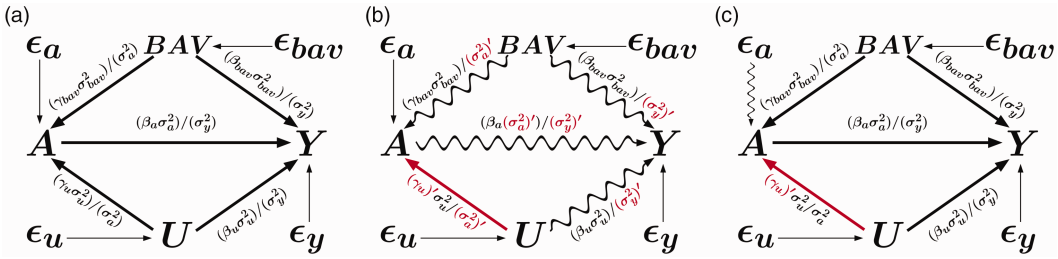

Figure 5. Causal diagrams where (a) represents the underlying causal structure. In addition to the usual causal pathways, we explicitly show the pathway of the independent error terms, ϵ_a, ϵ_u, ϵ_{bav}, ϵ_y. (b) The simulation experiment where we change strength of the edge $U \rightarrow A$ (shown in bold red) and do not renormalize the variance of the treatment (***A***). All edges which have been inadvertently modified are symbolized as squiggly arrows. All parameters which have been modified (inadvertently or intentionally) are shown in red along the edges. (c) Intervening on $U \rightarrow A$, where the variance of the treatment is fixed. To do so, we modify the variance of the noise term ϵ_a, visualized by the squiggly arrow. Notice no other edges are inadvertently modified.

Consider the implications of treating the coefficients themselves as the relevant measure of edge strength in a simulation trying to determine the effect of increasing the causal association along the path from $\boldsymbol{U}$ to $\boldsymbol{A}$. If we want to increase γ_u to $\gamma_u' > \gamma_u$ without changing any other parameters, we must also increase the total variance in the treatment, $\boldsymbol{A}$, since $\sigma_a^2 = \gamma_u^2\sigma_u^2 + \gamma_{bav}^2\sigma_{bav}^2 + \sigma_{\epsilon_2}^2$. A treatment with a larger variance is in some sense a different intervention, and thus this simulation is not compatible with the class of experiments which generated the original data with parameter γ_u. Further, from the previous sections we know this implies the total amount of variance explained from the bias amplifier $\boldsymbol{BAV}$ is reduced, since $\frac{\gamma_{bav}^2\sigma_{bav}^2}{(\sigma_a^2)'}$ has been reduced. Although we have not changed the parameter γ_{bav} we have decreased the extent to which $\boldsymbol{BAV}$ amplifies the bias as seen by examining equation (11). The increased variance in $\boldsymbol{A}$ in turn modifies the total variance of $\boldsymbol{Y}$. Therefore, the relative proportion of variance of $\boldsymbol{Y}$ that is explained by $\boldsymbol{BAV}$ is modified by changing the causal effect of $\boldsymbol{U} \rightarrow \boldsymbol{A}$, as are the measured proportion of variance of $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$, $\boldsymbol{A} \rightarrow \boldsymbol{Y}$, $\boldsymbol{U} \rightarrow \boldsymbol{Y}$, and $\boldsymbol{BAV} \rightarrow \boldsymbol{Y}$ and their associated covariance terms.

We can see in Figure (5(b)) that by modifying a single coefficient and leaving all other coefficients unchanged we have inadvertently modified the relative proportion of variance explained by the four other edges ($\boldsymbol{BAV} \rightarrow \boldsymbol{A}$, $\boldsymbol{BAV} \rightarrow \boldsymbol{Y}$, $\boldsymbol{U} \rightarrow \boldsymbol{Y}$, and $\boldsymbol{A} \rightarrow \boldsymbol{Y}$) represented by the wavy arrows. Data generated by the second set of structural equations are not compatible with the constraints of the experiment which generated the first data and by intervening on a single edge we have modified all of the competing effects of interest. Comparing the distribution of estimates produced under γ_u and γ_u' gives us a confounded and thus biased estimate of the impact of increasing the unmeasured confounding through its causal pathway to the treatment on the estimators or functions thereof. We will show that this bias can result in under-estimating the impact of bias amplifying variables.

In general, when we vary one of the regression coefficients along a causal pathway, this has upstream and downstream effects on the proportion of variance explained by all variables going into or out of the varied node. In order to keep the proportional effects of the other edges constant, we need to use the error terms of the structural equations (ϵ_1 and ϵ_2) to absorb the shocks to the marginal variances.

In Figure 5(c), if we change γ_u and simply adjust the structural error term ϵ_a such that the total variance in $\boldsymbol{A}$ remains constant, we can isolate the effect of modifying $\boldsymbol{U} \rightarrow \boldsymbol{A}$. In Figure 6(a) we visualize the consequences of failing to hold the variance of the treatment when we modify γ_u. In red, for Figure 6(a), we simulate bias amplification where $\gamma_u = 0.3$. In green, we simulate bias amplification where γ_u is increased to 0.55 holding all other parameters constant, thus allowing the total variance of the treatment to grow from 1 to 1.21. This has the downstream effect of also increasing the variance of the outcome from 1 to 1.02. This also then impacts the relative proportions of variance explained of the treatment and the outcome that are explained by $\boldsymbol{U}$ and $\boldsymbol{BAV}$, respectively. Notice that the bias increases from 20% ($\frac{0.36-0.3}{0.3}$) to 43% ($\frac{0.43-0.3}{0.3}$). In blue, we increase γ_u from 0.3 to 0.55, but re-normalize the variance in the treatment to remain constant at 1. The bias now increases further to 67% ($\frac{0.5-0.3}{0.5}$) with respect to the original simulation in red. We do this by decreasing the variance of the independent noise term, ϵ_a to ϵ_a', allowing it to absorb the increase in variation from $\boldsymbol{U}$. When we do not fix the variance, we underestimate the impact of the amplifier on both the bias and the variance *because the unfixed variance case simulates a different kind of intervention due to the change in variance of the treatment variable*. In the simulation

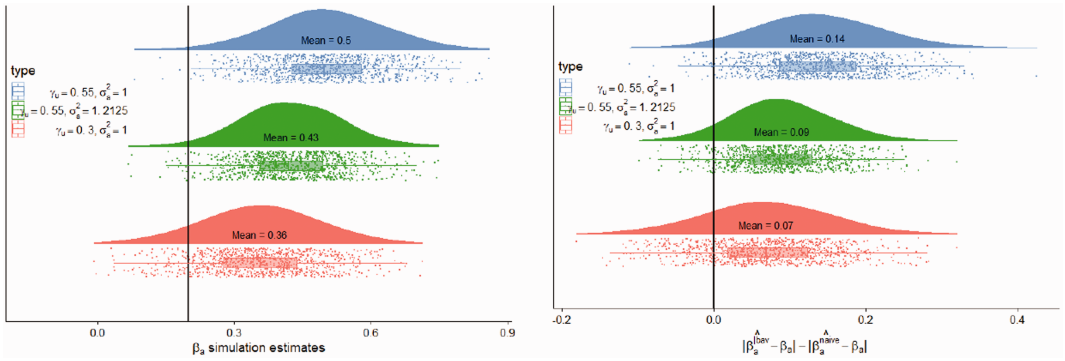

Figure 6. (a) Simulation results from the experiment of intervening on the edge $\boldsymbol{U} \rightarrow \boldsymbol{A}$. The ground truth, $\beta_a = 0.2$, is visualized by the black vertical line. In red we visualize the baseline bias amplification. Green shows the results of the conditional estimator in the case where we increase the weight of the edge, but fail to fix the variance of the treatment, allowing it to grow. In blue, we show the bias when we increase the weight of the edge but now hold the variance constant so the variances remain compatible with the original data. (b) Simulation results from the same experiment as (a) but the outcome is the difference in absolute bias between the conditional estimator $\widehat{\beta}_a^{bav\prime}$ and the naive estimator $\widehat{\beta}_a^{naive}$. A value greater than zero (black vertical line) indicates that the conditional estimator was more biased than the naive estimator. Parameter values: $\beta_a = 0.2$, $\beta_u = 0.3$, $\beta_{bav} = -0.05$, $\gamma_{bav} = 0.6$.

above, by not keeping the variance fixed in $\boldsymbol{A}$ we implicitly reduced the amount of variance that $\boldsymbol{BAV}$ accounts for in the treatment from 36% to 30%. In effect, we were comparing the distribution of

$$P(\widehat{\beta}_a|\gamma_u{}', \gamma_{bav}, \beta_u, \beta_a, \beta_{bav})$$

to

$$P(\widehat{\beta}_a|\gamma_u, \gamma_{bav}, \beta_u, \beta_a, \beta_{bav})$$

when a more fair causal counterfactual would be to compare the distribution of

$$P(\widehat{\beta}_a|(\boldsymbol{U} \rightarrow \boldsymbol{A})', \boldsymbol{U} \rightarrow \boldsymbol{Y}, \boldsymbol{BAV} \rightarrow \boldsymbol{A}, \boldsymbol{BAV} \rightarrow \boldsymbol{Y}, \boldsymbol{A} \rightarrow \boldsymbol{Y})$$

to

$$P(\widehat{\beta}_a|\boldsymbol{U} \rightarrow \boldsymbol{A}, \boldsymbol{U} \rightarrow \boldsymbol{Y}, \boldsymbol{BAV} \rightarrow \boldsymbol{A}, \boldsymbol{BAV} \rightarrow \boldsymbol{Y}, \boldsymbol{A} \rightarrow \boldsymbol{Y})$$

Therefore, our simulation *experiment* results in green are distorted because when we increased the unmeasured confounding through $\boldsymbol{U} \rightarrow \boldsymbol{A}$, we also decreased the strength of the bias amplifying variable through the pathway $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$. Notice that this bias will impact decisions and conclusions we might make about the merits of different estimators in this context. For example, below we compare the conditional estimator, $\widehat{\beta}_a^{bav\prime}$, to the naive estimator, $\widehat{\beta}_a^{naive}$ with respect to their bias in the same three simulation set ups.

In Figure 6(b) we show the direct comparison of the bias for the conditional and the naive estimators. When we increase the unmeasured confounding through γ_u but fail to renormalize the treatment variance, we do not capture the full extent to which the conditional estimator amplifies the bias. If we compare the green and red plot, it would seem that nearly doubling the unmeasured confounding coefficient only has a small impact on the relative bias of the naive and conditional estimator, since the relative bias only increased from 0.07 to 0.09 (29%). By comparing the green density plot to the blue, we see the relative bias doubles (from 0.07 to 0.14). Therefore, the decision to use the naive or conditional estimator is in fact much more sensitive to the amount of unmeasured confounding than it would appear under the improper simulation with floating variance. It is extremely important to do these kinds of simulations properly particularly in the context of sensitivity analysis where we are testing the performance of estimators with respect to untestable assumptions such as unmeasured confounding.

To properly simulate bias amplification and answer questions of clinical concern with respect to the merits of potential estimators, we must think of the structural equations as an interconnected system. While we typically specify such equations from the perspective of determining their conditional means, the structural equations along with our independence assumptions determine the variances of the variables in the system. Above, this necessitates increasing the strength of the edge $\boldsymbol{U} \rightarrow \boldsymbol{A}$ while holding all other edges constant, which requires us to re-normalize the variances to maintain the strength of the edge $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$.

In Appendix section A.10, we consider the properties of a simulation experiment aiming to vary the strength of the edge $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$. We show that in the case that we fail to fix the variance of the treatment that the bias of the conditional estimator $\widehat{\beta}_a^{bav\prime}$ is invariant to $\gamma_{bav}, \forall \gamma_{bav} \in (-\infty, \infty)$, but that the naive estimator is strictly increasing in γ_{bav}. It is clear from the theory we developed in section 3 that if we increase the edge from $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$ that amplification should strictly increase, but if we allow the variance in $\boldsymbol{A}$ to increase as the parameter increases, the amplification effect is precisely cancelled out.

In general terms, simulating linear systems of location-scale family random variables requires first fixing the variances of the variables in the DAG. The relevant quantity determining the strength of the various edges are ratios of variances and covariances of the upstream parent nodes to the variance of the child node in determining the edge's strength. Since the effects are relative, in a simulation context we can normalize the variances to 1 or set them to the expected/observed variances of the data in a particular context. For simplicity we will demonstrate the normalized approach. In Figure 3, this means that $\sigma_u^2 = \sigma_a^2 = \sigma_{bav}^2 = \sigma_y^2 = 1$.

The second step is to be explicit about independence and conditional independence assumptions. Given the independence assumptions, we can specify the covariance matrix of each child variable $\boldsymbol{Y_{child}}$ in terms of the matrix of k arbitrary parent variables which form the edges going into the child variable, $\boldsymbol{Y_{child}}$.

$$\boldsymbol{Y_{child}} = \boldsymbol{X_{parent}}\boldsymbol{\beta_{parent}} + \epsilon_{\boldsymbol{child}}$$

$$Var(Y_{child}) = \boldsymbol{\beta_{parent}^T}\boldsymbol{Var}(\boldsymbol{X_{parent}})\boldsymbol{\beta_{parent}} + Var(\epsilon_{\boldsymbol{child}})$$

$$\sigma_{y_{child}^2} = 1 = [\beta_1 \quad \beta_2 \quad \dots \beta_k] \begin{bmatrix} 1 & \sigma_{1,2} & \dots & \sigma_{1,k} \\ \sigma_{1,2} & 1 & \dots & \sigma_{2,k} \\ \vdots & \sigma_{j,2} & \ddots & \vdots \\ \sigma_{k,1} & \sigma_{k,2} & \dots & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \sigma_{\epsilon_c}^2$$

The diagonal of all the parent covariance matrices is 1 since we have normalized all variables pictured in the DAG. The covariances themselves will be determined by the independence assumptions, the edges connecting the child nodes, and their structural equations. Essentially we are choosing the proportion of the child variation that the variances and the covariances of the parent variances explain. The error terms, ϵ's are the only non-normalized variances, and they absorb the shocks when we increase and decrease the strength of the edges of the non-error variables. This maintains the strength of all other relations visualized on the DAG.

Since all variance terms must be non-zero (or equivalently that $\boldsymbol{\beta_{parent}^T}\boldsymbol{Var}(\boldsymbol{X_{parent}})\boldsymbol{\beta_{parent}} \leq 1 = \sigma_{child}^2$), the variance equations define bounds on the simulation parameter space. In the above example, conditional on holding the strength of the edges $\boldsymbol{U} \rightarrow \boldsymbol{Y}, \boldsymbol{BAV} \rightarrow \boldsymbol{A}, \boldsymbol{BAV} \rightarrow \boldsymbol{Y}, \boldsymbol{A} \rightarrow \boldsymbol{Y}$, $\gamma_u \in (-0.893, 0.893)$ defines the feasible range. That is, the edge $U \rightarrow A$ can explain up to 79.75% of the variation in $\boldsymbol{A}$ ($\frac{\gamma_u^2 \sigma_u^2}{\sigma_a^2} = \frac{\gamma_u^2}{1}$) since the edge $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$ explains 20.25% of the variation already. In general, the extent to which an edge can explain variation in the child node is constrained by the other child nodes and the covariance structure between those variables. A parameter, however, such as γ_u may be constrained by more than one set of inequalities. In this particular case, γ_u has to satisfy the following inequalities

$$|\gamma_u| \leq (1 - \gamma_{bav}^2)^{\frac{1}{2}}$$

$$\gamma_u \leq \frac{1 - \beta_a^2 - \beta_u^2 - \beta_{bav}^2 - 2\beta_a\beta_{bav}}{2\beta_a\beta_u}$$

where conditional on the strength of the particular edges ($\boldsymbol{U} \rightarrow \boldsymbol{Y}$, $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$, $\boldsymbol{BAV} \rightarrow \boldsymbol{Y}$, $\boldsymbol{A} \rightarrow \boldsymbol{Y}$) in the above simulation, only the first inequality was binding.

The nuance here is that the extent to which we can simulate unmeasured confounding depends upon not only how much amplifying we have simulated, but also on the true effect of the treatment on the outcome $\boldsymbol{A} \rightarrow \boldsymbol{Y}$. Since this is an interdependent system of equations, all of the parameters are competing for shares of fixed variances. If the treatment, independent of $\boldsymbol{U}$ and $\boldsymbol{BAV}$, explains the large majority of the outcome variance (i.e. the edge $\boldsymbol{A} \rightarrow \boldsymbol{Y}$), it means the weight of the edge $\boldsymbol{U} \rightarrow \boldsymbol{Y}$ must be relatively small, opposite signed, or the structural equations contain an effect modifier. This in turn constrains γ_u.

Consider again the above simulation experiment where we are interested in varying the strength of $\boldsymbol{U} \rightarrow \boldsymbol{A}$ conditional on all other pathways. Suppose that the pathway $\boldsymbol{A} \rightarrow \boldsymbol{Y}$ explains 64% of the variance in $\boldsymbol{Y}$, i.e. that $\beta_a = 0.8$. Now both constraints on γ_u are binding and the simulation parameter space is $\gamma_u \in (-0.893, 0.3916)$.

In summary, when simulating linear location-scale family systems of equations, we start by identifying the DAG and the independence assumptions between variables. Second, our simulation experiment should attempt to answer a causal question about how a proposed estimator behaves in response to an intervention on the weights of causal DAG. Just like experimental design, properly estimating the relevant counterfactual requires that the difference in distributions between our intervention(s) and the control is the effect of the intervention(s) themselves. As demonstrated in this section, simulating linear systems of equations requires varying one of the edges of the DAG holding all else constant, and matching the means and variances of the simulated variables with that of the target observational study we are trying to mimic. This allows us to generate simulations whose distributions are proper counterfactuals. Third, conditional on the other edges, the covariance matrices impose bounds for the parameter space that we can simulate and thus the extent to which we can vary the edge of interest. For a specific realization of the experiment and accompanying valid parameters, the variables are constructed in the downstream direction, that is from parent nodes to child.

In the example of simulating the proper intervention in Figure 5(c), we first simulate $\boldsymbol{U}$ and $\boldsymbol{BAV}$ independently with variance 1, respectively. Given γ_u and γ_{bav}, the variance of the error term ϵ_2 from equation (5) is implied and can be simulated. Having $\boldsymbol{U}$, $\boldsymbol{BAV}$ and ϵ_2 allows us to simulate the treatment $\boldsymbol{A}$. Conditional on the already simulated variables, their associated parameters, and β_a, β_{bav}, and β_u, the variance of the error term ϵ_1 is implied and can be simulated. Finally, since all of the child variables for the outcome have been simulated, we can simulate the outcome. To be clear, we can fix proportions of variance explained by each edge in any order we'd like as long as we respect the underlying constraints. However, given an admissible set of weights of the edges we must proceed from parent to child nodes to conduct the simulation.

While this method requires us to calculate inequalities and make explicit the implied variance formulas for our variables, the benefits are that we can view our simulation as a well-defined causal experiment matching the constraints of our target study and we get sets of parameter bounds. When we do not keep the variance fixed, there are no defined bounds beyond heuristics, and more importantly, we are no longer matching the data to our target observational study. In many small systems, such as the one in Figure 5(a), it is often computationally inexpensive to simulate a discretized approximation to all possible parameter configurations. In extremely large systems, we can use domain knowledge to make refinements on these bounds and simulate a reasonable subset of the parameter space. This method allows us to make refinements over edges with strong priors while simulating the entirety of edges with greater uncertainty.

An alternative, but equivalent way of developing a causal simulation for linear systems of equations would be to work with correlation (or covariance) matrices and variances directly as opposed to parameters. In the Appendix section A.5, we discuss representations of regression coefficients and linear structural parameters in terms of partial correlations, correlations, and coefficients of determination. In some cases, it may be easier to elicit domain knowledge using partial correlations or correlations, compared to the parameters themselves, depending on the clinician and the application at hand. There are still constraints on the system, which are implied by the constraints of a non-singular correlation matrix. One of the potential benefits of working directly with the correlation matrix is that, as shown in section A.5 much of the pairwise correlation matrix can be directly estimated form the observed data. Further, there exist methods to augment the estimable portion of the matrix to a non-singular full correlation matrix as well as methods to add noise to a valid correlation matrix to represent the uncertainty inherent in the estimation process. Additionally, sometimes causal DAGs may imply further restrictions on correlation matrices (or partial correlation matrices) and thus the underlying regression parameters and causal effects.

6 Simulating bias amplification from a real data set

Here we conduct a data simulation for an observational study. We want to consider a medical example with realistic amounts of variance in the treatment and the outcome. Further, we specifically consider the case of a binary treatment which is common in medical applications, biostatistics, and epidemiology. The difficulty, in general, when working with real data in the spirit of plasmode simulation[21,22] is that you do not know the true underlying parameter values. In this section, we start with a randomized controlled trial (RCT) and modify it appropriately, so that we can take the intention to treat (ITT) estimate as the true underlying effect for the foundation of our simulations.

In our simulation experiment, we keep the treatment data unchanged (thus fixing their variance), and then simulate unmeasured confounding ($\boldsymbol{U}$) and bias amplifiers ($\boldsymbol{BAV}$) in order to modify selected covariates ($\boldsymbol{X}$) and the outcome ($\boldsymbol{Y}$) to produce a synthetic observational experiment. In order to precisely control the relationships between the simulated variables and the real variables, we treat the binary treatment, $\boldsymbol{A}$, as though it comes from a latent probit model.

$$\begin{aligned}\boldsymbol{A} &= 1(\boldsymbol{A}^* > \boldsymbol{0}) \\ &= 1(\alpha_a + \boldsymbol{U}\gamma_u + \tilde{\boldsymbol{X}}\gamma_{\tilde{x}} + \epsilon_{\boldsymbol{2}} > \boldsymbol{0})\end{aligned}$$

where $\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\sigma'} + \boldsymbol{BAV}$, and σ' is a scaling variable such that $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ have the same population variance. All of the latent variables ($\boldsymbol{U}$, $\boldsymbol{BAV}_{\boldsymbol{n\times k}}$, $\epsilon_{\boldsymbol{2}}$, and hence $\boldsymbol{A}^* = \alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}} > \boldsymbol{0}$) are set to come from normal distributions. The details of the how the simulation is performed are in Appendix section A.11.

For this paper, we use data from Helle et al.,[23,38,39] a published RCT with 294 participants and relatively balanced distribution of covariates. While the researchers examined many outcomes, we will focus on the effects of an e-Health intervention in infants on child eating behaviors. The researchers gave the parents in the treatment group access to a "monthly age-appropriate video addressing infant feeding topics together with corresponding cooking films/recipes," and the outcome was eating habits of the child at a later point in time. In the observational study that we want to create (target observational study), we want to estimate the effect of the treatment on emotional overeating as measured by the Child Eating Behavior Questionnaire (CEBQ).

6.1 Unbiased ITT model

Our foundation is the unbiased ITT effect from the RCT data regressing the treatment on the outcome ($Y\sim A$) shown in the first column of Table 1.

In column 1 of Table 1, we see that the ITT estimate is 0.14. As this is an RCT, we do not expect baseline covariates [Child Food Neophobia Score ($\boldsymbol{CFNS}$), Child Feeding Questionnaire ($\boldsymbol{CFQ}$) subscale pressure, and Age of mother ($\boldsymbol{Age_{mother}}$)] to be associated with exposure. We thus assume that the experimental data are generated from the causal DAG in Figure 7(a), where $\boldsymbol{X}$ represents the matrix of all three covariates ($\boldsymbol{CFNS}$,$\boldsymbol{CFQ}$, and $\boldsymbol{Age_{mother}}$) after they have been individually standardized to have mean 0 and variance 1. To verify that these variables are not bias amplifiers, that is explain only a negligible proportion of the treatment variance, we also present the results of the regression of the treatment on the three covariates in column 3 in Table 1. We can see that jointly and individually the three covariates explain very little of the variance in the treatment, $\mathcal{R}^2 = 0.009$. This should be expected in a truly randomized experiment set-up since proper randomization breaks the causal association from the covariates to the treatment. Since these covariates do not cause $\boldsymbol{A}$ and we have assumed that the ITT estimator is unbiased, when we estimate $\boldsymbol{Y}\sim\boldsymbol{A} + \boldsymbol{CFNS_{score}} + \boldsymbol{CFQ_{pressure}} + \boldsymbol{Age_{mother}}$ the expectation and probability limit of $\widehat{\beta_a}$ remains unchanged regardless of the strength of association between the covariates and the outcome. However, actual results may vary due to final sample variation. In our RCT data, the unadjusted model estimates a treatment effect of 0.144 and the adjusted model estimates 0.137. Since simulation experiments performed in section 6.2 all condition on covariates, we consider the covariate adjusted results from the RCT as the gold standard for determining bias due to unmeasured confounding in our simulated data.

6.2 Biased model simulations

Our objective is to simulate data according to the DAG in Figure 7(b). To produce the simulations, we took 10,000 bootstrap replications of the original outcome, treatment and covariates. From each bootstrap sample of

Table 1. Clinical data regression table.

Model	ITT	ITT Cond.	*A*
A	0.144 (0.053)	0.137 (0.053)	–
CFNS	–	0.007 (0.005)	−005 (0.007)
CFQ	–	0.058 (0.036)	0.036 (0.040)
Age_{mother}	–	0.008 (0.006)	0.009 (0.007)
$\mathcal{R}^2$	0.018	0.045	***0.009***

Note: Each column represents a different regression. Column 1 is the unbiased Intention to Treat (ITT) model, regressing the outcome on the treatment. Column 2 represents the unbiased conditional ITT model with three covariates conditioned on. The third column represents the regression of the treatment on the three regressors. The row names are the independent variables, where *A* is the treatment, *CFNS* is the Child Food Neophobia Score, *CFQ* is the Child Feeding Questionnaire pressure subscale, and Age_{mother} is the age of the infant's mother. The relevant standard errors are displayed in brackets.

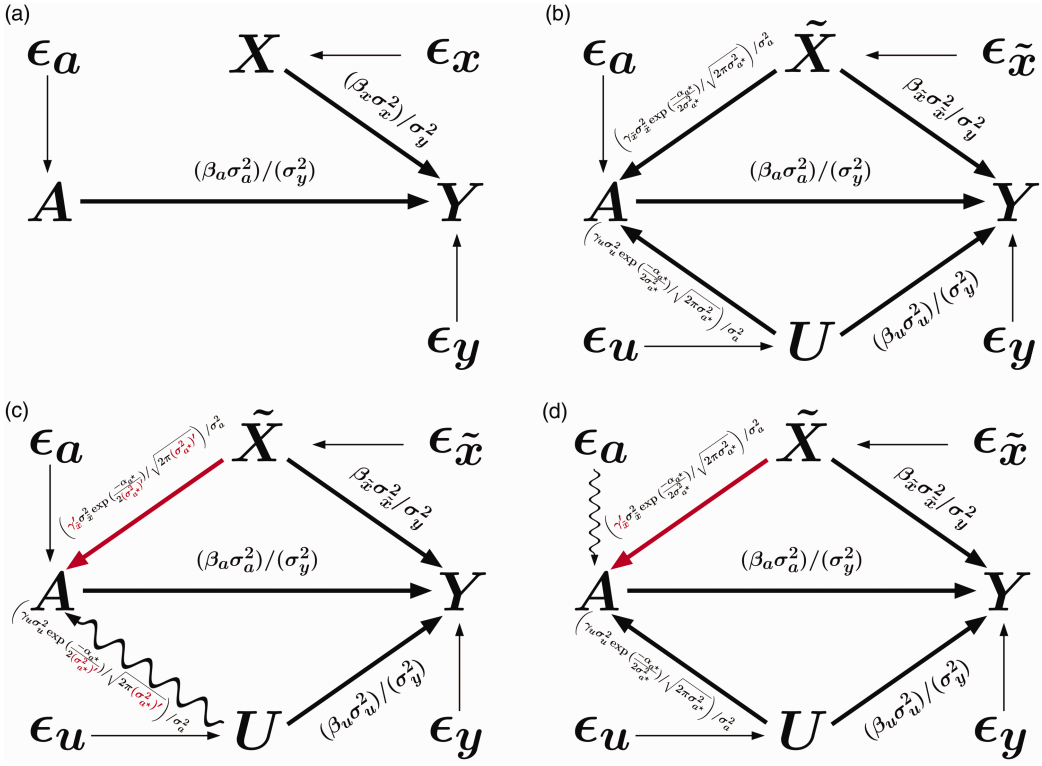

Figure 7. Causal diagrams. (a) DAG representing the original experiment data. (b) DAG with the modified data (see Appendix A.11 for details). (c) represents intervening on the causal DAG in (b) by changing the strength of the edge $\tilde{X} \to A$ without holding the latent treatment variance A^* constant. The edges which have been modified inadvertently are shown as squiggly arrows. Parameters which have been changed are shown in red. (d) The intervention on the edge $\tilde{X} \to A$ while holding the latent variance constant.

the treatment, $A_{bootstrap}$, of size $n = 294$ we simulated the latent variable A^* using the procedure outlined in the Appendix (section A.11). Next, conditional on the drawn latent samples of A^* and the bootstrapped covariates, we drew samples for the unmeasured confounding, U, and bias amplifying variable, BAV. The modified random control variables, $\tilde{X} = \frac{X}{\sigma'} + BAV$, were produced by adding the bias amplifying variables to a scaled version of the original control variables. Linear combinations of the unmeasured confounding and modified covariates were then added with reasonable values to the outcome such that the following DAG and equations hold (see simulation results).

$$\tilde{Y} = \alpha_y + A\beta_a + \tilde{X}\beta_{\tilde{x}} + U\beta_u + \epsilon_I \tag{22}$$

$$\boldsymbol{A}^* = \alpha_a + \boldsymbol{U}\gamma_u + \tilde{\boldsymbol{X}}\gamma_{\tilde{X}} + \epsilon_{\boldsymbol{2}} \tag{23}$$

$$\boldsymbol{A} = 1\{\boldsymbol{A}^* > \boldsymbol{0}\} \tag{24}$$

In section 6.1 we showed that the true treatment effect was 0.137 conditional on the covariates $\boldsymbol{X}$. In the bootstrap simulation pictured below, the unbiased model conditional on both the modified covariates, $\tilde{\boldsymbol{X}}$, and the unmeasured confounding $\boldsymbol{U}$ is 0.136 as expected. The naive model estimator had an average estimate of 0.234 in the simulations and thus an absolute estimated bias of 0.097, or a relative bias of 1.8 standard deviations ($\frac{0.234-0.137}{0.053}$) with respect to the unbiased estimate in section (6.1).

When we further condition on the modified covariates, the absolute bias ($E[|\widehat{\beta}_a^{|\tilde{x}} - 0.137|]$) more than doubles to 0.225, and the relative bias increases to 4.3 standard deviations ($\frac{0.36-0.137}{0.053}$) with respect to the unbiased estimate in section 6.1. The simulations confirm that bias amplification can be significant even when constrained to problems of realistic variance. Further, we see that bias amplification is potentially a problem for binary outcomes. This underscores the theoretical points made in sections (3) and (4) where we showed that the phenomenon behind bias amplification does not require specific distributional assumptions of the variables in the model.

More importantly, by combining the methodology outline in the appendix (see Appendix A.11) to simulate measured confounding using real data and the principles for simulating systems of equations in section 5, we can produce realistic and complete simulations of parameter spaces which match the underlying characteristics of the data. Investigators who choose covariates based on the assumption of no unmeasured confounding can now evaluate the amount of bias amplification that would occur if this assumption does not hold.

Finally, in the Appendix (section A.12) we consider an example of a causal simulation experiment with a binary treatment variable under the DAG in Figure 7(b) and structural equations (22) to (24). The experiment involves modifying the strength of the edge $\tilde{\boldsymbol{X}}_{\boldsymbol{1}} \rightarrow \boldsymbol{A}$ and evaluating the impact on the naive and conditional estimators. With binary treatment ($\boldsymbol{A}$), we show that if we fail to hold the variance of the latent treatment ($\boldsymbol{A}^*$) constant and increase $\gamma_{\tilde{x}_1}$, then it is possible to decrease the amount of observed treatment variance (σ_a^2) explained by $\tilde{\boldsymbol{X}}_{\boldsymbol{1}}$. Further, the increased treatment variance also decreases the strength of the edge $\boldsymbol{U} \rightarrow \boldsymbol{A}$. As a result of performing the causal simulation experiment improperly, it appears as though that varying the strength of the potential amplifiers has a negligible or negative impact on the resulting bias amplification. The improper and proper approaches to intervention are shown in Figure 7(c) and (d), respectively and the results from these simulations are visualized in Figure 10 of the Appendix section A.12. This of course leads to improper inferences regarding the relative merits of the naive and conditional estimators as well. This highlights once again the importance of comparing simulations with comparable properties and ensuring that when we intervene on the edges of our causal diagram that we are not inadvertently varying the edges we mean to keep fixed. Just as in the experimental context, our simulation results become muddled or meaningless if we are not evaluating well-articulated counterfactuals.

7 Discussion

Causal model selection techniques have largely been developed under the assumption that a sufficient set of variables is available to create ignorability. When a sufficient set is not available or when a causal variable selection technique does not correctly identify the sufficient set, we are at risk of bias amplification. In the first simulation in section 1, we showed that even under mild perturbations of the usual assumptions, conditioning on a set of jointly strong proxy variables for $\boldsymbol{A}$ in OLS led to a very biased estimator (0.73 standard deviations on average). Further, most current causal variable selection techniques are likely to include this set of variables since they are significant predictors of the outcome and the treatment as well as variables which cause large changes in estimates when included sequentially.

Under threat of bias amplification, treatment-oriented selection techniques for regression analyses using continuous exposure regimes should be used cautiously unless one has strong priors that a sufficient set is available and likely to be identified. We showed in section 3 that it is precisely the amount of variance in the treatment explained by the observables in our model which is responsible for bias amplification. Similarly, we can see that a significant change in estimate is not sufficient to suggest that overall bias is decreasing since this could be the result of further bias amplification.

These results call for new techniques to be developed for observational studies which can accommodate unmeasured confounding to help researchers choose reasonable and least-biased methods. We suggest to first

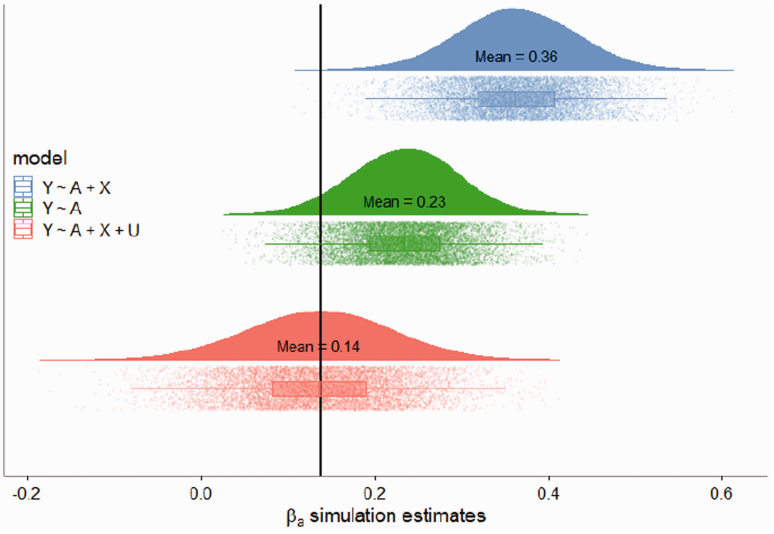

Figure 8. Here we compare three estimators for β_a from the structural equations in equations (22), (23), and (24). In red, the results of the unbiased and infeasible estimator are shown, centered at the true value $\beta_a = 0.137$ (shown by the vertical black line). In green, the replications for the naive estimator ($\widehat{\beta}_a^{naive}$) is shown and in blue the replications for the conditional estimator ($\widehat{\beta}^{|X}$) is shown. Simulation details: Bootstrap replications = 10,000. $\beta_a = 0.137$, $\beta_u = 0.15$, $\boldsymbol{\beta}\tilde{\boldsymbol{x}} = (0.10, -0.15, -0.10)$, $COV(\boldsymbol{A}, \boldsymbol{U}) = 0.25$, $COV(\boldsymbol{A}, \tilde{\boldsymbol{X}}) = (0.22, 0.15, 0.13)$.

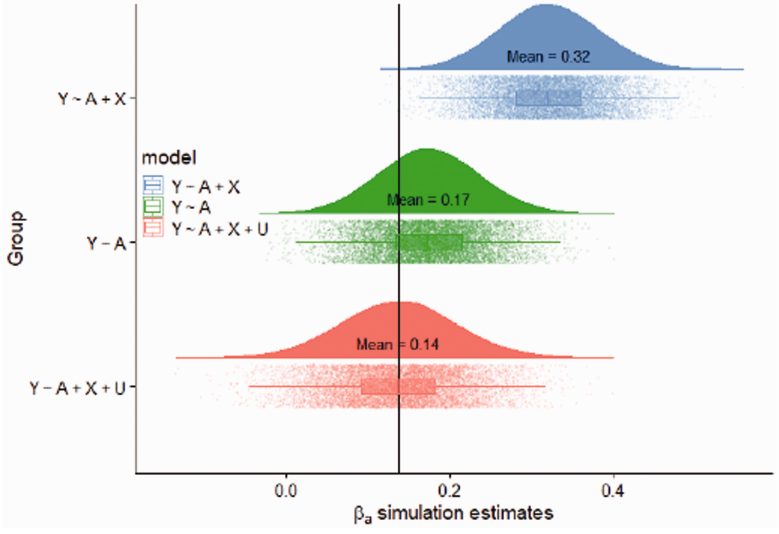

Figure 9. Control treatment estimators. Black line represents the true underlying parameter $\beta_a = .1377$. $N = 10{,}000$ simulation replications.

identify the most plausible causal DAG. From the DAG and basic structural equation assumptions, an expression for asymptotic bias can often be derived. Further, we suggest to estimate the always-identifiable amplification term in observational settings and to assess the risk of bias amplification. With a measure for amplification and a limiting bias expression, a sensitivity analyses can be performed. One reasonable sensitivity analysis approach would be to estimate the amount of unmeasured confounding required in the spirit of E-values[7] to determine the strength of confounding associations required to "explain away the treatment effect"[7] and to make principled inferences from the data. This would require, as we have shown, properly simulating the unmeasured confounder so that (1) the properties of the original data are respected and (2) other competing effects, i.e. edges of the DAG, are not inadvertently altered. In such a set-up, large effects when the controls jointly explain little of the treatment variance lend credibility to results as being robust to unmeasured confounding, particularly in cases when suitable

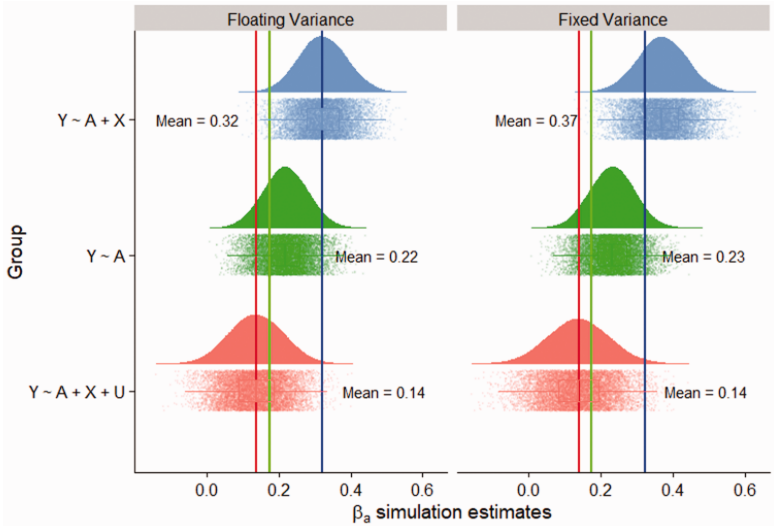

Figure 10. The vertical lines represent the means from the control experiment: red representing the mean of unbiased control estimator, green the mean of the naive estimator, and blue the mean of the amplified estimator.

priors can be placed on the variables along the unmeasured confounding pathway. Alternatively, one could follow the approach of Carnegie et al.[8] and use the underlying structural equations and the data to generate candidate values of the unmeasured confounding. As we showed in section 5, it is important that any such simulation method take into account the asymmetry of bias amplification with respect to the weight of the edge $\boldsymbol{U} \rightarrow \boldsymbol{A}$ and $\boldsymbol{U} \rightarrow \boldsymbol{Y}$.

Ultimately, simulation experiments must aim to produce data from which we can draw causal conclusions to questions about estimators or functions. This means having well-defined interventions on the edges of the causal graphs and holding the other edges constant. In linear systems of equations, this requires keeping the moments of the variables, in particular variance, fixed when modifying the weight of the DAG's edges. If we allow the treatment variance to vary incidentally as we increase confounding effects, the intervention arm of our simulations will no longer match the target observational study in the control arm. As a further consequence, the additional variance in the exposure may absorb much of the amplifying effect. This leads to systematic underestimation of bias amplification and may be an explanation for why the threat of bias amplification has not been appreciated as a concern for applied researchers.[15] Fixing the variance of the variables has the additional benefit of defining the feasible parameter space. By constraining the underlying parameters by the implied variance equations, it is computationally and conceptually easier to simulate the entire range of plausible treatment effects and biases. This leads to more representative simulations and more principled inferences.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors received funding from the Canadian Institute of Health Research (CIHR) through the Collaborative Health Research Projects (NSERC partnered) for the research and publication of this article (grant number: CPG-140204).

ORCID iD

Tyrel Stokes https://orcid.org/0000-0001-9305-1859

Supplemental Material

Supplementary material for this article is available online.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688.
2. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
3. Wooldridge JM. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press, 2010.
4. Witte J and Didelez V. Covariate selection strategies for causal inference: classification and comparison. *Biometric J* 2018; 61: 1270–1289.
5. Hernn MA, Hernndez-Daz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002; **155**: 176–184, https://doi.org/10.1093/aje/155.2.176
6. Greenland S and Pearce N. Statistical foundations for model-based adjustments. *Ann Rev Public Health* 2015; **36**: 89–108.
7. VanderWeele TJ and Ding P. Sensitivity analysis in observational research: introducing the E-value introducing the E-value. *Ann Intern Med* 2017; **167**: 268–274, https://doi.org/10.7326/M16-2607
8. Carnegie NB, Harada M and Hill JL. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J Res Educ Effectiveness* 2016; **9**: 395–420, https://doi.org/10.1080/19345747.2015.1078862
9. Talbot D and Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *Eur J Epidemiol* 2019; **34**: 725–730. Retrieved on 4 March 2021 from https://doi.org/10.1007/s10654-019-00529-y
10. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv*:**12033503**. 2012.
11. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol* 2011; **174**: 1223–1227.
12. Middleton JA, Scott MA, Diakow R, et al. Bias amplification and bias unmasking. *Political Analys* 2016; **24**: 307–323.
13. Wooldridge JM. Should instrumental variables be used as matching variables? *Res Econom* 2016; **70**: 232–237. Retrieved on 4 March 2021 from http://www.sciencedirect.com/science/article/pii/S1090944315301678
14. Ding P, Vanderweele T and Robins J. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* 2017; **104**: 291–302.
15. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol* 2011; **174**: 1213–1222. Retrieved on 4 March 2021 from http://dx.doi.org/10.1093/aje/kwr364
16. Velleman PF and Welsch RE. Efficient computing of regression diagnostics. *Am Stat* 1981; **35**: 234–242.
17. Chernozhukov V, Chetverikov D, Demirer M, et al. *Double/debiased machine learning for treatment and structural parameters*. Oxford, UK: Oxford University Press, 2018.
18. Robinson PM. Root-N-consistent semiparametric regression. *Econometrica: Econometric* 1988; 931–954.
19. Brockwell PJ, Davis RA and Fienberg SE. *Time series: theory and methods: theory and methods*. Berlin, Germany: Springer Science & Business Media, 1991.
20. Bowsher CG and Swain PS. Identifying sources of variation and the flow of information in biochemical networks. *Proc Natl Acad Sci* 2012; **109**: E1320–E1328. Retrieved on 4 March 2021 from https://www.pnas.org/content/109/20/E1320
21. Vaughan LK, Divers J, Padilla MA, et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Computat Statistics Data Analys* 2009; **53**: 1755–1766.
22. Cattell RB and Jaspers J. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavior Res Monograph* 1967; **67**: 1–212.
23. Helle C, Hillesund ER, Wills AK, et al. Replication data for examining the effects of an eHealth intervention from infant age 6 to 12 months on child eating behaviors and maternal feeding practices one year after cessation: the Norwegian randomized controlled trial Early Food for Future Health. *Data Set* 2019. Retrieved on 4 March 2021 from https://doi.org/10.18710/R2KJHK
24. Davidson R and MacKinnon JG. *Econometric theory and methods*. vol. **5**. New York, NY: Oxford University Press, 2004.
25. Chung KL. *A course in probability theory*. Cambridge, MA: Academic Press, 2001.
26. Whittaker J. *Graphical models in applied multivariate statistics*. Hoboken, NJ: Wiley Publishing, 2009.
27. Artner R, Wellingerhof PP, Lafit G, et al. The shape of partial correlation matrices. *Commun Stat-Theory Meth* 2020. DOI: 10.1080/03610926.2020.1811338.
28. Budden M, Hadavas P, Hoffman L. On the generation of correlation matrices. *Appl Math E-Notes* 2008; **8**: 279–282.
29. Hardin J, Garcia SR and Golan D. A method for generating realistic correlation matrices. *Ann Appl Stat* 2013; **7**: 1733–1762.
30. Vander Weele TJ and Shpitser I. A new criterion for confounder selection. *Biometrics* 2011; **67**: 1406–1413.
31. Vansteelandt S, Bekaert M and Claeskens G. On model selection and model misspecification in causal inference. *Stat Meth Med Res* 2012; **21**: 7–30.
32. Vander Weele TJ. Principles of confounder selection. *Eur J Epidemiol* 2019; **34**: 211–219. Retrieved on 4 March 2021 from https://doi.org/10.1007/s10654-019-00494-6
33. Helle C, Hillesund E, Omholt M, et al. Early food for future health: A randomized controlled trial evaluating the effect of an eHealth intervention aiming to promote healthy food habits from early childhood. *BMC Public Health* 2017; **17**: 1–12.

34. Helle C, Hillesund ER, Wills AK, et al. Examining the effects of an eHealth intervention from infant age 6 to 12 months on child eating behaviors and maternal feeding practices one year after cessation: The Norwegian randomized controlled trial Early Food for Future Health. *PloS One* 2019; **14**: e0220437.
35. Jiang L, Oualkacha K, Didelez V, et al. Constrained instruments and their application to Mendelian randomization with pleiotropy. *Genetic Epidemiol* 2019; **43**: 373–401.
36. Allen M, Poggiali D, Whitaker K, et al. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res* 2019; **4**.
37. Wickham H, Chang W and Wickham MH. GGplot2 package create elegant data visualisations using the grammar of graphics version. ■ 2016; **2**: 1–189.

Appendix 1

A.1 Matrix notation and FWL theorem

Throughout this paper, we make use of matrix notation to concisely represent estimates and as a way of considering the geometry of the least squares. Here is a quick guide for understanding the notation in this paper.

Let $\boldsymbol{A}$ be the $n \times p$ matrix of treatment variables. For illustrative purposes, consider that A is a single binary $n \times 1$ vector.

$$\boldsymbol{A_{n\times 1}} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

There are n rows of data, each with a 1 or 0 representing the observation being treated or not.

Another piece of notation that is used is annihilator and orthogonal projection matrices. Let $\boldsymbol{P_X}$ be the orthogonal projection matrix of $\boldsymbol{X}$, an $n \times k$ matrix, and $\boldsymbol{M_X}$ the annihilator or residual-making matrix of $\boldsymbol{X}$

$$\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T};$$

$$\boldsymbol{M_X} = \boldsymbol{I} - \boldsymbol{P_X} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}$$

A projection matrix maps each point to the nearest point in the subspace spanned by the columns in $\boldsymbol{X}$, $S(\boldsymbol{X})$. The annihilator matrix maps each point to the orthogonal complement of $S(\boldsymbol{X})$, $S^{\perp}(\boldsymbol{X})$. The predicted outcome in ordinary least squares is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} = \boldsymbol{P_X Y}$, which we can think of geometrically as "dropping a perpendicular"[24] from the outcome vector into the subspace spanned by the covariates in the regression. The orthogonal complement to the space spanned by the regressors $S^{\perp}(\boldsymbol{X})$ is where the fitted residual vector, $\widehat{\epsilon}$, lives. We can see that the residual vector $\widehat{\epsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{P_X Y} = (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y} = \boldsymbol{M_X Y}$ is just the projection of Y into the subspace orthogonal to $S(\boldsymbol{X})$.

By definition, we can always then decompose $\boldsymbol{Y}$ uniquely into its projection onto $S(\boldsymbol{X})$ and $S^{\perp}(\boldsymbol{X})$

$$\boldsymbol{Y} = \boldsymbol{P_X Y} + \boldsymbol{M_X Y}$$

Orthogonal projection matrices have two important properties, they are symmetric and idempotent. This means that $\boldsymbol{P_X} = \boldsymbol{P_X^T}$ and $\boldsymbol{P_X P_X} = \boldsymbol{P_X}$ and that these same two properties are equally enjoyed by $\boldsymbol{M_X}$. Further, any matrix in the subspace spanned by $\boldsymbol{X}$ is *annhilated* when operated on by $\boldsymbol{M_X}$, since it is by definition orthogonal to $S(\boldsymbol{X})$.

We also appeal to the Frisch-Waugh-Lovell (FWL) theorem to construct the matrix notation regression estimates as well as for visualizing the 2 dimensional plot of a single regression in the context of a multivariable regression. Suppose we construct an arbitrary partition of $\boldsymbol{X_{n\times k}} = [\boldsymbol{X_{1_{n\times k_1}}}, \boldsymbol{X_{2_{n\times k_2}}}]$, where $k_1 + k_2 = k$. The FWL

theorem states that the following two regressions, equations (25) and (27), produce numerically equivalent estimates of the vector $\widehat{\boldsymbol{\beta}_2}$ as well as numerically equivalent residuals.

$$\boldsymbol{Y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \epsilon \tag{25}$$

$$\boldsymbol{M_{X_1}Y} = \boldsymbol{M_{X_1}X_1}\boldsymbol{\beta_1} + \boldsymbol{M_{X_1}X_2}\boldsymbol{\beta_2} + \boldsymbol{M_{X_1}}\epsilon \tag{26}$$

$$= \boldsymbol{M_{X_1}X_2}\boldsymbol{\beta_2} + \boldsymbol{M_{X_1}}\epsilon \tag{27}$$

What this says in words is that it is numerically equivalent to regress $\boldsymbol{Y}$ on the columns of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ simultaneously as it is to first regress both $\boldsymbol{Y}$ and $\boldsymbol{X_2}$ on the columns of $\boldsymbol{X_1}$ separately, then save the respective residuals, $\boldsymbol{M_{X_1}Y}$ and $\boldsymbol{M_{X_1}X_2}$, and regress the former on the later. By simply pre-multiplying both sides of equation (27) by $\boldsymbol{X_2^T}$ and rearranging, we get the general matrix notation formulation for the vector of coefficient estimates $\widehat{\boldsymbol{\beta}_2} = (\boldsymbol{X_2M_{X_1}X_2})^{-1}\boldsymbol{X_2M_{X_1}Y}$. Using the idempotency and symmetry properties, we can rewrite the coefficient estimate

$$\widehat{\boldsymbol{\beta}_2} = (\boldsymbol{X_2^TM_{X_1}X_2})^{-1}\boldsymbol{X_2^TM_{X_1}Y} \tag{28}$$

$$= \frac{(\boldsymbol{M_{X_1}X_2})^{\boldsymbol{T}}(\boldsymbol{M_{X_1}Y})}{(\boldsymbol{M_{X_1}X_2})^{\boldsymbol{T}}(\boldsymbol{M_{X_1}X_2})} \tag{29}$$

$$= \frac{\widehat{\boldsymbol{\upsilon}}_{\boldsymbol{x_2}}\widehat{\boldsymbol{\upsilon}}_{\boldsymbol{y}}}{\widehat{\boldsymbol{\upsilon}}_{\boldsymbol{x_2}}^{\boldsymbol{2}}} \tag{30}$$

where $\widehat{\upsilon}_{x_2}$ and $\widehat{\upsilon}_y$ are the residuals from the regression of $\boldsymbol{X_2}$ and $\boldsymbol{Y}$ on $\boldsymbol{X_1}$, respectively.

A special case of the above result is when we have a $n \times k_1$ matrix of treatment variables, $\boldsymbol{A}$ and a $n \times k_2$ matrix of controlling variables. For example, we are trying to estimate the causal effect of the matrix $\boldsymbol{A}$ on the outcome $\boldsymbol{Y}$ using a selection on observables strategy by conditioning on $\boldsymbol{Z}$. The FWL theorem tells us that the estimates of the causal effect, $\widehat{\boldsymbol{\beta}_a}$ can be obtained by the two following regression equations

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta}_a + \boldsymbol{Z}\boldsymbol{\beta}_z + \boldsymbol{\upsilon_1}$$

$$\boldsymbol{M_ZY} = \boldsymbol{M_ZY} + \boldsymbol{\upsilon_1}$$

The second regression is a simple linear regression, with only one dependent variable $\boldsymbol{M_ZY}$ and a single regressor, $\boldsymbol{M_ZA}$. The error term remains unchanged by the projection into $S^{\perp}(\boldsymbol{Z})$ since it can be represented as $\boldsymbol{M_{A,Z}Y}$ which is already contained in the subspace $S^{\perp}(\boldsymbol{Z})$. Another way we can write the estimate $\widehat{\boldsymbol{\beta}_a}$ is to apply the well-known $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^TY}$ to the second regression equation above.

$$\widehat{\boldsymbol{\beta}_a} = \frac{(\boldsymbol{M_ZA})^{\boldsymbol{T}}(\boldsymbol{M_zY})}{(\boldsymbol{M_ZA})^{\boldsymbol{T}}\boldsymbol{M_ZA}}$$

$$= \frac{\frac{\boldsymbol{1}}{\boldsymbol{n}}(\boldsymbol{M_ZA})^{\boldsymbol{T}}(\boldsymbol{M_zY})}{\frac{\boldsymbol{1}}{\boldsymbol{n}}(\boldsymbol{M_ZA})^{\boldsymbol{T}}\boldsymbol{M_ZA}}$$

The numerator $\frac{1}{n}(\boldsymbol{M_ZA})^{\boldsymbol{T}}(\boldsymbol{M_zY})$ can be seen as the dot product of the residuals from the regression of the treatment on the control variables, $\boldsymbol{A}\sim\boldsymbol{Z}$, and the residuals from the regression of the outcome on the control variables, $\boldsymbol{Y}\sim\boldsymbol{Z}$, scaled by $\frac{1}{n}$. If a column of ones is included in the matrix $\boldsymbol{Z}$, both sets of residuals will be centered. We can then think of the dot product in the numerator as an estimator for the covariance of the two residuals,

$COV(\epsilon_a, \epsilon_y)$. In general terms, we will have bias due to unmeasured confounding if the covariance is a function of $\boldsymbol{U}$. The denominator can be seen as numerically equal to the sum of squared residuals.

An important special case of the annhilator matrix is $\boldsymbol{M_1}$, where $\boldsymbol{1}$ is a $n \times 1$ vector of ones. This is sometimes called the centering matrix because it de-means the matrix it operates on, since $\boldsymbol{M_1 X} = \boldsymbol{X} - \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T\boldsymbol{X} = \boldsymbol{X} - \frac{1}{n}\sum_{i=1}^{n} x_i = \boldsymbol{X} - \bar{\boldsymbol{X}}$

Using the symmetry and idempotency properties combined with the convergence in probability properties discussed in the next section, this implies

$$\frac{1}{n}\boldsymbol{X_1 M_1 X_2} = \frac{1}{n}(\boldsymbol{X_1} - \bar{\boldsymbol{X_1}})(\boldsymbol{X_2} - \bar{\boldsymbol{X_2}}) \tag{31}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{1_i} - \bar{\boldsymbol{X_1}})(x_{2_i} - \bar{\boldsymbol{X_2}}) \tag{32}$$

$$\xrightarrow{p} E[(\boldsymbol{X_1} - E[\boldsymbol{X_1}])(\boldsymbol{X_2} - E[\boldsymbol{X_2}]) \tag{33}$$

$$= COV(\boldsymbol{X_1}, \boldsymbol{X_2}) \tag{34}$$

When $\boldsymbol{X_1} = \boldsymbol{X_2}$, the last line becomes $\boldsymbol{Var}(\boldsymbol{X_1})$.

For a more complete and technical treatment of projection and annihilation matrices as it pertains to OLS, see *Econometric theory and methods* by Davidson and MacKinnon.[24]

A.2 Convergence in probability

Throughout the paper, we use the notation $\text{plim}_{\text{n}\to\infty}$ to mean the limit in probability. Specifically, if $\text{plim}_{\text{n}\to\infty}\boldsymbol{Y_n}=\boldsymbol{Y}$, then

$$lim_{n\to\infty} P(|\boldsymbol{Y_n} - \boldsymbol{Y}| > \epsilon) = 0, \forall \epsilon > 0$$

Alternatively we can write $\text{plim}_{\text{n}\to\infty}\boldsymbol{Y_n}=\boldsymbol{Y}$ as $\boldsymbol{Y_n}\xrightarrow{p}\boldsymbol{Y}$ or simply plim. Throughout this paper, all probability limits are as $n \to \infty$.

Below are a few important properties of Probability limits used throughout the paper. Suppose $\boldsymbol{X_n}\xrightarrow{p}\boldsymbol{X}$ and $\boldsymbol{Y_n}\xrightarrow{p}\boldsymbol{Y}$, then

$$\boldsymbol{X_n} + \boldsymbol{Y_n}\xrightarrow{p}\boldsymbol{X} + \boldsymbol{Y} \tag{35}$$

$$\boldsymbol{X_n Y_n}\xrightarrow{p}\boldsymbol{XY} \tag{36}$$

$$\frac{\boldsymbol{X_n}}{\boldsymbol{Y_n}} \xrightarrow{p} \frac{\boldsymbol{X}}{\boldsymbol{Y}} \tag{37}$$

where the third line is just a special case of the second and holds whenever the denominator is well defined. These properties are well known and follow from the Continuous Mapping Theorem.

Another useful theorem we use in this paper is the Weak Law of Large Numbers (WLLN). Here we consider a set of standard assumptions. Suppose we take the sample average of random variables $\{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}\}$, such that $\boldsymbol{X_i}$'s are independent and identically distributed (iid) and $E[\boldsymbol{X_i}] = m < \infty$, i.e. the expectation is finite then

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}\xrightarrow{p}E[\boldsymbol{X_i}] = m. \tag{38}$$

In the paper, whenever specified we assume that the error terms are coming from a normal distribution. In light of the WLLN, we can see that normal error terms are not required for the results to hold, and that in fact we just need the error terms to come from an identical and independent distribution. The above results can be weakened further such that we can replace the iid condition with pairwise independence (see Chung[25] for details).

Further, some probability limit results do not always have a closed-form expression, for example the probability limit of a sum of iid cauchy random variables since they do not have finite expectations. Sometimes we express the resulting limit as a function of random variables. These random variables tend to their respective probability limits, provided they exist.

A.3 Derivations continued

Below are derivations, extensions, proofs, and alternate forms of the equations presented in the main text.

A.3.1 Average causal and average partial effects

Throughout this paper, we consider linear models with continuous exposures and as such a natural causal estimand of interest is the average partial effect (APE). Under the linearity assumptions, the APE coincides with the ACE. Below we show the derivation of the APE under the various DAG and structural equation assumptions. Implicitly, we further assume standard regularity conditions such as existence and boundedness of the estimators in $\mathcal{L}_1$, so that the derivative operator can freely move inside the expectation integral.

Average partial effects for equations (4) and (5)

$$\begin{aligned}\text{APEs} &= \frac{\partial E[\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{U},\boldsymbol{BAV}]}{\partial \boldsymbol{A}} \\ &= \frac{\partial E[\alpha_y + \boldsymbol{A\beta_a} + \boldsymbol{U\beta_u} + \boldsymbol{BAV\beta_{bav}} + \boldsymbol{\epsilon_1}|\boldsymbol{A},\boldsymbol{U},\boldsymbol{BAV}]}{\partial \boldsymbol{A}} = \beta_a\end{aligned}$$

When we allow for $\boldsymbol{BAV}$ to be a $n \times k$ vector and for $\boldsymbol{\beta_{BAV}}$ to be potentially a zero vector, we can see that the above derivation holds for all of the DAG's and structural equations which assume there is no interaction term.

$\widehat{\beta}_a^{naive}$ in equation (7)

$$\widehat{\beta}_a^{naive} = \frac{\boldsymbol{A^T M_z Y}}{\boldsymbol{A^T M_z A}} \tag{39}$$

$$= \beta_a + \beta_u \frac{\boldsymbol{A^T M_1 U}}{\boldsymbol{A^T M_1 A}} + \beta_{bav} \frac{\boldsymbol{A^T M_1 BAV}}{\boldsymbol{A^T M_1 A}} + \frac{\boldsymbol{A^T M_1 \epsilon_1}}{\boldsymbol{A^T M_1 A}} \tag{40}$$

$$= \beta_a + \beta_u \frac{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})^2} + \beta_{bav} \frac{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})(bav_i - \bar{bav})}{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})^2} + \frac{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})(\epsilon_{1_i} - \bar{\epsilon_1})}{\frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})^2} \tag{41}$$

A.4 Derivation of Pearl (2011) result

Here we will explicitly follow Pearl's derivation,[10] to show the advantages of considering the probability limit over strictly expectations. We will derive the expectation for $\widehat{\beta}_a^{bav'}$ from estimating (9), which is the APE conditional on the treatment, $\boldsymbol{A}$, and the $\boldsymbol{BAV}$ variable.

$$\begin{aligned}E[\widehat{\beta}_a^{bav'}] &= \frac{\partial E[\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}} \\ &= \frac{\partial E[\alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_{bav} + \boldsymbol{\epsilon_1}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}} \\ &= \beta_a + \beta_u \frac{\partial E[\boldsymbol{U}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}}\end{aligned}$$

We must find the expectation of $\boldsymbol{U}$ conditional on $\boldsymbol{A}$ and $\boldsymbol{BAV}$. Pearl solves this challenge by supposing the true underlying relationship between $\boldsymbol{U}$, $\boldsymbol{A}$, and $\boldsymbol{BAV}$ is linear and writing this functional form as a linear regression equation

$$\boldsymbol{U} = \alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav} + \epsilon_{\boldsymbol{3}}$$

Using this equation in addition to the two structural equations for $\boldsymbol{Y}$ and $\boldsymbol{A}$, respectively, we can express the regression coefficients, ζ_a and ζ_{bav} in terms of the structural coefficients β_a, β_u, γ_a, γ_u by equating expressions for the covariances under the two sets of structural equations.

$$\begin{aligned}
&COV(\boldsymbol{U},\boldsymbol{A}) = E[\boldsymbol{AU}] - E[\boldsymbol{A}]E[\boldsymbol{U}] \\
&= E[(\alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}})\boldsymbol{U}] - E[(\alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}})]E[\boldsymbol{U}] \\
&= \gamma_u(E[\boldsymbol{U^2}] - E[\boldsymbol{U}]^2) \\
&= \gamma_u\sigma_u^2
\end{aligned}$$

Equivalently

$$\begin{aligned}
&COV(\boldsymbol{U},\boldsymbol{A}) = E[\boldsymbol{AU}] - E[\boldsymbol{A}]E[\boldsymbol{U}] \\
&= E[\boldsymbol{A}(\alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav})] - E[\boldsymbol{A}]E[(\alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav})] \\
&= \zeta_a(E[\boldsymbol{A^2}] - E[\boldsymbol{A}]^2) + \zeta_{bav}(E[\boldsymbol{ABAV}] - E[\boldsymbol{A}]E[\boldsymbol{BAV}]) \\
&= \zeta_a\sigma_a^2 + \zeta_{bav}(COV(\boldsymbol{A},\boldsymbol{BAV})) \\
&= \zeta_a\sigma_a^2 + \zeta_{bav}\gamma_{bav}\sigma_{bav}^2
\end{aligned}$$

where the last line follows analogously from our derivation of $COV(\boldsymbol{A},\boldsymbol{U})$. Putting these together, we have

$$\zeta_a = \frac{\gamma_u\sigma_u^2 - \zeta_{bav}\gamma_{bav}\sigma_{bav}^2}{\sigma_a^2}. \tag{42}$$

Similarly, putting the two steps together for succinctness

$$COV(\boldsymbol{U},\boldsymbol{BAV}) = 0 \text{ (independence)} \tag{43}$$

$$= E[\boldsymbol{UBAV}] - E[\boldsymbol{U}]E[\boldsymbol{BAV}] \tag{44}$$

$$= E[(\alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav})\boldsymbol{BAV}] - E[(\alpha_u + \boldsymbol{A}\zeta_a + \boldsymbol{BAV}\zeta_{bav})]E[\boldsymbol{BAV}] \tag{45}$$

$$= \zeta_a COV(\boldsymbol{A},\boldsymbol{BAV}) + \zeta_{bav}Var(\boldsymbol{BAV}) \tag{46}$$

$$= \zeta_a\gamma_{bav}\sigma_{bav}^2 + \zeta_{bav}\sigma_{bav} \tag{47}$$

Now we have two equations for the two new regression coefficients in terms of the structural equations. Combining

$$\zeta_a = \frac{\gamma_u\sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2} \tag{48}$$

$$\zeta_{bav} = \frac{-\gamma_u\gamma_{bav}\sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2} \tag{49}$$

Returning to the task of finding the partial effect of $\boldsymbol{A}$ on $E[\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{BAV}]$ and thus $E[\widehat{\beta}_a^{bav\prime}]$.

$$E[\widehat{\beta}_a^{bav\prime}] = \frac{\partial E[\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}} \tag{50}$$

$$= \frac{\partial E[\alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_{bav} + \epsilon_{\boldsymbol{1}}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}}$$

$$= \beta_a + \frac{\partial \beta_u E[\boldsymbol{U}|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}}$$

$$= \beta_a + \beta_u \frac{\partial E[(\alpha_u + \zeta_a \boldsymbol{A} + \zeta_{bav} + \epsilon_{\boldsymbol{1}})|\boldsymbol{A},\boldsymbol{BAV}]}{\partial \boldsymbol{A}}$$

$$= \beta_a + \beta_u \zeta_a$$

$$= \beta_a + \beta_u \frac{\gamma_u \sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2} \tag{51}$$

The approach of Pearl is limited in that it only works when the true underlying form of the conditional expectation, $E[\boldsymbol{U}|\boldsymbol{A},\boldsymbol{BAV}]$ is linear in both $\boldsymbol{A}$ and $\boldsymbol{BAV}$. As a result, the derivation is cumbersome and does not easily generalize to more complicated cases with more variables or different functional forms. Similarly, we can find the expectation for the naive estimator $\widehat{\beta}_a^{naive}$ from (6) using this method

$$E[\widehat{\beta}_a^{naive}] = \frac{\partial E[\boldsymbol{Y}|\boldsymbol{A}]}{\partial \boldsymbol{A}} \tag{52}$$

$$= \frac{\partial E[\alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_{bav} + \epsilon_{\boldsymbol{1}}|\boldsymbol{A}]}{\partial \boldsymbol{A}} \tag{53}$$

$$= \beta_a + \frac{\partial \beta_u E[\boldsymbol{U}|\boldsymbol{A}]}{\partial \boldsymbol{A}} + \frac{\partial \beta_{bav} E[\boldsymbol{BAV}|\boldsymbol{A}]}{\partial \boldsymbol{A}} \tag{54}$$

We assume that the true underlying relationship between $\boldsymbol{U}$ and $\boldsymbol{A}$ is linear, while also assuming a linear relationship between $\boldsymbol{BAV}$ and $\boldsymbol{A}$

$$\boldsymbol{U} = \alpha_u + \boldsymbol{A}\tau_a + \epsilon_{\boldsymbol{4}} \tag{55}$$

$$\boldsymbol{BAV} = \alpha_{bav} + \boldsymbol{A}\eta_a + \epsilon_{\boldsymbol{5}} \tag{56}$$

Using these two equations (55) and (56), we arrive at the following expressions for the $COV(\boldsymbol{A},\boldsymbol{U})$ and $COV(\boldsymbol{A},\boldsymbol{BAV})$

$$COV(\boldsymbol{A},\boldsymbol{U}) = \tau_a \sigma_a^2 \tag{57}$$

$$COV(\boldsymbol{A},\boldsymbol{BAV}) = \eta_a \sigma_a^2 \tag{58}$$

Following an analogous process, we can show that $COV(\boldsymbol{A},\boldsymbol{BAV}) = \gamma_{bav}\sigma_{bav}^2$ in terms of the original structural coefficients. Combining the four covariance expressions and solving for τ_a and η_a in terms of the structural equations yields

$$\tau_a = \frac{\gamma_u \sigma_u^2}{\sigma_a^2} \tag{59}$$

$$\eta_a = \frac{\gamma_{bav}\sigma^2_{bav}}{\sigma^2_a} \tag{60}$$

Substituting equations (55) and (56) in equation (54) and using equations (59) and (60) yields the following expectation for $\widehat{\beta}_a^{naive}$:

$$E[\widehat{\beta}_a^{naive}] = \beta_a + \beta_u \frac{\gamma_u \sigma^2_u}{\sigma^2_a} + \beta_{bav} \frac{\gamma_{bav}\sigma^2_{bav}}{\sigma^2_a} \tag{61}$$

Note in this derivation we needed to assume two linear relationships in order to derive the expectation, $E[\boldsymbol{U}|\boldsymbol{A}]$ and $E[\boldsymbol{BAV}|\boldsymbol{A}]$. These assumptions are not necessary when using probability limits to define limiting expressions in terms of the structural parameters.

A.5 Correlations, partial correlations, $\mathcal{R}^2$ and dimensionless quantities for sensitivity analysis and simulation

In section 3 we discussed the quantity $\eta = \frac{|\beta_{bav}\gamma_{bav}\sigma^2_{bav}|}{|\beta_u\gamma_u\sigma^2_u|}$ and mentioned that it is possible to express as a dimensionless quantity which is useful for sensitivity analysis and simulation studies.

$$\eta = \frac{|\beta_{bav}\gamma_{bav}\sigma^2_{bav}|}{|\beta_u\gamma_u\sigma^2_u|} \tag{62}$$

$$= \frac{\rho_{Y,BAV|A,U}\rho_{A,BAV|U}}{\rho_{U,Y|A,BAV}\rho_{A,U|BAV}} \frac{(1-\mathcal{R}^2_{Y|A,BAV})^{\frac{1}{2}}}{(1-\mathcal{R}^2_{Y|A,U})^{\frac{1}{2}}} \frac{(1-\mathcal{R}^2_{U|A,BAV})^{\frac{1}{2}}}{(1-\mathcal{R}^2_{BAV|A,U})^{\frac{1}{2}}} \frac{(1-\mathcal{R}^2_{A|U})^{\frac{1}{2}}}{(1-\mathcal{R}^2_{A|BAV})^{\frac{1}{2}}} \tag{63}$$

where a partial correlation is defined as the following for general variables Z_1, Z_2 and p dimensional X

$$\rho_{Z_1,Z_2|X} = \frac{E[\widehat{\epsilon_1}\widehat{\epsilon_2}]}{\sqrt{E[\widehat{\epsilon_1}^2]E[\widehat{\epsilon_2}^2]}} \tag{64}$$

and $\widehat{\epsilon_i} = \text{argmin}_{(\beta_0,\beta)\in\mathcal{R}^{(p+1)}} (Z_i - \beta_0 - X\beta)^2$, for $i \in \{1,2\}$. Partial correlations can be expressed as lower correlations by the following recursion formula.[26]

Consider the partition of $X_{n\times p} = [X_{1n\times(p-1)}, X_{2n\times 1}]$.

$$\rho_{Z_1,Z_2|X_1,X_2} = \frac{\rho_{Z_1,Z_2|X_1} - \rho_{Z_1,X_2|X_1}\rho_{X_2,Z_2|X_1}}{\sqrt{(1-\rho^2_{Z_1,X_2|X_1})(1-\rho^2_{X_2,Z_2|X_1})}} \tag{65}$$

Further, it can be shown that a non-singular partial correlation matrix uniquely determines the correlation matrix and vice versa.[27] To directly express the partial correlations in terms of the correlation matrix, the following formula is convenient, where $\tilde{R}$ is a partial correlation matrix, where element $\tilde{r}_{i,j} = \rho_{X_i,X_j|X\{X_i,X_j}$ is the partial correlation of two variables X_i, X_j partialling out all other variables. Analogously, let R be the correlation matrix, where elements correspond to the usual correlations.[27]

$$\tilde{R} = -D^{-1}_{R^{-1}} R^{-1} D^{-1}_{R^{-1}} \tag{66}$$

where $D^{-1}_{R^{-1}}$ is a diagonal matrix, where the elements of the diagonal equal the vector $\sqrt{diag(R^{-1})}$. Equivalently we can construct the correlation matrix from a specified partial correlation matrix.

$$R = D^{-1}_{-\tilde{R}^{-1}}(-\tilde{R})^{-1} D^{-1}_{-\tilde{R}^{-1}} \tag{67}$$

For example, using this relation, or iterative application of the recursion relation we can show that

$$\rho_{Y,BAV|A,U} = \frac{\left\{ \begin{array}{c} \rho_{A,U}\rho_{U,BAV}\rho_{Y,BAV} + \rho_{A,BAV}\rho_{U,BAV}\rho_{Y,U} - \rho_{A,U}\rho_{Y,U} \\ -\rho_{A,BAV}\rho_{Y,BAV} + \rho^2_{U,BAV}\rho_{Y,U} + \rho_{Y,A} - \rho^2_{U,BAV} \end{array} \right\}}{\sqrt{\left\{ \begin{array}{c} (1-\rho^2_{A,U} - \rho^2_{A,BAV} - \rho^2_{U,BAV} + 2\rho_{A,U}\rho_{A,BAV}\rho_{U,BAV})\times \\ (1-\rho^2_{Y,BAV} - \rho^2_{Y,U} - \rho^2_{U,BAV} + 2\rho_{U,BAV}\rho_{Y,U}\rho_{Y,BAV}) \end{array} \right\}}} \tag{68}$$

in the case the BAV is a single variable. When U and BAV are independent as is the case in the DAG considered in section 3, this simplifies

$$\rho_{Y,BAV|A,U} = \frac{\rho_{Y,A} - \rho_{A,U}\rho_{Y,U} - \rho_{A,BAV}\rho_{Y,BAV}}{\sqrt{(1-\rho^2_{A,U} - \rho^2_{A,BAV})(1-\rho^2_{Y,BAV} - \rho^2_{Y,U})}} \tag{70}$$

Similarly, when U and BAV are independent we have the following

$$\rho_{A,U|BAV} = \frac{\rho_{A,U}}{1-\rho^2_{A,BAV}} \tag{72}$$

$$\rho_{A,BAV|U} = \frac{\rho_{A,BAV}}{1-\rho^2_{A,U}} \tag{73}$$

On the other hand, $\rho_{U,Y|A,BAV}$ does not simplify nicely and is easier to handle numerically with the aid of the matrix relation rather than symbolically.

Two things of note. First, correlations and the variances of the variables determine the entire expression η. We showed this explicitly with the partial correlations, but it is also true that it is sufficient to fix a correlation matrix and the variances of the variables to determine the $\mathcal{R}^2$ values. It is sufficient to consider the space of valid correlation matrices, or equivalently a valid partial correlation matrix, and fix the variances of U and BAV (The variances for Y and A cancel) in order to simulate the entire parameter space for η. Moreover, if we have a real data set, many of the correlations are estimable and this greatly restricts the parameter space.

$$R = \begin{vmatrix} 1 & \boldsymbol{\rho_{Y,A}} & \boldsymbol{\rho_{Y,BAV}} & \rho_{Y,U} \\ \boldsymbol{\rho_{Y,A}} & 1 & \boldsymbol{\rho_{A,BAV}} & \rho_{A,U} \\ \boldsymbol{\rho_{Y,BAV}} & \boldsymbol{\rho_{A,BAV}} & 1 & \rho_{BAV,U} \\ \rho_{Y,U} & \rho_{A,U} & \rho_{BAV,U} & 1 \end{vmatrix}$$

The bolded correlations are strictly functions of observable and can be estimated from the data. By fixing the known values, or only considering variables within an estimated range, we can heavily restrict the space of plausible correlation matrices and thus the range of possible values of η. Similarly, we can estimate σ^2_{BAV} from the data.

Notice that the estimable correlations form a valid correlation matrix themselves, call this $R_{Y,A,BAV}$. Assuming the treatment and outcome are univariate, $R_{Y,A,BAV}$ is of dimension $(2+p)\times(2+p)$ where p is the dimension of the control variables p. It is well known how to estimate $\widehat{R}_{Y,A,BAV}$, and in order to do a proper simulation or sensitivity analysis it suffices to extend this correlation matrix to a valid $(2+p+1)\times(2+p+1)$. There are existing algorithms that can be implemented to extend $(n-1)$ dimension correlation matrices to n dimension correlation matrices[28] which can be used instead of solving parameter inequalities. Further, there is work looking at how to simulate realistic correlation matrices around a given structure[29] and finding ways to add noise to correlation matrices while ensuring the result is still positive definite. Such methods could allow for researchers to better represent the underlying uncertainty associated with the estimated matrix $\widehat{R}_{Y,A,BAV}$.

A.6 Partial correlations and regression coefficients

Here we show the relationship between partial correlations and regression coefficients. Like in the text we distinguish between true structural coefficients β_x and estimates from various regressions. First let us consider an estimate.

$$\widehat{\beta}_x = \underset{\beta \in \mathcal{R}^p}{\operatorname{argmin}} \, (Y - X\beta)^2 \tag{74}$$

$$\widehat{\beta}_{x_i} = \frac{\widehat{\epsilon}_{y \setminus \{x_i\}} \widehat{\epsilon}_{x \setminus \{x_i\}}}{\widehat{\epsilon}^2_{x \setminus \{x_i\}}} \tag{75}$$

where $\widehat{\epsilon}_{y \setminus \{x_i\}} = \operatorname{argmin}_{\beta \in \mathcal{R}^{p-1}} (Y - (X \setminus \{X_i\})\beta)^2$ and $\widehat{\epsilon}_{x \setminus \{x_i\}} = \operatorname{argmin}_{\beta \in \mathcal{R}^{p-1}} (X_i - (X \setminus \{X_i\})\beta)^2$ are the vector of residuals from the regression of Y on X excluding the variable X_i and the vector of residuals from the regression of X_i on X excluding X_i, respectively. This follows directly from the FWL theorem as discussed in section A.1.

$$\widehat{\beta}_{x_i} \xrightarrow{p} \frac{E[\widehat{\epsilon}_{y \setminus \{x_i\}} \widehat{\epsilon}_{x \setminus \{x_i\}}]}{E[\widehat{\epsilon}^2_{x \setminus \{x_i\}}]} \tag{76}$$

$$= \frac{\rho_{Y, X_i | X \setminus \{X_i\}} \sqrt{E[\widehat{\epsilon}^2_{y \setminus \{x_i\}}]}}{\sqrt{E[\widehat{\epsilon}^2_{x_i \setminus \{x_i\}}]}} \tag{77}$$

$$= \frac{\rho_{Y, X_i | X \setminus \{X_i\}} (1 - \mathcal{R}^2_{Y | X \setminus \{X_i\}})^{\frac{1}{2}} \sigma_y}{(1 - \mathcal{R}^2_{X | X \setminus \{X_i\}})^{\frac{1}{2}} \sigma_{x_i}} \tag{78}$$

So here we see that in the probability limit we can express any regression coefficient estimator in terms of population partial correlations and limits of coefficients of determinations. The sample regression coefficient can be expressed in terms of the appropriate sample statistics.

Similarly, if we have a data generating process that can be expressed as a linear function in parameters over the parameter space $\beta \in \mathcal{R}^p$, the we can similarly express the true regression coefficients in terms of population partial correlations and coefficients of determination. Let

$$Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon \tag{79}$$

be the true data generating process. As in the text we assume that $E[\epsilon | X] = 0$. Here we suppose that $[X_1, X_2]$ is an arbitrary partition of $X_{n \times p}$ such that X_1 is a vector and X_2 is $(p-1) \times n$ dimensional.

$$\beta_1 = \frac{\rho_{Y, X_1 | X_2} (1 - \mathcal{R}^2_{Y | X_2})^{\frac{1}{2}} \sigma_y}{(1 - \mathcal{R}^2_{X_1 | X_2})^{\frac{1}{2}} \sigma_{x_1}} \tag{80}$$

Notice that this does not depend in any way on the true data process for X_1.

If we think about X_1 as our treatment of interest, then we can see, using the tools from above, that we can express the treatment effect estimator $\widehat{\beta}_{x_1}$ directly in terms of partial correlation and coefficients of determinations. This could be further decomposed into a function of only pairwise correlations and variances, since the coefficients of determination can themselves be decomposed into further functions of correlations and variances. This tells us that a known covariance matrix is sufficient to completely determine any OLS estimator. As we say in section 3, we were able to express the relative bias of the naive and adjusted estimators entirely in terms of correlations and variances. This can be extremely useful for sensitivity analysis or for simulating estimators. As mentioned in the section above with a real data set, much of the matrix can be estimated and the structure

of admissible covariance and correlation matrices is well known and there exist many algorithms to simulate them and even priors such as the LKJ prior which can be placed over them. Correlation matrices especially may be easy to elicit domain knowledge which may further inform or restrict potential simulation studies or sensitivity analysis.

If the true outcome generating process is conditionally linear, we can similarly decompose and understand the true regression coefficient and this is independent of any assumptions on the treatment assignment mechanism and does not require any assumptions on independence between the variables. Of course, if the DAG or partial DAG is known and it implies independencies or conditional independences, this can help fill unknown partial correlations, correlations, or coefficients of determination. The ability to be able to easily convert correlations to partial correlations is particularly useful with DAGs which imply conditional independence, where there may be obvious restrictions on partial correlations which are harder to discern in terms of pairwise correlations. Here, the way that we model variables other than the outcome may be more important however, as partial correlations are defined in terms of linear projections which is only equivalent to conditioning on variables when the conditional expectations are linear, a special case of which is when the variables are multivariable normal. Here, careful asymptotic arguments may still be used at times to simplify the partial correlation structures in some non-linear or non-multivariate normal cases.

A.7 Probability limit calculations

Now we will show the generality of probability limits for generating meaningful expressions of estimator behavior and again we will use $\widehat{\beta}_a^{bav\prime}$ from equation (9), where $\boldsymbol{Z} = [\boldsymbol{1}, \boldsymbol{BAV}]$.

From the FWL theorem, $\widehat{\beta}_a^{bav\prime} = \frac{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{Y}}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}}$.

$$\widehat{\beta}_a^{bav\prime} = \frac{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{Y}}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} \tag{81}$$

$$= \frac{\boldsymbol{A}^T\boldsymbol{M}_z(\alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_{bav} + \epsilon_{\boldsymbol{1}})}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} \tag{82}$$

$$= \beta_a + \frac{\boldsymbol{A}^T\boldsymbol{M}_z(\boldsymbol{U}\beta_u + \epsilon_{\boldsymbol{1}})}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} \tag{83}$$

$$= \beta_a + \beta_u\frac{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{U}}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} + \frac{\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}}}{\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} \tag{84}$$

$$= \beta_a + \beta_u\frac{\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{U}}{\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} + \frac{\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}}}{\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}} \tag{85}$$

which follows by simply substituting in the true structural equations for $\boldsymbol{Y}$ and $\boldsymbol{A}$, 4 and 5, respectively, and then applying the annihilating properties of $\boldsymbol{M}_z$ to set linear combinations of constants and $\boldsymbol{BAV}$ to 0. Notice that we have not used any information about structural equation for the treatment. Thus the numerical form in equation (85) holds for any treatment structural equation, $\boldsymbol{A} = f(\boldsymbol{U}, \boldsymbol{BAV}, \epsilon_{\boldsymbol{2}})$. Further, since this is written in general matrix notation, $\boldsymbol{U}$ and $\boldsymbol{BAV}$ can be extended to be arbitrary p_1 and p_2 dimensional variables.

We can now solve for the probability limits of the three remaining expressions separately ($\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{A}$, $\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\boldsymbol{U}$, and $\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}}$) and combine them due to the properties of probabilities limits, namely equations (35), (36), and (37). We begin with deriving $\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}}$),

$$\frac{1}{n}\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}} = \xrightarrow{p} E[\boldsymbol{A}^T\boldsymbol{M}_z\epsilon_{\boldsymbol{2}}] \tag{86}$$

$$= E[\boldsymbol{A}^T\boldsymbol{M}_z E[\epsilon_{\boldsymbol{2}}|\boldsymbol{A}, \boldsymbol{Z}]] \tag{87}$$

$$= E[\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_z E[\epsilon_{\boldsymbol{2}}]] \tag{88}$$

$$= 0 \tag{89}$$

where the first line follows from the WLLN, line two from the Law of Iterated Expectations, and the third from the independence of $\epsilon_{\boldsymbol{2}}$ from $\boldsymbol{A}$ and $\boldsymbol{BAV}$. Next we consider $\frac{1}{n}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_z\boldsymbol{A}$

$$\frac{1}{n}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_z\boldsymbol{A} = \frac{1}{n}(\alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}})^T\boldsymbol{M}_z(\alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}}) \tag{90}$$

$$= \frac{1}{n}(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}})^T\boldsymbol{M}_z(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}}) \tag{91}$$

$$\xrightarrow{p} \operatorname{plim}\frac{1}{\mathrm{n}}(\boldsymbol{U}\gamma_{\mathrm{u}}+\epsilon_{\boldsymbol{2}})^{\mathrm{T}}\mathrm{M}_{\boldsymbol{1}}(\boldsymbol{U}\gamma_{\mathrm{u}}+\epsilon_{\boldsymbol{2}}) \tag{92}$$

$$= \operatorname{plim}\frac{1}{\mathrm{n}}(\boldsymbol{U}\gamma_{\mathrm{u}})^{\mathrm{T}}\mathrm{M}_{\boldsymbol{1}}(\boldsymbol{U}\gamma_{\mathrm{u}})+\operatorname{plim}\frac{1}{\mathrm{n}}2(\boldsymbol{U}\gamma_{\mathrm{u}})^{\mathrm{T}}\boldsymbol{M}_{\boldsymbol{1}}\epsilon_{\boldsymbol{2}}+\operatorname{plim}\frac{1}{\mathrm{n}}\epsilon_{\boldsymbol{2}}\boldsymbol{M}_{\boldsymbol{1}}\epsilon_{\boldsymbol{2}} \tag{93}$$

$$= \gamma_u^2 E[(\boldsymbol{U} - \bar{\boldsymbol{U}})^T(\boldsymbol{U} - \bar{\boldsymbol{U}})] + 2\gamma_u E[\boldsymbol{U}E[\epsilon_{\boldsymbol{2}}|\boldsymbol{U}]] + E[\epsilon_{\boldsymbol{2}}{}^T\epsilon_{\boldsymbol{2}}] \tag{94}$$

$$= \gamma_u^2\sigma_u^2 + 0 + \sigma_{\epsilon_2}^2 \tag{95}$$

$$= \sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2 \tag{96}$$

where line 92 follows from the fact that BAV is independent of both U and $\epsilon_{\boldsymbol{2}}$. Since M_z is a residual making vector, we can compare the residuals in the probability limit from the following two regressions

$$(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}}) = \alpha_{\boldsymbol{U}\gamma_u+\epsilon_{\boldsymbol{2}}} + \boldsymbol{BAV}\eta_{bav} + \boldsymbol{\upsilon_1} \tag{97}$$

$$(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}}) = \alpha_{\boldsymbol{U}\gamma_u+\epsilon_{\boldsymbol{2}}} + \boldsymbol{\upsilon_2} \tag{98}$$

Due to independence, $\widehat{\eta_{bav}} \xrightarrow{p} 0$ and thus the residuals from the two regressions will be equivalent asymptotically. Thus we can replace $\boldsymbol{M_z}$ with $\boldsymbol{M_1}$ in equation (92), which as the centering projection matrix enjoys favorable properties as discussed in section A.1.

Finally we need to find the probability limit of $\frac{1}{n}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_z\boldsymbol{U}$.

$$\frac{1}{n}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_z\boldsymbol{U} = \frac{1}{n}(\alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \epsilon_{\boldsymbol{2}})^T\boldsymbol{M}_z\boldsymbol{U} \tag{99}$$

$$= \frac{1}{n}(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}})^T\boldsymbol{M}_z\boldsymbol{U} \tag{100}$$

$$\xrightarrow{p} \frac{1}{n}(\boldsymbol{U}\gamma_u + \epsilon_{\boldsymbol{2}})^T\boldsymbol{M_1}\boldsymbol{U} \tag{101}$$

$$= \frac{1}{n}(\gamma_u(\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{M_1}\boldsymbol{U}) + \gamma_u\boldsymbol{U}\boldsymbol{M_1}\epsilon_{\boldsymbol{2}} \tag{102}$$

$$\xrightarrow{p} \gamma_u E[(U - \bar{U})^2] + E[(U - \bar{U})(\in_2 - \bar{\in}_2)] \tag{103}$$

$$= \gamma_u\sigma_u^2 \tag{104}$$

Putting this altogether, this implies

$$\widehat{\beta}_a^{bavl} \xrightarrow{p} \beta_u \frac{\gamma_u \sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2} \tag{105}$$

This is equivalent to the expectation in this case. The benefit is that it is more robust to functional form assumptions and by using properties (35) to (37) and the FWL theorem, we can find asymptotic bias expressions by partitioning the estimator into a series of functions of residuals from simpler regressions. Further, we can always find the limiting expression for the numerator and the denominator separately. Expectations cannot be split up in such a manner and ratios of variables can be very difficult to find closed-form expressions for the expectation without imposing restrictive assumptions.

A.8 Probability limit derivations for section 4

Derivation of bias amplification limits of estimators in equation (21) under the structural equations in equations (19) and (20). Here we assume without loss of generality that $E[\boldsymbol{f_1}(\boldsymbol{U})] = 0$ and $E[\boldsymbol{f_2}(\boldsymbol{BAV})] = 0$, since the intercept can absorb the means of such function if they are non-zero.

$$\widehat{\beta}_a^{naive} = \frac{\boldsymbol{A^T M_\iota Y}}{\boldsymbol{A^T \iota}} \tag{106}$$

$$= \frac{\boldsymbol{A^T M_\iota}(\alpha_y + \boldsymbol{A}\beta_a + f_1(\boldsymbol{U}) + f_2(\boldsymbol{BAV}) + \epsilon_{\boldsymbol{I}})}{\boldsymbol{A^T \iota}} \tag{107}$$

$$= \beta_a + \frac{\boldsymbol{A^T M_\iota f_1}(\boldsymbol{U})}{\boldsymbol{A^T \iota}} + \frac{\boldsymbol{A^T M_\iota} f_2(\boldsymbol{BAV})}{\boldsymbol{A^T \iota}} + \frac{\boldsymbol{A^T M_\iota} \epsilon_{\boldsymbol{I}}}{\boldsymbol{A^T \iota}} \tag{108}$$

$$\xrightarrow{p} \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U}))}{\sigma_a^2} + \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))}{\sigma_a^2} + \frac{E[\boldsymbol{A}\epsilon_{\boldsymbol{I}}]}{\sigma_a^2} \tag{109}$$

$$= \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U}))}{\sigma_a^2} + \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))}{\sigma_a^2} + \frac{E[\boldsymbol{A}E[\epsilon_{\boldsymbol{I}}|\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{BAV}]]}{\sigma_a^2} \tag{110}$$

$$= \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U}))}{\sigma_a^2} + \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))}{\sigma_a^2} \tag{111}$$

Since for example

$$\frac{\boldsymbol{A^T M_\iota f_1}(\boldsymbol{U})}{\boldsymbol{A^T \iota}} = \frac{\frac{1}{n}\sum_{i=1}^n (A_i - \bar{A})(f_1(U)_i - f_1(\bar{U}))}{\frac{1}{n}\sum_{i=1}^n (A_i - \bar{A})^2} \tag{112}$$

$$\xrightarrow{p} \frac{E[(\boldsymbol{A} - E[A])(f_1(\boldsymbol{U}) - E[f_1(\boldsymbol{U})])]}{E[(\boldsymbol{A} - E[\boldsymbol{A}])^2]} \tag{113}$$

$$= \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U}))}{VAR(\boldsymbol{A})} \tag{114}$$

Where the probability limit follows from the weak law of large numbers and the continuous mapping theorem. For the adjusted estimator $\widehat{\beta}^{|f_2(BAV)}$, let $\boldsymbol{Z} = [\boldsymbol{\iota}, \boldsymbol{f_2}(\boldsymbol{BAV})]$.

$$\widehat{\beta}^{|f_2(BAV)} = \frac{\boldsymbol{A^T M_Z Y}}{\boldsymbol{A^T M_Z A}} \tag{115}$$

$$= \frac{\boldsymbol{A^T M_Z}(\alpha_y + \boldsymbol{A}\beta_a + f_1(\boldsymbol{U}) + f_2(\boldsymbol{BAV}) + \epsilon_{\boldsymbol{I}})}{\boldsymbol{A^T M_Z A}} \tag{116}$$

$$= \beta_a + \frac{\boldsymbol{A^T M_Z} f_1(\boldsymbol{U})}{\boldsymbol{A^T M_Z A}} + \frac{\boldsymbol{A^T M_Z}\epsilon_{\boldsymbol{I}}}{\boldsymbol{A^T M_Z A}} \tag{117}$$

$$= \beta_a + \frac{\boldsymbol{A^T M_Z}(f_1(\boldsymbol{U})^T \boldsymbol{M_z})^T}{\boldsymbol{A^T M_Z A}} + \frac{\boldsymbol{A^T M_Z}\epsilon_{\boldsymbol{I}}}{\boldsymbol{A^T M_Z A}} \tag{118}$$

By idempotence of $\boldsymbol{M_Z}$. $\boldsymbol{A^T M_Z}$ is the residual from the regression of the treatment on $\boldsymbol{Z}$ and likewise $(f_1(\boldsymbol{U})^T \boldsymbol{M_z})^T$ is the transpose of the residuals from the regression of $f_1(\boldsymbol{U})$ on $\boldsymbol{Z}$. To find the residuals, it suffices to find the coefficients from the above regressions.

Let $\boldsymbol{A} = \boldsymbol{\iota}\alpha_1 + f_2(\boldsymbol{BAV})\zeta_{f_2(bav)_1} + \upsilon_1$ and $f_1(\boldsymbol{U}) = \boldsymbol{\iota}\alpha_2 + f_2(\boldsymbol{BAV})\zeta_{f_2(bav)_2}$ represent the regression equations. Following the FWL theorem once more, we have the following estimators

$$\widehat{\alpha_1} = \frac{\boldsymbol{\iota^T M}_{f_2(\boldsymbol{BAV})}\boldsymbol{A}}{\boldsymbol{\iota^T M}_{f_2(\boldsymbol{BAV})}\boldsymbol{\iota}} \xrightarrow{p} E[\boldsymbol{A}] - \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))}{Var(f_2(\boldsymbol{BAV}))} E[f_2(\boldsymbol{BAV})] = E[\boldsymbol{A}] \tag{119}$$

$$\widehat{\zeta_{f_2(bav)_1}} = \frac{f_2(\boldsymbol{BAV})^T \boldsymbol{M_\iota A}}{f_2(\boldsymbol{BAV})^T \boldsymbol{M_\iota} f_2(\boldsymbol{BAV})} \xrightarrow{p} \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))}{Var(f_2(\boldsymbol{BAV}))} \tag{120}$$

$$\widehat{\alpha_2} = \frac{\boldsymbol{\iota^T M}_{f_2(\boldsymbol{BAV})} f_1(\boldsymbol{U})}{\boldsymbol{\iota^T M}_{f_2(\boldsymbol{BAV})}\boldsymbol{\iota}} \xrightarrow{p} E[f_1(\boldsymbol{U})] - \frac{COV(f_1(\boldsymbol{U}), f_2(\boldsymbol{BAV}))}{Var(f_2(\boldsymbol{BAV}))} E[f_2(\boldsymbol{BAV})] = 0 \tag{121}$$

$$\widehat{\zeta_{f_2(bav)_2}} = \frac{f_2(\boldsymbol{BAV})\boldsymbol{M_\iota} f_1(\boldsymbol{U})}{f_2(\boldsymbol{BAV})\boldsymbol{M_\iota} f_2(\boldsymbol{BAV})} \xrightarrow{p} \frac{COV(f_2(\boldsymbol{BAV}), f_1(\boldsymbol{U}))}{Var(f_2(\boldsymbol{BAV}))} \tag{122}$$

Using the probability limits of the estimates, the continuous mapping theorem and the fact that $\boldsymbol{A^T M_Z A} \xrightarrow{p} (1 - \mathcal{R}^2_{A|Z})\sigma^2_a$ where $\mathcal{R}^2_{A|Z}$ is understood to be the probability limit of the $\mathcal{R}^2$ from the regression of the treatment on the columns of $\boldsymbol{Z}$, we get

$$\widehat{\beta}^{|f_2(BAV)} \xrightarrow{p} \frac{COV(\boldsymbol{A^T M_Z}, (f_1(\boldsymbol{U})^T \boldsymbol{M_Z})}{\sigma^2_a(1 - \mathcal{R}^2_{A|Z})} \tag{123}$$

$$= \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U})) - \frac{COV(\boldsymbol{A}, f_2(\boldsymbol{BAV}))COV(\boldsymbol{f_1(U)}, \boldsymbol{f_2(BAV)})}{Var(f_2(\boldsymbol{BAV})}}{\sigma^2_a(1 - \mathcal{R}^2_{A|Z})} \tag{124}$$

and the error term drops out since $E[\epsilon_{\boldsymbol{I}}|\boldsymbol{A}, \boldsymbol{\iota}, f_2(\boldsymbol{BAV})] = 0$.

A.8.1 Orthogonal semi-parametric models derivation

From the standard OLS formula we get

$$\widehat{\beta}^{semi} = \frac{\frac{1}{n}(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{Y} - \widehat{E}[\boldsymbol{Y}|\boldsymbol{BAV}])}{\frac{1}{n}(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{125}$$

$$\xrightarrow{p} \frac{E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{BAV}])]}{E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])]} \tag{126}$$

$$= \frac{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{BAV}])|\boldsymbol{BAV}]]}{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])]|\boldsymbol{BAV}]} \tag{127}$$

$$= \frac{E[COV(\boldsymbol{A}, \boldsymbol{Y}|\boldsymbol{BAV})]}{E(Var(\boldsymbol{A}|\boldsymbol{BAV})} \tag{128}$$

$$= \frac{E[COV(\boldsymbol{A}, \boldsymbol{Y}|\boldsymbol{BAV})]}{\sigma_a^2 - var(E[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{129}$$

This holds under any DAG and any set of structural equations. However, as discussed in section 4, without making restrictions on the underlying structural equations, the estimator may not be meaningful or easily comparable to a causal effect of interest. Under the structural equations in equations (19) and (20), the estimator probability limit is

$$\widehat{\beta}^{semi} \xrightarrow{p} \frac{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{BAV}])|\boldsymbol{BAV}]]}{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])]|\boldsymbol{BAV}]} \tag{130}$$

$$= \frac{\left\{ \begin{array}{c} E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T((\alpha_y + \boldsymbol{A}\beta_a + f_1(\boldsymbol{U}) + f_2(\boldsymbol{BAV}) + \epsilon_{\boldsymbol{I}}) \\ - E[((\alpha_y + \boldsymbol{A}\beta_a + f_1(\boldsymbol{U}) + f_2(\boldsymbol{BAV}) + \epsilon_{\boldsymbol{I}})|\boldsymbol{BAV}]|\boldsymbol{BAV}])] \end{array} \right\}}{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])]|\boldsymbol{BAV}]} \tag{131}$$

$$= \beta_a + \frac{E[E[(\boldsymbol{A} - E[\boldsymbol{A}|\boldsymbol{BAV}])^T(f_1(\boldsymbol{U}) - E[f_1(\boldsymbol{U})|BAV])}{\sigma_a^2 - var(E[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{132}$$

$$= \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U})|\boldsymbol{BAV})}{\sigma_a^2 - var(E[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{133}$$

In the special case that $\boldsymbol{U}$ and $\boldsymbol{BAV}$ are independent, as they would be under the DAG in Figure 3, this reduces to

$$\widehat{\beta}^{semi} \xrightarrow{p} \beta_a + \frac{COV(\boldsymbol{A}, f_1(\boldsymbol{U}))}{\sigma_a^2 - var(E[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{134}$$

A.9 Bias amplification under mispecification and unrestricted structural equations

A.9.1 Bias Amplification and mispecification of the control function

In section 4 of the main text, we considered the problem of bias amplification when $f_2(\boldsymbol{BAV})$ was specified correctly at least in the limit $n \to \infty$. Here we consider the additional possibility of misspecification. Suppose the structural equations were again those in equations (19) and (20) and instead of including the function $f_2(\boldsymbol{BAV})$ we include some other control function $h(\boldsymbol{BAV})$. For clarity, this control function could be just the linear term $\boldsymbol{BAV}$ or a polynomial function for example, but in general it might be anything the analyst believes is reasonable, but that is ultimately not quite correct. If $h(\boldsymbol{BAV})$ is unable to be written as a linear combination of $f_2(\boldsymbol{BAV})$ (i.e. $\nexists a, b \in \mathcal{R} : h(\boldsymbol{BAV}) = a + bf_2(\boldsymbol{BAV})$), then there may be mispecification bias in addition to unmeasured confounding.

The naive estimator ($\widehat{\beta}_a^{naive}$) and the estimator conditional on $h(\boldsymbol{BAV})$ will have the following probability limits.

$$\widehat{\beta}_a^{naive} \xrightarrow{p} \beta_a + (Cov(\boldsymbol{A}, f_1(\boldsymbol{U})) + Cov(\boldsymbol{A}, f_2(\boldsymbol{BAV}))\frac{1}{\sigma_a^2} \tag{135}$$

$$\widehat{\beta}^{|h(\boldsymbol{BAV})} \xrightarrow{p} \beta_a + (Cov(\boldsymbol{A^T M_z}, (f_1(\boldsymbol{U})^T \boldsymbol{M_z})^{\boldsymbol{T}}) + Cov(\boldsymbol{A^T M_z}, (f_2(\boldsymbol{BAV})^T \boldsymbol{M_z})^{\boldsymbol{T}}))\frac{1}{(1 - \mathcal{R}^2_{\boldsymbol{A}|\boldsymbol{h}(\boldsymbol{BAV})})\sigma_a^2} \tag{136}$$

$$= \beta_a + COV(\boldsymbol{A}, f_1(\boldsymbol{U})) - \frac{COV(\boldsymbol{A}, h(\boldsymbol{BAV}))COV(f_1(\boldsymbol{U}), h(\boldsymbol{BAV}))}{Var(h(\boldsymbol{BAV}))} \tag{137}$$

$$+COV(\boldsymbol{A}, f_2(\boldsymbol{BAV})) - \frac{COV(\boldsymbol{A}, h(\boldsymbol{BAV}))COV(h(\boldsymbol{BAV}), f_2(\boldsymbol{BAV}))}{Var(h(\boldsymbol{BAV}))}\frac{1}{1 - \mathcal{R}^2_{A|h(BAV)}} \tag{138}$$

When we compare the probability limit of the misspecified estimator ($\widehat{\beta}^{h(BAV)}$) in equation (136) and the adjusted estimator in equation (123), there is an additional term due to misspecification in the numerator ($Cov(\boldsymbol{A^T M_z}, (f_2(\boldsymbol{BAV})^T \boldsymbol{M_z})^{\boldsymbol{T}})$). This does not necessarily mean that it is more biased and will depend on the signs of the covariances between all of the relevant variables. The numerator of the bias term ($Cov(\boldsymbol{A^T M_z}, (f_1(\boldsymbol{U})^T \boldsymbol{M_z})^{\boldsymbol{T}}) + Cov(\boldsymbol{A^T M_z}, (f_2(\boldsymbol{BAV})^T \boldsymbol{M_z})^{\boldsymbol{T}})$) is still the covariance of the treatment and outcome through the confounding pathways $f_1(\boldsymbol{U}) \rightarrow \boldsymbol{A} \rightarrow \boldsymbol{Y}$ and the part of $f_2(\boldsymbol{BAV}) \rightarrow \boldsymbol{A} \rightarrow \boldsymbol{Y}$ which is not controlled by linear combinations of $h(\boldsymbol{BAV})$. The denominator is again always greater than 1 and contributing to amplification, but increases hyperbolically with respect to how well $h(\boldsymbol{BAV})$ linearly explains variance in the treatment rather than $f_2(\boldsymbol{BAV})$. Notice that even though $h(\boldsymbol{BAV})$ is misspecified with respect to the true outcome equation (19), this does not mean that the amplification term will be less severe. It is possible that $h(\boldsymbol{BAV})$ explains more of the treatment covariance since $\boldsymbol{A} = \alpha_a + g(\boldsymbol{U}, \boldsymbol{BAV}) + \epsilon_{\boldsymbol{2}}$ and $h(\boldsymbol{BAV})$ may better approximate $g(\boldsymbol{U}, \boldsymbol{BAV})$. Overall, misspecification does not qualitatively change the problem of bias amplification, but in the case that the analyst does not know $f_2(\boldsymbol{BAV})$ it may be more difficult to harness domain knowledge to accurately predict the sign and direction of the bias. The amplifying denominator can still be estimated using observed data and used in the process of model selection or sensitivity analysis.

A.9.2 Bias amplification under unrestricted structural equations, heterogeneous causal effects

In the main text, we restricted our focus to systems of structural equations which were partially linear in the treatment of interest such that the desired causal effect was a single parameter (β_a) in the true model. In practice, it may be that the causal effect of interest is heterogeneous and a more complicated functional and the treatment of this setting is beyond the scope of this manuscript. In the case that we assume that the effect of interest is β_a but it is not, we can say the following about the OLS and partially least squares estimators.

If we make no assumptions about the functional form, and allow for non-linearities, and interactions between all variables, we can show that the OLS adjusted estimator is always the expression below (see Appendix A.1)

$$\widehat{\beta_a^{|z}} = \frac{(\boldsymbol{M_z A})^{\boldsymbol{T}} \boldsymbol{M_z Y}}{(\boldsymbol{M_z A})^{\boldsymbol{T}} (\boldsymbol{M_z A})} \tag{139}$$

When $\boldsymbol{Z}$ includes an intercept column, both $(\boldsymbol{M_z A})^{\boldsymbol{T}}$ and $\boldsymbol{M_z Y}$ will have mean zero and thus we can think of the numerator as an empirical estimate of the covariance between the residuals from the regression of $\boldsymbol{A}$ on $\boldsymbol{Z}$ and the residuals from the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$. Unmeasured confounding bias in OLS occurs when after projecting out linear combinations of the controlling variables, $\boldsymbol{Z}$, there remain linear associations between the outcome and the treatment due to unobserved variables. The part of the bias due to unmeasured confounding is amplified whenever the control variables explain variance in the treatment. Holding all else constant, as the residuals from the regression of the treatment on $\boldsymbol{Z}$ decrease in magnitude, the absolute value of the estimator $\widehat{\beta_a^{|z}}$ will increase in magnitude. This is a general form of the result we showed in the previous section which is extremely powerful in that it encaptures a very large class of structural equations and DAGs. However, the cost of this generality is that

without making more specific assumptions about the particular form of the model, and in particular the outcome model, it becomes more difficult to incorporate the knowledge of the amplification factor into our model selection and thus a priori know which of the two estimators, $\widehat{\beta}_a^{naive}$ or $\widehat{\beta}_a^{bav\prime}$, will be less biased, especially when the nonlinearities imply heterogenous causal effects and thus that β_a itself is not the causal effect of interest. In particular, we cannot easily decompose the bias or relative bias into covariances through the confounding pathways and treatment variance. The amplifying denominator remains the same, but the numerator cannot be easily decomposed. Interaction terms, for example, are an additional difficulty. Pearl,[10] for example, showed that under a simple interaction effect between the unmeasured confounding and some function of a pure instrument, the adjusted estimator can be less biased than the naive case. To properly evaluate estimators in the context of bias amplification requires appropriate simulations.

Similarly, we can show that the partially linear estimator conditional on BAV under any structural equations is

$$\widehat{\beta_{partial}}^{|BAV} = \frac{(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{Y} - \widehat{E}[\boldsymbol{Y}|\boldsymbol{BAV}])}{(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])^T(\boldsymbol{A} - \widehat{E}[\boldsymbol{A}|\boldsymbol{BAV}])} \tag{140}$$

$$\xrightarrow{p} \frac{E[COV(\boldsymbol{A}, \boldsymbol{Y}|\boldsymbol{BAV})]}{(1 - \frac{VAR(E[\boldsymbol{A}|\boldsymbol{BAV}])}{\sigma_a^2})\sigma_a^2} \tag{141}$$

Much like the OLS estimator, the numerator of the partially linear estimator is an expectation of the covariance of the treatment and outcome once we have controlled for $\boldsymbol{BAV}$. The denominator, just like in the main text is proportional to the remaining treatment variance once we have controlled for fluctuations in $\boldsymbol{BAV}$. As $\boldsymbol{BAV}$ explains more of the treatment variation, the magnitude of the estimator increases hyperbolically. Again, however, without a priori knowledge of the structural form for the outcome and thus the form of the causal effect of interest, it is not clear if this will be closer or farther away from desired estimand. As an estimator, however, we can see that both of these orthogonalization estimators, OLS and partially least squares, have the property that explaining variance in the treatment increases the magnitude of the estimate hyperbolically.

A.10 Additional details simulation

Suppose we want to simulate a system of linear equations from equations (4) and (5) based on the DAG in Figure 3. Now suppose we are interested in assessing the effect of modifying the edge $\boldsymbol{BAV} \rightarrow \boldsymbol{A}$ on the conditional estimator $\widehat{\beta}_a^{bav\prime}$. If we incorrectly run this simulation simply by changing the parameter γ_{bav} to some (or some set of) $\gamma_{bav}{}'$ and fail to fix the variance of the treatment $\boldsymbol{A}$ as discussed in section 5, we can show that the bias of estimator $\widehat{\beta}_a^{bav\prime}$ will remain unchanged. In section A.14 we show that the variance of $\boldsymbol{A}$ in the above simulation design is equal to

$$\sigma_a^2 = \gamma_u^2\sigma_u^2 + \gamma_{bav}^2\sigma_{bav}^2 + \sigma_{\epsilon_2}$$

Further we showed that the expectation and probability limit of the estimator is

$$E[\widehat{\beta}_a^{bav\prime}] = \beta_a + \beta_u \frac{\gamma_u\sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2}$$

Thus, if we change γ_{bav} to $\gamma_{bav}{}'$ holding all other parameters constant, it can be shown that the resulting expectation is unchanged. This is because the increased amplification is precisely cancelled out by increasing the variance of the treatment.

$$\sigma_a{}' = \gamma_u^2\sigma_u^2 + \gamma_{bav}'^2\sigma_{bav}^2 + \sigma_{\epsilon_2}$$

$$= \sigma_a^2 - \gamma_{bav}^2\sigma_{bav}^2 + \gamma_{bav}'^2\sigma_{bav}^2$$

$$= \sigma_a^2 + (\gamma_{bav}'^2 - \gamma_{bav}^2)\sigma_{bav}^2$$

This implies the expectation of the estimator $\widehat{\beta}_a^{bav\prime}$ has the following expression

$$\begin{aligned} E[\widehat{\beta}_a^{bav\prime}]' &= \beta_a + \beta_u \frac{\gamma_u \sigma_u^2}{\sigma_a'^2 - \gamma_{bav}'^2 \sigma_{bav}^2} \\ &= \beta_a + \beta_u \frac{\gamma_u \sigma_u^2}{(\sigma_a^2 + (\gamma_{bav}'^2 - \gamma_{bav}^2)\sigma_{bav}^2) - \gamma_{bav}'^2 \sigma_{bav}^2} \\ &= \beta_a + \beta_u \frac{\gamma_u \sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2} \\ &= E[\widehat{\beta}_a^{bav\prime}] \end{aligned}$$

Thus the expectation of the estimator remains unchanged for any change of parameter, γ_{bav}. As a consequence, the comparison of this estimator with the naive estimator will seem favorable as the absolute magnitude of the parameter γ_{bav}' increases since the difference in absolute bias is

$$E[\widehat{\beta}_a^{naive}] - E[\widehat{\beta}_a^{bav\prime}] = \left| \frac{\beta_u \gamma_u \sigma_u^2}{\sigma_a^2} + \frac{\beta_{bav} \gamma_{bav} \sigma_{bav}^2}{\sigma_a^2} \right| - \left| \beta_u \frac{\gamma_u \sigma_u^2}{\sigma_a^2 - \gamma_{bav}^2 \sigma_{bav}^2} \right|$$

The bias of $\widehat{\beta}_a^{naive}$ is increasing in γ_{bav} for sufficiently large γ_{bav} and we showed above that the bias for $\widehat{\beta}_a^{bav\prime}$ is invariant to changes in γ_{bav} if we do not fix the variance of the treatment $\boldsymbol{A}$. Thus eventually the bias of the naive estimate is strictly increasing in γ_{bav} and will continue to appear worse relative to the conditional estimator. However, as discussed in section 5, this is a consequence of failing to conduct a proper causal simulation experiment comparing data sets plausibly generated from similar experiments and holding all other potentially confounding edges constant.

A.11 Real data simulation details

The goal of this section is to utilize the real randomized control trial data described in section 6, which comes from the DAG in Figure 7(a), and simulate modified covariates ($\tilde{\boldsymbol{X}}$) and a modified outcome ($\tilde{Y}$) such that they come from the DAG 7b and the equations (22), (23), and (24). This proceeds in two steps. First we need to simulate the latent variable $\boldsymbol{A}^*$ and then conditionally simulate $\boldsymbol{BAV}$ and $\boldsymbol{U}$.

For simplicity, we set $\boldsymbol{U}$ and $\boldsymbol{BAV}$ to be standard normal variables. The latent variable, $\boldsymbol{A}^*$, has variance of 1 but its mean, α_a is determined to ensure that $P(\boldsymbol{A}^* > 0) = P(\boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_b av + \epsilon_2 > \alpha_a) = p_a$, which is determined in our data by matching p_a to the observed quantity $\widehat{p_a} = \frac{1}{n}\sum_{i=1}^{n} A_i \approx 0.51$.

Under the assumptions above, this implies that $\alpha_a = -\Phi^{-1}(1 - p_a)$, where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution. As previously mentioned, the variance of the error term ϵ_2 is set precisely to ensure that variance of $\boldsymbol{A}^*$ is 1, $\sigma_{\epsilon_2}^2 = 1 - \gamma_u^2 - \gamma_{bav}^2$.

The first step is to use the observed $\boldsymbol{A}$ data to simulate the latent $\boldsymbol{A}^*$. Consider the CDF of the latent $\boldsymbol{A}^*$ conditional on $\boldsymbol{A} = 1$. Let $P(\boldsymbol{A} = 1) = p_a$, so that

$$\begin{aligned} P(\boldsymbol{A}^* \leq a^* | \boldsymbol{A} = 1) &= P(\boldsymbol{A}^* \leq a^* | \boldsymbol{A}^* \geq 0) \\ &= \frac{P(\boldsymbol{A}^* \leq a^*, \boldsymbol{A}^* \geq 0)}{P(\boldsymbol{A}^* \geq 0)} \\ &= \frac{P(0 \leq \boldsymbol{A}^* \leq a^*)}{P(\boldsymbol{A}^* \geq 0)} \end{aligned}$$

$$= \frac{P(\frac{-E[\boldsymbol{A}^*]}{\sigma^2_{a^*}} \leq \frac{\boldsymbol{A}^* - E[\boldsymbol{A}^*]}{\sigma^2_{a^*}} \leq \frac{a^* - E[\boldsymbol{A}^*]}{\sigma^2_{a^*}})}{P(\boldsymbol{A}^* \geq 0)}$$

$$= \frac{\Phi(\frac{a^* - E[\boldsymbol{A}^*]}{\sigma^2_{a^*}}) - \Phi(\frac{-E[\boldsymbol{A}^*]}{\sigma^2_{a^*}})}{p_a}$$

$$= \frac{\Phi(a^* - \alpha_a) - \Phi(-\alpha_a)}{p_a}$$

$$= \frac{\Phi(a^* - \alpha_a) - \Phi(\Phi^{-1}(1 - p_a))}{p_a}$$

$$= \frac{\Phi(a^* - \alpha_a) - (1 - p_a)}{p_a}$$

Since $P(\boldsymbol{A}^* \leq a^* | \boldsymbol{A} = 1)$ is the CDF of a continuous random variable, it is distributed uniformly between 0 and 1. Let $\boldsymbol{X} \sim U(0,1)$ be a uniform random variable with support $[0,1]$. Appealing to the probability inverse transform, X and $P(\boldsymbol{A}^* \leq a^* | \boldsymbol{A} = 1)$ are equivalent in distribution.

$$\Rightarrow \boldsymbol{X} \equiv \frac{\Phi(a^* - \alpha_a) - (1 - p_a)}{p_a}$$

$$\Rightarrow p_a \boldsymbol{X} + (1 - p_a) \equiv \Phi(a^* - \alpha_a)$$

$$\Rightarrow \Phi^{-1}(p_a \boldsymbol{X} + (1 - p_a)) + \alpha_a \equiv a^*$$

Similarly, it can be shown that conditional on $A = 0$

$$a^* \equiv \Phi^{-1}((1 - p_a)\boldsymbol{X}) + \alpha_a$$

Thus, in general

$$a^* \equiv \Phi^{-1}(p a^{\boldsymbol{A}} (1 - p_a)^{\boldsymbol{A}-1} \boldsymbol{X} + (1 - p_a)^{\boldsymbol{A}}) + \alpha_a$$

Therefore, by conditioning on $\boldsymbol{A}$ and simulating a uniform random variable, we can take draws from the unobserved latent variable $\boldsymbol{A}^*$. Once we have recovered the latent variable, we can jointly simulate $\boldsymbol{U}$ and $\boldsymbol{BAV}$ conditional on $\boldsymbol{A}^*$ and the observed covariates $\boldsymbol{X}$. The observed covariates are centered and asymptotically multivariate normal. Since $\boldsymbol{U}$, $\boldsymbol{BAV}$, and $\boldsymbol{A}^*$ are univariate or multivariate normal variables and $\boldsymbol{X}$ are asymptotically normal, the conditional distribution will be asymptotically multivariate normal and proportional to the joint density. From standard multivariate normal theory

$$\boldsymbol{U} = u, \boldsymbol{BAV} = bav | \boldsymbol{A}^* = a^*, \boldsymbol{X} = x \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$$

As stated above, the conditional distribution is proportional to the joint model. Thus, we will define Σ and $\boldsymbol{\mu}$ for the joint density.

$$\boldsymbol{U} = u, \boldsymbol{BAV} = bav, \boldsymbol{A}^* = a^*, \boldsymbol{X} = x \sim N(\boldsymbol{\mu}, \Sigma)$$

$$\Sigma = \begin{bmatrix} \Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{U},\boldsymbol{BAV})} & \Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{A}^*,\boldsymbol{X})} \\ \Sigma_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{U},\boldsymbol{BAV})} & \Sigma_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{A}^*,\boldsymbol{X})} \end{bmatrix}$$

$$\Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{U},\boldsymbol{BAV})} = \begin{bmatrix} \sigma_u^2 & 0 & 0 & 0 \\ 0 & \sigma_{bav_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{bav_2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{bav_3}^2 \end{bmatrix}$$

$$\Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{A}^*,\boldsymbol{X})} = \begin{bmatrix} \gamma_u \sigma_u^2 & 0 & 0 & 0 \\ \gamma_{\tilde{x_1}} \sigma_{bav_1}^2 & 0 & 0 & 0 \\ \gamma_{\tilde{x_2}} \sigma_{bav_2}^2 & 0 & 0 & 0 \\ \gamma_{\tilde{x_3}} \sigma_{bav_3}^2 & 0 & 0 & 0 \end{bmatrix}$$

$$\Sigma_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{U},\boldsymbol{BAV})} = \Sigma^{\boldsymbol{T}}_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{A}^*,\boldsymbol{X})}$$

$$\Sigma_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{A}^*,\boldsymbol{X})} = \begin{bmatrix} \sigma_{a^*}^2 & 0 & 0 & 0 \\ 0 & \sigma_{x_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{x_2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{x_3}^2 \end{bmatrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_u \\ \mu_{bav_1} \\ \mu_{bav_2} \\ \mu_{bav_3} \\ \mu_{a^*} \\ \mu_{x_1} \\ \mu_{x_2} \\ \mu_{x_3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \alpha_a \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Using standard multivariate normal theory and the matrices defined above, we can define the mean, $\boldsymbol{\mu}^*$ and variance, Σ^* of the conditional model

$$\Sigma^* = \Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{U},\boldsymbol{BAV})} - \Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{A}^*,\boldsymbol{X})} \Sigma^{-\boldsymbol{1}}_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{A}^*,\boldsymbol{X})} \Sigma_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{U},\boldsymbol{BAV})}; \text{ and}$$

$$\boldsymbol{\mu}^* = \Sigma_{(\boldsymbol{U},\boldsymbol{BAV}),(\boldsymbol{A}^*,\boldsymbol{X})} \Sigma^{-\boldsymbol{1}}_{(\boldsymbol{A}^*,\boldsymbol{X}),(\boldsymbol{A}^*,\boldsymbol{X})} ([\boldsymbol{A}^*, \boldsymbol{X}] - [\begin{smallmatrix} \mu_{a^*} \\ \boldsymbol{\mu_x} \end{smallmatrix}])^T$$

Using the conditional distribution, we can thus take draws of $\boldsymbol{U}, \boldsymbol{BAV}$ conditional on the particular values of $\boldsymbol{A}^*$ and $\boldsymbol{X}$. Using the simulated $\boldsymbol{BAV}$ we add it to the covariates $\boldsymbol{X}$ to form the modified covariates, $\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\sigma'} + \boldsymbol{BAV}$, where σ' is a scaling factor chosen simultaneously with σ_{bav} such that the variance of $\tilde{\boldsymbol{X}}$ is precisely equal to $\boldsymbol{\sigma_x} = 1$. This step is important if we would like to compare simulations with the modified and the unmodified covariates. In the particular simulations conducted in section 6, the scaling was chosen such that $Var(\frac{\boldsymbol{X_i}}{\sigma'}) = 0.01, i = 1, 2, 3$ and thus $\sigma_{bav_i}^2 = 0.99, i = 1, 2, 3$.

Now that the modified covariates have been constructed, the modified outcome can be constructed. The original RCT data coming from Figure 7(a) is assumed to come from the linear model

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{X}\beta_x + \epsilon_1$$

where β_a and β_x are estimated unbiasedly in section 6. Next, we add the unmeasured confounding, $\boldsymbol{U}\beta_u$ directly (where β_u is chosen) and then add $\boldsymbol{BAV}\beta_x + \tilde{\boldsymbol{X}}\beta_{adj}$, where $\beta_{adj} = \beta\tilde{x} - \beta_x$ where $\beta\tilde{x}$ is chosen to set the desired covariance between the modified covariates and the modified outcome

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_x + \tilde{\boldsymbol{X}}(\beta\tilde{x} - \beta_x)$$

$$= (\alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{X}\beta_x + \epsilon_{\boldsymbol{I}}) + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_x + \tilde{\boldsymbol{X}}(\beta\tilde{x} - \beta_x)$$

$$= \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + (\boldsymbol{X} + \boldsymbol{BAV})\beta_x + \tilde{\boldsymbol{X}}(\beta\tilde{x} - \beta_x) + \epsilon_{\boldsymbol{I}}$$

$$= \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + (\tilde{\boldsymbol{X}})\beta_x + \tilde{\boldsymbol{X}}(\beta\tilde{x} - \beta_x) + \epsilon_{\boldsymbol{I}}$$

$$= \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \tilde{\boldsymbol{X}}\beta\tilde{x} + \epsilon_{\boldsymbol{I}}$$

This is precisely the outcome equation in equation (22) in Section 6.2. Thus following this method we can use the real data to create a data simulation using the original treatment data and matching many of the characteristics of the real data, but precisely control the causal structure and correlations between the variables. As with the other simulations, there will still be restrictions on the parameters and correlations that we set such as positive definiteness of all the variance matrices in the above simulation.

A.12 Real data simulation comparison of estimators

Consider a causal simulation experiment coming from a DAG and system of equations identical to the one considered in section 6.2 as described by Figure 7(b) and the system of equations (22), (23), and (24). The experiment uses the real data described in section 6 and the procedure detailed in section A.11. The simulation experiment involves intervening on the edge $\tilde{\boldsymbol{X}}_{\boldsymbol{I}} \rightarrow \boldsymbol{A}$, that is increasing the covariance between $\tilde{\boldsymbol{X}}_{\boldsymbol{I}}$ and $\boldsymbol{A}$. As in section 5 we will explore the consequences of failing to properly hold all non-intervention edges of the DAG.

Simulation Parameters	$\gamma_{\tilde{X}}$	$\gamma_{\boldsymbol{u}}$	$\beta_{\boldsymbol{u}}$	$\beta_{\tilde{\boldsymbol{x}}}$	$\beta_{\boldsymbol{a}}$
Control	0.20, 0.38, 0.33	0.63	0.15	0.10, −0.15, −0.10	0.1377
Intervention	0.55, 0.38, 0.33	0.63	0.15	0.10, −0.15, −0.10	0.1377

Above, the parameters for the two simulation treatments are described. The only difference between the two is that in the control, $\gamma_{\tilde{x}_1} = 0.2$ and in the intervention $\gamma_{\tilde{x}_1} = 0.55$. Below we visualize the naive, adjusted, and unbiased estimators for the control treatment.

In the control treatment, we see that the unbiased estimator behaves as expected, centered on the true underlying parameter. The naive estimator $\widehat{\beta}_a^{naive}$ is only slightly biased, since some unmeasured biases due to the vector $\tilde{\boldsymbol{X}}$ and $\boldsymbol{U}$ happen to be of opposing signs and partially cancel each other out. If this is not the case, of course the naive estimator may be significantly more biased. The adjusted estimator behaves poorly with an average absolute bias of 0.18. Although the parameters in the latent space are relatively large, $\boldsymbol{\gamma}_{\tilde{\boldsymbol{x}}} = [0.2, 0.38, 0.33]$, the covariances in the observed space with respect to the treatment are relatively small, $COV(\boldsymbol{A}, \tilde{\boldsymbol{X}}) = [0.08, 0.15, 0.13]$, and yet the amplifying effect is quite large. In fact, the amplifying variables jointly explain only 18% of the variance of the treatment, but since the variance of the treatment was already quite small, $\sigma_a^2 \approx 0.25$, the amplifying variables had a more than proportional effect.

The bias attributed to the path $\boldsymbol{A} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{Y}$ for the naive estimator is $\frac{\beta_u COV(\boldsymbol{A},\boldsymbol{U})}{\sigma_a^2} = \frac{0.15 \times 0.25}{0.25} = \frac{0.0375}{25} = 0.0375 \times 4 = 0.15$, whereas for the amplified estimator it is $\frac{0.15 \times 0.25}{0.25 - (0.08^2 + 0.15^2 + 0.13^2)} = \frac{0.0375}{.183} = 0.0375 \times 4.88 = 0.183$. Since $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{MzA} \leq \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{M}_{\boldsymbol{\iota}}\boldsymbol{A}, \forall \boldsymbol{Z} : \boldsymbol{\iota} \subseteq \boldsymbol{Z}$, we can rewrite the bias due to bias amplification as $\frac{|\beta_u \times COV(\boldsymbol{A},\boldsymbol{U})|}{(1-c)\times\sigma_a^2}, c \in [0, 1]$, where c is the proportion of treatment variance explained by the bias amplifiers jointly

$$\frac{\partial^2\left(\frac{|\beta_u \times COV(\boldsymbol{A},\boldsymbol{U})|}{(1-c)\times\sigma_a^2}\right)}{\partial\sigma_a^2 \partial c} = \begin{cases} \dfrac{-\beta_u \times COV(\boldsymbol{A},\boldsymbol{U})}{(1-c)^2(\sigma_a^2)^2}, \beta_u \times COV(\boldsymbol{A},\boldsymbol{U}) > 0 \\ \dfrac{\beta_u \times COV(\boldsymbol{A},\boldsymbol{U})}{(1-c)^2(\sigma_a^2)^2}, \beta_u \times COV(\boldsymbol{A},\boldsymbol{U}) < 0 \end{cases}$$

The derivative above shows us that as the variance, σ_a^2, gets smaller, the marginal impact on absolute bias from an increase in the proportion of the variance explained by the amplifiers increases.

Now consider the intervention of increasing the proportion of variance explained by one of the potential amplifiers, $\tilde{\boldsymbol{X}}_{\boldsymbol{1}}$, by increasing $\gamma_{\tilde{x}_1}$ to 0.55. Again, we will consider the case of keeping all of the variances constant to the case where we simply change the parameter and allow the variances to float.

Notice in the left panel that although we have intentionally increased the amplification, the amplifying estimator has seemingly not changed. However, when we fix the variance, the bias amplification increases as we expected (the mean absolute bias increased from 0.18 to 0.23). The reason for this effect is that as the variance of the latent treatment $\boldsymbol{A}^*$ increases, the covariances of the variables of the unmeasured confounding and the treatment as well as the potential amplifiers, $COV(\boldsymbol{A}, \boldsymbol{U})$ and $COV(\boldsymbol{A}, \tilde{\boldsymbol{X}})$, decrease. It can be shown that when $\boldsymbol{A}^*, \tilde{\boldsymbol{X}}$, and $\boldsymbol{U}$ are normal or multivariate normal

$$COV(\boldsymbol{A}, \boldsymbol{U}) = \frac{1}{p_a} E[\boldsymbol{U}] + \frac{\gamma_u \sigma_u^2}{\sqrt{2\pi\sigma_{a^*}^2}} \exp\left(\frac{-\alpha_a^2}{2\sigma_{a^*}^2}\right) \tag{142}$$

$$COV(\boldsymbol{A}, \tilde{\boldsymbol{X}}) = \frac{1}{p_a} E[\tilde{\boldsymbol{X}}] + \frac{\boldsymbol{\gamma_{\tilde{x}} \sigma_{\tilde{x}}^2}}{\sqrt{2\pi\sigma_{a^*}^2}} \exp\left(\frac{-\alpha_a^2}{2\sigma_{a^*}^2}\right). \tag{143}$$

We can see in equation (142) that if we allow the variance of $\boldsymbol{A}^*$ to increase as $\gamma_{\tilde{x}_1}$ increases that $COV(\boldsymbol{A}, \boldsymbol{U})$ decreases. In the case of this simulation, the covariance decreased from 0.25 in the control treatment to 0.22, since $\sigma_{a^*}^2$ increased from 1 to $1 + (0.55^2 - 0.2^2) = 1.26$. Thus we have decreased the strength of the edge $\boldsymbol{U} \rightarrow \boldsymbol{A}$ incidentally. Further by considering equation (143), we can see that if we increase $\gamma_{\tilde{x}_1}$ we do not necessarily increase the amount of variance explained by $\tilde{\boldsymbol{X}}_{\boldsymbol{1}}$ since there are two opposing effects. First, consider the increase directly through $\gamma_{\tilde{x}_1}$ and the decrease through increasing $\sigma_{a^*}^2$. In the particular example, although our intended goal was to observe the effect of increasing the weight of the edge $\tilde{\boldsymbol{X}}_{\boldsymbol{1}} \rightarrow \boldsymbol{A}$, we have in fact inadvertently decreased the covariance from 0.2 to 0.196.

Again we can see that when we fail to hold the variances constant, we are no longer comparing a controlled intervention on the weight of a particular set of nodes, but have modified the edges into and out of the intervened upon edge. This example shows that this is true in cases beyond fully linear systems of equations explored in section 5. Examining the simulation results we can see that this might lead to inappropriate conclusions about the effects of our interventions and the relative merits of particular estimators in contexts of interest to us.

A.13 Simulation for Figure 4

The figure was simulated from the general structural equations (4) and (5) with the particular values below.

$$\boldsymbol{Y} = 2 + .2 \times \boldsymbol{A} + .5 \times \boldsymbol{U} + .05 \times \boldsymbol{BAV} + \boldsymbol{\upsilon_1}$$

$$\boldsymbol{A} = 1 + .3 \times \boldsymbol{U} + .75 \times \boldsymbol{BAV} + \boldsymbol{\upsilon_2}$$

$$\boldsymbol{BAV} \sim N(0,1), \quad \boldsymbol{U} \sim N(0,1)$$

$$\boldsymbol{\upsilon_1} \sim N(0, \sigma_{\upsilon_1})$$

$$\sigma_{\upsilon_1} = (\sigma_y^2 - (\beta_a^2\sigma_a^2 + \beta_u^2\sigma_u^2 + \beta_{bav}^2\sigma_{bav}^2$$

$$+2\beta_a\beta_u\gamma_u\sigma_u^2 + 2\beta_a\beta_{bav}\gamma_{bav}\sigma_{bav}^2\sigma_{\epsilon_1}^2))^{\frac{1}{2}}$$

$$= 0.906$$

$$\boldsymbol{\upsilon_2} \sim N(0, \sigma_{\upsilon_2})$$

$$\sigma_{\upsilon_2} = (\sigma_a^2 - (\gamma_u^2\sigma_u^2 + \gamma_{bav}^2\sigma_{bav}^2 + \sigma_{\epsilon_2}^2))^{\frac{1}{2}}$$

$$= 0.809$$

$\boldsymbol{\upsilon_1}$ and $\boldsymbol{\upsilon_2}$ had variances such that A and Y both have unit variance.

A.14 Variance derivations

A.14.1 Treatment variance for equation (5)

$$\boldsymbol{A} = \alpha_a + \boldsymbol{U}\gamma_u + \boldsymbol{BAV}\gamma_{bav} + \boldsymbol{\epsilon_2}$$

$$\Rightarrow Var(\boldsymbol{A}) = \gamma_u^2\sigma_u^2 + \gamma_{bav}^2\sigma_{bav}^2 + 2\gamma_u\gamma_{bav}COV(\boldsymbol{U}, \boldsymbol{BAV}) + Var(\boldsymbol{\epsilon_2})$$

$$\sigma_a^2 = \gamma_u^2\sigma_u^2 + \gamma_{bav}^2\sigma_{bav}^2 + \sigma_{\epsilon_2}^2 \tag{144}$$

A.14.2 Outcome variance for equation (4)

$$\boldsymbol{Y} = \alpha_y + \boldsymbol{A}\beta_a + \boldsymbol{U}\beta_u + \boldsymbol{BAV}\beta_{bav} + \boldsymbol{\epsilon_1} \tag{145}$$

$$\Rightarrow Var(\boldsymbol{Y}) = \beta_a^2 Var(\boldsymbol{A}) + \beta_u^2 Var(\boldsymbol{U}) + \beta_{bav} Var(\boldsymbol{BAV}) + \sigma_{\epsilon_2}^2 + \tag{146}$$

$$2\beta_a\beta_u Cov(\boldsymbol{A}, \boldsymbol{U}) + 2\beta_a\beta_{bav} Cov(\boldsymbol{A}, \boldsymbol{BAV}) + 2\beta_u\beta_{bav} Cov(\boldsymbol{U}, \boldsymbol{BAV}) \tag{147}$$

$$= \beta_a^2\sigma_a^2 + \beta_u^2\sigma_u^2 + \beta_{bav}^2\sigma_{bav}^2 + \sigma_{\epsilon_2}^2 + 2\beta_a\beta_u\gamma_u\sigma_u^2 + 2\beta_a\beta_{bav}\gamma_{bav}\sigma_{bav}^2 \tag{148}$$